Journal of Memory and Language xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

What is the scanpath signature of syntactic reanalysis?

Titus von der Malsburg *, Shravan Vasishth

Department of Linguistics, University of Potsdam, Germany

ARTICLE INFO

Article history:
Received 31 January 2010
revision received 17 February 2011
Available online xxxx

Keywords:
Reading
Syntactic reanalysis
Eye movements
Parsing
Individual differences
Scanpaths

ABSTRACT

Which repair strategy does the language system deploy when it gets garden-pathed, and what can regressive eye movements in reading tell us about reanalysis strategies? Several influential eye-tracking studies on syntactic reanalysis (Frazier & Rayner, 1982; Meseguer, Carreiras, & Clifton, 2002; Mitchell, Shen, Green, & Hodgson, 2008) have addressed this question by examining scanpaths, i.e., sequential patterns of eye fixations. However, in the absence of a suitable method for analyzing scanpaths, these studies relied on simplified dependent measures that are arguably ambiguous and hard to interpret. We address the theoretical question of repair strategy by developing a new method that quantifies scanpath similarity. Our method reveals several distinct fixation strategies associated with reanalysis that went undetected in a previously published data set (Meseguer et al., 2002). One prevalent pattern suggests re-parsing of the sentence, a strategy that has been discussed in the literature (Frazier & Rayner, 1982); however, readers differed tremendously in how they orchestrated the various fixation strategies. Our results suggest that the human parsing system non-deterministically adopts different strategies when confronted with the disambiguating material in garden-path sentences.

1. Introduction

Eye tracking is a very productive methodology in sentence processing research. Beginning with classic work such as Just and Carpenter (1980) and Frazier and Rayner (1982), reading studies involving eye tracking continue to provide a rich array of empirical evidence that inform competing theories of human sentence parsing. This entire body of work rests on some degree of belief on an assumption articulated first by Just and Carpenter (1980), the eye-mind assumption. As they put it (p. 331): "there is no appreciable lag between what is being fixated and what is being processed." Taken literally, this assumption is clearly false; this is evident from two facts: (i) preview effects which indicate that processing of a word can start even before the eyes fixate it for the first time (Rayner, 1998, 2009); (ii) the spillover-effects where processing initiated at one word can continue even after the eyes move to fixate another word (Rayner & Duffy, 1986). We can therefore safely assume that no eye tracking researcher believes in the strict formulation of the eye-mind hypothesis. At the other extreme, if we were to assume that the eye-mind assumption is completely false, then the eye movement record would be difficult to interpret because fixation durations would have no straightforward relationship with processing difficulty. Clearly, this extreme position is also untenable given the largely replicable findings in the sentence comprehension literature (cf. Clifton, Staub, & Rayner, 2007, for a review of the empirical results).

This leaves us with an intermediate version of the eye-mind assumption: fixation durations reflect processing difficulty, but lags in processing and constraints arising from oculo-motor control (Rayner, 1998) have the potential to complicate the interpretation of the eye movement record. Indeed, theories of eye movement control such as E-Z Reader 10 (Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Warren, & McConnell, 2009) standardly assume such a lag.

A particularly interesting situation arises when the eyes, instead of making the prototypical forward saccade,

* Corresponding author.
E-mail address: malsburg@gmail.com (T. von der Malsburg).

0749-596X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2011.02.004

carry out a regression. Such a regressive eye movement is interesting from the parsing perspective because it could be driven at least in part by parser actions that began when the launch site of the regression was fixated. The question then arises: to what degree are the eyes coupled to and directed by the parser's actions? Are they tightly coupled (as the strict form of the eye-mind assumption would assert), are they completely uncoupled, or is there a loose coupling (as the intermediate form of the eye-mind assumption would predict)? Moreover, the degree of coupling could be modulated and depend on, for instance, the particular type of processing difficulty encountered by the parser.

This question has been the focus of a trio of papers by Frazier and Rayner (1982), Meseguer et al. (2002) and Mitchell et al. (2008). Frazier and Rayner suggested that when the parser initiates a reanalysis action, detaching a constituent from the incremental tree built so far and attaching it to another part of the tree, the eyes carry out a regressive saccade to follow the parser's actions: as the parser intelligently searches for an alternative attachment site in the sentence, the eyes follow along. For example, in the sentence *Since Jay always jogs a mile seems like a very short distance to him* the noun phrase (NP) *a mile* is initially mis-attached to the verb *jogs* as its direct object; when the next word, *seems*, is processed, a reanalysis process begins whereby the NP is reattached as a subject of a main clause headed by the verb *seems*. Frazier and Rayner named this intelligent reanalysis process Selective Reanalysis, and argued that the eyes closely follow the parser's processing steps: "The selective reanalysis hypothesis predicts that eye movements should regress from the disambiguating region to the ambiguous region of the sentence" (p. 204).

Note, however, that Selective Reanalysis does not presuppose the strict form of the eye-mind assumption; it is consistent with the intermediate version of the eye-mind assumption, since the eyes could be following—with a lag—the parser's repair actions.

Twenty years after the Frazier and Rayner proposal, Meseguer et al. presented more evidence for Selective Reanalysis from Spanish. They examined garden-path sentences in Spanish where an adverbial phrase could attach high or low. They found evidence in favor of Selective Reanalysis (this work is discussed in detail below). In subsequent work, Mitchell et al. (2008) challenged the idea that the eyes regress in lock-step with the parser's actions.

Mitchell et al.'s alternative proposal was that the cognitive system that drives parsing may want to avoid moving forward to take in new information when reanalysis is triggered. It therefore takes a "time-out," which results in a regression (presumably because moving to the right would bring new information in, which is undesirable in the face of increased processing load).[1] In their analysis, Mitchell et al. were also concerned with eye movement patterns, but relied on "regression signatures" instead of qualitative assessments as Frazier and Rayner did; these are probabilities of the eye landing on any word preceding the word from which the regression started.

In these three papers, the crucial evidence for (and against) the coupling of eye movements with parsing actions dictated by Selective Reanalysis hinged on analyzing eye movement *patterns* rather than just fixation durations or regression probabilities; scanpaths are the principal object of inquiry. This makes sense because the question literally is: what patterns of regressive eye movements result when reanalysis begins? (Eye-tracking researchers often refer to 'patterns' of eye movements where they really mean fixation durations; in the present case, 'patterns' stands for scanpaths.)

Since there exists no suitable quantitative way to evaluate the similarity of one eye movement pattern with another, the three sets of authors mentioned above were forced to either look at scanpath patterns qualitatively (e.g. Frazier & Rayner, 1982, p. 196) or to reduce scanpaths to scalar duration measures and transition probabilities in order to derive conclusions about participants' behavior (e.g. Frazier & Rayner, 1982, pp. 199–200). Due to the unavailability of a method for quantitatively studying spatio-temporal fixation patterns, subjective descriptions were necessary. It is worth quoting one such description (Frazier & Rayner, 1982, pp. 196–197) to underline the fact that a major issue of interest is indeed eye movement *patterns*, i.e. scanpaths, and not only transitional probabilities:

> ...three or four patterns of eye movement behavior occurred which we shall attempt to characterize. In some cases, subjects read the ambiguous noun phrase and upon reading the disambiguating region made very long fixations. These long fixations were also accompanied by very short saccades ... Upon reading the end of the sentence, the subject then made a long regression to the beginning of the sentence and reread the sentence. The long fixations and short saccades in the disambiguating region and thereafter may also have been accompanied by short regressions, but the reader did not regress at that point back to the beginning of the sentence or to the ambiguous region. We shall characterize this behavior as *chaos* in that the reader apparently was having great difficulties understanding the sentence but seemed to have no insights as to what the nature of the processing difficulties were. This pattern of eye movements was particularly noticeable among three of the subjects and occurred less frequently with most of the other subjects.

Thus, it is clear that the debate about how eye movements are driven by a sequence of parsing actions needs a method for characterizing scanpath patterns and their relative similarities to each other or to theoretically proposed patterns of regressions.[2] In this paper, we provide such a method, along with freely available software for exploring eye movement patterns. We also reanalyze

[1] In order to prevent upcoming material from interfering with the processing of earlier material, the eyes could of course just stay on the word that caused the processing difficulty. Mitchell suggests (personal communication) that the oculo-motor system might have a strong drive to keep up the pace. Since the way forward is blocked, a random-walk on the previous part of the sentence ensues. This resulting scanpath would be influenced by the physical arrangement of the sentence on the screen but not by its linguistic structure.

[2] We use the term regression not only for a single regressive saccade but also to refer to a scanpath that starts with a regressive saccade and ends when the eyes return to the origin of this saccade.

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

Meseguer et al.'s data (which they generously provided to us) using this method to demonstrate the gain in information when we rely on a formal characterization of scanpath patterns rather than qualitative evaluations of scanpaths and regression signatures. One contribution of this paper is to provide such an additional analytical tool for directly evaluating scanpaths where these are at issue theoretically.

The present investigation was motivated by two main aspects of the above-mentioned studies on reanalysis. First, although the analytical methods used by Frazier and Rayner, Meseguer et al. and Mitchell et al. are undoubtedly informative, it is possible that aggregated eye-tracking measures are misleading because they could in principle arise from a blend of several classes of fixation patterns. Increased transition probabilities in one condition might be caused by changes in different underlying populations of scanpaths. For example, if there is a higher probability of transitions from region 9 to region 2 and from 2 to 7, we cannot decide if this was caused by patterns that went with transitions $9 \rightarrow 2 \rightarrow 7$ or by one class of patterns with occurrences of transition $9 \rightarrow 2$ and another with transitions $2 \rightarrow 7$. In principle, one could calculate transition probabilities that are conditional on the previous transitions. In a typical reading study, however, there are hardly enough data points to reliably estimate unconditioned transition probabilities, let alone conditional probabilities. In general, if the aggregated data stem from different populations, it can be very difficult to infer anything precise about the various strategies used by the reader as they relate to eye movement patterns.

Second, in both the Mitchell et al. and Meseguer et al. studies, many regressions ensued after reading the disambiguating material in the *non*-reanalysis condition. In fact Meseguer et al. recorded 700 regressions in the reanalysis condition and 667 in the non-reanalysis condition. This means that most regressions produced by participants cannot plausibly be explained by *any* of the competing theories, which predict that long-range regressions occur only in the garden-path condition. The observed regression patterns are, however, only problematic if we assume a strictly deterministic parser that always adopts the preferred structure during ambiguous segments of the sentence. In this framework garden-pathing takes place in only one condition; Frazier and Rayner (1982) made this assumption explicit by invoking the *late closure* and *minimal attachment* principles (Frazier, 1979) in order to explain the occurrence of regressive eye movement patterns in reanalysis conditions.

Evidence for Selective Reanalysis clearly exists in the above-mentioned data, but this evidence comes in the form of a slightly increased probability of regressions to particular words. However, if the eye-mind assumption—which is the basis for reasoning in favor of Selective Reanalysis—holds, then we have to ask what process is driving the numerous regressions in the non-reanalysis conditions. Even the Time-out hypothesis, which seems to partly reject the eye-mind assumption in the case of regressions, has not much to offer when it comes to explaining regressions in the non-reanalysis condition. Thus, the data suggest that there is a lot more to the issue of regressions than meets the eye.

2. Scanpath similarity and its application

In previous work (e.g. Brandt & Stark, 1997), one approach for quantifying scanpath similarity was to use a type of edit distance, e.g., the Levenshtein metric (Levenshtein, 1966). Edit distances define the similarity of two sequences of symbols as the minimal number of edit operations that have to be performed on one of these sequences in order to transform it into the other. The minimal number of edits can be computed using the Needleman–Wunsch algorithm which has been used in bioinformatics for analyzing DNA sequences (Needleman & Wunsch, 1970). The edit operations are usually *deletion*, *insertion*, or *substitution* of a symbol. This distance measure can be applied to eye movements by representing scanpaths as a sequence of symbols where the *n*th symbol specifies the region of interest (e.g., word) that was targeted by the *n*th fixation.

However, there are several problems with this measure when analyzing eye movements in reading. The first concerns the proper treatment of differences in fixation durations. In conventional methods, the only information that is available for each fixation is its target location; fixation durations play no role in the edit distance calculations. This is problematic because fixation durations are one of the most important sources of evidence in eye movements. The second problem with standard edit distances is that the spatial configuration of fixation targets is not taken into account; in standard approaches, fixations are only evaluated with respect to the identity of their targets. Therefore, the distance measure cannot penalize large spatial divergence between two scanpaths more than small ones. The third problem relates to regions of interest. A scanpath analysis based on regions of interest works only if those regions are sufficiently large, so that there is a substantial chance that fixations that functionally serve the same purpose are targeted at the same region. In reading, natural regions of interest such as words or phrases do usually serve this purpose. However, when the visual stimuli cannot be partitioned in such a natural way, or if one wants to retain information as to where in a region the fixation was, it would be more suitable to have a measure that does not necessitate discretization of the stimulus into regions.

In the remainder of this section we describe a new kind of edit distance that addresses these problems and, in doing so, provides a better tool for investigating regression patterns than visual inspection, conventional edit distance based metrics, and transition probabilities.

As already mentioned, the core idea of edit distances is to quantify the dissimilarity of two sequences as the minimal number of edit operations necessary to transform one sequence into the other. This involves finding and aligning parts of the two sequences that are similar already. Consider this example: if the letters of the alphabet stand for regions of interest in a visual stimulus, we can define a scanpath *s* as the sequence ABCF. This sequence means that the eyes first fixated on region A, then B, C, and finally at region F; every letter represents one fixation. Now, consider another scanpath, *t*, which is defined as ABCDEF. The scanpaths *s* and *t* differ in that *t* has fixations on D and E that are missing in *s*. Thus, an alignment that

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

minimizes the dissimilarities between s and t when comparing them letter by letter is:

Fixation:	1	2	3	4	5	6
s:	A	B	C	–	–	F
t:	A	B	C	D	E	F

The dissimilarity of s and t under the Levenshtein metric is then 2 because there are two positions, 4 and 5, in which s and t differ. Here is a more complex example: s' is defined as ABCDEFG, and t' as XBCDFGH. The optimal alignment is then:

Fixation:	1	2	3	4	5	6	7	8
s':	A	B	C	D	E	F	G	–
t':	X	B	C	D	–	F	G	H

The Levenshtein distance of s' and t' is then 3 because in scanpath s' we have to carry out a substitution in position 1, a deletion in position 5, and an insertion in position 8, in order to transform s' into t'. Our critique of the Levenshtein metric then amounts to stating that accounting a penalty of 1 for every edit operation is too crude. It neglects the durations of the affected fixations and, in the case of a substitution, the spatial distance of the replaced fixations from its replacement: if X is close to A, we should assign a smaller penalty than if X is far from A, because in the former situation the divergence between s' and t' is smaller. We simplify the matter by treating insertion and deletion as special cases of substitution where the substituted fixation or the substituting fixation are null-fixation with duration 0 ms. Our task is therefore to find a suitable cost function for calculating the penalty of a substitution of two fixations. This involves enriching the representation of scanpaths; letters are not sufficient because they do not encode the duration and precise position of the fixation. Instead of letters we will use tuples of fixation position and fixation duration in milliseconds. Note that when analyzing scanpaths, the substitution will be the default case because two fixations hardly ever have the exact same position and duration.

There are two extreme cases that we have to consider. One is that the position of two fixations f and g is the same. In this case, the penalty is straightforward and consists of the difference in the fixation durations since there is no other difference between them. The other case occurs when f and g are extremely far apart. Here, we might say that the longer each fixation is, the larger the penalty should be. On the other hand, if both fixations are short, the penalty should be small as well. In other words, it is not the difference between the fixation durations that counts but rather their sum. Intermediately, this leaves us with this penalty function for the substitution of f and g ($dur(f)$ is the duration of a fixation f):

$$d(f,g) = \begin{cases} |dur(f) - dur(g)| & \text{if } f \text{ and } g \text{ have the same location,} \\ dur(f) + dur(g) & \text{if } f \text{ is extremely far apart from } g. \end{cases} \quad (1)$$

The remaining question is: what should the penalty be in cases where f and g are only slightly apart? We need to define a smooth transition from one extreme case to the other, a means for calculating weights for the two terms in Eq. (1) that determine how much each of them contributes to the overall penalty.

There is one fundamental fact about human vision that has to be reflected by this transition: the much higher visual acuity in the center of the visual field, the fovea, and the low resolution in the periphery. The intuition is this: if a word is in the center of the fovea and we move it 5° away, this has a large effect on the word's readability. First, it is easy to read, but after moving it away reading it becomes quite hard. But when the word is already 10° away from the fovea, moving it 5° further away will not have a large effect on its readability; it is hard to read in both positions. Electrophysiological and psychophysical studies have shown that the drop in sensitivity from the fovea towards the periphery is roughly exponential (cortical magnification, Daniel & Whitteridge, 1961; Rovamo, Virsu, & Näsänen, 1978). A simple exponential function of the distance of f and g in the visual field can be used to approximate the drop in acuity:[3] $m^{distance(f,g)}$. Minimizing the squared deviations of this exponential function from the values measured by Rovamo et al. results in a value of 0.83 for m.

Can we derive the desired weights using this exponential function? These are the requirements: if the distance between f and g is 0, the weight for the first term in Eq. (1) should be 1 and the weight for the second term should be 0. If f and g are extremely far apart, the first term should have weight 0 and the second should have weight 1. Finally, if we are in a situation where f and g are close, we are in an intermediate situation where we want to have an influence of both terms. When the distance between f and g is 0, $m^{distance(f,g)}$ is 1. When the distance is large, the value of the function approaches 0. Hence, we can use $m^{distance(f,g)}$ for weighting the first term and $1 - m^{distance(f,g)}$ for the second term. The dissimilarity of f and g is then quantified as the sum of the weighted terms. We get the following penalty function for the substitution of fixations f and g:

$$d(f,g) = |dur(f) - dur(g)| \times m^{distance(f,g)} + (dur(f) + dur(g)) \times \left(1 - m^{distance(f,g)}\right) \quad (2)$$

Note that this function computes penalties for all four situations that can arise in an alignment of scanpaths: (i) substitution of dissimilar fixations, (ii) insertion and (iii) deletion, because they are treated as substitutions by or of a null-fixation which has duration 0, and lastly (iv) the no-edit situation: if there is no difference in position and duration, the result of the penalty function is 0, hence there is no penalty.

We calculate the overall dissimilarity of two scanpaths by aligning them with the Needleman–Wunsch algorithm and by summing the substitution penalties for the corresponding pairs of fixations. To summarize how our measure works:

[3] Fixations are usually described in terms of pixel coordinates on the screen. Conversion to visual field coordinates, i.e. latitude and longitude, can be achieved using the inverse gnomonic projection.

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

1. We use the Needleman–Wunsch algorithm to align similar parts in two scanpaths. Null-fixations are introduced where no matching fixations exit.
2. The similarity of two fixations is determined by the penalty function (Eq. (2)) which incorporates their positions and durations. If two fixations have the same position, their dissimilarity is given by the difference of their fixation durations. If they have extremely different targets, their dissimilarity is the sum of their fixation durations.
3. The Levenshtein metric operates in a binary fashion: if two fixations have exactly the same position, their difference is zero, otherwise it is a constant value no matter what their distance is. Our measure replaces this binary behavior with a smooth transition from one extreme case to the other. The shape of this transition mimics human cortical magnification.
4. The dissimilarity of two scanpaths is the sum of the dissimilarities of their matching fixations.

We shall call this measure *Scasim* for scanpath similarity. See Table 1 for two example calculations.

2.1. Discussion of Scasim

Fixation durations can be arbitrarily long and can therefore increase differences between two scanpaths indefinitely. Differences in position, however, are bounded because our visual field is limited. This is reflected in our similarity measure in an additive effect of differences in fixation durations and an asymptotic effect of differences in position. See Fig. 1 for an illustration.

Previously proposed approaches to analyzing scanpaths based on edit distances (e.g. Brandt & Stark, 1997; Cristino, Mathôt, Theeuwes, & Gilchrist, 2010) or transition probabilities (e.g. Meseguer et al., 2002; Salvucci & Anderson, 2001) require the definition of discrete regions of interest. This has two potential drawbacks: (i) spatial information is lost by this discretization as these approaches only retain information about which region was gazed at in a fixation, but not the precise location of the fixation, (ii) the definition of proper regions is to some extent arbitrary; in reading research some studies use words while others use phrases. Our measure, however, operates on the continuous coordinates of fixations and is therefore spatially more sensitive than measures that require discrete regions of interest. Deciding on reasonable definitions for those regions becomes a non-issue because none are needed. If a discrete model of space is desired, the coordinates of fixations can be mapped to the center of the region of interest enclosing them. If additionally the distances between the regions should not be taken into account, as in the Levenshtein metric, m can be set to 0 in Eq. (2). Then the measure distinguishes only between same and different fixation target but differs from the Levenshtein metric in that it takes fixation durations into account. Finally, when m is set to 1, scanpaths are evaluated only with respect to their temporal dynamics; in this case spatial information is completely ignored. Whether this mode of operation has useful applications remains to be seen (but this issue is orthogonal to the goals of the present paper).

It is reasonable to ask how well a much simpler measure would perform compared to Scasim. One obvious instance of a whole class of simple measures is the sum of the spatial distances of two scanpaths at each point in time. Example (i) in Fig. 2 illustrates this measure which quantifies the dissimilarity between scanpaths s and t as the area between the two scanpaths in the plot.[4] There are four problems with this measure: consider the two scanpaths in panel 1 in Fig. 2 which are perfectly similar except that in s reading starts immediately whereas in t the eyes spend additional 400 ms on the first word before they move on. The first problem is that t has a longer tail that does not have a counterpart in s. Second, a reasonable similarity value would consist of a penalty that is a function of the delay in t. If the delay is 0 ms instead of 400 ms, the two scanpaths are exactly the same; the longer the delay, the larger the dissimilarity. The simple measure, however, behaves very differently: it does not see the similarity of s and t after the delay, because the similar part is shifted in time, and the overall dissimilarity will be mainly determined by the length of s and t, not by the length of the delay. Specifically, we get this paradoxical result: the longer the similar part is, the smaller the similarity as seen by this measure. The problem is that it is not trivial to decide which part in t corresponds to which part in s. The Needleman–Wunsch algorithm offers a solution for this problem. Third, the simple measure compares s and t sample-by-sample of the eye movement record. In principle, we could apply the alignment detection to samples instead of fixations, but this would render the calculation of scanpath similarity very

Table 1
Scasim calculations for two pairs of scanpaths. Both pairs start out with fixations on the same locations. In the course of the following four fixations they diverge until their distance is 20° of the visual field. In the first pair, s and t, the fixation durations are similar. For both scanpaths they are sampled from a normal distribution with mean 200 ms and sd 30 ms. In the second pair, s' and t', the fixation durations are more different. In s' they have a mean of 200 ms, whereas in t' they are sampled from a normal distribution with mean 400 ms. Term 1 is the difference of the fixation duration of two corresponding fixations weighted by $m^{distance}$. Term 2 is the sum of the fixation durations weighted by $1 - m^{distance}$ (see Eq. (2)). The increased fixation durations in t' have an additive effect on the Scasim value. See also Fig. 1.

Similar fixation durations						
Fixation	1	2	3	4	5	
Distance	0°	5°	10°	15°	20°	
$m^{distance}$	1.00	0.39	0.16	0.06	0.02	
Durations s (ms)	190	203	211	201	216	
Durations t (ms)	209	220	212	185	157	
$m^{distance}\times$ term 1	19.0	6.7	0.2	1.0	1.4	
$m^{distance}\times$ term 2	0	256	357	362	364	
$\sum$	19	263	358	363	365	**Scasim = 1368**
Dissimilar fixation durations						
Durations s (ms)	208	216	217	206	156	
Durations t (ms)	361	394	423	341	410	
$m^{distance}\times$ term 1	153	70	32	8	6	
$m^{distance}\times$ term 2	0	370	541	514	552	
$\sum$	153	440	573	522	558	**Scasim = 2246**

[4] For convenience, we use scanpaths in which fixations vary only horizontally but not vertically. That allows us to plot position vs. time. The scanpaths in this section can be thought as being obtained in a reading task in which one sentence was presented on one line.

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

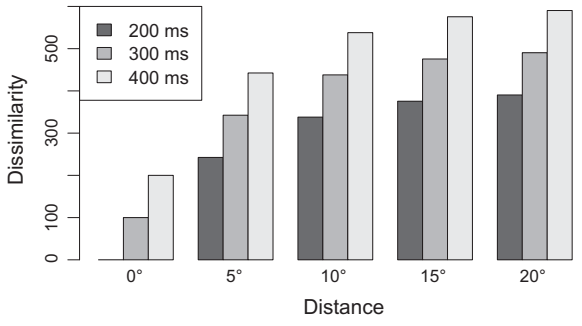

Fig. 1. Dissimilarities for pairs of fixations *f* and *g*. The duration of *f* is fixed to 200 ms. Increasing the duration of *g* has an additive effect on the dissimilarity. Increasing the spatial distance between *f* and *g*, however, has an effect that asymptotically approaches the sum of the fixation durations of *f* and *g* (term 2 in Eq. (2)).

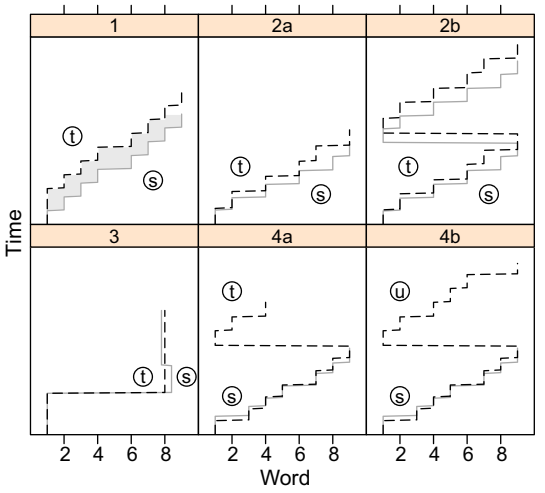

Fig. 2. Scanpaths as they typically occur in reading illustrating properties of Scasim. Panel 1: Highly similar scanpaths; in *t*, the first fixation is longer than in *s*. The size of the gray area between *s* and *t* can be used as a simple similarity measure. Panel 2a, 2b: Two pairs of scanpaths with similar patterns but different reading speed (*t* and *t*′ are slower). The raw Scasim score of *s* and *t* is half of the score of *s*′ and *t*′. Compared to that, the Scasim score per fixation is the same for both pairs, and therefore an indicator of speed differences that is not confounded by scanpath length. Panel 3: Two scanpaths in which the same time is spend on the same regions of the sentence. Scanpath *s*, however, has a refixation that is absent in *t*. Scasim is highly sensitive to these refixations. If this is not desired, adjacent fixations can be merged before applying a Scasim-based analysis. Panel 4a, 4b: Three scanpaths; *s* has no regression, *t* a short one, and *u* a regression that is twice as long as that in *t*. Scasim assigns a similar similarity score to *s* and *t* as to *t* and *u* because it does not make assumptions about the special theoretical status of regressions in reading. It is theory-agnostic because it is supposed to test theories about eye movements.

costly: the Needleman–Wunsch algorithm takes a number of processing steps that is proportional to $m * n$ where m and n are the numbers of items in the two sequences that are being compared. At a sampling frequency of 500 Hz and a duration of 10 s per scanpath, this amounts to a very large number of operations. When calculating the similarity of just two scanpaths this might not pose a problem, but many analyzes based on scanpath similarity require the calculation of all similarities among a large set of scanpaths. This has the consequence that one quickly reaches the limits of what is possible with current desktop computers. Hence, doing fixation detection and using a measure defined for fixation sequences instead of raw samples is basically a necessary optimization step. However, it complicates the measure somewhat because fixations, opposed to samples, differ in their duration. Fourth, the simple measure quantifies the distance between eye positions in veridical space. This means that a 1 cm difference in the middle of the visual field is treated the same as a 1 cm difference in the periphery. As discussed above, to us this does not seem to be a good design decision, since we know that human vision is highly sensitive in the fovea, but has a low resolution in the periphery. In sum, it seems that a simpler measure, like the one described above, would yield inferior or even incorrect similarity values.

In some situations, raw Scasim scores might not be the most suitable measure of similarity. Consider the pair of scanpaths *s* and *t* in panel 2a that are very similar, both reflect left-to-right reading of a sentence, except that the eyes progress faster in *s*. Another pair, *s*′ and *t*′ in panel 2b, consists of repetitions of *s* and *t*: reading the sentence left-to-right two times in a row. The Scasim score of *s*′ and *t*′ will then be twice the score of *s* and *t* because there is a penalty for the dissimilarity in every matching pair of fixations and there are twice as many such pairs in *s*′ and *t*′. However, we might say that the similarity of the scanpaths stays the same, no matter how often the pattern is repeated. The difference is rather given by the average reading speed in *s* vs. *t* which is independent of scanpath length. In general, it is trivial that scanpaths can be more different when they are longer. Hence, in many situations it will be much easier to detect interesting sources of variance by using a similarity measure that is not confounded by scanpath length. We obtain such a score by normalizing raw Scasim scores: we can divide them by either the number of fixations in both scanpaths, yielding similarity per fixation, or by the total duration of *s* and *t*, yielding similarity per unit of time.

How does Scasim deal with scanpaths that have the same trajectory but different numbers of fixations? One such situation arises when two scanpaths, *s* and *t* in panel 3, are similar because in both the gaze shifts from word 1 to word 8. They differ, however, in that there is an overshoot followed by a small correction saccade in scanpath *s* while in *t* the gaze precisely hits word 9. The fixation preceding the correction saccade in *s* lasts 60 ms and the fixation following it 120 ms, the corresponding fixation in *t* lasts 180 ms. In this case Scasim will match the 120 ms and the 180 ms fixations because they are more similar than the 60 ms and the 180 ms fixation. Now, the 60 ms fixation does not have a corresponding fixation in *t* and thus increases the dissimilarity of *s* and *t*. In some situations, this might not be desirable and one might prefer a measure that sees *s* and *t* as being almost perfectly similar because in both scanpaths the eyes spend 180 ms overall on word 8. However, whether or not this reasoning is valid depends on the question that is investigated. After all, it

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

means that saccades leading to refixations would be ignored, which would not be advisable when studying, e.g., oculo-motor control in reading, as refixations and correction saccades play a role in the relevant theories. Moreover, two subsequent fixations will never have exactly the same location and it is not clear which saccades qualify as small enough to be negligible. All in all, Scasim avoids making assumptions that would be ill-motivated in some situations. If, however, a researcher has reasons to treat subsequent fixations that are nearby as one long fixation, there is a simple way to achieve this: preprocessing the fixation data to merge subsequent, nearby fixations before applying a Scasim-based analysis. This way, all relevant parameters, like a threshold distance for merging, are under the control of the researcher. One possible scheme for such a merging procedure is this: (i) subsequent fixations that are closer than, e.g., 0.5° are combined into a new fixation; (ii) the fixation duration of this fixation is the sum of the fixation durations of the contributors; (iii) the position of the new fixation is a weighted average of the positions of contributing fixations, the weights of which are determined based on the respective fixation durations: long fixations have a stronger influence on the position than short ones.

Finally, we would like to discuss a concern about the validity of our measure in the context of reading research. Suppose we have three scanpaths, *s*, *t* and *u* in panels 4a and 4b. The first, *s*, consists of straight left-to-right reading of a sentence. In *t*, the eyes perform a short regression to the beginning of the sentence after a first pass through the sentence. In *u*, the regression trajectory following the first pass is twice as long as in *t*. It now appears that the dissimilarity of *s* and *t* according to Scasim is roughly half the dissimilarity of *s* and *u*. This is because the dissimilar part, the regression, is twice as long in pattern *u*. One reviewer of this article suggested that this behavior of the measure is not desirable because both patterns *t* and *u* exhibit a regression to the beginning of the sentence, whereas *s* does not have it. Consequently, the argument goes, patterns *t* and *u* should be more similar than patterns *s* and *t*, but according to Scasim they are not. However, in order to achieve this, a measure would need built-in knowledge of the special theoretical status of regressions in reading and of their functional significance. This knowledge would transform the measure into a theory of eye movements in reading of its own. While this might be desirable in some contexts, we consciously decided to design a measure for eye movements that is theoretically agnostic because it is supposed to test those very theories about eye movements. This discussion demonstrates that Scasim should be used while taking into careful consideration the research questions to be addressed. It should not be considered a one-size-fits-all tool.

2.2. Related work

While our scanpath measure makes no assumptions about the processes giving rise to the observed scanpaths, Salvucci and Anderson (2001) presented an intriguing approach coming from exactly the opposite direction (see also Salvucci, 1999). They assume a range of fully spelled-out theories of the processes underlying the eye movements in a given task. From each of these theories they derive a hidden Markov model that describes probability distributions over regions of interest indicating where the eye should fixate given a particular state of the cognitive system. Then, the likelihood of the competing hidden Markov models, given the observed eye movements, is used to decide which of the theories explains the data better. Reichle et al. (1998) compared E-Z Reader versions 3 and 5 on the basis of fixation durations and fixation probabilities but this analysis was not able to determine which version of the model was performing better. In the evaluation of their method, Salvucci and Anderson compared the two versions of E-Z Reader and showed that, with respect to fixation *patterns*, version 5 generates better predictions than version 3. This is an interesting result because it shows that, even in straight left-to-right reading of simple sentences, differences in scanpath patterns are informative.

Nevertheless, the method by Salvucci and Anderson has some limitations when we consider the goals of our investigation. First, it uses the location of fixations, the sequence in which they occur, but not their durations. Since we have no reason to exclude the possibility that fixation strategies might to some extent be distinguished by fixation durations, we would like to use a method which is sensitive to differences in durations. Second, it requires models making precise predictions about fixation patterns, something that does not exist in the case of syntactic reanalysis. The issue is further complicated by the Mitchell et al. finding that the visual presentation of a sentence influences the distribution of regressive eye movements; however, it is not well understood how precisely the layout influences regression guidance. This makes it difficult to apply the hidden Markov approach to the problem of regression patterns during syntactic reanalysis.

The main difference between Scasim and the method by Salvucci and Anderson relates to the type of questions that they can answer. If the question is: *which of those theories does a better job at explaining my data?*, the hidden Markov approach is able to answer that, provided the theories are explicit enough to allow the construction of working models. Scasim can produce an answer to this question as well, and in fact Salvucci and Anderson advocate the use of the Levenshtein metric for this purpose, but it can also be used to answer the following question: *what fixation strategies are present in this data set?* The literature has little to offer regarding the scanpath phenomena associated with syntactic reanalysis. Frazier and Rayner (1982) report a qualitative analysis, Meseguer et al. (2002) transition probabilities, and Mitchell et al. (2008) the distribution of landing sites of the first regressive saccade. Hence, we have little reason to expect a specific type of scanpath pattern and prefer to use a tool which allows us to explore the data without committing to any specific theories.

Another approach to analyzing scanpath patterns has recently been presented by Cristino et al. (2010, ScanMatch). Like Scasim, this approach uses a kind of edit distance. Despite some technical similarities, ScanMatch and Scasim are very different in a number of aspects. One is the way in which temporal information is treated. An-

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

other more important difference is that ScanMatch delivers a very different concept of similarity and there is perhaps little overlap in the potential applications of the two measures.

Cristino et al. (2010) also address the problem of gaze durations that were ignored in earlier proposals (e.g. Brandt & Stark, 1997) but they suggest a different solution than we do. In a first processing step, they detect fixations and saccades in raw eye movements. Then, however, they break fixations up into smaller temporal bins of, e.g., 50 ms. A scanpath in which the gaze trajectory went from region A to B to C with 80 ms fixation duration on A, 120 ms on B, and 130 ms on C, would then be represented as the sequence AABBCCC where each symbol accounts for 50 ms of gazing at the respective region. When this sequence is compared to the scanpath AABCCC, in which less time was spent on B, ScanMatch will account for a 50 ms difference in B. The temporal binning introduces some aliasing error and the temporal resolution is restricted to the size of the temporal bins. Reducing the bin size decreases this error but also increases the length of the symbol sequences and therefore the run time of the algorithm calculating the overall similarity of the two scanpaths. Cristino et al. (2010) note that run time costs can be alleviated by using the BLAST algorithm (Altschul et al., 1997) instead of the Needleman–Wunsch algorithm but warn that this algorithm does not guarantee to find the correct solution. Another potential shortcoming of this approach to accounting for time is that the symbol sequences on which ScanMatch operates only preserve information about the total time spend on regions but not about the fixations that occurred during that time. The sequence AAA could be the result of three short fixations on region A, or two fixations, or just one long fixation on A. In situations where fixations and refixations matter, this means a loss of relevant information. This is certainly the case in the research of oculomotor control in reading.

Apart from making the Levenshtein metric sensitive to gaze durations, Cristino et al. (2010) propose to make it more flexible by modifying the rules for penalizing differences between two scanpaths. The Levenshtein distance assigns a penalty of 1 for every missing, superfluous, or changed symbol. In ScanMatch, the penalty depends on the identity of the symbols. This has a variety of interesting applications, one is when fixations should not be compared with respect to their location on the screen but rather with respect to the type of object being looked at. For instance, when the task is to quickly find a tree in a complex visual scene, different scanpaths ending at different trees should be evaluated as being highly similar. Scanpaths that end on a bush should be evaluated as being less similar when compared to scanpaths aimed at trees, and scanpaths ending on very different objects, such as cars, should be even more dissimilar. In this case, the similarity of two fixations would not be given by their location on the screen but by the semantic or visual similarity of the material shown at that location. In ScanMatch, the similarity of two regions of interest is defined in a data structure called substitution matrix which can be tailored for a particular experiment to encode such relations between regions of interest. In the example above, the substitution matrix would specify that fixations on trees are equivalent, fixations on trees and bushes are somewhat similar, whereas fixations on trees and cars are very different. This gives a researcher enormous flexibility in designing similarity measures for scanpaths. Useful applications in psycholinguistics might include the analysis of eye movement data from visual world type experiments. However, designing workable substitution matrices is not entirely trivial: many algorithms that make use of similarities assume that the similarity measure is a metric, i.e. that certain mathematical constraints hold (reflexivity, symmetry, and subadditivity).[5] Some care must be taken when designing substitution matrices because a violation of these constraints could yield incorrect results depending on the type of analysis performed on the similarity values.

In principle, the flexibility of ScanMatch allows it to approximate the behavior of Scasim to some degree. Temporal and spatial binning, however would render this emulation of Scasim, which has a continuous model of space and time, coarse. Information about refixations within a region would be lost altogether.

In sum, we believe that the power of the ScanMatch approach lies its flexibility. ScanMatch is not a similarity measure but a powerful framework for building similarity measures. Defining a suitable measure for a particular task is, consequently, not necessarily a simple task. Scasim, on the other hand, commits to a very specific concept of similarity and offers only one degree of freedom: the parameter for spatial sensitivity, m, which rarely needs to be adjusted. As we explained above, in this investigation we need a scanpath measure that makes as few theoretical assumptions as possible. Also, we need a measure that is highly sensitive to fixation durations, which have been a prime source of evidence in reading research. Scasim evaluates scanpaths only with respect to the spatial and temporal properties of the fixations making them up and therefore fits this order.

2.3. An illustration using hypothetical scanpaths

This section will demonstrate a particularly useful way of putting Scasim to work: for a toy data set we will fit a map of scanpath space, i.e. a map on which scanpaths are represented as points that are located close to each other if they are similar. These maps provide a vector-representation of scanpaths that can be conveniently analyzed using a variety of standard statistical techniques. The data set consists of three scanpaths each for three virtual readers. These data were generated by modifying a scanpath recorded in an eye-tracking study where participants read one sentence per trial. In the trial from which we took the scanpath, a participant first read the sentence, regressed to the beginning of the sentence, and then reread the sentence skipping several

[5] This means the dissimilarity of an object to itself should be 0 (reflexivity), the dissimilarity of s and t should be the same as the dissimilarity of t and s (symmetry), and that the dissimilarity of s and t should be smaller than or equal to the sum of the dissimilarities of s and r and r and t (subadditivity/triangle inequality). The first two can easily be satisfied by restricting the set of substitution matrices to symmetric matrices that assign the highest score (indicating perfect similarity) to the diagonal.

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

words. Using this scanpath as a template, we created new scanpaths by shortening the regression and by dropping the regression altogether. A copy of the resulting three scanpaths was assigned to each virtual reader. Individual differences in reading speed were simulated by changing the fixation durations for two of the readers: for each fixation of reader 2 we sampled a value from a normal distribution centered at 1.3 with standard deviation 0.1 and multiplied the fixation duration with this factor (the choice of the sampling distribution parameters is arbitrary). The same procedure was followed for the reader 3, but this time 1.6 was used as the center of the normal distribution. Fig. 3 shows the resulting data set.

Despite the complex nature of scanpaths compared to fixation durations and regression probabilities we now have a simple scalar value to describe them: scanpath similarity as given by Scasim.[6] The question is, how can we make good use of these simple values? A similar problem arises when studying mental conceptual spaces, for instance: what are the dominant dimensions in people's mental representations of countries? Political alignment, economic development, religion, or geographic location? Asking subjects questions about those dimensions directly is problematic for many reasons. However, presenting pairs of countries and asking how similar they are is simple and does not bias participants towards certain answers. A productive technique for analyzing these empirically obtained similarity values has been Nonmetric Multidimensional Scaling (MDS, Kruskal, 1964). This method reconstructs the unobservable geometry of the conceptual space of countries from the pair-wise similarities. This is done by representing each country as a point on an *n*-dimensional map. Next, an iterative procedure optimizes the positions of the countries until their mutual distances on the map reflect the mutual similarities as specified by the participants of the experiment. It can then be tested if the dimensions of the map correspond to hypothesized dimensions such as political alignment. The same method can be used to reconstruct the unobservable geometry of scanpath space: if we interpret the similarities between scanpaths – normalized, as described above, to unit of time – as distances, we can fit maps on which each scanpath is represented as a point. On these maps distances between the points reflect the similarities of the corresponding scanpaths.

The goodness of fit of such maps can be quantified using a residual sum of squares called the stress of a map (c.f. Kruskal, 1964). Stress values are positive and small values are better, e.g. a stress <5% indicates a good fit, 10% is fair and 20% poor. Given a particular set of similarities, the stress also depends on the number of dimensions of the map. Higher dimensional maps have more degrees of freedom to place the items so that the similarities can be represented more accurately. The choice of the appropriate number of dimensions depends on several factors. One is the nature of the data. For instance, if the items are cities and the similarities their distances, two dimensions might be sufficient for reconstructing the map if the cities are not too far apart from each other. When the cities are scattered over the whole globe, two dimensions are not sufficient and high stress values will be the consequence. Stepping to a 3-dimensional map will then yield a precise representation and low stress values. However, in many applications it is not trivial to identify the appropriate number of dimensions. If the items are scanpaths, we do not know the underlying dimensionality of the data as in the case of locations on a sphere. In this situation, we have to select a number empirically so that we get a reasonably good fit of the map. A problem here is that for many data sets there exists some number of dimensions that allows an almost perfectly fitting map. However, if the dimensions to data points ratio is large, a map might not have a meaningful interpretation anymore. In this case the model represented by the map has too many degrees of freedom to be reliably estimated given a relatively small number of data points (c.f. Kruskal, 1964). Fortunately, in this investigation we get sufficiently good fits with only a few dimensions. Fig. 4 shows the map we get when applying MDS to the similarities obtained using Scasim. The stress of this map is 0.0071; in words: the variance in this synthetic data set with respect to spatio-temporal patterns is almost perfectly preserved by the map.[7]

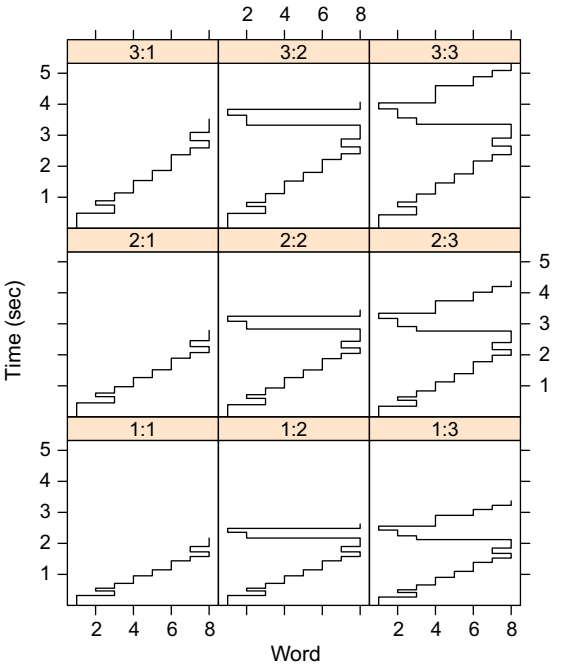

Fig. 3. Artificial scanpaths from a sentence reading task. Each reader produces a scanpath that involves performing a long regression, a short regression, and no regression at all. Readers differ in their reading speed: reader 1 is the fastest while reader 3 is the slowest.

Once we have such a map, we also have a vector-representation of scanpaths and almost the full array of statistical methods can be used to analyze the distribution of scanpaths in scanpath space. For example, a simple ap-

[6] From now on, we will, for convenience, talk about *similarities* although the numbers technically represent dissimilarities.

[7] We also tried Sammon's method for fitting maps instead of Kruskal's (Sammon, 1969). The Sammon maps look slightly different but the results of the analyzes of those maps were very similar.

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

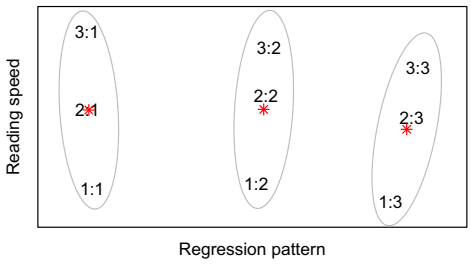

Fig. 4. A map of the artificial scanpaths shown in Fig. 3. Ellipses indicate clusters found using the *k*-means clustering procedure. The stars mark the locations of the centroids of these clusters. Variation in regression patterns is expressed along the *x*-axis, whereas the *y*-dimension captures differences in reading speed. The scanpaths are not located in a perfect grid because there was random variation in the fixation durations.

proach is to use the k-means clustering algorithm to check if there are subsets of scanpaths that have a distinctly increased mutual similarity compared to other subsets. Fig. 4 shows three such clusters that we identified using this method.

If there are strong sources of variance in a data set and if they are weakly correlated – as in our case reading speed and type of regression pattern – MDS organizes the items so that one dimension of the map is allocated to each source of variance. In other words, MDS identifies the latent dimensions inherent in the similarities. This way, most of the variance in the similarities can be preserved by the map. In our example the result is that reading speed is expressed along the vertical axis and type of regression pattern along the horizontal axis. By finding correlations of axis and properties of the scanpaths we can infer what properties are shaping the distribution of scanpath patterns.

The example demonstrates that this kind of analysis of scanpaths is sensitive to very different attributes of scanpaths: a regression pattern is a spatial property expressed locally, at a particular point in a scanpath, whereas reading time is a temporal property expressed globally, i.e. in every fixation.

Armed with this technique, we turn to a re-examination of the Meseguer et al. data.

3. A re-examination of the Meseguer et al. data set

Meseguer et al. recorded eye movements from 44 undergraduate students who read 48 experimental sentences intermixed with 96 filler items. Two sample sentences can be seen in Table 2. The experiment had a 2×2 design; one factor was the attachment site of an adverbial phrase (high or low), and the other factor manipulated the surface form of VP2. The second factor was included in order to check if phonological similarity of the VP2, which was always in subjunctive mood, influenced processing of the verb in the AdvP. This manipulation was made possible by the fact that in Spanish the subjunctive has two possible terminations. Since there was no interaction between the two factors we will not investigate the influence of the latter manipulation any further.

In half of the trials the participants had to answer true/false comprehension questions on which they performed with 92% mean accuracy rate. This indicates that they did not have any significant difficulty in comprehending the presented material. More information on the stimuli and the procedure is available from Meseguer et al. (2002).

Since we had to convert the data format used by Meseguer et al. to the format used by our software, we first calculated the means for the fixation durations reported by Meseguer et al. in order to verify that no errors were introduced in the process. We found the same pattern of results as Meseguer and colleagues report. From the full scanpaths we extracted regression patterns by selecting all fixations that followed the first regressive saccade after reading the disambiguating material (region 8). Therefore we are left with some regression patterns in which some earlier word was fixated (most of the time the disambiguating region), returned to region 9, and only then a long regression to early material occurred. While it is possible to treat these patterns as two separate regressions, we follow Meseguer et al. in assuming that the two subsequent regressions are not independent events and might in fact reflect one strategy or process. Therefore we treat them as one pattern.

In the data set that we obtained from Meseguer and colleagues, no coordinates are specified for fixations located outside the sentence. We removed these fixations because they cannot be analyzed with the approach used here. This affected 4.3% of the fixations most of which happened after reading of the sentence was finished. If a regression pattern had only one fixation left, we dropped the entire trial from the analysis. This resulted in the removal of 114 trials (8.3%).

When preprocessing eye movement data, one application of Scasim is the detection of trials in which something unusual has happened. These trials can be identified using two criteria. First, for each scanpath, calculate its average distance to all other scanpaths. A scanpath will be marked for deletion if this distance is more than two standard deviations larger than that of the other scanpaths. This criterion selects all trials that are obviously unusual: participant looks randomly around, reads the sentence several times, etc. This way however, we also select scanpaths that constitute the tail of a distribution. Since these do not qualify as outliers, we should retain them. Second, only those marked trials are dropped that have a distance to their nearest neighbor that is larger than two times the mean distance in the whole set. The whole procedure can be applied iteratively until it leads to no further change. In the data set examined here, this procedure selects 14 scanpaths. Although this is a small number, it can be advisable to remove them: they contribute, by definition, relatively large dissimilarity values, and therefore have a strong impact on the stress measure that guides map formation in Multidimensional Scaling. If we keep these outliers, the map might mainly depict differences between normal scanpaths and outliers and to a lesser extent differences among the normal scanpaths. Since the outliers most likely reflect processes unrelated to our investigation, it would be reasonable to exclude them from the data set.

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

Table 2
Sample item from the Meseguer et al. study. During the experiment, the line-breaks were in the same positions as in this table.

High attachment to VP1 (dispreferred)
El Profesor [VP1 *dijo* [CP *que los alumnos*
[VP2 *se levantaran del asiento*]] [AdvP *cuando los directores entraron en la clase.*]]
The teacher [VP1 said [CP that the students
[VP2 had to stand up from their seats]] [AdvP when the directors came INDIC into the room]]

Low attachment to VP2 (preferred)
El Profesor [VP1 *dijo* [CP *que los alumnos*
[VP2 se levantaran del asiento [AdvP cuando los directores entraran en la clase.]]]]
The teacher [VP1 said [CP that the students
[VP2 had to stand up from their seats [AdvP when the directors come SUBJ into the room]]]]

Regions:

El Profesor	dijo	que los alumnos						
1	2	3						
se levantaran	del asiento	cuando	los directores	entraran/entraron	en la clase.			
4	5	6	7	8	9			

However, in this investigation outliers were kept in the analysis because removing them did not change the results.

4. Results

4.1. Which fixation strategy do readers use when reanalyzing?

Participants executed on average 15 regressions from region 9 in the high-attachment condition and 14 in the low-attachment condition ($t(43) = 2.3$, $p < .05$). The average length of regressive saccades (measured in regions) from region 9 was 4.52 in the high condition vs. 4.51 in the low condition. The distribution of landing sites was not significantly different ($\chi^2 = 4.3$, $p = .75$, $df = 7$). This indicates that the effects of reanalysis are rather subtle, not strongly expressed in simple scalar measures, but buried in the spatio-temporal patterns formed by fixation sequences ensuing after the encounter of the disambiguating word. Meseguer et al. (2002) teased apart differences between the conditions using dependent measures tailored for this particular experiment. Scanpath similarity, however, provides a general solution for questions of the kind discussed here.

Since Selective Reanalysis predicts a characteristic fixation strategy for the garden-path condition, a straightforward way to test it is to check for clusters of regression patterns in the whole data set and to see whether (i) they are associated with one or the other condition and (ii) whether these patterns resemble a trajectory consistent with predictions of the Selective Reanalysis Hypothesis. The underlying reasoning is: clusters correspond to parsing strategies and strategies occurring more often in one condition address the peculiar problem presented by that condition, in this case reanalysis. Strategies that occur equally often in the two conditions cannot plausibly be attributed to reanalysis but might reflect other processes. Therefore, we are mainly interested in clusters of regression patterns that have more instances in the high-attachment condition.

All data analysis was done in GNU-R (R Development Core Team, 2009). Maps of scanpaths were fit on similarity per fixation scores using the function isoMDS from the package MASS. For the first cluster analysis, we fitted a 2-dimensional map of all 1253 regression patterns. The stress of this map was 22%. Clusters were detected on this map by fitting mixture of Gaussians models using entropy maximization (Fraley & Raftery, 2002). Calculation of the mixture of Gaussian models was performed using the mclust package (Fraley & Raftery, 2007). The optimal model cannot be chosen using only the likelihood of the candidate models because they differ in the number of free parameters, and adding free parameters can only improve the fit of a model. Hence, the model with the most parameters would always be the best. Therefore, a Bayesian information criterion (BIC) was used. It consists of the maximized log likelihood of a model minus a penalty for model complexity (Schwarz, 1978). This penalty is the number of parameters times the log of the number of data points. The free parameters of the cluster models were the position of Gaussians, their dispersion in each dimension, and their rotation. Only cluster models with 1 up to 20 clusters were examined. The BIC reached a maximum for relatively small numbers of clusters.[8] See Fraley and Raftery (1998, 2002) for details of the method.

[8] Although Fraley and Raftery (2002) propose a Bayesian information criterion for model selection, we also tried the Akaike information criterion (AIC, Akaike, 1974). This results in much larger numbers of clusters because then the penalty for free parameters is only two times the number of parameters. Fitting a cluster model for the 7-dimensional map of scanpaths (see below) with the AIC leads to a model with as many as 60 clusters (14 with BIC). Although this model gives us a much more detailed description of the eye movement phenomena occurring in the data set, we are running a very real risk of overfitting the data, i.e. many of those clusters might not reflect underlying populations but might merely be statistical flukes. Moreover, given the small numbers of members in those clusters (mean: 21, sd: 8.5), deriving statistically reliable statements about the influence of garden-pathing on those clusters seems futile. As we have no theoretical reasons to prefer the AIC we stick with the BIC. Still, using the AIC might be useful when exploring a set of scanpaths. In R, this can be achieved by overwriting the function `bic` from the package Mclust with a function that calculates the AIC.

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

If there is prior knowledge about the structure of cluster models, it might make sense to impose constraints on the set of possible models. For instance, one might restrict the search space to models in which all Gaussians have the same orientation, the same or even a predetermined dispersion. In our case, there is no such a priori knowledge and we allowed the parameters of the Gaussians to vary freely and independently.

The advantage of mixture of Gaussians modeling over other clustering procedures is that they can identify clusters even if they intersect or overlap. This is particularly important in this investigation because the Time-out Hypothesis predicts random walks, a fixation strategy that is not characterized by a common spatio-temporal fixation pattern. On a map of scanpaths this would result in a large cloud of scanpaths that would overlap other categories of scanpaths that do exhibit a common pattern. This means that by combining Scasim with mixture modeling, we can not only detect categories of similar-looking scanpaths but also heterogeneous categories in which the scanpaths are characterized by their mutual dissimilarity. A drawback of mixture modeling is, however, that we have to make assumptions about the distribution of scanpaths in scanpath space. In the present case, we assume that clusters are well-described by Gaussians, which may not necessarily be appropriate.

The optimal mixture model for the 2-dimensional map consisted of three clusters. From each cluster a prototype was selected; this was the pattern that maximized the similarity to all other members of the cluster, i.e. the one in the center of gravity of the cluster. Fig. 5 shows the map of scanpaths and Fig. 6 shows the prototype of each cluster. In cluster A, the eyes move back to the beginning of the sentence and then reread the whole sentence or parts of it.[9] In cluster B, the eyes perform a single saccade to early material, while in cluster C the gaze shifts to the nearby disambiguating region. Only cluster A had significantly more members from the garden-path condition (Cluster A: $\chi^2 = 5.2$, $p < .05$, $df = 1$, cluster B: $\chi^2 = 0.10$, $p = .75$, $df = 1$, cluster C: $\chi^2 = 0.36$, $p = .55$, $df = 1$).[10] However, this pattern frequently occurred in both conditions: 136 instances in the high-attachment condition vs. 101 in the low-attachment condition (cluster B: 310/302, cluster C: 208/196). The rank order of categories was B (612), C (404), A (237).

Regressions are common even during reading of simple sentences. An analysis of the Potsdam Sentence Corpus (Kliegl, Nuthmann, & Engbert, 2005), a German eye tracking corpus containing 144 simple everyday language sentences read by 222 readers each, shows that regressions occur frequently, 8.5% of the saccades are regressive, and that more than 80% of them are short and targeted at the directly preceding word. Models of oculo-motor control in reading predict these regressions as a consequence of premature forward saccades occurring before lexical access of the current word has finished. This raises the question: can we safely attribute regressions in cluster C to the particular syntactic phenomenon studied here? In order to answer this, we counted, for all regions, the occurrences of regressions to previous regions to determine whether region 9 had unusually many such regressions. Region 4 at the beginning of line 2 had only 28 instances. Apparently readers do not like to regress if that involves changing the line (c.f. Mitchell et al., 2008). Region 9 had 517 and the other regions between 200 and 330 instances, except region 1 which cannot have any regressions because there is nothing to regress to. While this suggests that region 9 had more than the usual rate of regressions to the previous region, this result is at odds with the above finding that cluster C was not modulated by garden-pathing. We might conclude that pattern C regressions were induced by the temporary ambiguity but do not necessarily reflect reanalysis.

A scanpath is potentially a complex thing that has many degrees of freedom. Although the stress of the two-dimensional map (22%) is relatively high, it is surprising that two dimensions are sufficient to explain a large part of the variance in the spatio-temporal fixation patterns. However, given the large number of available data points,

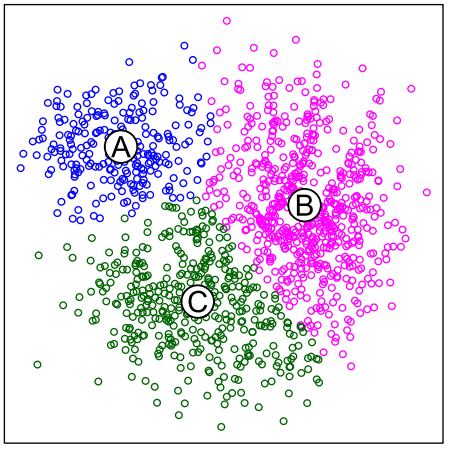

Fig. 5. Map of all regression patterns in the data set originating from region 9. Colors indicate clusters that were found using mixture of Gaussian modeling. The letters mark the positions on the centroids these clusters.

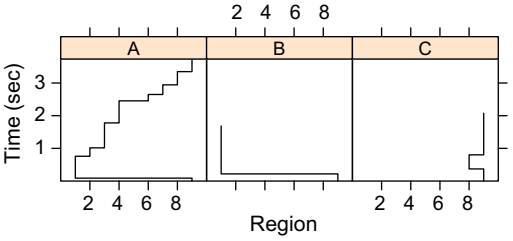

Fig. 6. The regression patterns that were closest to the centroids of the clusters identified on the 2-dimensional map of all regressions from region 9 (see Fig. 5). We call these the prototypical patterns for a cluster.

[9] It looks as if region 5 was skipped but this is not necessarily real skipping because region 5 (modification of V2) was not present in all items.

[10] Please note that these three χ^2-tests are not independent. If pattern A occurs more often in the garden-path condition, there are fewer trials left in which B and C could possibly occur. In fact, if you know the numbers of patterns A and B and the numbers of trials without any regression for the two conditions, the counts for pattern C are fully determined.

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

we can calculate robust maps for higher dimensions. We fitted maps for 2 up to 15 dimensions. For these maps we calculated cluster models using the procedure described above. Fig. 7 shows the stress of those maps and the number of clusters obtained as a function of the number of dimensions. As we add more dimensions, the maps represent more variance in regression patterns and more structure emerges that is conflated when only two dimensions are available. The number of clusters peaks at seven dimensions and only little additional variance is explained when more dimensions are added. A knee in the stress curve can hint towards the "true" dimensionality of the data (cf. Kruskal, 1964) but the curve in Fig. 7 does not offer any such clues. Therefore, we arbitrarily chose seven dimensions for further analysis in order to contrast our simple two-dimensional model with a more complex one. The main results come out similarly when we analyze maps with other numbers of dimensions. Fig. 8 shows projections of the seven-dimensional map and Fig. 9 shows the prototypical scanpaths of the detected clusters.

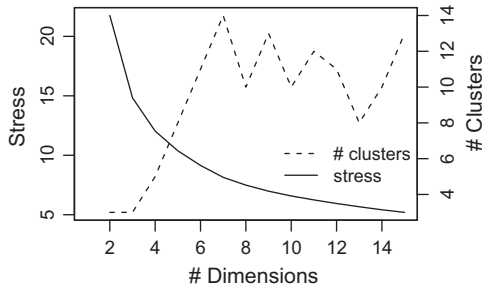

Fig. 7. Stress values and numbers of clusters for increasing numbers of map dimensions. As the number of dimensions goes up, the stress of maps decreases, i.e. more variance is explained by higher-dimensional maps. The number of clusters detected on those maps reaches a plateau at six dimensions.

In seven dimensions (stress: 8.1%), strategy A from the two-dimensional map is preserved as cluster D. Cluster B breaks up into clusters F–M which differ mostly in the precise landing position of the regressive saccade. Strategy C (inspection of the critical word in region 8) splits into N, O, and P. These differ in whether or not the gaze returns

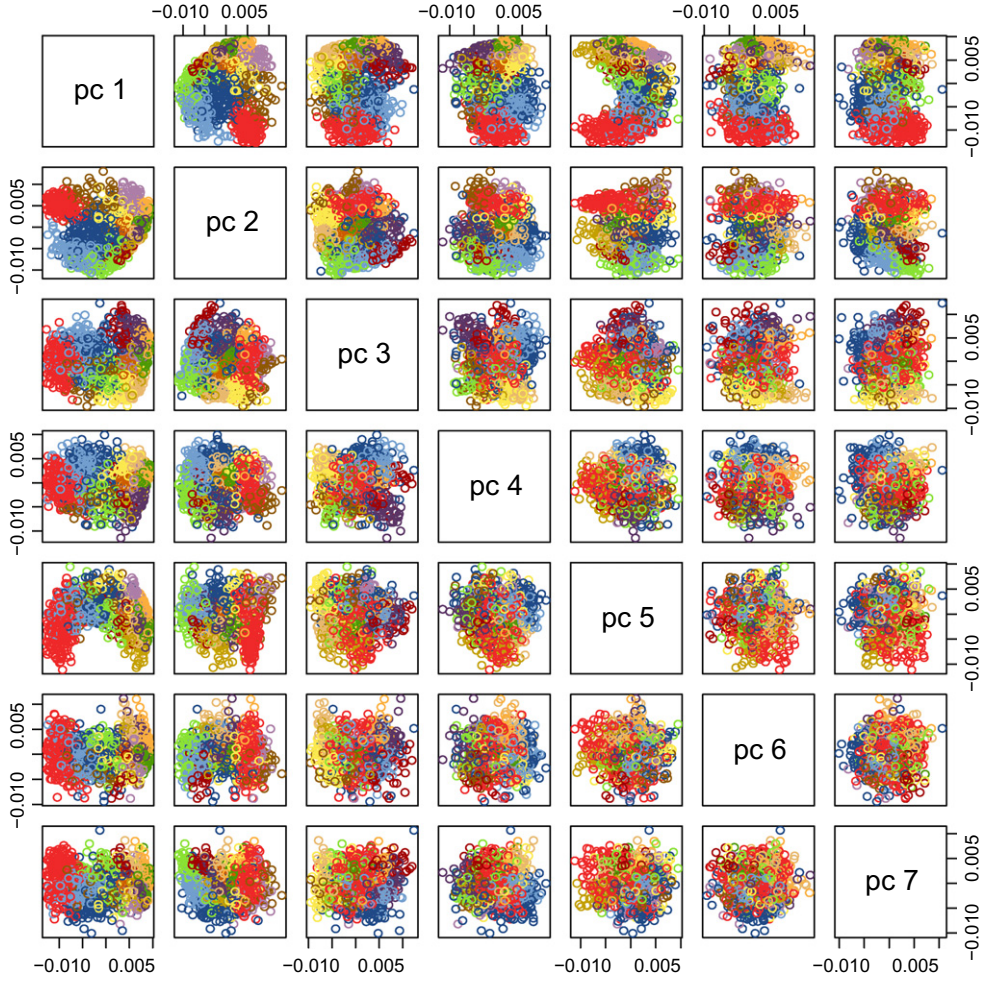

Fig. 8. Projections of the 7-dimensional map of all regressions from region 9. Color indicates cluster membership.

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

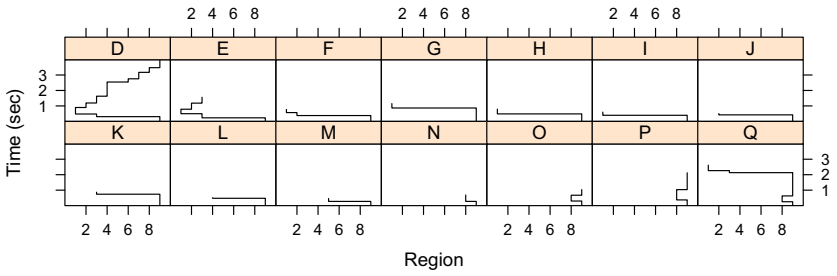

Fig. 9. Prototypical regression patterns of the clusters on the 7-dimensional map (Fig. 8).

to the origin of the regression and in how long region 8 is fixated. Two new categories that were not detected on the low-dimensional map are Q, a blend of B and C, and E in which only the first line (region 1–3) is reread. However, again only rereading of the whole sentence was significantly more frequent in the garden-path condition ($\chi^2 = 3.9$, $p < .05$, $df = 1$).

Fig. 10 shows a scatter plot of the sizes of clusters D–Q for the garden-path vs. the non-garden-path condition. The farther away a cluster is located from the central line, the more it had a tendency to occur in one of the two conditions.

We also performed a similar analysis of the 154 regression patterns that originated on region 8, the disambiguating word. On a 2-dimensional map (stress: 15.71%) two clusters were detected, one containing regressions to the previous word (size: 117) and one heterogenous set of very different patterns (size: 37). Neither the clusters nor the whole set of regressions from region 8 were modulated by garden-pathing. Apparently the scanpath effects of disambiguation show up only on the following region.

4.2. Do readers differ in their fixation patterns?

All regression patterns in the data set were classified according to the three main strategies: rereading (A), going to the beginning of the sentence (B), and rechecking of the disambiguating word (C). Fig. 11 shows the counts of instances of the main strategies that each reader contributed. For every pattern there are readers that had a preference for it. Many readers show a strong preference for a particular pattern, mostly B or A, while others produced all patterns equally often. Some readers did not use patterns A or C but pattern B was found in all participants.

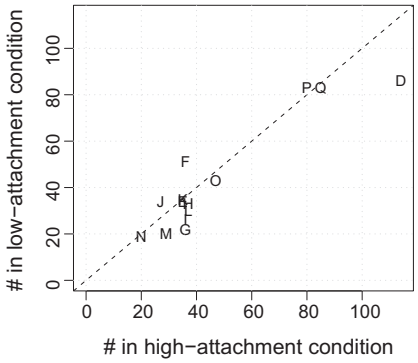

Fig. 10. Counts of the instances in each cluster on the 7-dimensional map in the high- vs. low-attachment condition. Pattern D had the most instances and the most reliable bias towards the high-attachment condition.

Out of 44 participants 24 had a distribution of patterns that was significantly different from the overall distribution across participants (χ^2, $df = 2$, $p < .05$). These are, of course, many tests, some of which might have resulted in false alarms. An examination of Spearman rank correlations of counts of patterns per participant revealed that patterns A and C ($\rho = 0.36$, $p < .05$) and patterns B and C ($\rho = -0.3$, $p < .05$) were significantly correlated but not patterns A and B ($\rho = -0.23$, $p = .13$). In other words, readers who produced more type A regressions produced more type C regressions, while readers who had more type C regressions had fewer type B regressions. Note that the occurrence of patterns A, B, and C are not independent events. If a reader performs many type A regressions, there are just not many trials left in which patterns B and C could occur. So if there is no functional relationship between patterns, we would expect a slightly negative correlation between patterns. This weakens the reliability of the correlation between B and C but strengthens that of the positive correlation of A and C.

When we did the same analysis for items instead of participants, there were only 4 items that elicited a distribution of patterns that was significantly different from the overall distribution. At an alpha level of 0.05, this is roughly the expected rate of false positives. We conclude that the items presented in this experiment largely elicited the same fixation patterns. This also constitutes further evidence that the inter-individual variation is not due to random fluctuations but systematic, which in turn shows that the signal extracted by our method is not spurious but constitutes genuine and interpretable variance in eye movements.

We also identified prototypical scanpaths for each participant. A cluster analysis of a 2-dimensional map of those 44 regression patterns resulted in the same three categories as the cluster analysis of the two-dimensional map of all scanpaths.

5. Discussion

Our main results are the following: in a 2-dimensional cluster model of regression patterns we found three fixation strategies (Fig. 6): regressions to the beginning of

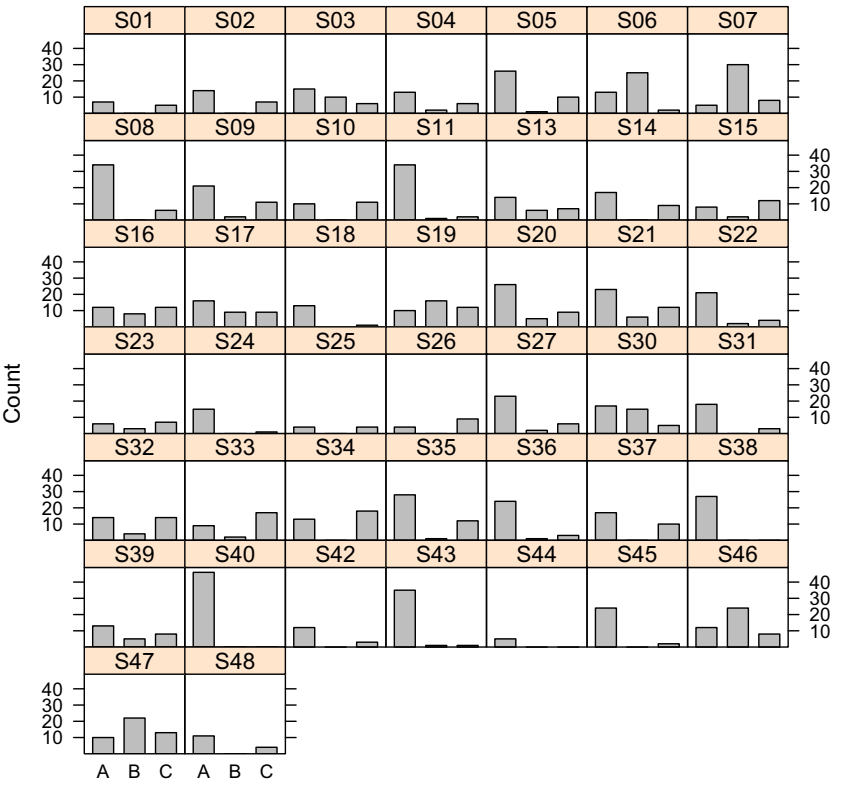

Fig. 11. Individual differences in regression patterns: The graph shows how many instances of the main patterns (A, B, and C, see Fig. 6) each reader produced. Many readers have a strong tendency to either A or B, whereas only few perform all patterns equally often.

the sentence followed by rereading (A), short regressions to beginning of the sentence (B), and short regressions to disambiguating material (pattern C). The same patterns were detected when we clustered only the preferred regression patterns of the participants. Only rereading occurred significantly more often in the garden-path condition. An analysis of a 7-dimensional model revealed more structure. 14 fixation patterns were identified (Fig. 9): short regressions to early, intermediate or late material, rereading of the first line (regions 1–3), and rereading of the whole sentence. Again, only rereading of the whole sentence was significantly modulated by attachment site.

The analysis of individual readers showed that there were strongly expressed differences in how participants orchestrated the different fixation strategies. Within participants, patterns A and C had moderately positive correlation while B and C were negatively correlated.

5.1. How do our results relate to the results of Meseguer et al. (2002)?

One of the main results of the original study by Meseguer et al. was that transitions from region 9 to the main verb (region 2) and to the adverb (region 6) were more likely in the high-attachment condition. We found four clusters of regressions that start with fixations to those regions (patterns E, F, G, and Q in Fig. 9). F and G had more instances in the garden-path condition (48 vs. 39, 50 vs. 39) but these differences were not significant according to Pearson's χ^2 tests. This apparent contradiction is explained by the fact that regression patterns that start with a fixation to region 2 or 6 diverge in the course of the following fixations. Consequently they do not form clusters of their own under our analysis.

One difficulty when testing Selective Reanalysis is that it is not a fully formalized theory that produces precise predictions for the eye movement record. Meseguer et al. considered the possibility of a specific fixation pattern that would ensue when the need to reanalyze arises. Mitchell et al. offered a weaker interpretation of Selective Reanalysis; according to them, relevant words exert some attraction on the eyes, but this does not entail a characteristic fixation signature because other properties of the text also influence the sequence of fixations. A signature scanpath, where certain words are fixated in a certain order, should be detectable with our method, which was designed to find such patterns. However, while we in fact found clearly separated patterns, none of them fit the description of Selective Reanalysis.

The weaker form of Selective Reanalysis would be harder to detect using our method, and therefore we argue that our results neither rule out nor support this variant of the idea. A certain amount of attraction has been shown to occur by Meseguer et al. and Mitchell et al. However, we

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

could provide clear evidence that there is a class of fixation sequences that is indicative of an alternative mechanism for dealing with garden-paths: rereading, presumably while inhibiting unwanted aspects of the original parse.

One obvious question is then: why did Meseguer et al. not find an increased number of regressions to region 1 in the garden-path condition? The simple explanation is that, in the rereading cluster, the eyes often undershoot on their way to the first word where rereading begins (see Figs. 6 and 9). For example, 37 regression scanpaths in the rereading cluster (A) first visit region 3 before shifting to region 1 (region 4 was visited only eight times in such regressive patterns). It seems that the eyes first change the line – there were two lines, and region 3 occurred at the end of line one – and then skip to the beginning of the line. The idea that the eyes initially undershoot receives further support from the fact that we see the same pattern in cluster B (short regression to the start of the sentence), where region 3 was the most frequent stepping stone on the way to region 1. This interpretation is consistent with Mitchell et al. (2008) insofar as they show an influence of the layout of the text, particularly of the line-breaks, on regressive fixation patterns. This suggests that saccade programming is indeed influenced by low-level properties of the text while a loosely coupled linguistic system is guiding the overall shape of the trajectory.

Does the fact that we found clear patterns entirely rule out the possibility, suggested by Mitchell et al. (2008), that the eyes perform a random-walk on previous material during reanalysis in order to buy time for the parser? Not entirely: while the detected categories of patterns occur quite often, there are also trajectories that fit less well into those categories. A more in-depth analysis is required in order to clarify whether these patterns constitute just the tails of the detected Gaussians, or whether they form a separate population that is not identified by a common spatio-temporal fixation pattern.

We see more pattern C regressions than we can explain by oculo-motor constraints. Hence, pattern C could be seen as indicative of a parser that operates in a fashion predicted by Selective Reanalysis: the parser, and the eyes, target a critical word whose status is decisive for the success of reanalysis. In this case, however, it is odd that this pattern is not modulated by the attachment site.

Our findings also show that (a) transition probabilities alone can render a misleading picture of what is going on in the individual trials, and (b) by taking the context of fixations into account, we can distinguish functionally different transitions between regions, for instance: transition to region 3 as an intermediate step on the way to region 1 and eye movements which aim for region 3 in the first place.

5.2. Implications for theories of parsing

As discussed by Frazier and Rayner (1982), two alternatives to Selective Reanalysis are forward reanalysis and backward reanalysis. According to forward reanalysis, the parser returns to the beginning of the sentence whenever an error is encountered and starts re-parsing it. Backward reanalysis assumes a stepwise and successive undoing of the parse from the current word to the previous one.

Assuming (following Frazier & Rayner, 1982) that the eye movement record reflects parsing processes, we found evidence showing that, consistent with forward reanalysis, rereading is indeed a strategy readers use to cope with garden-path material. Such a strategy is also theoretically well-motivated; for example, as Lewis (1998) points out:

> ...the backtracking strategy that places the least demands on memory is forward reanalysis by overt recomprehension from the beginning of the sentence. This strategy requires no memory for input or prior parsing states—it needs just enough memory for the parser to remember not to continue going down the same path. The drawback of this method is time, but it is a reliable strategy when all else fails.

Despite the relative efficiency of this approach, forward reanalysis seems unlikely to be the only or the default strategy in general parsing. After all, temporary ambiguities are almost ubiquitous, and rereading in all those cases would be very costly. However, Grodner, Gibson, Argaman, and Babyonyshev (2003) argue that re-parsing of earlier material using first-pass parsing processes might in fact be the only form of reanalysis. At least they show that it is sufficient to explain self-paced reading data from their own study and that reported in Sturt, Pickering, and Crocker (1999). A reasonable assumption is that the most common temporary ambiguities can be reanalyzed in-place, without the eyes having to reread the sentence. In the Meseguer et al. experiment, this was apparently often not possible and the alternative strategy was, we speculate, constructing a new parse from scratch. Priming and residual activation should, however, facilitate re-parsing of material that does not need revision.

The reason for the difficulty of in-place reanalysis might be that in this study the constituents affected by the reanalysis (V1, V2, AdvP) were scattered over a long sentence. This might have induced a significant amount of retrieval difficulty. Interesting questions for future research are therefore: under which circumstances is in-place reanalysis impossible so that the parser has to resort to rereading? Is the strength of the preference for one structure the key? Or the distance between the affected constituents? Or both?

We found no evidence for backward reanalysis, the second alternative to Selective Reanalysis. There are only rare cases of scanning backwards and they do not form a homogeneous cluster.

If we assume that rereading reflects reanalysis and furthermore that the parser deterministically selects a preferred structure during first-pass reading, rereading should only occur in the garden-path condition. However, none of the patterns that we identified occurs only in one condition. Hence, the frequent occurrence of rereading in the non-garden-path condition could mean two things: (a) although participants built the correct parse, they reread the sentence in order to reconfirm that they built the correct structure; (b) readers employ a parsing strategy that non-deterministically builds either structure during

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

first-pass reading of the ambiguous material; here, non-deterministically structure building is intended to mean that non-syntactic sources of information (such as lexical, semantic, and discourse factors) can differentially impact the decision about which structure to build.

The first possibility, (a) above, that participants mainly reread to make sure they can answer the comprehension question, could be tested with an extended experimental design that includes a baseline condition that does not have any temporary ambiguity: if readers just reread to check, they should also reread in the non-ambiguous condition. In the present study, we cannot exclude such an explanation.

Regarding the second possibility, (b), there is independent support for the idea that the parser is not restricted to only syntactic information when making a decision for a particular structure (see Tanenhaus & Trueswell, 1995, for a review). In the light of these findings it might seem plausible that garden-pathing would not only occur in the high-attachment condition. After all, the material used here had only a weak preference for low-attachment, which might give other factors a better chance to override a syntactic bias. The syntactic information available during the ambiguity might just not be very predictive of the true attachment site of the adverbial phrase.[11] A problem with this view is that it attributes variance in structure selection to differences of the experimental items (lexical, semantic, etc.). In the discussion above, however, we showed that the observed eye movement phenomena occurred at similar frequencies across items.

In sum, the rereading patterns suggest that readers may be carrying out forward reanalysis, but this is only one of several strategies readers use. Finally, determining the underlying reason for the adoption of forward reanalysis requires further study; it could be the result of a reconfirmation process, or it could be due to the parser engaging in non-deterministic parsing.

5.3. What can individual differences tell us about parsing?

Our analysis of individual differences in regression patterns shows that the preferred regression patterns of readers are categorizable into the three main classes that we also identified independently when analyzing the whole set of regressions irrespective of the readers that produced them. However, readers differ considerably in the degree to which they draw on those patterns. Rereading (pattern A), which is the only detected pattern that is clearly associated with reanalysis, is produced in more than two thirds of the trials by some readers while other participants did not produce this pattern at all. Occurrences of pattern A and C are positively correlated within participants, which could mean that they serve a similar purpose or at least that they are explained by a common cause. Since psychometric measures for the participants of this experiment are not available, we cannot tell at this point what these differences reflect. However, working memory seems to be a candidate that is worth testing. In any case, the pronounced inter-individual differences provide a potentially rich source of information about determinants in parsing and eye movement control and models of the related processes could gain a lot by seeking to explain them (c.f. Underwood, 1975).

5.4. The function of regressive eye movements

The statement of the Selective Reanalysis Hypothesis in Frazier and Rayner (1982) is not explicit with respect to whether regressive eye movements are driving the parsing process or if they are just epiphenomenal to parsing. The latter possibility is motivated by the following dilemma: after seeing the disambiguating material, if the parser already knows what the correct attachment site of the adverbial phrase is, what is the use then of looking at it? On the other hand, if this information is not available, how can the parser guide the eyes to the relevant material? What kind of information is the parser seeking? In the strict version of the Time-out hypothesis, the function of regressive eye movements seems to be to keep upcoming material from interfering with ongoing processing. However, it remains unclear whether and how the visual input afforded by regressive eye movements interacts with reanalysis.

Pattern A, rereading of the sentence, begins with a saccade to an intermediate region, which is perhaps better explained by undershooting of saccades and layout constraints á la Mitchell et al. (2008) than by a need of the parser to pick up information at these positions. During first-pass reading, visual input is necessarily driving the parsing process. Since the sequence of fixations in pattern A looks very much like first-pass reading, we speculate that similar parsing processes are at work during those regression paths: the parser is building structure guided by visual input.

Pattern B, regression to early material, was the most frequent pattern and is negatively correlated with pattern C. The correlation of B and A was also negative but not significant ($\rho = -0.23$, $p = .13$). This means that participants who had few instances of patterns A and C (regressing to the disambiguating material), which presumably reflect language processing, regressed more often to the beginning of the sentence and finished the trial once they arrived there. One possible explanation lies in a detail of the procedure of the experiment: after finishing the sentence, participants were not required to look to a corner of the screen in order to terminate the trial as is commonly done in many experiments. Instead they just had to press a button. Hence, at least some of the regressions to the beginning of the sentence might be spurious and reflect anticipation of the next sentence or comprehension question; readers who quickly grasp the sentence may be more likely to anticipate the next trial when they reach the end of the sentence.

Since transitions from region 9 to 1 in clusters A and B are distinguished by subsequent fixations, it would be very difficult to identify these negatively correlated classes of transitions if one were to only analyze transition probabil-

[11] A corpus analysis of the relative frequencies could shed light on this but this would require an annotated corpus of Spanish containing a large number of such temporary ambiguities; we could not obtain such a corpus for this study.

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004

ities. Hence, scanpaths seem to be a useful notion and object of investigation when researching questions related to strategies in cognitive processing.

The relatively large number of extra regressions to region 8, the disambiguating material, suggests that cluster C consists of at least two separate populations: regressions due to oculo-motor constraints and regressions triggered by processes addressing the resolution of the ambiguity. Since the difference between the two conditions is located in region 8 and consists of just one letter (*entraron* vs. *entraran*) it is likely that readers revisit this region to verify what they think they saw there.

In sum, we argue that the purpose of regressions in the Meseguer et al. experiment mainly was to seek out additional information required by the parser in order to finish the sentence: in one case, readers wanted to make sure they got the disambiguating verb right and it was easy to target because it was close. In the other case, restructuring of the derived parse might have demanded more resources than available (or the existing memory representation of the parse had deteriorated) and the parser had to resort to rereading from scratch while inhibiting aspects of the original parse. To which extent these results generalize to other experiments must be left for future work.

6. Conclusions

In previous studies (Frazier & Rayner, 1982; Meseguer et al., 2002; Mitchell et al., 2008), analyses of regressive saccades from the disambiguation region have shown that linguistically relevant material attracts the eyes during syntactic reanalysis. From these results it was concluded that the parser guides the eyes towards relevant material, suggesting an intelligent repair process. However, at least in the Meseguer et al. data set, the attraction effect was weak, and an attempt to identify a signature scanpath of syntactic reanalysis did not produce conclusive evidence. We developed a novel method for analyzing sequences of fixations that attacks the problem of spatio-temporal fixation patterns in a more direct way. Our re-examination of the Meseguer et al. data set using the proposed method showed that a substantial part of the regression trajectories triggered by the disambiguating material follow patterns that cannot plausibly be explained by Selective Reanalysis. These scanpath patterns were not detected using aggregates of the traditional eye-tracking measures, showing that our method provides novel information from eye movement data.

In one pattern, participants reread the whole sentence. This pattern was more frequent in the garden-path condition and is predicted by forward reanalysis, i.e., reanalysis by application of first-pass structure building processes (Frazier & Rayner, 1982; Grodner et al., 2003; Lewis, 1998). In a second pattern, the eyes regressed to the beginning of the sentence. We believe that this pattern was not related to reanalysis but reflects anticipation of the next trial. The third pattern, rechecking the disambiguating material, might be indicative of diagnosis (Fodor & Inoue, 2000). The frequent occurrence of the rereading pattern suggests that, for the material studied here, targeted repair of a parse was often not viable. The fact that all regressive fixation strategies were common in both experimental conditions suggests that the parser did not deterministically select the low-attachment interpretation of the sentence in the first pass.

This leaves us with the possibility of a parsing system that makes strategic decisions when it encounters the disambiguating material, the alternatives being: (i) Selective Reanalysis, which is neither supported nor ruled out by our results, (ii) re-parsing, and (iii) diagnosis presumably followed by covert reanalysis if necessary. The evidence for the parsing system making such decisions is that these strategies seem to exclude each other. Rereading either occurs immediately or not at all. There is no indication that it serves as a fall-back when other strategies failed.

Our analysis of individual differences showed that readers differ tremendously in how they orchestrate the various fixation strategies. Given the present data set, we can only speculate about the reasons of these differences, because psychometric measures were not available for the participants of the experiment. Theories of parsing and oculo-motor control might gain a lot by tapping into this rich source of variance.

Finally, we introduced a new dependent variable for eye movement research: Scasim. We demonstrated a general method for analyzing spatio-temporal patterns in eye movements based on this measure. Since the results of analyses of the proposed type consist of categories of fixation patterns and prototypical patterns, interpretation is straightforward and not prone to pitfalls associated with analyses of aggregates of eye-tracking measures. However, our measure is not intended to replace analyses of duration measures and transition probabilities but rather to complement them. The proposed method is not restricted to the analysis of regressions in reading, but is applicable in eye movement research in general because it comes without assumptions specific to reading. Other applications include evaluation of computational models of oculo-motor control (Reichle et al., 1998; Engbert, Nuthmann, Richter, & Kliegl, 2005; Reichle et al., 2009) and visual attention (Itti, Koch, & Niebur, 1998), analysis of eye movements in the visual world paradigm, diagnosis of reading impairments, and usability research. A package for the GNU-R system for statistical computing that implements Scasim, along with various convenience functions, is freely available on the web.[12]

Acknowledgments

We are grateful to Enrique Meseguer and colleagues for releasing their data. This paper has benefitted from comments by Reinhold Kliegl, Keith Rayner, Don Mitchell, one anonymous reviewer, and from audience members at various conferences (European Conference on Eye Movements 2007, 2009, and the CUNY conference on human sentence processing 2008, 2009, 2010). Comments may be sent to the first author at malsburg@gmail.com.

[12] http://www.ling.uni-potsdam.de/~malsburg/scasim.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research, 25*(17), 3389.

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*, 27–38.

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*. Amsterdam, Netherlands: Elsevier Science Ltd..

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

Daniel, P. M., & Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys. *The Journal of Physiology, 159*, 203–221.

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

Fodor, J. D., & Inoue, A. (2000). Garden path repair: Diagnosis and triage. *Language and Speech, 43*(3), 261.

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 98195-4322, USA.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal, 41*(8), 578–588.

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

Frazier, L. (1979). On comprehending sentences: Syntactic parsing strategies. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

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

Grodner, D., Gibson, E., Argaman, V., & Babyonyshev, M. (2003). Against repair-based reanalysis in sentence comprehension. *Journal of Psycholinguistic Research, 32*, 141.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.

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

Kliegl, R., Nuthmann, A., & Engbert, R. (2005). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology, 135*, 12–35.

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

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

Lewis, R. L. (1998). Reanalysis and limited repair parsing: Leaping off the garden path. *Journal of Psycholinguistic Research*, 247–284.

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

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*, 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.

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

Rayner, K. (2009). Eye movements in reading: Models and data. *Journal of Eye Movement Research, 2*(5), 1–10.

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*, 191–201.

R Development Core Team (2009). R: A language and environment for statistical computing [computer software manual], Vienna, Austria (ISBN 3-900051-07-0). <http://www.R-project.org>.

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

Rovamo, J., Virsu, V., & Näsänen, R. (1978). Cortical magnification factor predicts the photopic contrast sensitivity of peripheral vision. *Nature, 271*(5640), 54–56.

Salvucci, D. D. (1999). Mapping eye movements to cognitive processes. Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA, USA.

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

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers, 18*(5), 401–409.

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

Sturt, P., Pickering, M. J., & Crocker, M. W. (1999). Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language, 40*, 136–150.

Tanenhaus, M., & Trueswell, J. (1995). Sentence processing. In J. L. Miller & P. D. Eimas (Eds.). *Speech, language, and communication* (Vol. 1, pp. 217–262). San Diego: Academic Press.

Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *American Psychologist, 30*, 128–134.

Please cite this article in press as: von der Malsburg, T., & Vasishth, S. What is the scanpath signature of syntactic reanalysis?. *Journal of Memory and Language* (2011), doi:10.1016/j.jml.2011.02.004