Eye Movements during Reading Proverbs and Regular Sentences:

The Incoming Word Predictability Effect

Gerardo Fernández*, Diego E. Shalom*, Reinhold Kliegl, & Mariano Sigman

* These authors contributed equally to this work.

Running head: Incoming word predictability effect.

Keywords: eye movements, reading, proverbs, incoming word predictability effect.

Abstract

Reading is an everyday activity requiring the efficient integration of several central cognitive subsystems ranging from attention and oculomotor control to word identification and language comprehension. Effects of frequency, length, and cloze predictability of words on reading times reliably indicate local processing difficulty of fixated words; also a reader's expectation about an upcoming word apparently influences fixation duration even before the eyes reach this word. Moreover, this effect has been reported as non-canonical (i.e., longer fixation durations on word N when word N+1 is of high cloze predictability; Kliegl, Nuthmann, & Engbert, 2006). However, this effect is difficult to observe because in natural sentences the fluctuations in predictability in content words is very small.. To overcome this difficulty we investigated eye-movements while reading proverbs as well as sentences constructed for high and low average cloze predictability. We also determined for each sentence a word at which predictability of words jumps from a low to high value. Fixation durations while reading proverbs and high-predictable sentences exhibited significant effects of the change in predictability along the sentence (when the successive word is more predictable than the fixated word). Results are in agreement with the proposal that cloze predictability of upcoming words exerts an influence on fixation durations via memory retrieval.

Eye Movements during Reading Proverbs and Regular Sentences:

The Incoming Word Predictability Effect

During fluent reading, the duration of a fixation on a word is influenced by lexical properties of the word. For example, fixation durations reliably decrease with frequency and predictability from prior sentence context. These conclusions derived from two strands of experimental research. One focuses on the analyses of fixation durations on one or two target words per sentence (e.g., Ehrlich & Rayner, 1981; Inhoff & Rayner, 1986; Rayner & Duffy, 1986), the other on multivariate analyses of fixation durations including fixations on all words of the sentences (e.g., Kennedy & Pynte, 2005; Kliegl, Grabner, Rolfs, & Engbert, 2004; Reichle, Pollatsek, Fisher, & Rayner, 1998; Schilling, Rayner, & Chumbley, 1998).

There is also agreement that information about word length (e.g., Juhasz, White, Liversedege, & Rayner, 2008; McConkie & Rayner, 1975), orthography (e.g., Rayner, 1975; White, 2008), and phonology (e.g., Pollatsek, Lesch, Morris, & Rayner, 1992) of the upcoming word is available during fixations on prior words. Indeed, some of this information is deemed necessary for programming saccades. There is, however, controversy about whether, in addition to these so-called low-level influences, high-level lexical properties (such as word frequency or predictability) of parafoveal words also influence fixation durations before the eyes reach these words (*arguments for positive evidence*: e.g., Kennedy & Pynte, 2005; Kennedy, Pynte, Murray, & Paul, 2012; Kliegl, 2007; Kliegl, Nuthmann, & Engbert, 2006; Vitu, Brysbaert, & Lancelin,

2004; *arguments for absence of evidence*: e.g., Rayner, Pollatsek, Drieghe, Slattery, & Reichle, 2007; Reichle, Liversedge, Pollastek & Rayner, 2009). In two recent reviews of this research, Drieghe (2011) and Hyönä (2011) concluded that the evidence about these issues is mixed.[1]

In the present study, we provide new tests of the effects of frequency and cloze predictability of the fixated word as well as of the corresponding effects of its left and right neighbors during reading of Spanish sentences. Cloze predictability is the probability that the next word in a sentence is guessed, given only the prior words of the sentence (i.e., incremental Cloze Task procedure; Taylor, 1953). Kliegl et al. (2006) demonstrated that fixation durations on word N decrease with increasing cloze predictability of word N (as expected), but increase with cloze predictability of word N+1, irrespective of whether the next word was fixated or skipped. The direction of this N+1-predictability effect is surprising, because usually high predictability covaries negatively with fixation duration. Consequently, the result was met with skepticism (Rayner et al., 2007). In a follow-up analysis, Kliegl (2007) reported that the effect was significantly positive in each of the nine subsamples comprising Kliegl et al.'s (2006) corpus and was moderated by the lexical status of words N and N+1 (i.e., the effect was stronger when either word N or word N+1 were function words). Recently, Kennedy et

[1] While there are studies reporting absence of evidence for parafoveal-on-foveal effects, we do not agree that these studies challenge the results of studies providing positive evidence for parafoveal-on-foveal effects. Obviously, the necessary conditions for parafoveal-on-foveal effects are not completely identified yet.

al. (2012) replicated the effect for English with the constraint that fixations had to be on content words; there were also quite a few differences in cloze-predictability task and details of statistical analysis between Kliegl et al. (2006) and Kennedy et al. (2012).[2]

Obviously, predictability is an important factor during fluent reading (Rayner, Ashby, Pollatsek, & Reichle, 2004). The consistency of the counterintuitive N+1-predictability effect across nine samples of readers is a strong argument for its statistical reliability. Of course, its validity (i.e., its theoretical status) is a different question and its establishment remains a challenge to be met. Kliegl et al. (2006) proposed that it is not the effect of the parafoveal visual presence of the word N+1 that increases the duration of the fixation to word N. Instead, it is its likelihood of appearance determined by the regularities of the sentence that evoke memory retrieval mechanisms prior to the initiation of the saccade. With enough long-term memory support for an upcoming word, readers may start to process this word before their eyes move to it. In support of this argument, they argue that, by definition, cloze probability is actually a measure of the evidence for a word *in the absence* of any parafoveal visual information. Thus, in principle, the effect may have very little to do with *visual* parafoveal processing, but instead reflect a contribution of long-term memory that facilitates comprehension during reading. The moderation of the effect by lexical status of word N and word N+1 is in agreement with this retrieval

[2] By now, positive predictability effects of word N+1 have also been presented, but not yet published, for reading English as well as Simplified and Traditional Chinese sentences (Kliegl, 2012).

interpretation (Kliegl, 2007). Alternatively, Rayner et al. (2007) showed that, in principle, a positive N+1-predictability effect could be the consequence of an ignored covariate (i.e., a confound) in the reading material. However, they did not spell out which covariate this might be.

In the present study we tested the proposal that positive N+1-predictability effects on fixation duration are linked to memory retrieval. The main prediction was that the positive N+1-predictability effect would increase with overall average cloze predictability. To this end, we had subjects read proverbs (i.e., arguably, sentences for which we expected a maximum of average cloze predictability) and regular sentences; regular sentences were constructed to be of either high or low average cloze predictability. Obviously, proverbs and high-predictable sentences were expected to contain a substantial number of content words with high cloze predictability. Therefore, with these sentences we also address a limitation of Kliegl et al.'s (2006) sentences in which high values of cloze predictability were primarily associated with function words. We hypothesize that when a sentence is being read, expectations about the next incoming words are incrementally generated and confirmed. Proverbs and high-predictable sentences of the present study should yield a stronger signal for the pattern of hypothesized negative N-predictability and positive N+1-predictability effects than low-predictable sentences.

In addition to the manipulation of overall cloze predictability (proverbs vs. average high-predictable vs. average low-predictable sentences), using proverbs as reading material presents a unique

opportunity to examine the hypothesis of the involvement of memory retrieval in the predictability effect (Katz & Ferretti, 2001). When reading a proverb, there is typically a word at which not only the next word but the entire sentence becomes available. In order to capture this sharp transition in predictability in which a subject matches an entire sentence being read to one held in memory, we determined the word with the maximum change in cloze predictability relative to the previous word in a given sentence. In the context of reading proverbs we will refer to this word as the *Eureka* word (borrowing from Ahissar & Hochstein, 1997). On the basis of this word we defined the binary variable *maxjump* assigning the value 0 to the prior words of the sentence and including this word and the value 1 to the remaining words. We expect that the effect of cloze predictability of individual words on fixation durations will differ for these two regions of sentences. It is an empirical question whether high-predictable sentences exhibit a pattern similar to proverbs, but we would not expect a sharp transition in low-predictable sentences, because the maximum change in predictability is likely to occur at different words for different subjects.

In the first part of the present paper we explore the cloze-predictability distribution across proverbs, high-predictable sentences and low-predictable sentences. In the second part of the paper we use (a) a *baseline* linear mixed model to test a standard set of fixed effects (i.e., length, frequency and predictability of word N and word N+1 by type of sentence) and (b) a *maxjump* linear mixed model including also the new *maxjump* variable and its interaction with terms of the *baseline* model as

fixed effects. In both models we also estimated variance components for partially crossed random factors of subjects, sentences, and words.

METHODS

Participants

Forty-one graduate and undergraduate students (all native Spanish speakers) of Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales participated in the experiment.

Corpus of sentences

The sentence corpus is composed of 184 sentences (1422 words). The corpus comprises (1) 64 proverbs (e.g., "A bird in the hand is worth two in the bush"), (2) 45 high-predictable sentences, which are sentences with well-defined semantic regularities (e.g. "Pinocchio's nose grows every time he lies"), and (3) 75 low-predictable sentences (e.g., "Yesterday I talked to Laura about her daughter"). Discarding the first and the last word of each sentence left a total of 1054 words based on 482 different words. A full list of all sentences used in this experiment (in Spanish) is included in Supplementary Table 1.

Sentence and word lengths

Sentences ranged from a minimum of 5 words to a maximum of 14 words. Words ranged from 1 to 14 letters. Mean length of sentences was 7.3 words (SD=1.9) for proverbs, 7.6 words (SD=1.5) for high-predictable sentences, and 8.1 words (SD=1.4) for low predictable sentences. Mean word lengths were 4.0, 4.1, and 4.6 (SD=2.0, 2.3, and 2.5) for proverbs, high-predictable, and low-predictable sentences, respectively.

Word frequencies

Word frequencies for all words were taken from the Spanish Lexical *Léxesp CORCO* (Sebastián-Gallés, Martí, Cuetos, & Carreiras, 1998). They range from 1 to 264721 per million. We transformed frequency to $\log_{10}$(frequency). Mean $\log_{10}$(frequency) was 3.47 (SD=1.36) for proverbs, 3.45 (SD=1.51) for high-predictable sentences, and 3.41 (SD=1.38) for low-predictable sentences. A one-way ANOVA showed no significant differences between sentence types ($f_{2,1419}$=0.29, p=0.75).

Cloze predictability of words

Cloze predictability was measured in an independent experiment with 18 graduate and undergraduate students (all native Spanish speakers) of Facultad de Ciencias Exactas of Universidad Nacional del Sur; each of them generated a prediction for every word of the complete sentence corpus. Cloze predictability was measured as the probability of predicting a word given knowledge of the preceding part of the sentence (Taylor, 1953). Sentences were presented in a random order for each participant. The trial begun with participants guessing the first word of the unknown original sentence (responses were collected with a keyboard response). Then the computer presented the first word of the original sentence on the screen and participants entered their guess for the second word. This procedure continued until the end of the sentence. When subjects correctly predicted the word, it remained in the screen. Participants of the Cloze task were between 25 and 40 years old, and did not participate in the reading experiment. The academic background of

the reading and the Cloze Task groups was similar; all participants had completed high-school. The average predictability measured from the Cloze Task was transformed using a logit function:

$$\text{Logit}(\text{pred}) = 0.5 \ln\left(\frac{\text{pred}}{1-\text{pred}}\right).$$

The term between parentheses $\text{pred}/(1-\text{pred})$ is called the odds. To avoid taking the log of zero values or dividing by zero, we replaced values of zero predictability in the Cloze Task with $\text{Logit}(1/2N_p) = -1.77$, where $N_p = 18$ represents the number of complete predictability protocols (18 subjects completing all the sentences). Analogously, values with perfect predictability in the Cloze Task were replaced with $\text{Logit}(1 - 1/2N_p) = 1.77$.

With this transformation (Cohen & Cohen, 1975), the odds of guessing a word with predictability .50 are one and, therefore, the log odds of guessing the word are zero. Thus, words with predictability larger than .50 yielded positive logits, and those with predictabilities smaller than .50 yielded negative logits. Further procedural details on the norming study are provided in Kliegl et al. (2004). Mean logit predictability was 0.08 (SD=1.23) for proverbs, -0.08 (SD=1.29) for high-predictable sentences, and -0.98 (SD=0.94) for low predictable sentences. A one-way ANOVA showed significant differences between sentence types ($f_{2,1419}$=134.65, p<0.0001). Post-hoc t-tests only showed significant differences between low-predictable sentences and the other two sentence types (low-predictable vs. proverbs: t(1078)=15.97,

p<0.00001; low-predictable vs. high-predictable: t(951)=12.27, p<0.00001; proverbs vs. high-predictable: t(809)=1.78, p=0.076).

Maxjump

For each sentence, we determined the word with the largest difference in predictability relative to the predictability of the previous word according to the following equation:

$$jump\ word = \max\left[Logit\left(pred_{N+1}\right) - Logit\left(pred_{N}\right)\right].$$

Jump word separates the sentence in two regions. The variable *maxjump* is assigned a value of 1 for each word after the jump word and a value of 0 to all words prior to and including the jump word. With this *maxjump* variable we test the contextual word predictability effect due to memory retrieval. Obviously, we expected that the jump word should occur earliest in proverbs and earlier in high-predictable than low-predictable sentences. The expectation was also that fixation durations are shorter after than before *jump* words. This expectation is based in the assumption that after the jump word, less processing would be required since words have already been recovered from memory. We had no clear expectations about how this change would interact with sentence type and cloze predictabilities for word N and word N+1.

Apparatus and eye movement data

Single sentences were presented on the center line of a 19-inch Monitor (1024 x 768 pixels resolution; frame rate 85 Hz, font: regular; New Courier; 12 point, 0.5º in height). Participants were seated in front of the monitor with the head positioned on a chin rest at a distance of 60 cm from the monitor. Eye movements were recorded with an EyeLink

2K Desktop Mount (SR Research) eyetracker, with a sampling rate of 1000 Hz and an eye position resolution of 20-s arc. All recordings and calibration were binocular. Only right eye data was used for the analysis.

Eye movement data from 41 participants reading 184 sentences were cleaned from blinks and track losses. Fixations shorter than 51 ms and longer than 750 ms and fixations on first and last word of each sentence were removed for the analysis. This left 19.550 fixations in the data set, which corresponded to first-pass fixations, and were included in the statistical analyses.

Procedures

Participants were calibrated with a standard 13-point grid for both eyes. After validation of calibration, a trial started with the presentation of a fixation point at the position of the first letter of the sentence. The sentence was presented as soon as both eyes were detected within a 1º radius from the fixation spot. After reading the sentence, participants looked at a dot in the lower right corner of the screen. When the gaze was detected there, the trial ended. On 20% of the trials, a three alternative multiple-choice question about the current sentence was presented. Participants answered the question moving a mouse, and choosing the response with a mouse click. Overall mean accuracy was 96.5% (SD=3.0%). Then, the next trial started with the presentation of the fixation spot. The experimenter carried out an extra calibration if the eye tracker did not detect the eye at the initial fixation point within 2 s.

Linear mixed models (LMMs)

We used the *lmer* program of the *lme4* package (version 0.999999-0) (Bates, 2010; Bates, Maechler, & Bolker, 2012) for estimating fixed and random coefficients. This package is supplied in the *R* system for statistical computing (version 2.15.1; R Core Team, 2012). We used maximum likelihood (ML) statistics for model comparisons with different fixed effects and identical random effects and restricted maximum likelihood (REML) statistics for estimation of fixed and random effects in the final model. For assessment of (differences in) goodness of fit, the *lmer* program provides the Akaike Information Criterion (AIC; decreases with goodness of fit), the Bayesian Information Criterion (BIC; decreases with goodness of fit), the log likelihood (logLik; increases with goodness of fit), and, in the case of model comparisons, the likelihood ratio. The AIC (= -2 logLik + 2 n_{param}) and BIC (= -2 logLik + n_{param}log N_{obs}) values correct the log-likelihood statistic for the number of estimated parameters and the number of observations, to avoid overfitting during the process of model selection.

The dependent variable was log of single fixation duration. The critical factors of this experimental design are sentence type and predictabilities. We specified two *a priori* contrasts for the three levels of sentence type, (1) proverbs vs. high-predictable sentences and (2) high-predictable vs. low-predictable sentences. They inform about different aspects of the relevance of long-term memory. The main hypothesis is that the effect of predictability, especially the effect of predictability of word N+1, depends on the overall level of cloze predictability, as

represented in the three types of sentences. Therefore, interactions between the two fixed effects coding the two contrasts with predictabilities of word N-1, N, and word N+1 were the focus of the *baseline* LMM. Of course, these interactions may also depend on how quickly in a sentence the jump in predictability occurred. Therefore, interactions between *maxjump* and the fixed effects coding the two contrasts of sentence type and between *maxjump* and predictability of words N-1, N, and N+1 were the focus of the *maxjump* LMM. Contrasts do not inform about the significance of predictability for a given condition. When this information was needed for the interpretation of an interaction, we tested the effect in a post-hoc LMM, using the specific level of sentence type as the reference category in a treatment contrast.

The LMMs included a number of other covariates, which are known to affect fixation durations. Launch site (i.e., the distance of the last fixation from the beginning of the current word) is known to be one of the strongest predictors of fixation durations (Heller & Müller, 1983; Pollatsek, Rayner, & Balota, 1986). We also included its interactions with the contrasts for sentence types. A traditional variable in coding predictability in ERP research has been the ordinal position of a word in the sentence (e.g., Van Petten & Kutas, 1990; Dambacher, Kliegl, Hofmann, & Jacobs, 2006). Kuperman, Dambacher, Nuthmann, and Kliegl (2010) reported that word number of sentence is a significant predictor in addition to predictability. Given the focus on *maxjump,* it was important to control for this effect as well. Finally, we also included lengths and frequencies of words N-1, N, and N+1 as covariates to

reduce differences between our models and those reported by Kliegl et al. (2006). All covariates were centered such that the intercept estimated the mean log fixation duration.

For the LMMs we report regression coefficients (*b*s), standard errors (*SE*s) and *t*-values (*t*=*b*/*SE*). There is no clear definition of "degree of freedom" for LMMs and therefore precise *p*-values cannot be estimated. In general, however, given the large number of observations, subjects, sentences, and words entering our analysis and the comparatively small number of fixed and random effects estimated, the *t*-distribution is equivalent to the normal distribution for all practical purposes (i.e., the contribution of the degrees of freedom to the test statistic is negligible). Our criterion for referring to an effect as significant is $t = b/SE > 2.0$. The significance of fixed effects was checked with Markov Chain Monte Carlo methods. Specifically, we generated 10000 samples from the posterior distribution of the fitted model parameters and constructed the highest posterior density (HPD) intervals covering 95% of the empirical cumulative density function for model parameters. In 56 of 59 tests both statistics led to the same decision; we will mention the disagreements, all of which were minor.

We specified both LMMs such that they yielded estimates for three variance components associated with intercepts for subjects, sentences, and words. The models account for dependencies between fixations due to the clustering associated with these three partially crossed random factors.

RESULTS

Word predictability distribution across sentences

We first verify whether the proverbs, high-predictable sentences and low-predictable sentences have different distributions of predictability. For each word in the sentence we calculate the mean predictability, averaging across all subjects who participated in the Cloze Task. To align predictability uniformly across all sentences, we normalized the data, assigning to each word the relative position in the sentence (its word number divided by the sentence length). In this representation, the first word of a sentence is indexed 0 and the last word is indexed 1 regardless of sentence length. Both representations consistently show an increase of predictability of all types of sentences (Figure 1a,b). The rate of increase is larger for high-predictable sentences and proverbs than for low-predictable sentences.

(FIGURE 1 about here)

When collapsed across all positions in the sentence, the distributions of word predictabilities show very different patterns: while in low-predictable sentences only a small proportion of words are highly predictable, high-predictable sentences and proverbs show relatively symmetric distributions with comparable numbers of high predictable and low predictable words (Figure 1c).

We reasoned that a critical parameter for memory retrieval during fluent reading is transitions in predictability, i.e. the moment in which a sentence becomes predictable and is retrieved from memory. Hence, we characterize the transitions in predictability in the different kinds of sentences. For proverbs and high-predictable sentences, predictability of

the first few words is low (when the proverb has not yet been detected), in a critical stage in which the proverb is detected predictability grows rapidly and from this moment on, words remain highly predictable until the end of the sentence. This effect is revealed by a diagram in which predictability for each word of each sentence is coded in a grey scale (Figure 2a) which shows that this pattern is actually observed in the vast majority of proverbs and high-predictable sentences.

(FIGURE 2 about here)

This structured distribution of predictability is further confirmed by the autocorrelation function, which shows a very strong lag-1 predictability autocorrelations (> 0.55, $p<0.0001$). In fact the autocorrelation function decreases very slowly as a function of lag (Supplementary Figure 1). In contrast, the highly predictable words in low-predictable sentences occur sporadically, typically reflecting very constrained grammatical structures which occur discretely throughout the sentence (Figure 2c). The lag-1 predictability autocorrelation of low-predictable sentences is $R=0.080$ ($p=0.065$), considerably lower than proverbs and vanishes very rapidly (Supplementary Figure 1). In summary, the analyses revealed first an expected decrease of average cloze predictability for the comparison of proverbs, high-predictable sentences, and low-predictable sentences. More importantly, it shows that predictability in proverbs reflects higher cloze values for subsequent words and not merely the occurrence of an individual word with high predictability.

Baseline Linear Mixed Model (LMM)

N+1-predictability and type-of-sentence effects. LMM-based test statistics are summarized in Table 1. A plot of standardized model residuals over fitted values did not reveal any problems with outliers, heteroskedastic error variance, or nonlinearity (see Supplementary Figure 2a). Fixation durations increased significantly with the predictability of word N+1 (N+1-predictability effect; t = 3.00), replicating the non-canonical direction of this effect, interpreted as evidence for memory retrieval of predictable words (Kliegl et al., 2006). As expected, the N+1-predictability effect was significantly stronger in high-predictable than low-predictable sentences (t = -2.66); the N+1-predictability effect was not significantly different for proverbs vs high-predictable sentences (t = 1.27). In a post-hoc LMM with low-predictable sentences as the reference category for a treatment contrast, the N+1-predictability effect was not significant in low-predictable sentences (t = 0.17). LMM-based partial effects (top row) and observed, unadjusted results (bottom) are shown in the right panels of Figure 3.

(TABLE 1 and FIGURE 3 about here)

N-1- and N-predictability and type-of-sentence effects. N-1-predictability and N-predictability effects were not significant as main effects, but their numerical trends were in the expected negative direction (t-values of -1.68 and -1.89, respectively; according to the 95% HPD interval the N-predictability effect was significant: -0.0129, -0.0005). Moreover, there were unexpected, but significant interactions between N-1-predicability and the two contrasts for sentence type (N-1-predictabily

x proverb vs. high-predictable sentences: t = -3.93; N-1-predictability x high- vs. low-predictable sentences: t = 3.31) and a marginally significant interaction between N-predictability and the contrast between high- and low-predictable sentences (t = -1.98; the interaction was significant for the HPD interval: -0.0257, -0.0004). The two significant interactions involving the predictability of word N-1 were linked to a significantly negative predictability effect restricted to high-predictable sentences (post-hoc LMM: t=-4.10). The N-predictability effect (middle panel) was significantly negative only for low-predictable sentences (post-hoc LMM: t=-3.07). Neither N-1- nor N-predictability effects were significant for proverbs. We return to this unexpected constellation of results in the Discussion, also after we report results relating to the *maxjump* variable.

Effects of other covariates. We also included a number of additional covariates with reference to earlier research. With respect to word frequency, the effects were significant for word N-1 (t = -4.56), word N (t = -6.94) and word N+1 (t = -2.73). Fixation durations increased with the length of word N-1 (N-1-length effect; t = 4.39) and decreased with the length of word N (N-length-effect; t = -2.61); the N+1-length effect was not significant. With the exception of the significant inverse N-length effect (i.e., fixation durations decrease with increasing length of the fixated word), the pattern of frequency and length effects replicates Kliegl et al. (2006). An inverse N-length effect was reported by Kliegl et al. (2006) for the subset of fixations where the last fixation was on word N and the next fixation was on word N+1 (i.e.,

triplet-constrained single fixations) and was traced to a suppressor constellation triggered by ignoring skippings of word N-1 (Kliegl et al., 2006, Appendix). Fixation durations also significantly increased with word number (t = 3.73; replicating Kuperman et al., 2010). Finally, as in previous research, the largest effect was associated with launch site: The larger the distance between the last fixation location and the beginning of the fixated word, the longer the duration (t = 33.61). This effect was significantly stronger in high-predictable sentences than proverbs (t = 4.09 for the interaction).

***Maxjump* Linear Mixed Model (LMM)**

Model comparison. Adding the *maxjump* variable and 14 interaction terms associated with it to the *baseline* LMM significantly improved the goodness of fit as assessed with the likelihood ratio statistic; $\chi^2(15\ df) = 56.4$, $p < 0.00001$, and a decrease of AIC from -2287 to -2313. However, according to BIC, which takes into account the increase in model complexity due to the additional model parameters, the goodness of fit was not significantly better (i.e., increase of BIC from -2082 to -1990). We could resolve the ambiguity by removing a few non-significant higher-order interaction terms, but the overall pattern of results was not affected by such model tuning. A plot of standardized model residuals over fitted values did not reveal any problems with outliers, heteroskedastic error variance, or nonlinearity (see Supplementary Figure 2b). Therefore, we stayed with the *maxjump* LMM as reported in the right part of Table 1.

Effect of maxjump and interactions with other factors. Fixation

durations before and after maxjump were differed by only 2 ms (before jump word: 185±49; after jump word: 187±51 SD). In the LMM, the main effect of maxjump was significant. However, this main effect relates to the difference given statistical control of the other variables in the model, that is this main effect is a partial effect that cannot be interpreted independently of several higher-order interactions involving maxjump.

We observed significant three-factor interactions, involving the maxjump variable, the contrast between high-predictable and low-predictable sentences. This shows that maxjump variable changes the dependency of fixation times with predictability. Significant differences were observed in the low vs high predictability contrast but not in the high-predictability vs proverbs (see Table 1). This is expected since the pattern of predictability is very similar between proverbs and high-predictability (see Figure 2) and shows a very different pattern for low-predictable sentences.

(FIGURE 4 about here)

The primary source of the first interaction is a negative N-1-predictability effect for low-predictable sentences before *maxjump* (Figure 4, top left panel) and a negative N-1-predictability effect for high-predictable sentences after *maxjump* (Figure 4, bottom left panel). A parsimonious explanation of this interaction relates to differences in the number of fixations observed before and after the predictability jump for low- and high-predictable sentences. The N-1-predictability effect was significant when the number of fixations was large (i.e., reliability was

high), that is before the predictability jump for low-predictable sentences and after the predictability for high-predictable sentences. Thus, with enough statistical power, we observe significant predictability-related spillover effects from word N-1.

Aside from fixed effects relating to the contrasts between sentences and to predictabilities (described above), lengths and frequencies of words N-1, N, and N+1 as well as launch site and word number exhibited the pattern of significant effects on fixation duration already reported for the *baseline* LMM.

In summary, with the exception of predictability-related spillover effects from word N-1 to word N, the distinction between fixations before and after the maximum jump in predictability accounted for positive predictability effects associated with word N+1 observed in the *baseline* model. This result is consistent with the hypothesis that words were retrieved from memory and hence no further effect of cloze-predictability was expected with this variable in the LMM.

Gaze duration Linear Mixed Models

Studies of reading have relied on different measures of fixation durations (Rayner, 1998, 2009). All the analyses reported so far were based on single-fixation durations. Here we investigated the robustness of this effect by repeating the LMM analysis with gaze durations, which result from accumulating all the fixation times to a given word during first-pass reading. Results on both *baseline* models on gaze duration were similar to those obtained on single-fixation durations (Table 2), There was coincidence (either the effect was significant for both measures or

for none) for 19 out of 22 effects. The three effects which changed significance for gaze and single-fixation durations showed very similar patterns and where close to the boundary of significance. Interestingly, when considering the effect of *maxjump,* while the trends were very similar, gaze data showed overall higher patterns of significance. This is evident when comparing the lower panels of Table 1 and 2, with most of the interactions reaching significance in the analysis of gaze durations. Hence the conclusions derived form single-fixation durations were even more pronounced for the analyses of gaze durations.

DISCUSSION

We reported an experiment in which we (a) examined the distribution of cloze predictability of individual words across proverbs, high-predictable sentences, and low-predictable sentences, and (b) tested the effects of cloze predictabilities of the words surrounding a fixation on the duration of this fixation for these three types of sentences. Our experiment was motivated by previous findings which had suggested that the successor word N+1 may have an effect on fixation duration on word N via memory retrieval (Kliegl et al., 2006). According to this interpretation, it is not the effect of the parafoveal presence of the word N+1 itself that increases the duration of the fixation to word N. Instead, it is its likelihood of appearance determined by the regularities of the sentence which evoke memory retrieval mechanisms prior to the initiation of the saccade.

With the construction of the three types of sentences, we could explicitly vary the dynamics of memory retrieval during sentence

reading. Thus, our first contribution is purely methodological, that is the construction of a calibrated corpus of sentences with properties suitable for this purpose. We found that predictable words in proverbs and high-predictable sentences are clustered in a sentence instead of being isolated moments of highly regular fragments of a sentence (prepositions, articles). As a consequence, predictability is uncorrelated from other main factors governing fixation duration. Large values of predictability in sentences with a low average cloze predictability correspond to shorter words, seemingly more related to the grammatical structure of the sentence than to its semantic contents, as connectors, prepositions, articles. Highly predictable words in proverbs and sentences with a high average cloze predictability relate to semantic content and hence might provide a test bed for the investigation of the specific effect of memory retrieval and the subsequent facilitated incoming word reading process.

The second contribution is an operationalization of the word at which readers becomes aware of the entire sentence they are reading. This is simply the word with the largest difference in cloze predictability relative to the previous word. Inclusion of the distinction between fixations before and after the maximum predictability jump in the LMM revealed that the positive N+1-predictability effect, which we already saw for high-predictable sentences in the *baseline* LMM, was solely due to fixations before *maxjump*; there was no evidence for this effect for fixations after *maxjump*. Thus, the interpretation is fairly straightforward. First, the positive incoming predictability effect needs at least a moderate average level of predictability, as available for our high-predictable

sentences. In this case, the predictability of the upcoming word affects fixation durations until the complete sentence is retrieved from memory. After this *Eureka* event, predictability loses its relevance as an indicator of cognitive effort.

Why did we fail to find this pattern for proverbs and low-predictable sentences? We suspect that for proverbs the *Eureka* event simply came too quickly. Thus, there are too few fixations before *maxjump* for a reliable assessment of the incoming N+1-predictability effect (i.e., there is a "predictability ceiling" after the *Eureka* word). For low predictable sentences the *Eureka* event came too late or not at all; the generally low level of predictability of individual words did not provide enough cues for the effect to kick in (i.e., there is something like a "predictability floor"). Thus, despite the large number of fixations before the "jump" word, a restriction in predictability range may have worked against finding the positive N+1-predictability effect in these sentences. The significant interaction involving *maxjump,* the contrast between low- and high-predictable sentences, and the predictability of word N-1 also suggests that low predictable words exert their influence on fixation durations only with some delay, that is after the eyes have already moved on to the next word. Overall, the *maxjump* effect offers a perspective that maintains the viability of the interpretation of the contextual incoming word predictability effect as a consequence of anticipatory retrieval of word meaning from memory up to point of complete retrieval.

The third contribution is a follow-up of some controversial results about the effects of the properties of words in the neighborhood of a

fixated word on fixation durations. Specifically, Kliegl et al. (2006) reported length, frequency, and predictability effects for the previous and the next word for reading of German sentences. Rayner et al. (2007) critically commented and discussed these results (see Kliegl, 2007, for a reply). In general, the present results from reading Spanish sentences replicate the most controversial effects of this earlier exchange: positive N+1-predictability and negative N+1-frequency effects. This constellation of a canonical negative N+1-frequency effect and a positive N+1-predictability effect is remarkable because frequency and predictability are positively correlated with each other. Thus, their opposite relation with fixation duration on the prior word may prove very diagnostic about the integration of higher-order memory and lower-order visual processes.

There is probably broad agreement that cloze predictability captures a highly relevant aspect of reading behavior. Unfortunately, cloze predictability is a labor-intensive statistic to collect. Several computational alternatives have been proposed: transition probabilities (Frisson, Rayner, & Pickering, 2005; Keller & Lapata, 2003; McDonald & Shillcock, 2003), latent semantic analysis (Landauer & Dumais, 1997, Ong & Kliegl, 2008, Pynte, New & Kennedy, 2009), surprisal (Demberg & Keller, 2008; Boston, Hale, et al., 2008; Boston, Hale, Vasishth, & Kliegl, 2011). As far as we know, when including these measures in LMMs along with cloze predictability and frequency, they accounted for a significant amount of unique variance in fixation durations, but they were more closely related to frequency than to cloze predictability.

Indeed, none of these alternatives rendered cloze predictability irrelevant. In this respect, the *maxjump* variable may move us a step closer. Although the current version is based on cloze predictability, conceivably alternative and hopefully less labor-intensive determinations of the jump word could be developed, ideally guided by theories of sentence comprehension, and might be used in place of cloze predictability. The need for such a measure is evident from its use in computational models of eye-movement control in reading.

In general, our results support a theoretical perspective of reading which favors a model of distributed processing of words across fixation durations and challenges psycholinguistic immediacy-of-processing and eye-mind assumptions. Distributed processing effects tied to properties of upcoming words may exert an influence on fixation duration not only with respect to visual processing in the perceptual span, but may indicate whether an accurate representation of the sentence has already been achieved by relying on memory retrieval for the prediction of incoming words. Most importantly, probing online comprehension processes, as operationalized with the identification of a *Eureka* word in proverbs and high-predictable sentences, and tracing their effects to fixation durations might facilitate a productive exchange between theories about text comprehension and eye-movement control during reading.

REFERENCES

Ahissar, M. & Hochstein, S.(1997). Task difficulty and the specificity of perceptual learning. *Nature* **387**, 401 - 406

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R.* New York: Springer. Prepublication version at: http://lme4.r-forge.r-project.org/book/

Bates, D. M., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-0.[Computer software]. http://CRAN.R-project.org/package=lme4.

Binder, K. S., Pollatsek, A., & Rayner, K. (1999). Extraction of information to the left of the fixated word in reading. *J Exp Psychol Hum Percept Perform, 25*(4), 1162-1172.

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

Boston, M.F., Hale, J.T., Vasishth, S., & Kliegl, R. (2011). Parallelism and syntactic processes in reading difficulty. *Language and Cognitive Processes, 26,* 301-349.

Cohen, J., & Cohen, P. (1975). *Applied Multiple Regression and Correlation Analysis for the Behavioral Sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

Drieghe, D. (2011). Parafoveal-on-foveal effects on eye-movements during reading. In: S.P. Liversedge, I.D. Gilchrist, & S. Everling (eds.), *The Oxford Handbook of Eye Movements*(pp. 839-855). Oxford: Oxford University Press.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word recognition and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20, 641-655.Henderson, J. M., & Ferreira, F. (1990). Effects of foveal processing difficulty on the perceptual span in reading: implications for attention and eye movement control. *J Exp Psychol Learn Mem Cogn, 16*(3), 417-429.

Frisson, S., Rayner, K., & Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Ex- perimental Psychology: Learning, Memory, and Cognition*, 31(5), 862–877.

Heller, D., & Müller, H. (1983). On the relationship of saccade size and fixation duration in reading. In R. Groner, C. Menz, D. F. Fisher & R. A. Monty (Eds.), *Eye movements and psychological functions: International views* (pp. 287–302). Hillsdale, NJ: Erlbaum.

Henderson, J. M., & Ferreira, F. (1993). Eye movement control during reading: fixation measures reflect foveal but not parafoveal processing difficulty. *Can J Exp Psychol, 47*(2), 201-221.

Hyönä, J. (2011). Foveal and parafoveal processing during reading. In: S.P. Liversedge, I.D. Gilchrist, & S. Everling (eds.), *The Oxford Handbook of Eye Movements*(pp. 819-838). Oxford: Oxford University Press.

Inhoff, A. W., Pollatsek, A., Posner, M. I., & Rayner, K. (1989). Covert attention and eye movements during reading. *Q J Exp Psychol A, 41*(1), 63-89.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: effects of word frequency. *Percept Psychophys, 40*(6), 431-439.

Inhoff, A. W., Starr, M., & Shindler, K. L. (2000). Is the processing of words during eye fixations in reading strictly serial? *Percept Psychophys, 62*(7), 1474-1484.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychol Rev, 87*(4), 329-354.

Katz, A. N., & Ferretti, T. R. (2001). Moment-By-Moment Reading of Proverbs in Literal and Nonliteral Contexts. *Metaphor and Symbol, 16*(3), 193 - 221.

Keller, F., & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3), 459– 484.

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

Kennedy, A., Pynte, J., Murray, W.S., & Paul, S.-A. (2012). Frequency and predictability effects in the Dundee corpus. *Quarterly Journal of Experimental Psychology*. doi.10.1080/17470218.2012.676054(available online 15 Mar 2012).

Kliegl, R. (2007). Toward a perceptual-span theory of distributed processing in reading: A reply to Rayner, Pollatsek, Drieghe, Slattery, and Reichle (2007).*Journal of Experimental Psychology: General, 136*(3), 530-537.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *Journal of Cognitive Psychology, 16*(1), 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. *J Exp Psychol Gen, 135*(1), 12-35.

Kuperman, V., Dambacher, M., Nuthmann A., Kliegl, R., (2010). The effect of word position on eye-movements in sentence and paragraph reading. *Quartely Journal of Experimental Psychology*, 63:9, 1838-1357.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240.

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

McDonald, S. A., & Shillcock, R. C. (2003). Eye move- ments reveal the on-line computation of lexical prob- abilities during reading. *Psychological Science*, 14, 648–652.

Ong, J. K. Y., & Kliegl, R. (2008). Conditional co-occurrence probability acts like frequency in predicting fixation durations. *Journal of Eye Movement Research*, 2(1):3, 1-7.

Pollatsek, A., Rayner, K., & Balota, D. A. (1986). Inferences about eye movement control from the perceptual span in reading. *Perception & Psychophysics, 40*, 123–130.

Pollatsek, A., Lesch, M., Morris, R.K., & Rayner, K. (1992). Phonological codes are used in integrating information across saccades in word identification and reading. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 148-162.

Pynte, J., New, B., & Kennedy, A. (2009). A multiple regression analysis of syntactic and semantic influences in reading normal text. *Journal of Eye Movement Research*, 2(1:4), 1-11.

R Core Team (2012). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. [Computer software]. ISBN 3-900051-07-0, URL http://www.R-project.org/.

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

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychol Bull, 124*(3), 372-422.

Rayner, K. (2009). Eye Movements in Reading: Models and Data. *Journal of Eye Movement Research, 2*(5), 1-10.

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: implications for the E-Z Reader model. *J Exp Psychol Hum Percept Perform, 30*(4), 720-732.

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

Reichle, E. D., Liversedge, S. P., Pollastek, A., & Rayner, K. (2009). Encoding multiple words simultaneously in reading is implausible. *Trends in Cognitive Sciences*, 13. 115-119.

Schroyens, W., Vitu, F., Brysbaert, M., & d'Ydewalle, G. (1999). Eye movement control during reading: foveal load and parafoveal processing. *Q J Exp Psychol A, 52*(4), 1021-1046.

Sebastián-Gallés, N., Martí, M. A., Cuetos, F., & Carreiras, M. (1998). *LEXESP: Léxico informatizado del español*. Barcelona: Ediciones de la Universidad de Barcelona.

Starr, M. S., & Rayner, K. (2001). Eye movements during reading: some current controversies. *Trends Cogn Sci, 5*(4), 156-163.

Taylor, W.L. (1953). "Cloze procedure: A new tool for measuring readability." *Journalism Quarterly*, 30, 415–433.

Vitu, F. o., Brysbaert, M., & Lancelin, D. (2004). A test of parafoveal-on-foveal effects with pairs of orthographically related words. *Journal of Cognitive Psychology, 16*(1), 154 - 177.

White, S.J. (2008). Eye movement control during reading: Effects of word frequency and orthographic familiarity. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 205-233.

Table 1.*Baseline* and *maxjump*LMMs for first fixations.

	FIRST FIXATION DURATION					
	BASELINE			MAXJUMP		
FIXED EFFECTS	*M*	*SE*	*t-value*	*M*	*SE*	*t-value*
Mean First Fixation duration (log)	5'181	0.02	**296.05**	5'186	0.02	**291.95**
Launch site	0.023	0	**33.61**	0.023	0	**31.70**
Word number	0.006	0	**3.73**	0.005	0	**2.87**
Sentence contrasts (* Launch site)						
High-pred vs Proverb	0.014	0.01	**1.31**	0.027	0.01	**1.88**
High-pred vs Proverb * Launch site	0.007	0	**4.09**	0.006	0.03	**3.29**
Low-pred vs High-pred	0.015	0.01	**1.31**	-0.014	0.01	**-1.04**
Low-pred vs High pred * Launch site	-0.003	0	**-1.74**	-0.003	0	**-1.66**
Lengths						
Word N	-0.113	0.04	**-2.61**	-0.113	0.04	**-2.64**
Word N-1	0.08	0.02	**4.39**	0.072	0.02	**3.87**
Word N+1	0.034	0.02	**1.81**	0.032	0.02	**1.72**
Frequencies(log)						
Word N	-0.033	0.01	**-6.94**	-0.032	0.01	**-6.79**
Word N-1	-0.013	0	**-4.56**	-0.011	0	**-3.85**
Word N+1	-0.008	0	**-2.73**	-0.007	0	**-2.53**
Predictabilities(logit)						
Word N	-0.006	0	**-1.68**	-0.014	0	**-3.26**
Word N-1	-0.006	0	**-1.89**	-0.002	0	**-0.61**
Word N+1	0.009	0	**3.00**	0.006	0	**1.82**
Predictabilities(logit) * Sentence contrasts						
Word N * High-pred vs Proverb	0	0.01	**0.01**	0.014	0.01	**1.43**
Word N * Low-pred vs High-pred	-0.013	0.01	**-1.98**	-0.019	0.01	**-2.13**
Word N-1 * High-pred vs Proverb	-0.026	0.01	**-3.93**	-0.025	0.01	**-2.74**
Word N-1 * Low-pred vs High pred	0.021	0.01	**3.31**	0.001	0.01	**0.11**
Word N+1 * High-pred vs Proverb	0.009	0.01	**1.27**	0.011	0.01	**1.45**
Word N+1 * Low-pred vs Hiigh-pred	-0.018	0.01	**-2.66**	-0.012	0.01	**-1.78**
Maxjump				0.036	0.01	**3.62**
Maxjump * Launch site				0	0	**-0.01**
Maxjump * Sentence contrasts						
Maxjump * High-pred vs Proverbs				-0.024	0.03	**-0.94**
Maxjump * High-pred vs Proverbs * Launch site				0.005	0	**1.46**
Maxjump * Low-pred vs High-pred				0.023	0.02	**0.97**
Maxjump * Low-pred vs High-pred * Launch site				-0.003	0	**-1.08**
Maxjump * Predictabilities(logit)						
Maxjump * Word N				0.009	0.01	**1.21**
Maxjump * Word N-1				-0.012	0.01	**-1.79**
Maxjump * Word N+1				-0.01	0.01	**-1.81**
Maxjumps * Predictabilities(logit) *						
Sentence contrasts						
Maxjump * Word N * High-pred vs Proverb				-0.019	0.02	**-1.02**
Maxjump * Word N * Low-pred vs High-pred				0.014	0.02	**0.81**
Maxjump * Word N-1 * High-pred vs Proverb				-0.008	0.02	**-0.46**
Maxjump * Word N-1 * Low-pred vs High-pred				0.064	0.02	**4.00**
Maxjump * Word N+1 * High-pred vs Proverb				-0.024	0.02	**-1.66**
Maxjump * Word N+1 * Low-pred vs High-pred				0.027	0.01	**2.04**
VARIANCE COMPONENTS	*Variance*	*SD*		*Varian*	*SD*	
Groups						
Word (n=482)	0.005	0.07		0.005	0.07	
Sentence (n=184)	0.002	0.04		0.002	0.04	
Subject (n=41)	0.011	0.11		0.011	0.11	
Residual (n=19950)	0.049	0.22		0.049	0.22	

Table 2.*Baseline* and *maxjump*LMMs for gaze data.

	GAZE DURATION					
	BASELINE			MAXJUMP		
FIXED EFFECTS	*M*	*SE*	*t-value*	*M*	*SE*	*t-value*
Mean gaze duration (log)	5.230	0.020	**258.47**	5.288	0.021	**257.78**
Launch site	0.028	0.001	**36.16**	0.027	0.001	**33.52**
Word number	0.005	0.002	**2.44**	0.001	0.002	**0.51**
Sentence contrasts (* Launch site)						
High-pred vs Proverb	0.011	0.013	**0.87**	0.019	0.017	**1.11**
High-pred vs Proverb * Launch site	0.007	0.002	**3.89**	0.006	0.002	**3.09**
Low-pred vs High-pred	0.006	0.013	**0.48**	-0.040	0.017	**-2.35**
Low-pred vs High pred * Launch site	0.000	0.002	**-0.14**	0.001	0.002	**0.38**
Lengths						
Word N	-0.298	0.050	**-5.95**	-0.272	0.056	**-4.89**
Word N-1	0.080	0.080	**3.74**	0.079	0.023	**3.47**
Word N+1	0.022	0.022	**0.28**	0.020	0.023	**0.85**
Frequencies(log)						
Word N	-0.045	0.006	**-8.07**	-0.047	0.006	**-7.52**
Word N-1	-0.010	0.003	**-3.11**	-0.011	0.003	**-3.06**
Word N+1	-0.006	0.003	**-1.91**	-0.006	0.003	**-1.77**
Predictabilities(logit)						
Word N	-0.006	0.004	**-1.63**	0.007	0.005	**1.42**
Word N-1	-0.010	0.004	**-2.70**	0.015	0.004	**3.42**
Word N+1	0.010	0.004	**2.61**	0.007	0.004	**1.82**
Predictabilities(logit) * Sentence contrasts						
Word N * High-pred vs Proverb	-0.003	0.008	**-0.32**	-0.019	0.010	**-1.83**
Word N * Low-pred vs High-pred	-0.016	0.008	**-2.03**	-0.029	0.010	**-2.94**
Word N-1 * High-pred vs Proverb	-0.025	0.007	**-3.31**	-0.038	0.010	**-3.92**
Word N-1 * Low-pred vs High pred	0.018	0.008	**2.43**	0.003	0.009	**0.37**
Word N+1 * High-pred vs Proverb	0.014	0.009	**1.60**	-0.006	0.009	**-0.60**
Word N+1 * Low-pred vs Hiigh-pred	-0.018	0.008	**-2.36**	-0.002	0.008	**-0.29**
Maxjump				0.025	0.009	**2.92**
Maxjump * Launch site				0.000	0.001	**-0.29**
Maxjump * Sentence contrasts						
Maxjump * High-pred vs Proverbs				0.099	0.022	**4.50**
Maxjump * High-pred vs Proverbs * Launch site				0.007	0.004	**1.82**
Maxjump * Low-pred vs High-pred				-0.068	0.021	**-3.71**
Maxjump * Low-pred vs High-pred * Launch site				-0.009	0.003	**-2.67**
Maxjump * Predictabilities(logit)						
Maxjump * Word N				-0.113	0.007	**-16.11**
Maxjump * Word N-1				-0.012	0.006	**-10.45**
Maxjump * Word N+1				-0.023	0.006	**-3.93**
Maxjumps * Predictabilities(logit) *						
Sentence contrasts						
Maxjump * Word N * High-pred vs Proverb				0.007	0.018	**0.38**
Maxjump * Word N * Low-pred vs High-pred				0.069	0.017	**4.08**
Maxjump * Word N-1 * High-pred vs Proverb				-0.015	0.017	**-0.84**
Maxjump * Word N-1 * Low-pred vs High-pred				0.010	0.016	**0.65**
Maxjump * Word N+1 * High-pred vs Proverb				0.005	0.016	**0.32**
Maxjump * Word N+1 * Low-pred vs High-pred				0.037	0.014	**2.63**
VARIANCE COMPONENTS	*Variance*	*SD*		*Variance*	*SD*	
Groups						
Word (n=482)	0.007	0.085		0.010	0.098	
Sentence (n=184)	0.002	0.044		0.004	0.063	
Subject (n=41)	0.011	0.120		0.014	0.120	
Residual (n=21203)	0.071	0.266		0.068	0.260	

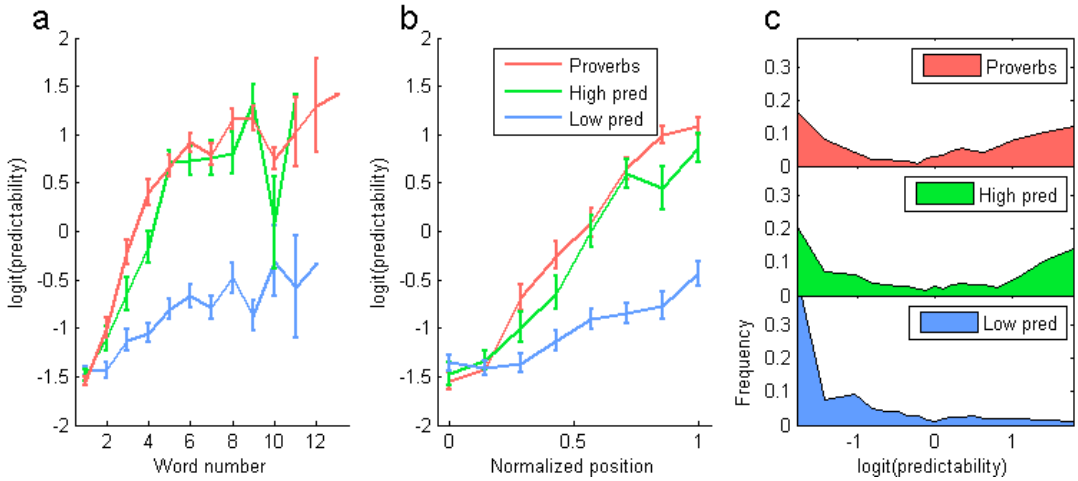


Figure 1. a- Averaged predictability of each sentence type as a function of word number.
Error bars correspond to standard errors. b- Predictability as a function of normalized word position. c- Distribution of predictability values for each type of sentence.

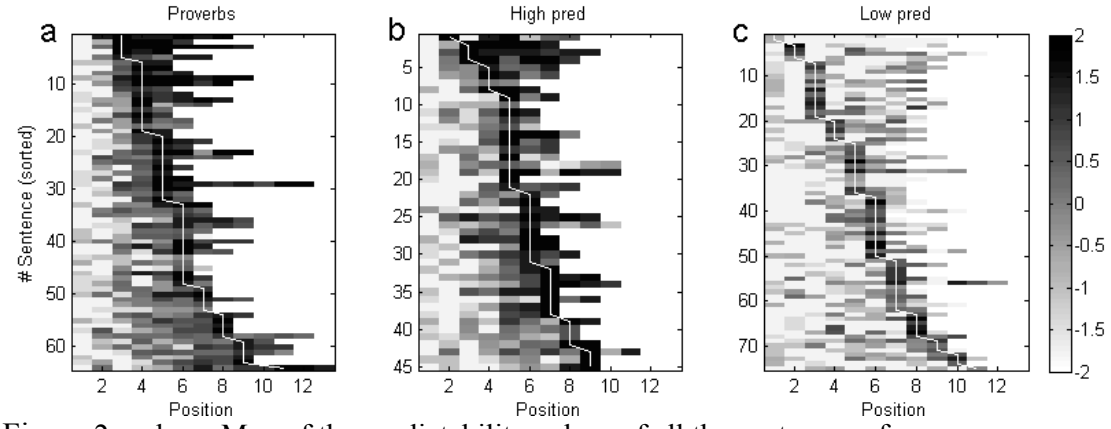


Figure 2. a- b- c- Map of the predictability values of all the sentences of each type, sorted by the word number at which the maximum value is reached. Darker gray scales correspond to higher values of predictability. White lines indicate to jump word on each sentence.

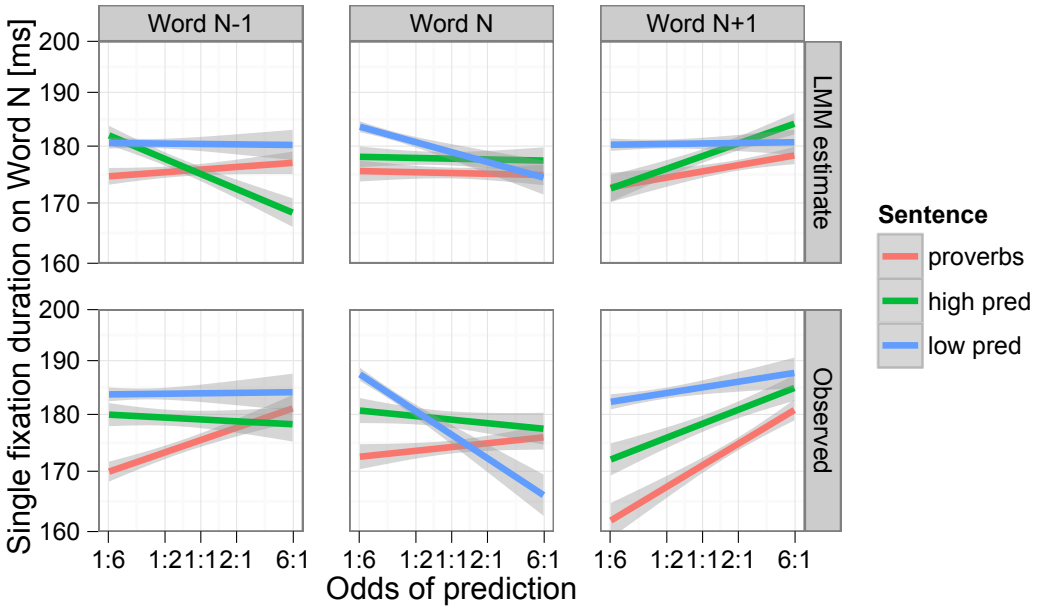


Figure 3. Predictability effects of word N-1 (left), word N (middle), and word N+1 (right) on single-fixation durations on word N, broken down by for proverbs, high-predictable sentences, and low-predictable sentences. Panels in top row reflect regression of single fixation durations on word N on respective logits of predictability; panels in bottom row show corresponding partial effects of *baseline* LMM (i.e., after removal of other fixed effects and variance components for mean fixation durations of subjects, sentences, and words). Shaded areas are 95% confidence intervals; fixation duration is plotted on a log scale for correspondence with the LMM.

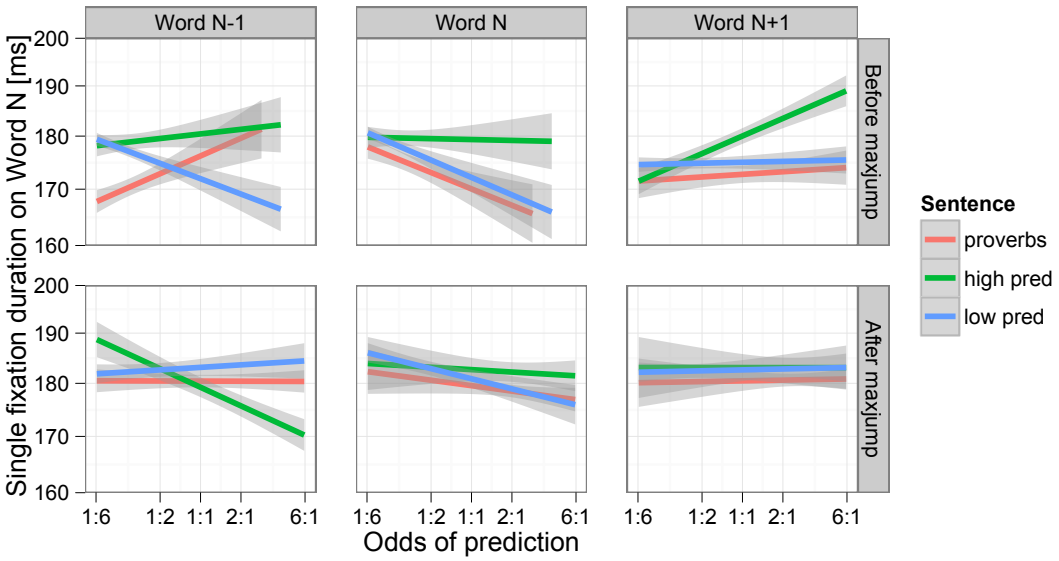
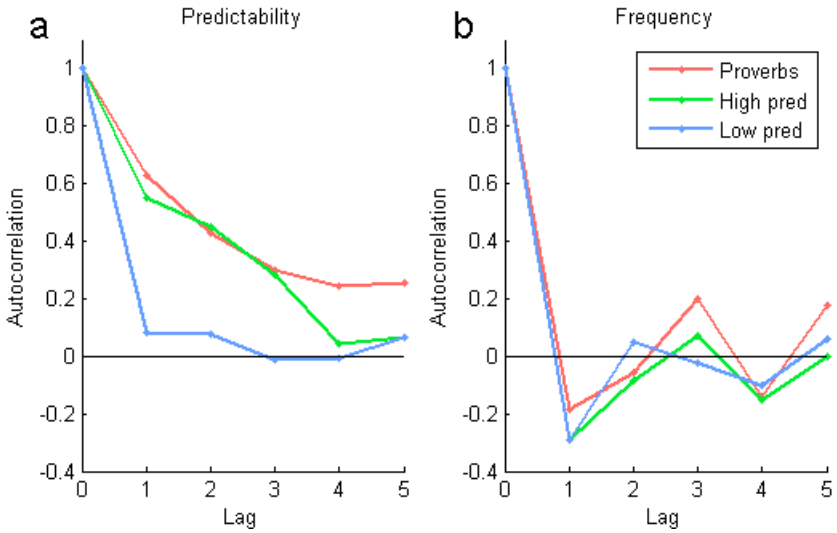
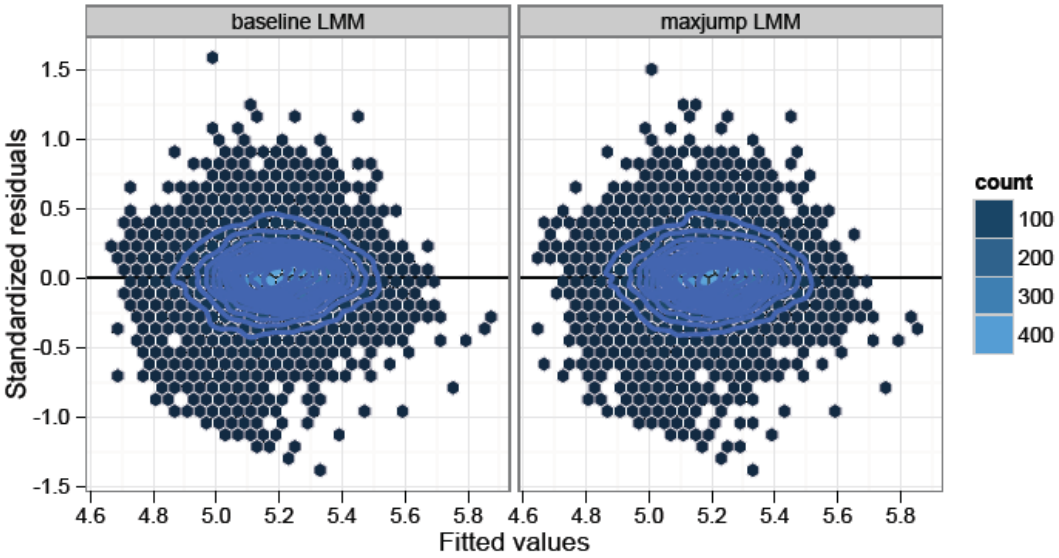


Figure 4. Interactions of *maxjump* (top panels correspond to *maxjump*=0, before jump word; bottom panels correspond to *maxjump*=1, after jump word), type of sentence (high predictable sentences vs. proverbs; low vs. high predictable sentences), and predictability of word N-1, N, and N+1 for single fixation durations on word N. Panels reflect partial effects of *maxjump* LMM (i.e., after removal of other fixed effects and variance components for mean fixation durations of subjects, sentences, and words). Shaded areas are 95% confidence intervals; fixation duration is plotted on a log scale for correspondence with the LMM.

SUPPLEMENT



Supplementary Figure 1. a- Predictability autocorrelation, b- Frequency autocorrelation.



Supplementary Figure 2. Diagnostics based on standardized residuals plotted over fitted values for *baseline* (left) and *maxjump* (right) linear mixed models.