COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 39 (2015) 1675–1703

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12208

Determinants of Scanpath Regularity in Reading

Titus von der Malsburg,[a] Reinhold Kliegl,[b] Shravan Vasishth[c]

[a]*Department of Psychology, University of California, San Diego*
[b]*Department of Psychology, University of Potsdam*
[c]*Department of Linguistics, University of Potsdam*

Received 16 January 2012; received in revised form 16 September 2014; accepted 16 September 2014

Abstract

Scanpaths have played an important role in classic research on reading behavior. Nevertheless, they have largely been neglected in later research perhaps due to a lack of suitable analytical tools. Recently, von der Malsburg and Vasishth (2011) proposed a new measure for quantifying differences between scanpaths and demonstrated that this measure can recover effects that were missed with the traditional eyetracking measures. However, the sentences used in that study were difficult to process and scanpath effects accordingly strong. The purpose of the present study was to test the validity, sensitivity, and scope of applicability of the scanpath measure, using simple sentences that are typically read from left to right. We derived predictions for the regularity of scanpaths from the literature on oculomotor control, sentence processing, and cognitive aging and tested these predictions using the scanpath measure and a large database of eye movements. All predictions were confirmed: Sentences with short words and syntactically more difficult sentences elicited more irregular scanpaths. Also, older readers produced more irregular scanpaths than younger readers. In addition, we found an effect that was not reported earlier: Syntax had a smaller influence on the eye movements of older readers than on those of young readers. We discuss this interaction of syntactic parsing cost with age in terms of shifts in processing strategies and a decline of executive control as readers age. Overall, our results demonstrate the validity and sensitivity of the scanpath measure and thus establish it as a productive and versatile tool for reading research.

Keywords: Eye movements; Reading; Scanpaths; Language understanding; Oculo-motor control; Individual differences; Aging; Development

1. Scanpaths in reading

Scanpaths—sequences of fixations of the eyes—have been a central concept in early classic work on eye movements. For example, Alfred Yarbus (1967) reported a seminal

Correspondence should be sent to Titus von der Malsburg, Department of Psychology, University of California, 9500 Gilman Drive, La Jolla, CA 92093-0109. E-mail: malsburg@ucsd.edu

study in which he had several participants look at the same picture; what differed was the task that they had to perform (pp. 171–196). The main finding was that these tasks elicited characteristic trajectories of the gaze that traced the attention shifts in the underlying cognitive processes. Yarbus concluded (p. 196):

> . . . the distribution of the points of fixation on an object, the order in which the observer's attention moves from one point of fixation to another, the duration of the fixations, the distinctive cyclic pattern of examination, and so on are determined by the nature of the object and the problem facing the observer at the moment of perception.

This work demonstrated that scanpaths can serve as a window into mental processes that are not directly observable. Interestingly, one of the most influential early eyetracking studies on sentence processing and reading also found that fixation sequences can be informative about cognitive processing. In 1982, Frazier and Rayner presented a pioneering experiment that investigated how readers recover from incorrect analyses in temporarily ambiguous sentences such as (1).

(1) Sally found out the answer was in the book.

This sentence is temporarily ambiguous because the complementizer "*that*" is elided. As a consequence, the noun phrase "*the answer*" is initially interpreted as the object of "*found out*." When the verb of the subordinate clause "*was*" is read, this analysis has to be discarded because the grammar of English does not allow a verb in this position under the currently maintained main clause interpretation. Thus, the current interpretation has to be replaced by another in which "*the answer*" is the subject of the subordinate clause. Frazier and Rayner discussed three competing hypotheses about this reanalysis process. The predictions of these hypotheses differ with respect to the sequences of fixations occurring after the disambiguating material is read. The first, forward reanalysis, assumes that, as soon as the incorrect interpretation is detected, a new interpretation is built from scratch; this is assumed to be accompanied by a gaze pattern in which the eyes return to the beginning of the sentence and start to reread. The second hypothesis, backward reanalysis, assumes that the detection of a parse error triggers a step-wise undoing and reevaluation of earlier interpretive decisions (Kaplan, 1972). The eyes follow this process, which results in a gaze pattern that could be described as reverse reading. The third hypothesis, selective reanalysis, posits that the parser corrects the defective interpretation by selectively editing the affected parts (Winograd, 1972). This recovery mechanism predicts regressive eye movements targeted at particular, linguistically relevant words. Frazier and Rayner (1982) analyzed the observed eye movement patterns qualitatively and ultimately concluded that the observed eye movement patterns favor the selective reanalysis hypothesis.

Given the important role of scanpaths in early eyetracking studies, one might think that this work should have sparked an intense interest in eye movement patterns. However, it is striking how little attention scanpaths received in later research. Most studies investigating eye movements employ relatively simple dependent measures such as fixation

durations or the probability of looking at a region or to regress from it. One reason for this state of affairs may be that scanpaths are complex. They are rich in structure and may contain a lot more information than a duration measure or a binary variable indicating whether or not an event occurred. In the simplest case, a scanpath consists of just two fixations. The maximal number of fixations is, however, only bounded by the time given to participants for examining the stimulus. Since fixations are typically short, a scanpath can become fairly complex in just a few seconds of viewing time. Additionally, each fixation in a scanpath has to be described in three dimensions: the location of the fixation in the visual field, given for example as latitude and longitude coordinates, and its duration. The resulting degrees of freedom may seem overwhelming and render scanpaths a comparatively unwieldy object for mathematical analysis. Statistical tests commonly used in cognitive psychology are well suited for analyzing univariate fixation-based measures, but they are rarely directly applicable to more complex representations such as those needed for accurate descriptions of the spatial and temporal properties of eye movement trajectories. This difficulty of analyzing scanpaths directly was probably one reason that led many researchers to use simplifying dependent variables even when the object of interest really was the pattern of fixations. Since a lot of information is lost by the reduction to scalar measures and probabilities, it is important to tailor measures that capture the variance of interest as well as possible. Today, a conventionally agreed-upon set of eyetracking measures is used for data analysis, and experiments are designed to elicit effects in these measures (see Clifton, Staub, & Rayner, 2007, for a review).

This approach has been tremendously successful and is perfectly valid in situations where the effect of a manipulation is focused on a small critical region. Nonetheless, there are also situations where the canonical measures are less suitable. The investigation of reanalysis strategies discussed above is one obvious example. The gaze trajectories predicted by the three competing hypothesis about reanalysis simply cannot be distinguished using measures such as gaze duration or regression probability; all three hypotheses predict the same amount of regressions, and they do not make any predictions about reading times. However, there are also less obvious situations where scanpaths may play a role. Parafoveal preview, for instance, can cause effects even before the eyes land on a critical word (Rayner, 1975). Similarly, the effects of a word can often be observed not just on the word itself but also on the following words (spillover; Rayner & Duffy, 1986). Thus, an analysis of classical eyetracking measures calculated only for the critical word takes into account only partial information and may be misleading. One remedy could be to increase the size of the region of interest for which eyetracking measures are calculated but that may introduce more problems than it solves: for example, regressions between words within that region would not be recognized and differential reading time effects on the words in that region may cancel each other out. Compared to that, a scanpath approach may be better suited for analyzing the perturbation in the gaze trajectory caused by a manipulation. Long effects that are spread over several words can be captured in whole by a scanpath analysis and do not have to be chopped up into pieces. For a further discussion, see Vasishth, von der Malsburg, and Engelmann (2013) and von der Malsburg, Vasishth, and Kliegl (2012).

One of the few studies in psycholinguistic research that investigated scanpath phenomena in reading quantitatively was presented by Meseguer, Carreiras, and Clifton (2002). The purpose of their study was to rigorously evaluate the selective reanalysis hypothesis. To do this, Meseguer and colleagues used a very clean experimental design where two types of sentences (example 2) were distinguished by only one letter that determined whether a decisive word was in indicative or in subjunctive mood ("*entraron*" vs. "*entraran*").

(2) a. *El profesor* **dijo** *que los alumnos se levantaran del asiento*
The teacher said that the students had to stand up from their seats
[$_{AdvC}$ *cuando los direciores* **entraron**$_{IND}$ *en la clase*].
[$_{AdvC}$ when the directors came into the class room].

b. *El profesor dijo que los alumnos* **se levantaran** *del asiento*
The teacher said that the students had to stand up from their seats
[$_{AdvC}$ *cuando los directores* **entraran**$_{SUBJ}$ *en la clase*].
[$_{AdvC}$ when the directors come into the class room].

This small difference leads to a change in the syntactic structure of the sentence. In (2a), the verb of the adverbial clause ("*entraron*") is in indicative mood, which entails that the adverbial clause ("*cuando los directores*. . .") modifies the main verb of the sentence ("dijo"), whereas the verb "entraran" in (2b) indicates that the adverbial clause modifies the embedded verb ("levantaran"). Since modification of the embedded verb is preferred in Spanish, the human sentence processor initially attaches the adverbial clause to the embedded verb but has to revise this decision in (2a) at "*entraron*." Any systematic difference in scanpaths observed for these two types of sentences can thus be attributed to the difference in the syntactic structure that is built because the visual layout of the two sentences is the same.

Meseguer and colleagues analyzed conventional duration measures but also devised a new dependent variable that measured the proportion of regressive saccades that land on a particular word. This measure was used to show that words that are linguistically relevant to the reanalysis *(*"*dijo*" and "*cuando*") were more likely to be targeted by regressions (see also Mitchell, Shen, Green, & Hodgson, 2008, for a similar approach). Additionally, they tried to identify a characteristic signature scanpath of selective reanalysis. This was done by calculating transition probabilities of the gaze between all regions of the sentence. It was found that some transitions were more likely in the condition that required reanalysis. However, it remained unclear in which order these transitions happened and whether they came from one or several types of scanpath patterns. Thus, the identification of a signature scanpath of syntactic reanalysis was ultimately not successful.

Another area in which scanpaths have played a role, albeit more implicitly, is the research on age differences in reading. Aging affects many if not all levels of processing involved in reading: visual perception (Fozard & Gordon-Salant, 2001), oculo-motor control (Abel, Troost, & Dell'Osso, 1983), lexical processing (Lima, Hale, & Myerson, 1991), and discourse processing (Stine-Morrow, Gagne, Morrow, & DeWall, 2004; see

Laubrock, Kliegl, & Engbert, 2006, for a review). The work by Stine-Morrow and colleagues is particularly interesting because it suggests that older readers may use different reading strategies than young readers. They found that older readers allocated more resources to building a situation model in the first pass over a text than young readers. The authors interpret this as showing that older readers are focusing more on the "holistic structures of discourse." While this study used only reading times, it seems likely that the different processing strategy employed by older readers should also be reflected in scanpath patterns.

In related work, a corpus analysis showed that older readers have longer fixation durations, larger word-frequency effects in duration measures, and higher re-inspection probabilities (Kliegl, Grabner, Rolfs, & Engbert, 2004; see also Wotschack & Kliegl, 2013). Moreover, older readers respond differently to differences in the predictability of words: Young readers tend to skip highly predictable words more often, whereas older readers show a decreased rate of refixations compared to less predictable words. Similar results were reported by Rayner, Reichle, Stroud, Williams, and Pollatsek (2006), who also show a generally higher skipping probability in older readers.

Solan, Feldman, and Tujak (1995) showed that training can improve the reading efficiency of older readers, which in their study was expressed in lower rates of regressions, fewer fixations per word, higher rate of reading (words per minute), and increased span of recognition. Effectively, the training made their eye movements resemble those of young readers.

The reasons for differences between old and young readers are a matter of debate. Two groups investigated this question using model simulations. Based on simulations using the SWIFT model of reading (Engbert, Nuthmann, Richter, & Kliegl, 2005), Laubrock et al. (2006) argued that visual acuity and lexical processing are two factors determining age differences in reading. Rayner et al. (2006) conducted similar tests using the E-Z Reader model (Reichle, Pollatsek, Fisher, & Rayner, 1998) and concluded that older readers use a more "risky" reading strategy under which words are sometimes guessed in advance and therefore skipped. This strategy could compensate for an age-related slowdown in general processing but would sometimes lead to disruptions when guesses are wrong. Incorrect guesses would then result in regressive eye movements back to the skipped words. Another explanation for age differences in reading was proposed by Wotschack and Kliegl (2013), who attributed the increased skipping rate in older readers to lapses of attention ("mindless reading," Reichle, Reineberg, & Schooler, 2010; Schad & Engbert, 2012).

In sum, these studies indirectly suggest that young readers produce more regular left-to-right directed scanpaths, while old readers exhibit more irregular gaze trajectories. These differences were captured in skipping, refixation, and regression probabilities. However, it is possible that these measures furnish an incomplete picture of the underlying scanpath phenomena. To our knowledge, a direct analysis of scanpath differences between old and young readers has so far not been conducted.

Apart from sentence processing difficulty and age, a third factor that is known to determine the regularity of scanpaths in reading is the length of words. In the literature on

oculo-motor control in reading, it is well known that short words are skipped considerably more often than long words (Brysbaert & Vitu, 1998; Drieghe, Rayner, & Pollatsek, 2005; Kliegl et al., 2004). This effect is usually explained in terms of parafoveal processing: Readers can start processing of the next word even before it is fixated. If the processing of the next word can be finished quickly, it may be unnecessary to fixate it and it can be skipped. Short function words, for instance, are skipped in roughly two-thirds of the cases. A frequent consequence of skipping is a subsequent regressive saccade (Engbert et al., 2005; Vitu & McConkie, 2000). Because these regressions occurred more often when the skipped word was low frequency, Vitu and McConkie argued that they are a consequence of premature skipping. The function of this type of regression would then be to fixate the skipped word for closer examination. This explanation is of course not easily reconciled with the assumption that words can only be skipped when they were successfully identified. This dilemma can be resolved by assuming that the decision to skip a word is made before it is fully processed but when the probability is reasonably high that it can be sufficiently processed before the eyes start the saccade to the next word. Such a mechanism would lead to some moderate amount of premature skippings but also to good overall performance. Two models that assume such a trade-off between speed and accuracy are the SWIFT model of oculo-motor control in reading (Engbert et al., 2005) and the Bayesian model by Bicknell and Levy (2010). Whatever the underlying mechanism might be, sentences with many short words should elicit more irregular scanpath patterns than sentences with long words—a prediction that we can test using the scanpath approach by von der Malsburg and Vasishth (2011).

All in all, the effects of syntactic reanalysis, age, and word length demonstrate that the gaze does not monotonously jump from one word to its successor; it rather follows complex trajectories that reflect various aspects of the underlying perceptual, cognitive, and motor processes. However, a more thorough investigation of these scanpath phenomena is impeded by the relative difficulty of analyzing fixation sequences discussed above. What is needed is a method that allows us to attack scanpath phenomena like those discussed by Frazier and Rayner (1982) in a more direct way than through simplifying measures like fixation durations and regression probabilities.

2. Analyzing scanpaths in reading

One attempt at a more scanpath-oriented method has been made by Salvucci and Anderson (2001). Their goal was to evaluate scanpath predictions of models of cognitive processing such as the E-Z Reader model of eye movement control in reading. Reichle et al. (1998) compared various versions of that model using a set of classical eyetracking measures like gaze duration (the duration from entering a word for the first time until leaving it) and skipping probability. While this technique successfully ruled out some versions of the model, it could not distinguish between the two best-performing models, version 3 and 5. One difference between these models relates to assumptions about how words in the periphery of the visual field are processed and is mainly expressed in

scanpath patterns. Version 3 and 5 of E-Z Reader could not be distinguished because the word-based measures used by Reichle et al. (1998) did not capture these scanpath phenomena well enough. To address this problem, Salvucci and Anderson developed a new approach for comparing the predictions of models of eye movement control.

The idea behind this approach is as follows: First, run a model (e.g., a version of the E-Z Reader model) repeatedly and collect a database of eye movement predictions. Next, calculate two sets of probabilities for this database: (i) probabilities about where the eyes should be during a particular state of the model; and (ii) transition probabilities between these states of the system. Using these probabilities, a hidden Markov model can be constructed that can answer the following question: What is the probability of seeing a set of empirically observed eye movements assuming that the hidden Markov model is an accurate description of the underlying cognitive process? More technically, the hidden Markov model allows us to calculate the likelihood of a model of eye movements, given a particular data set. If we have such a hidden Markov model for each of two versions of E-Z Reader, we can calculate likelihoods for both models and identify the better model by its higher likelihood. The benefit of this approach is that it does not evaluate fixations in isolation, as analyses of eyetracking measures usually do. Instead, it evaluates fixations conditional on the preceding fixations, which introduces some sensitivity to scanpath phenomena. For instance, such a hidden Markov model can express the fact that a certain word is likely to be skipped but only when the previous word has not been skipped.

Using this approach, Salvucci and Anderson (2001) could show that version 5 of E-Z Reader produces more accurate predictions with respect to scanpath patterns than version 3, which is an encouraging result because Reichle et al. (1998) favored version 5 for its higher psychological plausibility. However, there are restrictions to this method: The first is that fixation durations are not taken into account. In the comparison of the two versions of E-Z Reader, this was not an issue because both models made very similar predictions with respect to fixation durations. In general, however, it is desirable to have a method that accounts for differences in spatial *and* temporal patterns. After all, a large share of evidence in reading research comes from duration measures. Also, spatial divergences between scanpaths should have more weight if they last longer. Another restriction is that the hidden Markov approach proposed by Salvucci and Anderson cannot be applied when a detailed process model such as E-Z Reader is not available, or when an exploration of scanpath phenomena in a data set is desired as was the case in the study by Frazier and Rayner (1982).

A method that is suitable for the latter goal is the analysis of Markov matrices. These matrices describe the transition probabilities of the gaze between the regions of the visual stimulus (Hacisalihzade, Stark, & Allen, 1992). If there are 10 regions in the stimulus, the matrix contains 10^2 such transition probabilities. While this yields a mathematically tractable representation of scanpath patterns, it has a serious drawback: The transition probabilities depend only on the current state, the region currently fixated on. The earlier history of fixations in a scanpath is ignored. In principle this can be addressed by constructing higher-order Markov matrices taking into account previous fixations, but this leads to an explosion of the number of probabilities that have to be estimated and would require huge amounts of data. This restriction was the reason why Meseguer et al. (2002)

could not draw definite conclusions about the scanpath trajectories that go along with syntactic reanalysis.

Hacisalihzade et al. (1992) also propose the analysis of scanpath patterns using edit-distance measures such as the Levenshtein metric (Levenshtein, 1966). These measures quantify the similarity of two symbol sequences as the number of edit-operations (typically insertion, deletion, and substitution of a symbol) that are needed to transform one sequence into the other. This metric has been used in diverse fields, for example as a mathematical device in information theory, for the analysis of nucleotide sequences in bioinformatics, and for the automatic correction of typing errors. This metric can also be used for analyzing scanpaths because they can, with some loss of information, be represented as sequences of symbols: When each region of interest is labeled with a symbol (say a letter), we can represent a scanpath as a sequence of letters. The n-th letter in that sequence would indicate the region in which the n-th fixation occurred. The amount of letters that have to be changed in one scanpath in order to transform it into another can then be used as a measure of the dissimilarity of the two scanpaths. Similarity scores for pairs of scanpaths obtained with such a measure can be analyzed directly (Brandt & Stark, 1997; Feusner & Lukoff, 2008) or can serve as the basis for cluster analyses (Josephson & Holmes, 2002). Salvucci and Anderson (2001) also used edit-distances as an alternative method to determine which model makes the best predictions, given the recorded scanpaths: They calculated the similarities of the scanpaths predicted by the various models to the observed scanpaths and identified the model making the most similar predictions.

A potential problem with these methods is the division of the stimulus into regions of interest. The results of the analysis depend to some degree on the definition of these regions but, as Hacisalihzade et al. (1992) point out, it is not clear how to define them and there is often no uniquely correct solution for this problem. Moreover, these approaches do not account for fixation durations because the symbolic representation of scanpaths is stripped of this information. A final problem is that the Levenshtein metric evaluates spatial differences between fixations only in the most simplest way: Either two compared fixations occurred on the same region or on a different region. How far away they were does not matter. In reading, reasonable regions of interest are words or phrases. If words are used, the Levenshtein distance treats the small distance between the last character of a word and the first of the next word the same as any other distance between words, no matter how large. If phrases are used as regions of interest, we lose information about where in that region a fixation occurred because all fixations within these regions are evaluated as being the same.

For these reasons, it is desirable to have a similarity measure for scanpaths with two properties: First, it should treat space in a continuous fashion without requiring discrete regions of interest. Second, the measure should have fine-grained sensitivity to temporal differences. See also Mathôt, Cristino, Gilchrist, and Theeuwes (2012) for a discussion of desirable properties for scanpath measures.

Recently, a range of interesting new scanpath measures have been proposed in the area of scene perception research (Coco & Keller, 2012; Cristino, Mathôt, Theeuwes, & Gilchrist, 2010; Jarodzka, Hohnqvist, & Nyström, 2010; Mathôt et al., 2012). However,

these measures have been developed with scene perception data in mind and their suitability for reading data has so far not been investigated. A brief discussion of the measure by Cristino et al. can be found in von der Malsburg and Vasishth (2011).

In sum, while there have been various interesting methodological proposals for the analysis of scanpaths in reading, it seems that none of the available approaches is fully suitable. We see the main problems as being a too coarse-grained treatment of spatial information and no sensitivity for temporal information.

Recently, von der Malsburg and Vasishth (2007, 2011) presented a new similarity measure for scanpaths that addresses the shortcomings of earlier approaches. This similarity measure is based on the same logic as the Levenshtein metric (see illustration in Fig. 1): Two scanpaths are similar if only few modifications are necessary to transform one into the other. However, where the Levenshtein metric counts all edit-operations (deletions, insertions, and substitutions) equally, our measure uses a function that weights these operations depending on spatial and temporal properties of the fixations involved in an edit-operation. If a fixation is short, deleting it or inserting it leads to a smaller dissimilarity than when it is long. Specifically, the dissimilarity contributed by that fixation is simply its duration measured, for example, in milliseconds. If one fixation needs to be replaced by another, the dissimilarity depends on the duration of the two fixations and on their spatial distance. If they have the same location, the dissimilarity is the difference of their fixation durations because there is no other difference between them. If the two fixations are extremely far away from each other, the dissimilarity score is the sum of their fixation durations. The rationale is this: If the two fixations are very short, the overall dissimilarity that is added by them is little. If they are both long, this means that the part where the two scanpaths diverge is long and the dissimilarity should then be larger. In other words, when two fixations are far apart, the difference of their fixation durations is not decisive; what counts is their total duration. When there is a medium distance between two fixations, the dissimilarity score of the substitution is given by a weighted sum of the difference and the sum of the two fixation durations. The smaller the distance between the fixations, the more the result is determined by the difference in durations. The larger the distance, the stronger the impact of the sum of the durations. The transition from one extreme case to the other is determined by a smooth function that mimics the exponential drop in acuity of the human visual system when moving away from the fovea toward the periphery of the visual field: Slightly changing the distance of two fixations that are close to each other has a strong impact on the weighting, whereas making equally small changes to the distance of fixations that are far apart has little effect. Deletion and insertion can be handled as special cases of substitution, namely as a substitution with (or of) a fixation with zero duration.[1] The overall dissimilarity of two scanpaths is then calculated by matching pairs of fixations (this is done using the Needleman–Wunsch algorithm; Needleman & Wunsch, 1970), calculating dissimilarity scores for these pairs, and by summing the dissimilarity scores of all pairs. While this may sound fairly abstract, there is a simple and concrete intuition explaining what this measure is doing: given two scanpaths, the measure essentially quantifies how much time was spent looking at different things. See von der Malsburg and Vasishth (2011) for a precise specification

and extensive discussion of this measure. Some of the key properties of our measure are as follows: (a) it has fine-grained sensitivity to temporal and spatial differences (the measure is completely continuous); (b) arbitrary regions of interest are not needed, because the measure operates on the coordinates of fixations; and (c) dissimilarity scores can be computed efficiently.[2] Implementations of this scanpath measure for the GNU-R system and for the Python programming language are freely available from the first author.

The goal of the study by von der Malsburg and Vasishth (2011) was to identify the scanpath correlates of syntactic reanalysls—the signature scanpath that Meseguer et al. (2002) tried to identify. To this end, von der Malsburg and Vasishth reexamined the data collected by Meseguer and colleagues using the scanpath measure described above. A cluster analysis based on that measure identified three distinct categories of scanpaths that occurred after the critical word in the sentence was read. One pattern suggested that rereading was a common strategy to recover from misanalyses and therefore supported the forward reanalysis hypothesis. This reading pattern had gone unnoticed in earlier studies that examined only transition probabilities and word-based eyetracking measures. In a follow-up study, von der Malsburg and Vasishth (2013) reproduced their earlier findings with modified stimuli, an improved procedure, and a diverse subject population, thus showing the stability of the observed scanpath phenomena. This second experiment also

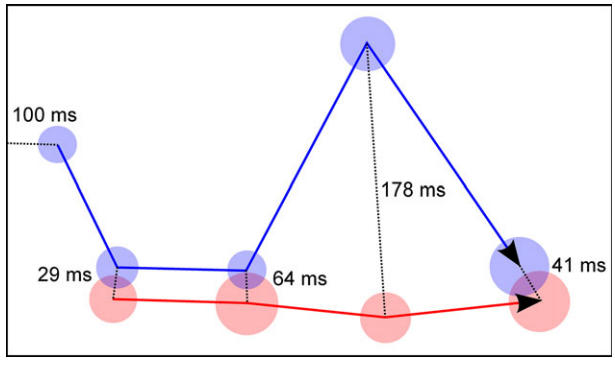

Fig. 1. An illustration of our measure for the similarity of two scanpaths. The red scanpath at the bottom has four fixations and the blue scanpath at the top five. Fixations are represented by circles and the size of the circles indicates the duration of the fixations. The Needleman–Wunsch algorithm is used to find the optimal alignment of the fixations in the two scanpaths (the dotted lines in the graph indicate alignment). The algorithm attempts a one-to-one alignment but leaves superfluous fixations unaligned. In order to calculate the goodness of an alignment, a score is used that quantifies the similarity of the aligned fixations. The alignment is optimal if the average similarity of the aligned fixations is maximal. If two aligned fixations are close to each other, this difference score mainly reflects the difference of their fixation durations (see the first, second, and forth pair of fixations from the left). If two aligned fixations are far away from each other, the difference score will be determined mainly by the sum of their fixation durations: A large spatial distance between two fixations should yield a bigger difference if it lasts long (see third pair from the left). If a fixation has no counterpart in the other scanpath, the difference score for that fixation is given by its fixation duration (see the first fixation in the blue scanpath). This difference score is the same as if the fixation was aligned with another fixation of duration 0. The overall dissimilarity of the two scanpaths is simply the sum of all difference scores (100 ms + 29 ms + 64 ms + 178 ms + 41 ms = 412 ms).

showed that the scanpath patterns are moderated by individual differences in working memory capacity: Some patterns occurred more often in individuals with a high working memory capacity than in readers with a low capacity.

This sensitivity of the scanpath measure to effects of syntactic reanalysis as well as to differences in working memory capacity suggests that it may also be effective when studying the effect of more general syntactic processing difficulty and effects of age in reading patterns. In the present study, we will therefore use the measure to investigate the scanpath phenomena associated with differences in age, word length, and syntactic processing difficulty. One goal in doing so is to determine whether the measure can reveal scanpath effects due to factors other than syntactic reanalysis. If scanpath effects can be picked up in relatively difficult garden-path sentences as those analyzed in von der Malsburg and Vasishth (2011, 2013), this may not be too surprising because these sentences were designed to cause strong effects. If, however, scanpath effects can also be found in a corpus of relatively innocuous, that is not psycholinguistically contrived sentences, this would suggest a much wider applicability of our scanpath measure in eye movement research.

3. Investigating effects of syntactic structure, age, and word length in scanpaths

The data set that we analyzed was the Potsdam Sentence Corpus, a database of eye movements recorded for 230 participants who read a set of 144 German sentences each (Kliegl et al., 2004). The sentence material consists of short sentences (ranging from 5 to 11 words; mean: 7.9 words) that do not involve any specific psycholinguistic manipulation but represent a broad spectrum of linguistic phenomena. These sentences were specifically constructed for the corpus and were presented in isolation during data collection. This means that scanpaths for these sentences are likely simpler and less variable than those observed in longer texts such as newspaper articles (e.g., in the Dundee Corpus; Kennedy & Pynte, 2005).

As discussed above, the literature on effects of age, word length, and syntactic processing difficulty suggests that the relevant dimension along which scanpaths differ may be the irregularity of the gaze trajectory. While young readers move their eyes relatively monotonically from left to right, skipping and regression probabilities hint toward more unpredictable gaze trajectories in older readers. Simple sentences are typically read from left to right, but more difficult material may elicit regressions and therefore more irregular scanpaths. Thus, the main issue to be solved before we can start is the selection of a suitable dependent variable for the statistical analysis that captures these differences in scanpath regularity.

Mitchell et al. (2008) showed that when readers encounter the disambiguating material in a garden-path sentence, they sometimes regress back not just to a specific word; instead, any of the preceding words can be the target of regressions. In other words, the scanpath pattern is much less predictable than in non-garden-path sentences where the eyes presumably move mostly to the next word on the right. In the Potsdam Sentence

Corpus, we find that 50% of the saccades target the next word in a sentence; in 19% of the saccades, the next word is skipped; 17% of the saccades result in refixations of the current word; and 8% are regressive saccades landing on the word directly preceding the current word. Other saccade targets are rare. A consequence of such a high predictability of saccade targets is that scanpaths in the Potsdam Sentence Corpus should be relatively similar to each other. A low predictability of saccade targets, on the other hand, is expected to lead to more varied fixation patterns that should be relatively less similar to each other. This association between predictability of saccade targets and scanpath similarity (highly predictable targets result in similar patterns) allows us to link scanpath similarity, which is a property of pairs of scanpaths, to scanpath regularity, a property of individual scanpaths.

In order to make use of this link, we will leverage the notion of maps of scanpaths that we introduced in von der Malsburg and Vasishth (2011, 2013). On these maps, every scanpath is represented as a point. The mutual distances between these points reflect the similarities of the corresponding scanpaths according to our scanpath measure. This means that on such maps similar scanpaths are located close to each other, and that groups of highly similar scanpaths emerge as dense clusters of points. Hence, the density on the map at a particular point can be used as a measure of the regularity of the scanpaths located at that point.

The above-mentioned statistics about saccade targets in the Potsdam Sentence Corpus show that the next word is the most likely saccade target. Hence, trajectories with only a few deviations from a strictly left-to-right movement should be the most common patterns and should densely populate a relatively small area on a scanpath map. The more a gaze trajectory deviates from this default pattern, for example, by having unusually many regressions and instances of skipping, the further away from the dense center this trajectory will be located on the map. Thus, for the present purpose, we equate scanpath density with scanpath regularity. The prediction for age differences is then that older readers produce more irregular scanpaths, while young readers produce more regular scanpaths. Similarly, sentences with a low average word length are predicted to elicit more irregular scanpaths than sentences with longer words.

Deriving predictions for the effect of syntactic processing difficulty on scanpaths is a bit more involved. Unlike in an experimental setting, where one condition is designed to be more difficult to process than the other, we have a set of 144 sentences and no obvious independent criterion for separating them into easy and difficult. Therefore, we need a principled measure for the expected difficulty of processing the syntax of these sentences. Two measures that are established in the psycholinguistic literature are the surprisal and the retrieval cost of a word. For both measures, earlier studies have shown that they are reliable predictors of sentence processing difficulty in both corpus analyses of the Potsdam Sentence Corpus (Boston, Hale, Kllegl, Patil, & Vasishth, 2008; Boston, Hale, Vasishth, & Kliegl, 2011), and in the modeling of planned experiments (Lewis & Vasishth, 2005). In independent work, Demberg and Keller (2008) have also demonstrated the efficacy of surprisal in predicting reading times in the Dundee Corpus of English (Kennedy & Pynte, 2005). Since both metrics, surprisal and retrieval cost, have been

shown to make independent contributions to the explanation of sentence processing difficulty (Boston et al., 2011; Vasishth & Drenhaus, 2011), we used both as predictors for scanpath irregularity.[3] Specifically, we quantified the processing difficulty of a sentence as the average surprisal and average retrieval cost of the words in that sentence. In the following paragraphs, we briefly introduce the theoretical background of surprisal and retrieval and outline how they are computed.

Surprisal is an information-theoretic measure of sentence processing difficulty (Hale, 2001; Levy, 2008). It posits that continuations of a sentence that are less likely to occur are harder to process, which should be reflected in increased reading times on those continuations. This notion is mathematically captured by the probability of seeing a word (or one of its properties, e.g., its part-of-speech category) given the previous words in the sentence. The surprisal of a word is then the negative logarithm of that probability. In principle, surprisal need not be a purely syntactic measure because other cues like discourse information can also shape our expectation of what the next word in a sentence might be. In practice, however, surprisal calculations are often restricted to the syntactic level simply because more comprehensive models taking into account other types of information are currently technically not feasible. The identity of the next word in the sentence *Peter gave the...* depends on many aspects and deep discourse understanding might be required to guess it. However, it is relatively easy to computationally determine how likely it is that the next word is going to be a noun.

A common vehicle for calculating surprisal scores have been parsers for probabilistic context-free grammars (two other possibilities are n-gram models, Smith & Levy, 2008; and simple recurrent neural networks, Frank, 2009). Given an incomplete sentence string (such as *Peter gave the...*), such a parser generates all possible syntactic continuations of a sentence at that point and assigns to each continuation a probability that is derived from the probabilities of the individual grammar rules that have been used. This has been interpreted to imply that the human language system performs an exhaustive analysis of all possible interpretations of a sentence fragment. Boston et al. (2011) argue that this assumption is not plausible because such a parser would potentially consume enormous processing resources. They propose a variant of surprisal that is calculated by a parser that explores only a limited number of high-probability syntactic interpretations at each word. Boston et al. derived these candidate interpretations using an incremental probabilistic parser for dependency grammars. A parameter in that parser controls how many interpretations are pursued. A comparison of this flavor of surprisal with the original version by Hale (2001) has not been reported, but Boston et al. showed that their metric is a reliable predictor for a set of standard eyetracking measures in the Potsdam Sentence Corpus.

One central question asked by Boston et al. (2011) was how the degree of parallelism in their parser, that is, the number of interpretations pursued, affects the predictive power of the resulting surprisal scores. Their analysis found that surprisal scores were good predictors for eyetracking measures in the Potsdam Sentence Corpus when a parser was used that maintained only one possible interpretation at each time. However, when they used a parser that developed up to one hundred interpretations in parallel, the effect of surprisal

was considerably stronger. Since the limit on the number of interpretations is hypothesized by Boston et al. to reflect limitations in processing resources, it is likely that this limit is subject to individual variation, differences in processing load (e.g., induced by secondary tasks), and that it changes as language comprehenders age (Dobbs & Rule, 1989; Waters & Caplan, 2005). Age-related decline in cognitive processing might, for example, go along with a reduction in the maximal number of interpretations that the language system can handle at once. In any case, the variable degree of parallelism opens many opportunities for empirical evaluation and thus makes the Boston et al. type of surprisal particularly interesting. For this reason, we used their variant of surprisal for the present investigation. Another benefit was that these surprisal scores were already available for the Potsdam Sentence Corpus.

While surprisal relates to processes concerned with upcoming material, the other metric of sentence processing difficulty considered here, retrieval cost, relates to material that has already been processed. The integration of a new word into the sentence context involves establishing its relations to the previous material and this process requires that material be retrieved from memory. The cue-based parsing theory (Lewis & Vasishth, 2005) makes two crucial assumptions about how items are stored in and retrieved from memory:[4] First, the activation of an item in short-term memory is a function of usage—the further back in time an item is accessed, the harder is its retrieval—and frequent retrievals boost its activation, making subsequent retrievals easier. Second, if other items are present in memory that share features with an item that is targeted for retrieval, these distractors cause interference, which slows down the retrieval process and increases the likelihood of misretrievals (similarity-based interference). Integrating the second assumption into their dependency parser allowed Boston et al. (2011) to generate predictions for each word in the corpus about how much time is needed to retrieve earlier material needed for processing that word. The evaluation showed that retrieval scores generated with a serial parser (which maintains only one interpretation at each time) were not predictive of eyetracking measures. However, when the parser was allowed to entertain several interpretations in parallel, retrieval was a reliable predictor. Refer to Boston et al. (2011) for a detailed description of their parsing system and the surprisal and retrieval calculations.

Having laid out the theoretical framework that allows us to quantify the expected processing difficulty for the sentence material studied here, we turn to the predictions that our analysis will evaluate.

Our predictions for scanpath regularity as a function of sentence difficulty are as follows: Sentences with high average surprisal should elicit more irregular scanpath patterns. The same should hold for sentences with high average retrieval costs. Any age differences found in scanpaths can be due to adaptations of reading strategies to refined language and world knowledge, or to degraded processing resources in older readers, or both. Whatever the driving force of these changes might be, these shifts in strategies may also be expressed in the effects of surprisal and retrieval difficulty. If, for example, the parser used by Boston et al. (2011) to calculate surprisal and retrieval costs is a better model for language processing strategies in young readers, the effects of surprisal and

retrieval should be weaker in old readers; in other words, there may be an interaction of age with surprisal and with retrieval.

The Boston et al. (2011) variant of surprisal allows us to explore these ideas even further. Boston et al. showed that a parser with a higher degree of parallelism produces surprisal and retrieval scores that are better predictors of eyetracking measures than surprisal and retrieval scores computed using a serial parser. If we assume that older readers have less cognitive resources available than younger readers (Dobbs & Rule, 1989), then scanpath regularity of older readers should be better modeled by surprisal and retrieval scores calculated by a parser with a lower degree of parallelism. More technically, the prediction is that when we use surprisal and retrieval scores from a parser with a lower degree of parallelism, there should be interactions of age with surprisal and with retrieval. An alternative explanation for such interactions would be that older readers can rule out more candidate structures early on in the sentence using their more elaborate language and world knowledge. In other words, their experience allows them to preserve resources. In any case, the presence of such interactions would suggest that the degree of parallelism is a variable that can explain age-related differences in sentence processing.

4. Method and materials

The sentences in the Potsdam Sentences Corpus have lengths ranging from 5 to 11 words (mean: 7.9 words). The readers come from varied socioeconomic backgrounds and include teenagers, university students, and pensioners. The sentences were presented individually on a single line on a 21″ computer screen. After 27% of the sentences, participants had to answer easy multiple-choice comprehension questions. See Kliegl et al. (2004) for more details.

Surprisal and retrieval scores for the sentences in the Potsdam Sentences Corpus were provided to us by Marisa Ferrara Boston and colleagues. Both scores quantify the difficulty of a word in its sentential context. For this analysis, however, we need measures for the difficulty of whole sentences. Therefore, we calculated for each sentence the average surprisal and the average retrieval difficulty of the words in the sentence. We used the scores that were calculated with a parser with beamwidth 100 (i.e., a parser that entertained up to 100 interpretations concurrently) because these scores were the best predictors for eye movements in the analysis by Boston et al.

Calculation of the similarity scores of the scanpaths contained in the Potsdam Sentence Corpus was done using a software package for the R system (available from the first author). Maps of scanpaths were fit using multidimensional scaling (Kruskal, 1964) as implemented in the function isoMDS in the R package MASS (Venables & Ripley, 2002). For calculating density scores we used the package Mclust, which provides functions for fitting mixture of Gaussians models (Fraley & Raftery, 2002, 2007). Linear mixed models were fit with the function lmer from the package lme4 (Bates, 2005).

5. Results

5.1. Fitting maps of scanpaths

For each sentence, we calculated the pair-wise similarities of all scanpaths recorded for that sentence. To factor out trivial effects of sentence length, the similarity of a pair of scanpaths was divided by the total number of fixations in that pair; this yields a score quantifying the similarity per fixation. Next, we calculated a map of scanpaths for every sentence. The number of dimensions of those maps was set to 4, which lead to a reasonably faithful preservation of similarity scores. The maps had an average stress of 14.03% (*SD*: 1.22). Stress quantifies the percentage of the overall variance that could not be explained by the a map. The smaller the stress, the more accurate is the map representation of the scanpath variance.

To get a sense of the range of scanpath phenomena in the eyetracking corpus, we examined a simple statistic of the maps of scanpaths: the spread of the points representing the individual scanpaths. We calculated the spread for each map as the average similarity of the scanpaths on that map. This gives us a measure of the variety of scanpath phenomena that were recorded for a sentence. Fig. 2 shows the scanpaths for the sentence whose map had the lowest spread (3) and Fig. 3 those for the sentence with the highest spread (4). The maps of scanpaths for these two sentences can be seen in Fig. 4.

(3)	*Wolfgangs*	*Töchter*	*studieren*	*Literatur*	*und*	*Maschinenbau.*
	Wolfgang's	daughters	study	literature	and	engineering.

(4)	*Den*	*Ton*	*gab*	*der*	*Künstler*	*seinem*	*Gehilfen.*
	The	clay	gave	the	artist	to his	apprentice.

"The artist gave the clay to his apprentice."

The sentence in (3) has canonical word order (subject, verb, object) and consists of relatively long words. Sentence (4), on the other hand, has noncanonical word order (object, verb, subject), relatively short words, and a lexical ambiguity: The word *Ton* can mean sound (common) or clay (less common); here the correct meaning is the less common clay. Because of these properties, sentence (4) is expected to be harder to process, and the fact that it elicited more diverse scanpaths is a first hint that our scanpath measure may be sensitive to the phenomena targeted in this study.

5.2. Calculating scanpath regularity

The dependent variable that we used in the following analyses is the regularity of a scanpath. As discussed above, we operationalize the regularity of a scanpath as the amount of similar scanpaths. This can in turn be quantified as the density on the map of scanpaths at the location of the scanpath in question. We calculated this density by fitting

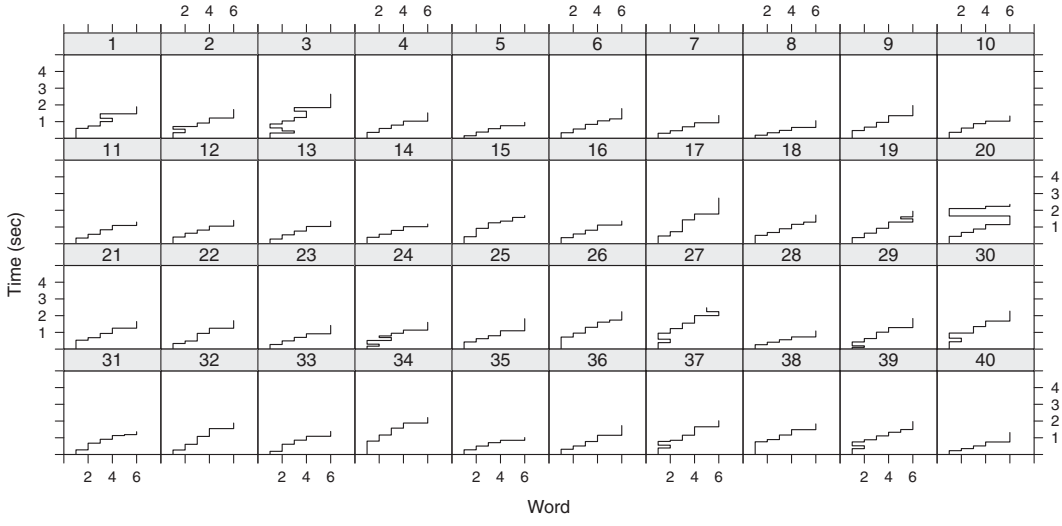

Fig. 2. A random sample of the scanpaths recorded for the sentence that elicited the least diverse eye movement patterns (minimal average mutual similarity): "*Wolfgangs Töchter studieren Literatur und Maschinenbau*" ("Wolfgang's daughters study literature and engineering").

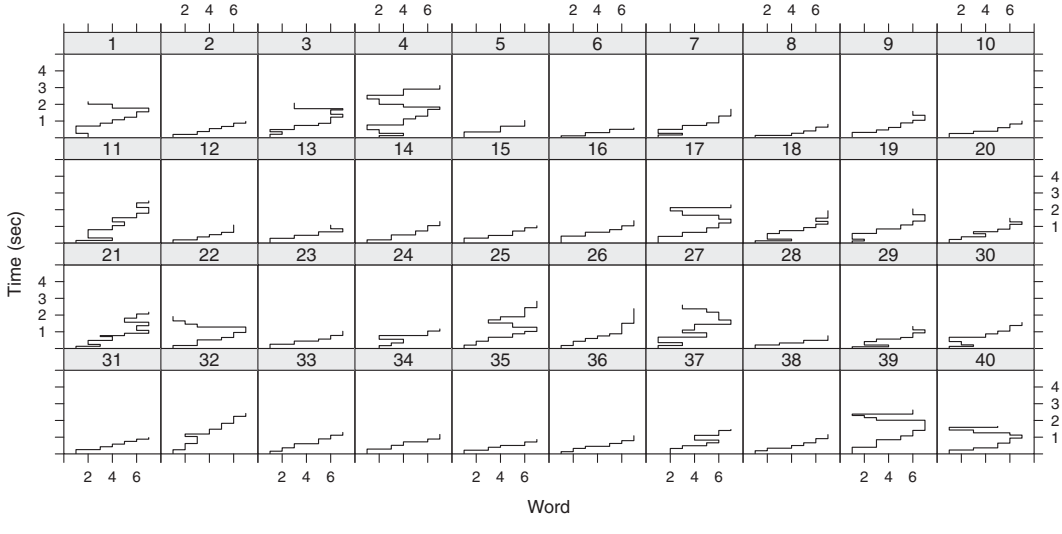

Fig. 3. A random sample of the scanpaths recorded for the sentence that elicited the most diverse eye movement patterns: "*Den Ton gab der Künstler seinem Gehilfen*" ("The artist gave the clay to his apprentice"). This sentence has noncanonical word order and a lexical ambiguity (*Ton* can mean sound or clay).

mixture of Gaussian models for each map (Fraley & Raftery, 2002). A mixture model explains the potentially complex distribution of data points—in this case scanpaths on a map—as the sum of several multivariate Gaussians. These Gaussians differ in location, shape, and rotation. The parameters of a mixture model with a fixed number of Gaussians

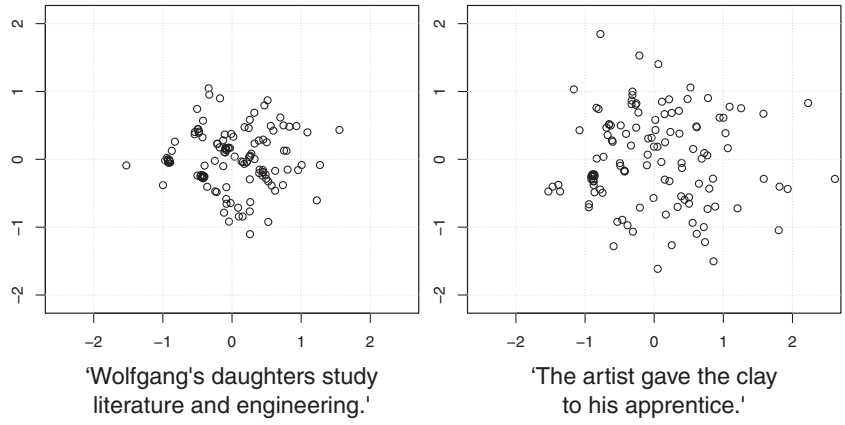

Fig. 4. Two maps of scanpaths: left, the map for the sentence that elicited the smallest scanpath variance ("*Wolfgangs Töchter studieren Literatur und Maschinenbau*"). On the right, the map for the sentence with the most varied scanpath patterns ("*Den Ton gab der Künstler seinem Gehilfen*"). Each point represents a scanpath recorded for the respective sentence. If two scanpaths are located close to each other on a map, that is because they are similar according to our measure; if they are far apart, they are dissimilar. The first sentence has canonical word order, while the second has noncanonical word order and a lexical ambiguity that can lead to garden-pathing ("*Ton*" can mean sound or clay). See Figs. 2 and 3 for plots showing a sample of the recorded scanpath patterns for these two sentences.

can be determined using entropy maximization. The optimal number of Gaussians was identified by fitting 20 mixture models consisting of 1 to 20 Gaussians. A Bayesian information criterion was used to select the model that described the distribution on the map optimally (Schwarz, 1978). This criterion favors models that explain the data well but punishes models that need many degrees of freedom to do so, that is, models consisting of many Gaussians. Using the best fitting model, it is possible to calculate the scanpath density at every position on the map. This density at the location of a scanpath was used as the scanpath's measure of regularity.

5.3. Scanpath effects of syntactic structures, age, and word length

There is a potential issue when using map density as a proxy variable for scanpath regularity: If a data set contains more young readers than old readers, those areas on scanpath maps that contain the scanpaths of old readers will trivially have lower density. It is therefore important to analyze a data set that contains a similar number of young and old readers. In the Potsdam Sentence Corpus, the age of the 230 readers ranges from 16 to 84 years (median: 24). There was one large group of readers around the age 22 years and another smaller group around 68 years. To balance the number of young and old readers, we compiled a data set consisting of the first 72 young readers and all 72 old readers, resulting in 144 readers overall.

We tested the predictions about the influence of age, word length, surprisal, and retrieval cost on scanpath regularity by fitting a linear mixed model (Bates, 2005). This model

had the logarithm of scanpath density as the dependent variable. The log-transform was applied because it yielded a distribution of residuals that was much closer to a normal distribution than the distribution of residuals we obtained when raw density scores were used. The model had age, word length, surprisal, and retrieval scores as fixed effects. Age was coded as a continuous variable. Additionally there were terms for the interaction of age with surprisal and age with retrieval. All fixed effects were centered at zero and scaled to have a standard deviation of one. The model also had random intercepts for sentences and for readers. Adding random slopes for age, word length, surprisal, and retrieval cost did not qualitatively change the results.

Table 1 lists all parameter estimates for this model. Effects larger than twice their standard error were interpreted as significant and are marked with an asterisk in the table. Fig. 5 shows a plot of the coefficients with 95% highest posterior density intervals (10,000 MCMC runs using the R function mcmcsamp). There was an effect of age in the predicted direction: Older readers produced more irregular eye movement trajectories than young readers ($\hat{\beta} = -0.41$, $\hat{SE} = 0.07$, $t = -5.71$). Sentences with high average surprisal elicited more irregular scanpath patterns ($\hat{\beta} = -0.3$, $\hat{SE} = 0.08$, $t = -3.71$). Sentences

Table 1
Parameter estimates for the linear mixed model

	Coef.	SE	*t*
Age	−0.41	0.07	−5.71*
Surprisal	−0.30	0.08	−3.71*
Retrieval	−0.12	0.07	−1.68
Word length	0.40	0.07	5.47*
Age × surprisal	0.06	0.02	2.95*
Age × retrieval	0.06	0.02	2.97*

Note. (stars indicate significance)

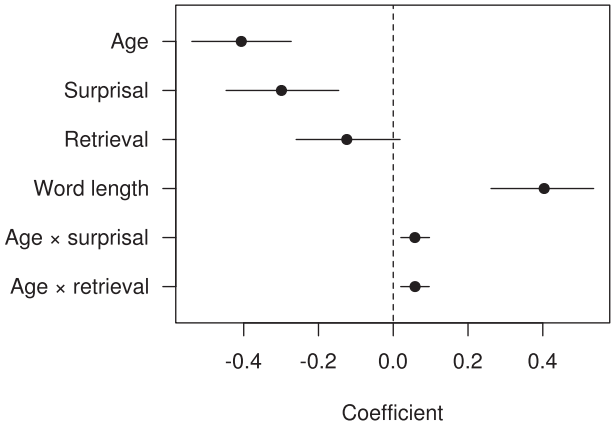

Fig. 5. Coefficients of the linear mixed model that models the irregularity of scanpaths. Error bars show 95% highest posterior density intervals (10,000 MCMC runs).

with high average retrieval cost also elicited more irregular scanpaths, but this effect was only marginally significant ($\hat{\beta} = -0.12$, $\hat{SE} = 0.07$, $t = -1.68$). Both surprisal and retrieval interacted with age ($\hat{\beta} = 0.06$, $\hat{SE} = 0.02$, $t = 2.96$; $\hat{\beta} = 0.06$, $\hat{SE} = 0.02$, $t = 2.97$): The older a reader, the smaller the effects of surprisal and retrieval. Sentences with longer words elicited more regular scanpaths ($\hat{\beta} = 0.4$, $\hat{SE} = 0.07$, $t = 5.47$).

To make sure that these results do not only hold for the particular subset of young readers, we repeated the analysis 1,000 times with different random samples of young readers. The coefficient of the retrieval effect changed sign in 2.1% of the repetitions and the coefficient of the interaction of age with surprisal changed sign in 2.7% of the cases. All other coefficients (age, surprisal, word length, age × retrieval) had the same sign in all repetitions.[5] These results show that the observed effects are very stable across various samples of young readers.

During the analysis, we noticed that the surprisal scores of words are highly correlated with the position of the word in the sentence. The further a word is in the sentence, the higher its surprisal according to the definition used by Boston et al. (2011).[6] Consequently, the average surprisal of a sentence is correlated with the length of the sentence ($\rho = 0.76$). This potentially poses a problem for the interpretation of surprisal effects because these effects could be trivial effects of sentence length. When we calculated similarity scores for pairs of scanpaths, we divided them by the total number of fixations in the pair. This excludes some trivial effects of sentence length but perhaps not all. To make sure that the effects of surprisal are in fact related to syntactic expectation, we re-fitted the linear mixed model reported above, but this time using data only from sentences with seven words, the most common sentence length in the corpus. If there is still an effect of surprisal, it cannot be explained away by sentence length because length was constant. Although sentences with seven words constituted only 37% of the corpus, the effects were largely the same as for the model fit with the full data set: particularly the main effect of surprisal was qualitatively the same as before ($\hat{\beta} = -0.66$, $\hat{SE} = 0.31$, $t = -2.11$). Not significant anymore were the effect of word length ($\hat{\beta} = 0.05$, $\hat{SE} = 0.17$, $t = 0.31$) and the interaction of surprisal and age ($\hat{\beta} = 0.11$, $\hat{SE} = 0.09$, $t = 1.26$). The directions of the effects remained the same and the lack of significance may be due to low statistical power.

5.4. Influence of degree of parallelism

Finally, we evaluated the hypothesis that age differences in effects of syntactic processing cost can be modeled in terms of the beam size of the parser (the number of interpretations pursued concurrently). The analyses reported above show that syntax effects on scanpath regularity were smaller in older readers. This might be the case because the massively parallel parser used to calculate surprisal and retrieval is a better model for sentence processing in young readers than in older readers. If that was the case, we should find a reversed interaction of age and the syntactic measures when using surprisal and retrieval scores obtained with a parser that maintains fewer interpretations concurrently: Younger readers should show smaller effects of surprisal and retrieval than older

readers. We tested this prediction by fitting linear mixed models like the one reported above. The only difference was that we used surprisal and retrieval values calculated with increasingly restricted parsers: We fitted models for parsers which maintained 100, 50, 25, 20, 15, 10, 5, and 1 interpretations. Fig. 6 shows the relevant parameter estimates as a function of the number of maintained interpretations. The estimates of the other factors in the model (age, word length) remained largely the same across models. As the number of maintained interpretations decreased, the predictive power of surprisal and retrieval scores diminished. Also the interactions of the syntactic measures with age became non-significant for smaller beam sizes. These results suggest that more limited parsers are inadequate models for young *and* old readers; in other words the hypothesis that beam size models age differences in working memory resources was not confirmed.

6. Discussion

Based on observations reported in the literature on sentence processing and oculomotor control in reading, we predicted that three factors contribute to scanpath irregularity in reading: (a) average word length in a sentence, (b) age of reader, and (c) syntactic difficulty of the sentence according to the surprisal and retrieval cost metrics. All these

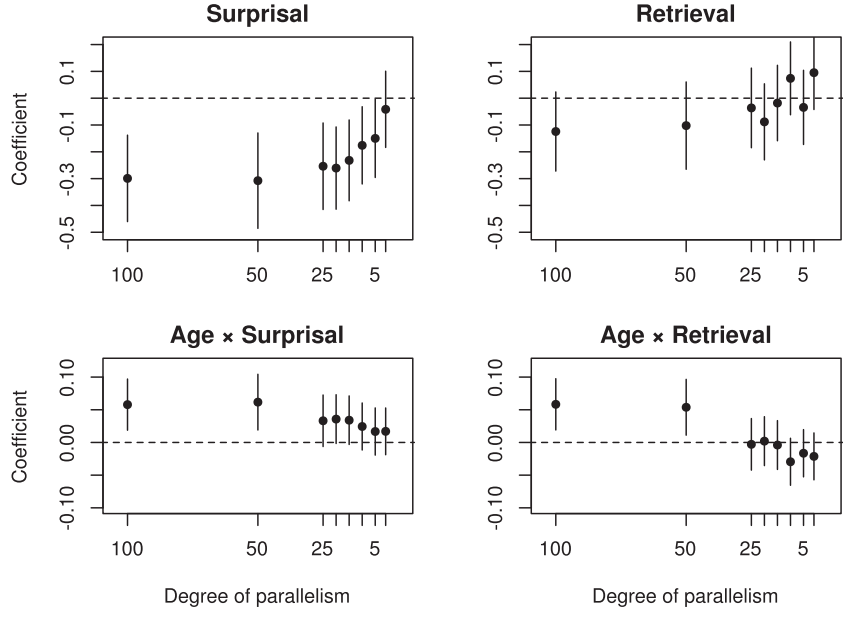

Fig. 6. Coefficients for linear mixed models of scanpath irregularity. The models differed in the surprisal and retrieval values that were used as predictors. These values were generated with parsers that maintained varying numbers of interpretations ranging from 1 to 100 (*x*-axis). As the number of interpretations decreases, the effects of surprisal and retrieval and their interactions with age diminish. Intervals are 95% confidence intervals.

predictions were confirmed in the analysis of scanpaths: The shorter the words in a sentence, the more irregular the scanpaths elicited by that sentence. The older the reader, the more irregular their scanpath patterns. If the average surprisal in a sentence was high, indicating many syntactically unexpected continuations, the scanpaths were more irregular. Finally, if the sentence had high expected retrieval cost, the scanpaths were more irregular (marginally significant). Beyond these predicted effects, we also found that surprisal and retrieval both interact with age: They explain less scanpath variance in older readers than in young readers. To our knowledge, such interactions of age with surprisal and retrieval cost have not been reported in the earlier literature. The significant interaction of age and retrieval also shows that retrieval really does play a role in shaping scanpath patterns even though the main effect of retrieval was only marginally significant. Finally, we tested the theoretically attractive hypothesis that the degree of parallelism of the parser used to calculate surprisal and retrieval is a variable that can be used to model age differences in sentence processing, but there was no evidence for this idea. In the following, we will discuss some of these findings in more detail.

6.1. Effect of age

The effect of age is the strongest effect on scanpath irregularity that we found. It is even stronger than the effect of word length, which is known to heavily influence whether words are fixated or not. While the effect of word length is relatively well understood, there is no generally accepted explanation for the more irregular scanpaths produced by older readers. The proposals for processing deficits in older readers comprise reduced perceptual acuity (Fozard & Gordon-Salant, 2001), increased susceptibility to interference (Hasher & Zacks, 1988), reduced working memory capacity (Dobbs & Rule, 1989; Waters & Caplan, 2005), and slowed processing (Salthouse, 1996). These changes are believed to cause disruptions in processing and to induce shifts toward processing strategies that compensate for these deficits. One very concrete proposal has been made by Rayner et al. (2006), who argue that slowed lexical processing and reduced parafoveal processing in older readers (see also Rayner, Castelhano, & Yang, 2009; Risse & Kliegl, 2011) leads them to adopt a more risky reading strategy that relies on partial visual information and increased guessing of words. This allows the eyes to proceed through the text at a faster pace because words can be skipped more often. Since this strategy is less reliable, the rate of misindentifications is increased, which in turn results in remedial regressions (cf. Vitu & McConkie, 2000). Rayner et al. investigated this idea by modifying the E-Z Reader model to implement such a risky strategy. They found that the simulations conducted with this modified model indeed predicted a pattern of reading measures that resembled that found in the experimental data (slowed reading, longer saccades, increased skipping, and regression probability). In related work, Bicknell and Levy (2010) presented simulations using a Bayesian model of reading showing that risky reading strategies can outperform safer strategies in which the eyes only move forward when a word was reliably identified. The speed-up gained by aggressively moving the eyes forward can apparently outweigh the cost of regressions occurring when words have been misidentified.

If we assume the strategy proposed by Rayner et al. (2006), it may seem surprising that we did not find an interaction of word length and age.[7] Processing a word without looking at it directly may be viable for short, highly predictable words, but it will almost certainly fail for long words. Hence, scanpaths should be particularly irregular in old readers when a sentence with short words is read, but this interaction of word length and age was not found. Bicknell and Levy (2011) reported a corpus analysis suggesting that there might be another effect of word length that counteracts and perhaps cancels out this interaction of age and word length: The probability of a regression occurring on a word is higher when the previous word was long and not skipped (regressions following skipped words were not analyzed in their study). Bicknell and Levy argue that this effect is also predicted by a risky reading strategy: If the word is long, it is harder to identify and an early forward saccade may more often lead to identification failure than when the word is short and predictable. Under the assumption that the oculo-motor system is using a prior on fixation durations, the chance of leaving a long and unpredictable word too early would in fact be higher than the chance of leaving a medium-length word too early. Such a prior would allow the system to make a decision about saccade programming relatively early on, which would be crucial for a reading strategy optimized for speed. The current version of the model by Bicknell and Levy (2010) does not use such a prior on fixation durations. However, such extensions follow naturally from the basic assumptions made in the model and will potentially be added in the future.

It is likely that a shift to a more risky reading strategy in response to changes in lexical processing and perceptual span is not the only factor contributing to the more irregular reading patterns observed in older readers. However, the risky-processing account is attractive because it is well motivated and supported by experimental data and model simulations using the E-Z Reader model. The simulations reported by Bicknell and Levy (2010) are also encouraging because they not only capture the phenomenology of the risky strategy but also explain from which underlying principles aggressive reading behavior emerges (optimization of a speed-accuracy tradeoff).

Another account of the differences in eye movements between young and older readers has been proposed by Wotschack and Kliegl (2013). According to these authors, the observed eye movement behavior in older readers is not due to a shift in reading strategies but may be due to difficulties with the execution of reading strategies; these difficulties could be caused by problems with executive control and visual perception. The two explanations, shifts in strategies and degraded execution of strategies, do not necessarily exclude each other and may jointly contribute to age effects in reading. Clearly, more research is needed to elucidate the reasons underlying age differences in reading.

6.2. *Effects of syntactic processing difficulty*

Surprisal and retrieval cost both play a role in shaping scanpaths. The more difficult to process the words in a sentence were, the more irregular was the scanpath pattern. This is consistent with the findings of Boston et al. (2011), who reported higher rates of

regressions and longer fixation durations for words that were difficult according to these measures. Hence, our results show that syntactic effects that can be found in classical eyetracking measures can also be recovered with the scanpath measure proposed by von der Malsburg and Vasishth (2011); this is an important finding for the questions posed originally by Frazier and Rayner (1982) regarding the effect of parsing difficulty on eye movement patterns. In previous work, von der Malsburg and Vasishth (2011, 2013) showed that the scanpath measure can help to clarify one specific problem in psycholinguistic research: the function of regressive eye movement patterns in relatively difficult garden-path sentences. However, the present results go beyond this finding because they show that scanpaths can be informative even when the sentences are syntactically relatively simple and when the eyes move from left to right in the majority of cases. The present study therefore delivers an important validation of the scanpath measure as a tool for reading and sentence processing research.

While the main effects of surprisal and retrieval cost were expected, we also found two new effects: Scanpath effects of surprisal and retrieval were weaker in older readers than in young readers. These effects are somewhat surprising because a plausible case could rather be made for the opposite effect. For example, work by Speranza, Daneman, and Schneider (2000) shows that older adults rely more on context and expectation than young adults when they have to recognize words that are presented with visual noise (see also Stine & Wingfield, 1994, and Pichora-Fuller, Schneider, & Daneman, 1995). This suggests that older readers employ a more top-down reading strategy. Consequently, surprisal should be a particularly good predictor in older readers because it is precisely the violation of top-down generated expectations that explains reading performance in a surprisal-based theory. With respect to retrieval costs, we could also expect that the effect increases with age. The model by Boston et al. (2011) quantifies retrieval cost as a function of interference with respect to syntactic features: Retrieval of an item with a certain part-of-speech category is more difficult if there are other items in the sentence with that category. Since older adults are believed to be more susceptible to interference (Hasher & Zacks, 1988; Lustig, May, & Hasher, 2001), this measure should predict the performance of older adults particularly well.

One possible explanation for weaker effects of syntactic in older readers is that sentence processing in older readers might be less syntax-driven. If that's the case, the predictive power of surprisal and retrieval should be diminished in older readers, because the surprisal and retrieval measures used here are purely syntactic. The interaction of age with surprisal and retrieval would then reflect a shift to a processing strategy that is driven more by world knowledge and discourse. Some evidence for this hypothesis comes from a study by Christianson, Williams, Zacks, and Ferreira (2006), who probed the performance of young and older readers on comprehension questions about garden-path sentences. In their experiment, older readers maintained the original, incorrect interpretation of a garden-path sentence more often than younger readers. Christianson et al. interpreted this result as showing that older adults are using a more heuristic interpretation strategy that may not make full use of syntactic cues.

7. Conclusions

We presented the first comprehensive investigation of spatio-temporal fixation patterns in reading. The main purpose of the study was to subject the scanpath measure by von der Malsburg and Vasishth to a benchmark test. In earlier work, we had already demonstrated that the measure is sensitive to effects in reading patterns that were missed in analyses of the traditional eyetracking measures. However, in these studies, the scanpath effects were strong because the sentence material was designed to be difficult to process. Here, we asked: Can the scanpath measure recover effects even in simple sentences which are typically read effortlessly from left to right? Our results show that the scanpath measure is in fact sensitive to these relatively subtle effects. The measure recovered three kinds of effects on scanpath patterns that the earlier literature suggested: effects of oculo-motor constraints, effects of syntactic processing difficulty, and effects of age. This demonstrates not only the sensitivity and validity of the measure, it also shows that the measure can be used to study processes at various levels of processing. Thus, these findings establish the scanpath method by von der Malsburg and Vasishth as a viable tool for investigating eye movements in reading. In addition, we reported an effect that was, to our knowledge, previously unknown: Surprisal and retrieval difficulty, despite being conceptually very different measures of sentence processing difficulty, both had attenuated effects on the eye movements of older readers. This attenuation of syntax effects as a function of age may reflect a shift toward less syntax-driven processing strategies or increased disruptions of processing due to problems with executive control. Planned experiments are needed to discriminate between these two hypotheses.

Acknowledgments

We are grateful to Marisa Ferrara Boston and colleagues for releasing the surprisal and retrieval scores. We thank Klinton Bicknell, Roger Levy, Erik Reichle, and two anonymous reviewers for insightful comments. Titus von der Malsburg was supported by a grant of the International Graduate Program for Experimental and Clinical Linguistics at the University of Potsdam and by the research group Mind and Brain Dynamics (FOR 868) funded by the German Research Foundation.

Notes

1. Other edit-distance measures use an arbitrary weighting of the three edit-operations (substitution, deletion, and insertion) and yield different results depending on how the weights are set. In our measure, the weighting is not arbitrary but principled because deletion and insertion are nothing else but special cases of substitution.

2. A parameter in our measure allows switching to a mode of operation where discrete regions of interest are used instead of fixation coordinates. The difference to the Levenshtein metric is then that our measure has sensitivity to fixation durations. This mode may be useful when the precise locations of fixations do not matter, as for example in experiments employing the visual world paradigm.
3. Based on their analysis of the Dundee Corpus, Demberg and Keller (2008) also argued that surprisal and linguistic memory cost make independent contributions to the overall processing difficulty. However, these authors quantified memory cost as proposed by Dependency Locality Theory (DLT, Gibson, 1998, 2000). DLT is conceptually related to the retrieval theory by Lewis and Vasishth (2005) but differs in the specific assumptions and precise predictions.
4. These assumptions are inherited from the ACT-R framework (Anderson, 1990) in which the cue-based parsing model was implemented.
5. Can we conclude that the retrieval effect was significant because it was negative on 97.9% of the models? No, because these models did not use independent data sets. They used the same set of 72 old readers and overlapping sets of young readers.
6. According to Marisa Boston and John Hale (personal communication), there are two reasons for this correlation. One is that, at the beginning of the sentence, only a few structures can be built. As more words come in, more candidate interpretations are possible, all of which are retained initially. When the limit on their number is reached, the most implausible candidates are discarded. This filling up of the available memory may give rise to systematic effects of sentence length. The other reason may be related to the fact that, at the end of the sentence, there are more possibilities to build dependencies between incoming words and previous words. This drives the probability of each possible dependency down and, consequently, surprisal increases.
7. When we added such an interaction term to the linear mixed model presented above, the estimate of the coefficient was 0.01 and the *t*-value 0.48.

References

Abel, L. A., Troost, B. T., & Dell'Osso, L. F. (1983). The effects of age on normal saccadic characteristics and their variability. *Vision Research*, *23*(1), 33–37.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Bates, D. (2005). Fitting linear mixed models in R. *R News*, *5* (1), 27–31.

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

Bicknell, K., & Levy, R. (2011). Why readers regress to previous words: A statistical analysis. In L. Carlson, Ch. Hölscher, T. Shipley (Eds.), *Proceedings of the 33rd annual meeting of the Cognitive Science Society* (pp. 931–936). Boston, MA: Cognitive Science Society.

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.

Brandt, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, *9* (1), 27–38.

Brysbaert, M., & Vitu, F. (1998). Word skipping: Implications for theories of eye movement control in reading. In G. D. M. Underwood (Ed.), *Eye guidance in reading and scene perception* (Chap. 6, pp. 125–148). Oxford, England: Elsevier.

Christianson, K., Williams, C. C., Zacks, R. T., & Ferreira, F. (2006). Younger and older adults' "good-enough" interpretations of garden-path sentences. *Discourse Processes*, *42*(2), 205–238.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. V. Gompel (Ed.), *Eye movements: A window on mind and brain* (Chap. 15, pp. 341–374). Amsterdam, the Netherlands: Elsevier Science Ltd.

Coco, M. I., & Keller, F. (2012). Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive Science*, *36*(7), 1204–1223.

Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. D. (2010). ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, *42* (3), 692.

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

Dobbs, A. R., & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and Aging*, *4* (A), 500–503.

Drieghe, D., Rayner, K., & Pollatsek, A. (2005). Eye movements and word skipping during reading revisited. *Journal of Experimental Psychology: Human Perception and Performance*, *31* (b), 954–969.

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

Feusner, M., & Lukoff, B. (2008). Testing for statistically significant differences between groups of scan patterns. In K.-J. Raiha, & A. T. Duchowski (Eds.), *Proceedings of the 2008 symposium on eye-tracking research & applications* (pp. 43–46). Savannah, GA: Association for Computing Machinery.

Fozard, J. L., & Gordon-Salant, S. (2001). Changes in vision and hearing with aging. In K. W. Schaie & S. L. Willis (Eds.), *Handbook of the psychology of aging* (pp. 241–266). Louden, UK: Academic Press Inc.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97* (458), 611–632.

Fraley, C., & Raftery, A. E. (2007). MCLUST version 3 for R: Normal mixture modeling and model-based clustering (tech. rep. No. 504). Department of Statistics, University of Washington, Seattle, WA.

Frank, S. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 1139–1144). Amsterdam, 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. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68* (1), 1–76.

Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium* (Chap. 5, pp. 95–126). Cambridge, MA: MIT Press.

Hacisalihzade, S. S., Stark, L. W., & Allen, J. S. (1992). Visual perception and sequences of eye movement fixations: A stochastic modeling approach. *IEEE Transactions on Systems, Man and Cybernetics*, *22*(3), 474–481.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In A. Kehler, L. Levin, & D. Mareu (Eds.), *Proceedings of NAACL 2001* (pp. 1–8). Pittsburgh, PA: Association for Computational Linguistics.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 193–225). San Diego, CA: Academic Press.

Jarodzka, H., Hohnqvist, K., & Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. In H. Istance, A. Hyrskykari, & Q. Ji (Eds.), *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 211–218). New York: ACM.

Josephson, S., & Holmes, M. E. (2002). Attention to repeated images on the World-Wide Web: Another look at scanpath theory. *Behavior Research Methods*, *34*(4), 539–548.

Kaplan, R. M. (1972). Augmented transition networks as psychological models of sentence comprehension. *Artificial Intelligence*, *3*(1–3), 77–100.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*(2), 153–168.

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*(1/2), 262–284.

Kruskal, J. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Laubrock, J., Kliegl, R., & Engbert, R. (2006). Swift explorations of age differences in eye movements during reading. *Neuroscience & Biobehavioral Reviews*, *30*(6), 872–884.

Levenshtein, V. (1966). Binary codes capable of correcting deletions and insertions and reversals. *Soviet Physics Doklady*, *10* (8), 707–710.

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

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 1–45.

Lima, S. D., Hale, S., & Myerson, J. (1991). How general is general slowing? Evidence from the lexical domain. *Psychology and Aging*, *6*(3), 416.

Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General*, *130*(2), 199–207.

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

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

von der Malsburg, T., Vasishth, S., & Kliegl, R. (2012). Scanpaths in reading are informative about sentence processing. In P. B. Michael Carl & K. K. Choudhary (Eds.), *Proceedings of the first workshop on eye-tracking and natural language processing* (pp. 37–53). Mumbai, India: The COLING 2012 Organizing Committee.

von der Malsburg, T., & Vasishth, S. (2007). A time-sensitive similarity measure for scanpaths. In *Proceedings of the European conference on eye movements* (pp. 101). Potsdam, Germany: University of Potsdam.

Mathôt, S., Cristino, F., Gilchrist, I. D., & Theeuwes, J. (2012). A simple way to estimate similarity between pairs of eye movement sequences. *Journal of Eye Movement Research*, *5*(1), 1–15.

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.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453.

Pichora-Fuller, M., Schneider, B., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *Journal of the Acoustical Society of America*, *97*(1), 593–608.

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

Rayner, K., Castelhano, M. S., & Yang, J. (2009). Eye movements and the perceptual span in older and younger readers. *Psychology and Aging*, *24*(3), 755–760.

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

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging*, *21*(3), 448–465.

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

Reichle, E. D., Reineberg, A. E., & Schooler, J. W. (2010). Eye movements during mindless reading. *Psychological Science*, *21*(9), 1300–1310.

Risse, S., & Kliegl, R. (2011). Adult age differences in the perceptual span during reading. *Psychology and Aging*, *26*(2), 451–460.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*(3), 403–428.

Salvucci, D. D., & Anderson, J. R. (2001). Automated eye-movement protocol analysis. *Human-Computer Interaction*, *16*(1), 39–86.

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.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: A formal model and empirical investigation. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 595–600). Austin, TX: Cognitive Science Society.

Solan, H. A., Feldman, J., & Tujak, L. (1995). Developing visual and reading efficiency in older adults. *Optometry & Vision Science*, *72*(2), 139.

Speranza, F., Daneman, M., & Schneider, B. A. (2000). How aging affects the reading of words in noisy backgrounds. *Psychology and Aging*, *15*(2), 253–258.

Stine, E. A., & Wingfield, A. (1994). Older adults can inhibit high-probability competitors in speech recognition. *Aging, Neuropsychology, and Cognition*, *1*(2), 152–157.

Stine-Morrow, E. A., Gagne, D. D., Morrow, D. G., & DeWall, B. H. (2004). Age differences in rereading. *Memory & Cognition*, *32* (b), 696–710.

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

Vasishth, S., von der Malsburg, T., & Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(2), 125–134.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. New York: Springer.

Vitu, F., & McConkie, G. W. (2000). Regressive saccades and word perception in adult reading. In A. Kennedy, R. Radach, D. Heller, & J. Pynte (Eds.), *Reading as a perceptual process* (Chap. 12, pp. 301–326). Amsterdam, the Netherlands: Elsevier.

Waters, G., & Caplan, D. (2005). The relationship between age, processing speed, working-memory capacity, and language comprehension. *Memory*, *15*(3–4), 403–413.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, *3*(1), 1–191.

Wotschack, C., & Kliegl, R. (2013). Reading strategy modulates parafoveal-on-foveal effects in sentence reading. *The Quarterly Journal of Experimental Psychology*, *66*(3), 548–562.

Yarbus, A. L. (1967). Eye movements During Perception of Complex Objects. In *Eye Movements and Vision* (pp 171–211). New York: Springer.