View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Journal of Eye Movement Research

Journal of Eye Movement Research
2(1):4, 1-11

A multiple regression analysis of syntactic and semantic influences in reading normal text

Joel Pynte
Université Paris-Descartes, CNRS

Boris New
Université Paris-Descartes, CNRS

Alan Kennedy
University of Dundee and
Université Paris-Descartes, CNRS

Semantic and syntactic influences during reading normal text were examined in a series of multiple regression analyses conducted on a large-scale corpus of eye-movement data. Two measures of contextual constraints, based on the syntactic descriptions provided by Abeillé, Clément and Toussenel (2003) and one measure of semantic constraint, based on Latent Semantic Analysis, were included in the regression equation, together with a set of properties (length, frequency, etc.), known to affect inspection times. Both syntactic and semantic constraints were found to exert a significant influence, with less time spent inspecting highly constrained target words, relative to weakly constrained ones. Semantic and syntactic properties apparently exerted their influence independently from each other, as suggested by the lack of interaction.

Keywords: Reading, eye movements, Latent Semantic Analysis, syntactic constraint, semantic constraint.

Introduction

Although eye movements are routinely used as an index of syntactic processing in psycholinguistic experiments, syntax plays no explicit role in current models of eye movement control in reading (see Clifton, Staub & Rayner, 2007, for a discussion). A measure of ”predictability” (e.g., as assessed by the classical Cloze task, Taylor, 1953) is an important parameter in models like E-Z Reader (Pollatsek, Reichle, & Rayner, 2006) and SWIFT (Engbert, Nuthmann, Richter, & Kliegl, 2005), but this is treated as a property intrinsic to individual words; the contribution of sources of influence involving several words, and in the effect of syntax on inspection time has never been fully acknowledged. For many theorists, syntax is primarily seen as a source of occasional local disruption to the normal course of the reading process (e.g. when a syntatic difficulty is encountered). As Reichle, Pollatsek, Fisher and Rayner (1998) put it, ” these [syntactic and discourse] processes are usually too slow to be the usual signal to move forward and are better used as an occasional signal to stop lexical access and sort things out” (p. 150).

The present study was based on the assumption that, contrary to this view, syntactic processing exerts an immediate influence during reading. Since syntactic integration probably occurs on-line, the syntactic structure of the sentence in which a word is embedded, or some derived property associated with it, can be expected to modulate inspection time. Such on-line influence should be observed on a regular rather than occasional basis. Moreover, we are assuming that syntax directly contributes to predictability effects, and can thus be a source of processing facilitation (e.g., by increasing the probability that a given word will be encountered at a given position in a sentence). As noted by Frisson, Rayner and Pickering (2005), predictability is a composite factor, combining several sources of influence, ranging from discourse coherence to inter-word association, and it seems reasonable to assume that part of the facilitation brought by contextual constraints may be syntactic in nature.

In a recent multiple regression analysis conducted on the French part of the Dundee Corpus, the time spent inspecting a given target word was found to depend on the depth of its embedding in the syntactic structure of the carrying sentence: the deeper the embedding, the shorter the gaze duration (Pynte, New & Kennedy, 2007, in press). This effect was independent from the degree of semantic constraint exerted by the prior sentence fragment, as assessed by Latent Semantic Analysis (Landauer, & Dumais, 1997; Landauer,

We are grateful to S. Vasishth and an anonymous reviewer for helpful comments on an earlier version of the paper. This research was supported by grant No R560.805823 from the UK Economic and Social Research Council.

DOI 10.16910/jemr.2.1.4

This article is licensed under a
Creative Commons Attribution 4.0 International license.

ISSN 1995-8692

Foltz , & Laham, 1998). However, it remains unclear whether this facilitation effect is related to on-line syntactic integration. On the one hand, it could be argued that, for a given target word, the deeper its embedding, the more syntactic constraints will bear on it. From this point of view, the observed decrease in inspection time could be attributed to the on-line operation of syntactic constraints. On the other hand, it could equally be argued that deeply embedded words are frequently in a position of modifier. They are more likely to function as members of a prepositional phrase, an adjectival phrase, a relative clause, etc., and will be, by definition, less central to the main topic of the sentence than less deeply embedded words. For this reason, they may have received less attention, with less time devoted to semantic integration processes.

On-line syntactic influences have been examined for English, using the English part of the Dundee corpus (Demberg, & Keller, 2007), and for German, on the Potsdam sentence corpus (Boston, Hale, Kliegl, Patil, & Vasishth, 2008). In both studies the syntactic index employed was surprisal (Hale, 2001; 2006), a notion that relies on the total probability of parses disconfirmed at the moment of one transition from word n-1 to word n (see Hale, 2001, and Boston et al., 2008, for a formal account of the notion of surprisal). It should be noted that the surprisal effect reported by Boston et al. (2008) was expressed in terms of processing difficulty: the higher the surprisal value at a word (the more possible parses eliminated), the longer the inspection time. However, this can be reconciled with the notion that contextual constraint translates to processing facilitation if it is accepted that the higher the degree of contextual constraint, the lower will be the surprisal value associated with a given word. The more constraining the prior sentence fragment, the less will be the uncertainty regarding the way an incoming word can be integrated, and the lower the surprisal value associated with it. These results are thus consistent with the notion that syntactic effects may act to speed up reading on a word to word basis. Importantly, Boston et al., found an effect for measures of single-fixation duration and gaze duration, suggesting that syntax operates at a quite early stage in the word-recognition process.

A first aim of the present study was thus to model the effects of different syntactic indexes on measured eye-movements. Together with depth of embedding, each word in the Dundee Corpus was associated with a new measure of syntactic constraint[1], based on the proportion of possible continuations at the moment when the incoming word is to be integrated[2]. For example, in examples 1a-c there is an uncertainty as to whether the Noun Phrase (NP) and/or the Adjectival Phrase (AP) must be closed before integrating the incoming Preposition (P). The number of possible continuations is lower in examples 2b-c, where the only uncertainty concerns the closing of the NP. Another important factor concerns the probability of each continuation. For example, 2b is more probable than 2c, and its uncertainty is thus lower. By contrast, 1b and 1c are more or less equi-probable (because of the presence of the [AP] which makes the presence of an additional modifier in 1b less probable). Further details are provided in the Method section.

```
1a   [NP [AP     + P    ->     [NP [AP [PP
1b   [NP [AP     + P    ->     [NP [AP] [PP
1c   [NP [AP     + P    ->     [NP [AP]] [PP

2b   [NP     + P    ->     [NP [PP
2c   [NP     + P    ->     [NP] [PP
```

If depth of embedding is related to on-line syntactic integration, both measures can be expected to operate in the same way, in terms of temporal locus (early vs. late measure of visual inspection), class of words involved (content vs. function words) and direction of the effect (if any). As both depth of embedding and this new syntactic index are likely to vary as a function of the position of the target word relative to the beginning of the sentence, Position in sentence was also entered as a potential predictor of inspection time. Equally, a measure of the degree of semantic constraint exerted on a given target word was added to the model. Latent Semantic Analysis (Landauer, & Dumais, 1997; Landauer, Foltz ,& Laham, 1998) was used to estimate this. In the LSA framework, word meanings are represented as vectors in a high dimensional space, with the distance between two meanings expressed as a numerical value. Importantly, the meaning of a sequence of words, whatever its length, can also be represented as a vector in the same high dimensional space. This allows for the meaning of a sentence fragment to be directly compared to the meaning of a single word. The distance between the vector representing the target word and the vector representing the prior sentence fragment provides an estimate of the amount of semantic constraint exerted at the sentence level: the closer the two vectors, the more semantically constrained the target word.

The analyses were conducted in the linear-mixed effects model framework (e.g., Pinheiro & Bates, 2000). Two dependent variables were examined, namely single fixation duration and first-pass gaze duration (the

[1] In this paper, the notion of syntactic constraint is empirically defined as the number of ways that an incoming word can be integrated to the current structure

[2] This notion is also present in the definition of surprisal. If, in a transition from word n-1 to word n, we eliminate a huge probability mass of possible parses, then there will be only a few possible continuations, and surprisal will be high. However, it should be noted that the measure used in the present study only relies on how many ways there are for the current word to be integrated with the prior context. It does not take the probability mass of possible continuations into account.

This article is licensed under a
Creative Commons Attribution 4.0 International license.

sum of all fixations between the moment when the eyes entered the target word and the moment when they left it for the first time). Contextual influences were assessed by comparison to a baseline model comprising seven well known predictors of inspection time in reading (length, frequency, length and frequency of the prior word, size of the entering saccade, landing position, etc.). Indices of syntactic and semantic constraint were successively added as new predictors to the regression equation, and their contribution to the goodness-of-fit of the resulting models was tested.

Method

Materials

The analyses were conducted on the French part (52,173 tokens and 11,321 types) of the Dundee Corpus (Kennedy, Hill & Pynte, 2003) which is based on extended articles taken from the French language newspaper Le Monde. Over a number of testing sessions, ten French-speaking participants read the texts presented at a viewing distance of 500 mm from a display screen, five lines at a time. The set of articles presented to participants was selected from those used by Abeillé, Clément and Toussenel (2003) to construct their French tree-bank, and the syntactic-constraint indexes used in the present study were based on the syntactic descriptions provided by these authors.

Syntactic-constraint scores

An example sentence, with its associated (simplified) syntactic description is presented in the right part of Figure 1. Each word was subsequently associated with a vector (left part of the figure), corresponding to the eight sub-structures that may or may not be open when an incoming word is integrated into the current tree (Verbal Nucleus: VN, Relative Clause: Sr , Subordinate Clause: Ss, Preposition Phrase: PP, Noun Phrase: NP, Adjectival Phrase: AP, non-finite Clause: VP, Coordinated Phrase: CO). For example, the preposition au (14th word) is associated with the vector <0 1 0 1 1 0 0 0> , meaning that a relative clause, a PP and an NP are still open at that point. The vector associated with the previous word corresponds to the phrase structure of the prior sentence fragment (i.e. the set of phrases that have been opened), at the moment when the preposition is encountered, that is, the phrase structure into which the preposition must be integrated. In this example, integrating a preposition into <0 1 0 2 4 0 0 1> produces <0 1 0 1 1 0 0 0>, corresponding to the fact that a new PP had to be opened, and that that new PP was attached as a complement of the verb of the relative clause, which involved closing three NPs, two PPs and one COORD. The first syntactic index used in the present study (SYN hereafter) corresponds to the number of cases in the corpus where <0 1 0 2 4 0 0 1> + P indeed produced <0 1 0 1 1 0 0 0>, divided by the total number of occurrences of <0 1 0 2 4 0 0 1> + P, whatever the post-integration vector[3]. SYN thus provides an index of the degree of constraint exerted by the prior syntactic context upon the incoming word, given its syntactic category. If the prior sentence fragment is highly constraining, i.e., if there are only a few ways the incoming word can be integrated, most of the post-integration vectors will be identical to the one actually obtained, and the ratio will be close to 1. In contrast, if the incoming word can be integrated in many different ways, the specific post-integration vector obtained will be only one among many, and the ratio will be close to 0. In other words, SYN refers to how many possible ways there are (based on the Treebank) to integrate the incoming word. The higher the SYN ratio, the greater the certainty about how that word would be integrated to the current incomplete tree, with certainty ranging from 0 to 1.

Depth of embedding and position in sentence

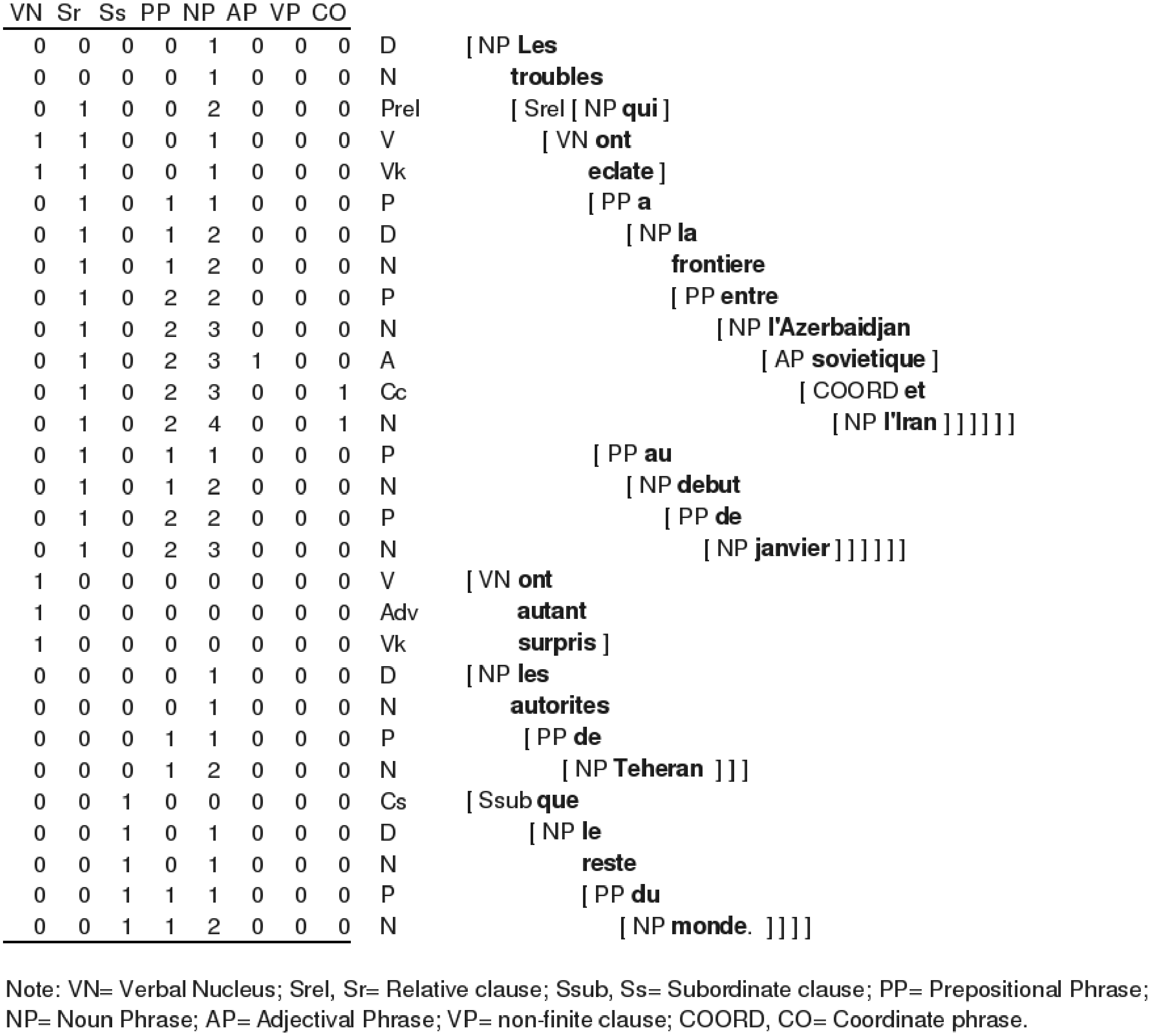
The depth of embedding of target words (defined here as the total number of open brackets at that point, EMB hereafter) is the second syntactic index to be examined in the present study. EMB was obtained by counting the total number of syntactic brackets open at that point in the sentence where the target word was encountered, minus the number of closing brackets (ending a constituent) encountered since the beginning of the sentence. In the example sentence presented in Figure 1, the depth of embedding of the preposition au would be defined as 3. Since both the SYN and EMB measures are likely to interact with the physical position of the target word, relative to the sentence beginning, a measure of the position of the target word in the sentence was included in the analysis as a control. This Position measure (POS hereafter) was obtained by simply counting the number of words separating the target word from the beginning of the sentence.

Semantic-constraint scores

Each word in the eye-movement corpus was also associated with a measure of the degree of semantic constraint exerted by the prior sentence fragment (SEM hereafter). A large corpus of novels (14.7 millions words) and film dialogues (16.6 millions words) was first used to obtain co-occurrence data (based on small paragraph) : if two words are present in the same paragraph, they may be thought to share some semantic property. The co-occurrence data were then reduced to a 300-dimension LSA space in which the semantic content of words and sentence fragments of the Dundee

[3] The syntactic descriptions used in the present study (Abeillé et al., 2003) were fully disambiguated, with only one attachment site provided at any given position. As a consequence, there was only one possible vector representation at any particular stage in sentences.

This article is licensed under a
Creative Commons Attribution 4.0 International license.



VN	Sr	Ss	PP	NP	AP	VP	CO		
0	0	0	0	1	0	0	0	D	[NP **Les**
0	0	0	0	1	0	0	0	N	**troubles**
0	1	0	0	2	0	0	0	Prel	[Srel [NP **qui**]
1	1	0	0	1	0	0	0	V	[VN **ont**
1	1	0	0	1	0	0	0	Vk	**eclate**]
0	1	0	1	1	0	0	0	P	[PP **a**
0	1	0	1	2	0	0	0	D	[NP **la**
0	1	0	1	2	0	0	0	N	**frontiere**
0	1	0	2	2	0	0	0	P	[PP **entre**
0	1	0	2	3	0	0	0	N	[NP **l'Azerbaidjan**
0	1	0	2	3	1	0	0	A	[AP **sovietique**]
0	1	0	2	3	0	0	1	Cc	[COORD **et**
0	1	0	2	4	0	0	1	N	[NP **l'Iran**]]]]]]
0	1	0	1	1	0	0	0	P	[PP **au**
0	1	0	1	2	0	0	0	N	[NP **debut**
0	1	0	2	2	0	0	0	P	[PP **de**
0	1	0	2	3	0	0	0	N	[NP **janvier**]]]]]]
1	0	0	0	0	0	0	0	V	[VN **ont**
1	0	0	0	0	0	0	0	Adv	**autant**
1	0	0	0	0	0	0	0	Vk	**surpris**]
0	0	0	0	1	0	0	0	D	[NP **les**
0	0	0	0	1	0	0	0	N	**autorites**
0	0	0	1	1	0	0	0	P	[PP **de**
0	0	0	1	2	0	0	0	N	[NP **Teheran**]]]
0	0	1	0	0	0	0	0	Cs	[Ssub **que**
0	0	1	0	1	0	0	0	D	[NP **le**
0	0	1	0	1	0	0	0	N	**reste**
0	0	1	1	1	0	0	0	P	[PP **du**
0	0	1	1	2	0	0	0	N	[NP **monde**.]]]]

Note: VN= Verbal Nucleus; Srel, Sr= Relative clause; Ssub, Ss= Subordinate clause; PP= Prepositional Phrase; NP= Noun Phrase; AP= Adjectival Phrase; VP= non-finite clause; COORD, CO= Coordinate phrase.

Figure 1. Vector representation of an example sentence.

Corpus could be represented as vectors. The cosine of the angle between the vector associated with a given target word in the eye-movement corpus and the vector associated with the prior sentence fragment was then computed (the higher the value, the more similar the meanings). All words, at all steps of the procedure were submitted to lemma transformation. Unlike in our prior study, the context taken into account for the measure of the SEM index consisted of all the words located between the previous sentence terminator (e.g., full stop, question mark, etc) and the target word, including the immediately prior word. SEM scores, as all other Independent Variables (except those expressed in terms of numbers of characters) were log transformed.

Selection criteria

For selection in the present analyses, a word (word n or target word hereafter) must have been reached by a saccade launched from the immediately prior word. Indeed, as predictability effects are likely to affect both parafoveal processing and skipping probability (Ballota, Pollatsek, & Rayner, 1985), it was important to equalize possible preview benefit in order to obtain reasonably stable data. Function words were excluded from the Latent Semantic Analysis: SEM scores are of little interest in the case of high-frequency function words such as determiners, prepositions, pronouns, etc., simply because such words can be found in any context.

This article is licensed under a Creative Commons Attribution 4.0 International license.

Procedure

The influence of syntactic and semantic constraints on single-fixation and first-pass gaze duration was assessed via a series of regression analyses. Measures of contextual constraint were successively added to a baseline model comprising a set of predictors known to influence inspection time. The contribution of these contextual properties to the goodness of fit of the model was evaluated. The analyses were conducted in the linear-mixed effects model (lme) framework, using the lme4 package (Bates, 2007) in the R system for statistical computing (R Development Core Team, 2006). Both readers and words were treated as random factors. Syntactic-constraint, semantic-constraint and position effects were estimated as varying across readers.

Baseline Model

In addition to the length and frequency of the target word and its prior word, the baseline model comprised three predictors whose purpose was to account for variation in inspection time arising as a function of landing position and preview benefit. These were: the size of the saccade entering the word, its relative landing position (landing position divided by word length), and the square of this latter measure (quadratic trend). To maintain compatibility with previous analyses (Pynte & Kennedy, 2006; 2007), measures of lexical frequency were based on the texts used in the Dundee Corpus and were submitted to log transformation.

Correlation between predictors

The correlation matrix is provided as an Appendix. The highest values (+.58 and +.55 for content and function words, respectively) are observed for Position and EMB. This corresponds to the fact that depth of embedding almost inevitably increases with the number of words encountered. It is important to note, however, that the SYN index does not correlate with POS (or any other predictor) as EMB does.

Results

We start with the analysis of content words, and first of all with a brief description of a baseline model, comprising seven well known predictors of inspection time in reading. The corresponding regression equation is subsequently enriched by successively adding new predictors, to examine the contribution of contextual constraints[4].

Baseline model

In the baseline model, the time spent inspecting a given target word is accounted for in terms of its own length and frequency, the length and frequency of the prior word, the size of the incoming saccade, the landing position of this saccade (relative to word length) and the square of this latter measure. Increasing word length by 1 character led to a 4.3 ms increase in single-fixation duration and to a 13.4 ms increase in gaze duration, t=20.73 and 47.49, respectively1. Each log frequency unit increment lead to a 5 ms decrease in single-fixation duration and to a 7.9 ms decrease in gaze duration , t = -14.63 and -16.31 respectively. A spillover effect of prior-word frequency was also present, with regression coefficients of -2.3 and -2.8, for single-fixation and gaze durations respectively, t = -8.17 and -7.25. A spillover effect of prior-word length was present for the measure of single-fixation duration (B= 1.11, t =4.89), but not for gaze durations (t = -.83, n.s.). Both the linear and quadratic trends of landing position produced significant effects (t = 10.59 and -6.09, respectively for single-fixation duration; t = -20.98 and 20.60, respectively for gaze duration), thus confirming the role of landing position as a major determinant of visual inspection time. Longer gaze durations were also associated with longer incoming saccades, with a 1.1 ms increase in single-fixation duration and a 0.6 ms increase in gaze duration for each character increase in saccade size, t = 10.83 and 4.85, respectively.

Contextual influences

Contextual properties were subsequently added to the regression equation. As syntactic and semantic influences may vary as a function of position in the sentence (that is, whether the target word was near to or far away from the sentence beginning), POS was included first. An improvement in the goodness of fit was obtained for the analysis of gaze durations ($\chi^2(2)$ = 7.70, p=0.02), but not for the analysis of single fixation durations ($\chi^2(2)$ = 1.88, n.s.). Subsequent steps included EMB, SYN and SEM successively, in that order. An improvement in the goodness of fit of the model was obtained at each step of the analysis: $\chi^2(2)$ = 4.95 and 20.38; 10.79 and 15.67; and 8.84 and 88.62, for single-fixation and gaze durations respectively, $p<.03$.

The resulting model is presented in Table 1. The regression coefficients presented in this table are for the model after inclusion of POS, EMB, SYN and SEM. The values for the seven predictors of the baseline model may thus be slightly different from the description provided in the previous section. As can be seen in Table 1, SEM affected both single-fixation and gaze durations. Each log increment decreased single-fixation duration by 2.1 ms and gaze duration by 8.7 ms, t = -2.97 and -4.11, respectively. In contrast, EMB produced a significant effect for gaze duration only, with a regression coefficient value of -5.9, t = -4.15. As for SYN, no main effect showed up. The influence of this latter predictor varied as a function of the position of the target word

[4] Similar results were obtained, whether or not the dependent variables were log transformed. The no-transformation analyses are reported here.

DOI 10.16910/jemr.2.1.4

This article is licensed under a
Creative Commons Attribution 4.0 International license.

ISSN 1995-8692

in the sentence. A significant POS-by-SYN interaction was present for both single fixation duration (t =2.64) and gaze duration (t = 2.90). Adding the corresponding interaction term to the regression equation improved the goodness of fit of the model for both measures, χ^2 (2) = 6.98 and 8.39, respectively, p<0.003. As can be seen in figure 2, the influence of the SYN index is no longer visible for words located close to the end of the sentence. The independence of syntactic influence, relative to semantic relatedness was also tested. Neither the corresponding EMB-by-SEM nor SYN-by-SEM interaction terms reached significance (t = -1.08 and 0.81, respectively for single-fixation duration; t = -1.67 and 1.29, respectively for gaze duration)

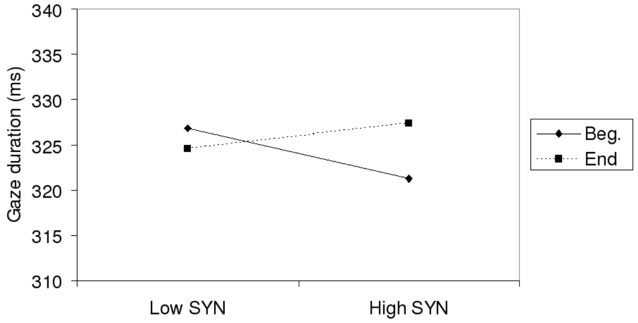


Figure 2. SYN effect for gaze duration as a function of position in sentence.

Table 1
Regression coefficients with associated standard errors from the analysis of content words.

	Variance	
Random effects	Single fixation	First-pass Gaze
itm (Intercept)	362.24	714.12
sub (Intercept)	215.87	808.20
sub Sem	0.00	34.85
sub Syn	6.78	16.53
sub Pos	0.00	0.02
Residual	4576.50	11551.00

	Estimate (Std. Error)	
Fixed effects	Single fixation	First-pass Gaze
(Intercept)	195.86 (5.14)	324.70 (9.40)
Saccade	1.10 (0.10*)	0.65 (0.14*)
Landing	79.99 (7.54*)	-203.19 (9.73*)
Landing^2	-38.12 (6.22*)	168.47 (8.22*)
Freq. n-1	-2.25 (0.28*)	-2.65 (0.39*)
Length n-1	1.10 (0.23*)	-0.22 (0.31)
Frequency	-5.23 (0.35*)	-8.65 (0.49*)
Length	4.49 (0.21*)	13.81 (0.29*)
Pos	0.06 (0.05)	0.18 (0.08*)
Emb	-1.98 (1.25)	-5.86 (1.41*)
Syn	-0.83 (1.05)	-1.68 (1.57)
Sem	-2.05 (0.69*)	-8.70 (2.12*)

	Estimate (Std. Error)	
Interactions	Single fixation	First-pass Gaze
Emb:Pos	0.02 (0.06)	0.07 (0.08)
Syn:Pos	0.16 (0.06*)	0.25 (0.09*)
Sem:Pos	-0.06 (0.06)	-0.07 (0.09)
Emb:Sem	-1.41 (1.30)	-3.06 (1.83)
Syn:Sem	0.67 (0.83)	1.54 (1.19)

Note: Asterisks correspond to significant effects (t>2)

Function words

The analyses presented so far only concerned content words. In this section we present the equivalent data for function words (mean length = 3.45 characters, sd = 2.16), restricted to the analysis of the baseline model and syntactic influence. As already mentioned, the method used for assessing semantic constraints, namely Latent Semantic Analysis, is not appropriate for function words. The results are summarized in Table 2. Regarding the baseline model, it should be noted that, apart from landing position and prior-word length, all properties elicited significant effects in the same direction as those reported for content words. Depth of embedding did not lead to any improvement in the goodness of fit of the model, χ^2 = 0.36 and 3.43, n.s. for single-fixation and gaze durations, respectively. By contrast, adding SYN to the regression equation did improve the model for single-fixation duration , χ^2 = 6.31, p = 0.043, each log increment corresponding to a 2 ms decrease in single fixation duration, t = -2.52 (There was no interaction with Position-in- Sentence). No improvement was obtained for gaze duration, however, χ^2 = 3.43, n.s. A similar analysis was conducted for first-fixation durations (i.e., including both single fixations and first fixations obtained in the several-fixation case). The results obtained in the single-fixation analysis were replicated. Again, SYN improved the goodness of fit of the model, χ^2 = 6.07, p = 0.048, each log increment corresponding to a 1.9 ms decrease in first-fixation duration, t = -2.47.

Discussion

The degree of contextual constraint exerted on a given target word is known to affect its reading time. Numerous eye-tracking experiments have shown first fixation and gaze durations to vary as a function of target word predictability , e.g., as assessed by the classical Cloze task (Altarriba, Kroll, Sholl, & Rayner, 1996; Ashby, Rayner, & Clifton, 2005; Ballota, Pollatsek, & Rayner, 1985; Binder & Rayner, 1998; Calvo & Meseguer, 2002; Drieghe, Brysbaert, Desmet, & De Baecke, 2004; Erlich & Rayner, 1981; Inhoff, 1984; Kliegl, Grabner, Rolfs, & Engbert, 2004; Kliegl, Nuthmann, & Engbert, 2006; Lavigne, Vitu, & dYdewalle, 2000; Rayner, Ashby, Pollatsek, & Reichle, 2004; Rayner

This article is licensed under a
Creative Commons Attribution 4.0 International license.

Table 2
Regression coefficients with associated standard errors from the analysis of function words.

	Variance	
Random effects	Single fixation	First-pass Gaze
itm (Intercept)	765.04	856.78
sub (Intercept)	175.09	291.00
sub Syn	0.00	1.71
sub Emb	0.00	19.83
sub Pos	0.02	0.01
Residual	4751.70	6974.30

	Estimate (Std. Error)	
Fixed effects	Single fixation	First-pass Gaze
(Intercept)	182.35 (6.72)	230.06 (7.81)
Saccade	1.92 (0.14*)	1.91 (0.17*)
Landing	24.44 (16.69)	-97.42 (18.08*)
Landing^2	-3.05 (12.02)	87.03 (13.12*)
Freq. n-1	-1.66 (0.46*)	-1.85 (0.53*)
Length n-1	-0.28 (0.32)	-0.36 (0.37)
Frequency	-3.53 (0.69*)	-2.64 (0.78*)
Length	3.75 (0.65*)	10.06 (0.70*)
Pos	0.00 (0.08)	-0.01 (0.09)
Emb	-0.51 (1.55)	-0.76 (2.26)
Syn	-2.05 (0.81*)	-1.29 (1.02)

	Estimate (Std. Error)	
Interactions	Single fixation	First-pass Gaze
Emb:Pos	0.00 (0.10)	0.11 (0.11)
Syn:Pos	0.04 (0.08)	0.08 (0.09)

Note: Asterisks correspond to significant effects (t>2)

& Well, 1996).

As noted by Frisson, Rayner and Pickering (2005), participants in a Cloze task may rely on several sources to complete a sentence fragment, and disentangling these various sources of influence over measured predictability is an important issue that has recently emerged in the reading literature. The present study, conducted on the French part of the Dundee Corpus, was based on the assumption that syntactic constraints constitute one such source of influence, and that syntactic integration processes can thus be associated with on-line facilitation effects.

The influence of syntactic and semantic constraints on inspection time, together with the possible contribution of the position of the target word relative to sentence beginning and the depth of its embedding in the syntactic structure, were assessed in a series of regression analyses. These four properties were successively added to a baseline model comprising a set of properties whose impact on visual inspection is well documented in the literature. Separate analyses were conducted for two dependent variables, namely single-fixation and gaze durations, and for two classes of words (content vs. function words). The study addressed two questions. First, whether there are effects of syntactic structure on eye movements that can be dissociated on the one hand from effects of semantic constraints and on the other hand from the measure of depth of embedding. Second, how early in processing terms do these properties have their effect? Our assumption was that different sources of influence might operate in different ways in terms of their temporal locus (early vs. late measure of visual inspection), and class of word affected (content vs. function words).

Syntactic constraints (SYN), depth of embedding (EMB) and semantic relatedness (SEM) were all found to exert a significant influence (see details below), with less time spent inspecting highly constrained target words, relative to weakly constrained ones. There was no statistically reliable evidence for an interaction. Moreover, semantic and syntactic constraints could not be differentiated as early or late in processing terms. These issues, therefore, must remain open for now. The SEM index affected both single-fixation and gaze durations. It may be important to note that SYN also exerted an influence (via an interaction with POS) on both measures in the analyses involving content words.

The finding that both semantic and syntactic effects kick in even for single-fixation durations suggests that there is little motivation for a sequential separation, distinguishing syntactic structure building and semantic processing in terms of the point in time at which each operates. Rather, the finding is consistent with the view that syntactic and semantic operations proceed more or less concurrently. From this point of view, our results seem inconsistent with the syntax-first idea pursued by Frazier and colleagues (see Frazier & Clifton, 1998), and more in line with a constraint-satisfaction approach in which all sources of information apply simultaneously in the course of online parsing. This result must be treated with caution, however. The SEM index used in the present study did not control for the degree of semantic relatedness between the target word and the word located immediately to its left. In a previous study (Pynte, New & Kennedy, in press), this factor was found to be responsible for part of an observed facilitation effect, possibly via some kind of inter-lexical priming mechanism.

Another important aspect of our results concerns the distinction between content and function words. Although SYN exerted its influence on both classes of word, a significant main effect was only obtained for function words. The effect was restricted to words located at the sentence beginning, in the case of content words. This pattern of results might reflect different parsing operations. The fact that function words were sensitive to the SYN index seems consistent with the importance of these words in determining syntactic function and attachment sites. The suggestion that function and content words might be submitted to different types of processing operation during reading has

DOI 10.16910/jemr.2.1.4
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

emerged from the extensive work on letter detection (letter-detection errors occur disproportionately on frequent function words, see Greenberg, Healy, Koriat & Kreiner, 2004, for a discussion). For example, Koriat and Greenberg (1994) argued that function words are monitored on the basis of a shallow and rapid initial analysis that paves the way for the semantic integration of content words, in line with the syntax-first approach. Further investigations will clearly be necessary in order to disentangle the various sources of semantic and syntactic influence at work during reading.

Regarding the SYN vs. EMB contrast, a first important difference concerned the class of words affected. A significant EMB effect was only found for content words. In contrast, as already mentioned, SYN exerted its influence on both classes of word, with a significant main effect in the analysis of single fixations for function words and a significant interaction with POS in the analyses of single-fixation and gaze durations for content words. SYN and EMB also contrasted regarding the moment in time when an influence could be observed. The EMB effect for content words was only apparent in the analysis of gaze durations, which suggests that this predictor tapped primarily late integration processes. By contrast, SYN exerted its influence as early as on first fixation durations, suggesting an immediate influence of parsing operations.

One possible interpretation of the EMB effect might be found in the effort required to maintain memory traces of ”open” maximal projections (i.e., an account in terms of ”storage cost”). The EMB value at a word is indeed higher if there are more open brackets (maximal projections), but the interpretation seems at odd with the direction of the observed effect (the greater the EMB value, the faster the gaze duration). An alternative interpretation can be proposed in terms of reading strategy. As mentioned in the introduction, deeply embedded words are frequently in a position of modifier. They are more likely to function as members of a prepositional phrase, an adjectival phrase, a relative clause, etc., and will, by definition, be less central to the main topic of the sentence than less deeply embedded words. For this reason, they may receive less attention, with less time devoted to semantic integration processes.

To summarize, both semantic and syntactic constraints were found to exert an immediate and independent influence on inspection time, suggesting that the eye-movement control system may be sensitive to both semantic-integration and syntactic-parsing processes. It is important to note that all the effects mentioned above correspond to a decrease in the time spent inspecting the target word, associated with an increase of contextual constraints. This is in line with the literature on predictability effects, and suggests that the semantic and syntactic indexes used in the present study capture part of contextual constraints responsible for predictability effects. This is not to say that syntactic parsing and semantic integration may not be a source of processing difficulty. Our suggestion is that contextual constraints, including syntactic constraints, usually function as a source of facilitation (e.g. by increasing the probability of occurrence of a given word at a given position in the sentence), although inhibition (e.g. increased inspection time) may occasionally occur, for example when a specific difficulty is encountered.

References

Abeillé, A., Clément, L., & Toussenel, F. (2003). *Building a treebank for French*. In A. Abeillé (Ed.) Treebanks, Kluwer, Dordrecht.

Altarriba, J., Kroll, J., Sholl, A., & Rayner, K. (1996). The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. *Memory & Cognition, 24*, 477-492.

Ashby, J., Rayner, K., & Clifton, C. (2005). Eye movements of highly skilled and average readers: differential effects of frequency and predictability, *The Quarterly Journal of Experimental Psychology, 58A*, 1065-

Balota, D. A., Pollatsek, A. & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology, 17*, 364-390.

Bates, D.M. (2007). *lme4: Linear mixed-effect models using S4 classes*. R package version 0.995-2.

Baayen, R.H. (2008). *Analyzing linguistic data: A practical introduction to statistics*. Cambridge: Cambridge University Press.

Binder, K., & Rayner, K. (1998). Contextual strength does not modulate the subordinate bias effect: Evidence from eye fixations and self-paced reading. *Psychonomic Bulletin & Review, 5*, 271-276.

Boston, F.M., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing cost as predictors of reading difficulty: an evaluation using the Potsdam sentence corpus. *Journal of Eye Movement Research, 2*, 1-12.

Calvo, M.G., & Meseguer, E. (2002). Eye movements and processing stages in reading: relative contribution of visual, lexical and contextual factors. *The Spanish Journal of Psychology, 5*, 66-77.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movement in reading words and sentences. In R. V. Gompel, M. Fisher, W. Murray, & R.L. Hill (Eds.), *Eye Movements: A Window in Mind and*

DOI 10.16910/jemr.2.1.4
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

Brain (pp. 341-372). Elsevier.

Demberg, V., & Keller, F. (2007). *Eye-tracking corpora as evidence for theories of syntactic processing complexity*. Cognition, submitted.

Drieghe, D., Brysbaert, M., Desmet, T., & De Baecke, C. (2004). Word skipping in reading: On the interplay of linguistic and visual factors. *European Journal of Cognitive Psychology*, 16, 79-103.

Ehrlich, S.F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading.*Journal of Verbal Learning and Verbal Behavior, 20*, 641-655.

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

Frazier, L., & Clifton, C. (1998). *Construal*. Cambridge, MA: MIT press.

Frisson, S., Rayner, K. & Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements.*Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 862-877.

Greenberg, S.N., Healy, A.F., Koriat, A., & Kreiner, H. (2004). The GO model: A reconsideration of the role of structural units in guiding and organizing text on line.*Psychonomic Bulletin & Review, 11*, 428-433.

Hale, J. (2001) A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*. Pittsburgh, PA.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science, 30*(4). 609-642.

Inhoff, A.W. (1984). Two stages of word processing during eye fixations in the reading or prose.*Journal of Verbal Learning and Verbal Behavior, 23*, 612-624.

Kennedy, A., Hill, R., & Pynte, J. (2003). *The Dundee corpus*. Poster presented at ECEM12: 12th European Conference on eye movements., Dundee, August 2003.

Kliegl, R. (2007). Towards a perceptual-span theory of distributed processing in reading: a reply to Rayner, Pollatsek, Drieghe, Slattery, & Reichle (2007). *Journal of Experimental Psychology: General, 138*, 530-537.

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

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

Koriat, A., & Greenberg, S.N. (1994). The extraction of phrase structure during reading: Evidence from letter detection errors. *Psychonomic Bulletin & Review, 1*, 345-356.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis . *Discourse Processes, 25*, 259-284.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Lavigne, F., Vitu, F., & dYdewalle, G. (2000). The influence of semantic context on initial landing sites in words. *Acta Psychologica, 104*, 191-214.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics, 28*, 661-677.

Pinheiro, J.C., & Bates, D.M. (2000). *Mixed-effects models in S and S-Plus*. New York: Springer.

Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the E-Z reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology, 52*, 1-56.

Pynte, J., New, B., & Kennedy, A. (2007). *A regression analysis of syntactic influence in reading normal text.* Poster presented at ECEM, Potsdam, and AMLaP, Turku, August 2007.

Pynte, J., New, B., & Kennedy, A. (2008). Contextual influences during reading normal text: a regression analysis. *Vision Research*, in press.

R Development Core Team (2006). *R: A language and environment for statistical computing*. (version 2.3.1). R Foundation for Statistical Computing, Vienna, Austria

DOI 10.16910/jemr.2.1.4
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin* & Review, 3, 504-509.

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E.D. (2004). The effect of frequency and predictability on eye fixations in reading: Implications for the E-Z Reader model. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 720-732.

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

Taylor, W.L. (1953). Cloze procedure: a new tool for measuring readability. *Journalism Quarterly, 30*, 415-4

This article is licensed under a
Creative Commons Attribution 4.0 International license.

Appendix

Table A1
Correlation between predictors.

	Length	Freq.	SEM	SYN	POS	EMB	Lgth n-1	Freq n-1
Length		-0.30	0.30	-0.03	0.06	0.13	0.07	0.00
Freq.	-0.41		-0.16	-0.06	0.00	-0.04	0.06	-0.05
SEM				0.00	0.22	0.18	-0.01	0.02
SYN	-0.04	-0.07			0.12	0.06	-0.07	0.01
POS	-0.02	0.03		0.11		0.58	-0.05	0.04
EMB	-0.07	0.13		0.17	0.55		-0.14	0.15
Lgth n-1	0.01	0.05		-0.16	-0.02	-0.08		-0.56
Freq n-1	-0.03	-0.02		0.12	0.04	0.16	-0.50	

Note: the top-right corner of the table corresponds to content words, whereas the bottom-left corner corresponds to function words.

This article is licensed under a
Creative Commons Attribution 4.0 International license.