Cognitive Psychology 88 (2016) 22–60

Contents lists available at ScienceDirect

Cognitive Psychology

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

Limits on lexical prediction during reading

Steven G. Luke [a,*], Kiel Christianson [b]

[a] Brigham Young University, United States
[b] University of Illinois at Urbana-Champaign, & Beckman Institute for Advanced Science and Technology, United States

ARTICLE INFO

Article history:
Accepted 15 June 2016
Available online 1 July 2016

Keywords:
Prediction
Reading
Word recognition
Eye movements
Cloze probability

ABSTRACT

Efficient language processing may involve generating expectations about upcoming input. To investigate the extent to which prediction might facilitate reading, a large-scale survey provided cloze scores for all 2689 words in 55 different text passages. Highly predictable words were quite rare (5% of content words), and most words had a more-expected competitor. An eye-tracking study showed sensitivity to cloze probability but no mis-prediction cost. Instead, the presence of a more-expected competitor was found to be facilitative in several measures. Further, semantic and morphosyntactic information was highly predictable even when word identity was not, and this information facilitated reading above and beyond the predictability of the full word form. The results are consistent with graded prediction but inconsistent with full lexical prediction. Implications for theories of prediction in language comprehension are discussed.

1. Introduction

In recent years, psycholinguists have devoted considerable attention to the prediction of upcoming input by language comprehenders. This body of research has shown that highly predictable words are easier to recognize (Fischler & Bloom, 1979; Kutas & Hillyard, 1984; Schwanenflugel & LaCount, 1988; Schwanenflugel & Shoben, 1985) and to produce (Piai, Roelofs, & Maris, 2014). Research has suggested that highly constraining sentences permit the generation of highly specific predictions that include

[*] Corresponding author at: Department of Psychology, Brigham Young University, 1062 SWKT, Provo, UT 84602-5543, United States. Fax: +1 801 422 0602.

E-mail address: steven_luke@byu.edu (S.G. Luke).

http://dx.doi.org/10.1016/j.cogpsych.2016.06.002
0010-0285/© 2016 Elsevier Inc. All rights reserved.

not just semantic content (Federmeier & Kutas, 1999; Federmeier, McLennan, Ochoa, & Kutas, 2002), but also morphosyntax (Luke & Christianson, 2015; Otten, Nieuwland, & Van Berkum, 2007; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Bates, Moreno, & Kutas, 2003), phonology (DeLong, Urbach, & Kutas, 2005), orthography (Laszlo & Federmeier, 2009), letter position (Luke & Christianson, 2012), and visual features of objects, such as shape (Rommers, Meyer, Praamstra, & Huettig, 2013). Predictions generated from spoken language guide eye movements when looking at images (Altmann & Kamide, 1999, 2007; Kamide, Altmann, & Haywood, 2003; Staub, Abbott, & Bogartz, 2012). Further, eye-tracking studies have shown that the predictability of a word strongly influences reading times on that word (Ashby, Rayner, & Clifton, 2005; Balota, Pollatsek, & Rayner, 1985; Ehrlich & Rayner, 1981; Kliegl, Grabner, Rolfs, & Engbert, 2004; Rayner, Slattery, Drieghe, & Liversedge, 2011; Rayner & Well, 1996; Zola, 1984; for a recent review see Staub, 2015). Thus, it appears that not only are detailed and specific predictions generated in highly constraining contexts, but that these predictions facilitate auditory and visual word and sentence comprehension. As a result, many different theoretical accounts have incorporated strong predictions as an essential component of language comprehension (Christiansen & Chater, 2016; Dell & Chang, 2014; Pickering & Garrod, 2007, 2013).

In spite of the accumulated evidence that the language processor engages in prediction, there are still many questions that remain unanswered about the nature of predictive processes in language comprehension. Huettig (2015) poses four questions about prediction in language processing, of which the present paper will focus on two: When does prediction occur, and what is predicted? Traditional views of language comprehension held the role of prediction to be limited, at best, and language processing was conceptualized as a largely bottom-up process. DeLong, Troyer, and Kutas (2014; Kutas, DeLong, & Smith, 2011) outline several reasons why this might be the case, the most prominent of which is that any given linguistic input can unfold in innumerable ways in terms of lexical content and, to a lesser extent, syntactic structure (Jackendoff, 2002). Prediction should therefore be practically limited in most contexts, irrespective of theoretical predispositions. A slew of past studies have examined this question with regard to reading (Finn, 1977; Gough, 1983; Gough, Alford, & Holley-Wilcox, 1981; Nicholson & Hill, 1985; Perfetti, Goldman, & Hogaboam, 1979; Rubenstein & Aborn, 1958; Schatz & Baldwin, 1986). Using a cloze task as an estimate of predictability, in which a group of participants are given the context up to but not including the target word and are asked to provide the next word (Taylor, 1953), these researchers generally found low average predictability (0.2–0.3), and Gough (1983) found that the distribution of predictability was strongly skewed, with relatively few highly predictable words and many unpredictable ones. Based on these observations, these early researchers identified two significant problems with prediction, as least in reading. First, most words are not predictable from context. "Relatively few words could be successfully predicted on the basis of one trial and from a knowledge of the preceding context alone" (Rubenstein & Aborn, 1958, pg. 31). Second, when a prediction was made it was likely to be incorrect. "[T]he effect of . . . context cannot be mediated by a reader's . . . predictions because, if they are precise enough to help, they are wrong too often to do so" (Gough et al., 1981, pg. 101). These concerns, and others, have led some researchers to question the necessity of predictive processes in language comprehension (Huettig & Mani, 2016).

On the surface, it appears difficult to reconcile the large body of evidence that prediction can occur with the relative unpredictability of most words in reading or listening. Is prediction a central feature of the language processor, or is it a peripheral one, a bonus that perhaps proves useful in a few specific situations? Unfortunately, language researchers have not been precise in their use of the term *prediction*. DeLong et al. (2014) note many different ways that the word *prediction* has been used by researchers (see also Staub, Grant, Astheimer, & Cohen, 2015). It can be defined as "the all-or-none process of activating a linguistic term (a word) in advance of perceptual input" (DeLong et al., 2014, p. 632), a sense that we will term *lexical prediction*. Lexical prediction is conceived of as an active process that can be facilitative if the correct word is predicted. Full lexical predictions might also be expected to incur a processing penalty when the prediction is incorrect (DeLong et al., 2014; Kutas et al., 2011; Posner & Snyder, 1975), although this may depend on whether the processor keeps multiple representations active or 'commits' to a single prediction (Kuperberg & Jaeger, 2016). Importantly, lexical prediction is unlikely to occur unless context is sufficiently constraining. A sentence fragment like ''I saw a ___'' does not provide enough constraint to permit full lexical prediction. If a prediction were formed for this sentence, it would most likely be wrong. This type of prediction therefore is subject to the objections raised by early researchers that most words are not predictable and that most predictions would be incorrect. Even so, there is evidence that prediction, as defined above, does occur, including many of the studies cited in the first paragraph. Lexical predictions involving a single word can be made, with little or no facilitation observed for other, non-predicted words (Kleinman, Runnqvist, & Ferreira, 2015, using a picture naming task) and such predictions produce different ERP responses when compared to contextual facilitation in the absence of specific, explicit prediction (Brothers, Swaab, & Traxler, 2015).

Prediction has also been used to mean broad activation of linguistic features in advance of the bottom-up input. This type of prediction, which we will term *graded prediction*, contrasts strongly with lexical prediction in many ways. While lexical prediction is active, all-or-nothing, potentially costly, and restricted to highly constraining contexts, graded prediction is conceived of as passive, diffuse, cost-free, and ubiquitous. Given these differences, graded prediction is not necessarily subject to the same objections as lexical prediction is, and Staub (2015) argues that the weight of the eye-tracking evidence from reading studies supports this definition of prediction. However, little is known about when graded prediction might occur and what the content of such predictions might be, since most research has, implicitly if not explicitly, defined *prediction* as full lexical prediction. Huettig and Mani (2016) point out that the majority of published studies investigating prediction focus exclusively on highly predictable content words for which conscious prediction is possible and likely does occur. For example, in one of the earliest eye-tracking studies to investigate the influence of predictability on reading, Ehrlich and Rayner (1981) compared highly predictable words (mean cloze score = 0.93) with non-predictable words (mean cloze score = 0.15). Other studies have used more lenient criteria for high predictability, but typically words are not considered as ''predictable'' unless their close score is higher than about 0.67. The selection of such stimuli can be interpreted as an implicit endorsement of all-or-nothing lexical prediction as the mechanism used by the language comprehension system. If prediction is not all-or-nothing, than pre-activation of linguistic features should not be restricted to highly constraining context but should be observed across the full range of contextual constraint. To effectively distinguish between lexical prediction and graded prediction it is necessary to investigate intermediate levels of constraint as well.

There is some evidence suggesting that effects of predictability may be observed even in moderate or low constraint contexts. In an eye-tracking study, Rayner and Well (1996) observed differences in reading times between low- and medium-constraint words, as defined by cloze scores, as well as between low- and high-constraint words. Other eye-tracking studies that have addressed this issue have reported mixed results (Lavigne, Vitu, & d'Ydewalle, 2000; Rayner, Li, Juhasz, & Yan, 2005; Rayner, Reichle, Stroud, Williams, & Pollatsek, 2006), but Staub (2015) notes that the overall trend is for reading times for the moderate-constraint group to fall in between the high and low constraint groups, suggesting a continuous effect of predictability. Using transitional probability as a measure of predictability, Smith and Levy (2013) observed a continuous logarithmic influence of predictability on reading times in the Dundee eye-tracking corpus (Kennedy, Hill, & Pynte, 2003). Further, ERP studies examining the effect of cloze scores on the N400 component typically observe a linear decrease in N400 amplitude as cloze scores increase (e.g. Kutas & Hillyard, 1984; Wlotko & Federmeier, 2013; although this linearity may be the result of summing different hemispheric responses). Overall, though, more research is needed to determine whether prediction has a continuous influence across the entire predictability span and, if so, what shape this function takes.

If graded prediction does occur, then it should be possible to generate partial predictions about upcoming input that do not include the full word form. The few studies that have investigated this issue suggest that such partial predictions can be made. Szewczyk and Schriefers (2013) showed that Polish readers appear to make predictions regarding the animacy of a word even when the identity of the word is not predictable from context. Roland, Yun, Koenig, and Mauner (2012) investigated the effect of semantic similarity on word reading times, and showed that when the word encountered is similar in meaning to the expected word, reading times are faster. These studies suggest that predictions can contain some semantic information even in less constraining contexts. Predictions in these contexts may include morphosyntactic information as well. Dikker and colleagues (Dikker, Rabagliati, Farmer, & Pylkkänen, 2010; Dikker, Rabagliati, & Pylkkänen, 2009) showed that the language processing system detects syntactic violations (e.g., a noun is encountered when a verb is expected) at an extremely early stage, and argue that this may be because word class is predicted by the processor. Luke and Christianson (2015) provide evidence that the morphological structure of English verbs (e.g., the presence of a suffix) can also be predicted from syntactic context even when the verb itself is not predictable. Taken together, these studies suggest that, consistent with graded prediction, other information besides the identity of an upcoming word is predictable from context, even when that context is not constraining enough to permit prediction of a particular word.

The study reported here explores prediction in the context of reading. The goal of the study was to look for evidence of prediction in reading, and to determine whether that prediction is lexical or graded in nature. We hypothesized that if language processing relies on lexical prediction, then predictions will be formed only when context is sufficiently constraining to permit them. Further, when context is misleading, processing costs might be incurred when a prediction is likely to be wrong. Finally, predictions should be highly detailed, typically specifying a single word. On the other hand, if prediction is graded then even mildly constraining contexts can influence processing and predictions can vary in their specificity in a graded fashion. To this end, we developed what is currently the largest corpus of cloze probabilities for words in extended texts, and subsequently recruited separate participants to read these texts on an eye-tracker. The cloze probabilities were then used as predictors of reading (eye movement) patterns within those same texts. Our corpus differs from other, similar corpora that contain both predictability and eye-movement data in significant ways. The Dundee corpus (Kennedy et al., 2003) is quite large, but cloze probabilities are only available for a subset of words (Kennedy, Pynte, Murray, & Paul, 2013). While cloze scores for every word in the Potsdam corpus are available (Kliegl, Nuthmann, & Engbert, 2006; Kliegl et al., 2004), it is a corpus of sentences, rather than texts. Our corpus consists of multi-sentence texts, and cloze scores were obtained for every word in each text, excluding the first word.

Analyses of the cloze procedure data showed that, consistent with past research (Gough, 1983; Gough et al., 1981; Rubenstein & Aborn, 1958), highly predictable words are quite rare in connected texts, and that for most words in a given text readers would expect a word other than the one that is actually present. Further, the eye-tracking data showed robust effects of cloze probability, even for low- and moderately constrained contexts, and no evidence for processing costs from misprediction. At the same time, partial semantic and syntactic information was often readily and consistently available from context even when word identity was not. The predictability of this semantic and syntactic information influenced reading times above and beyond cloze scores, suggesting that predictions generated from context are most often graded rather than all-or-nothing. These results are consistent with graded prediction but not with lexical prediction.

2. Method

The present study involved two stages. First, cloze scores were collected via an online survey for each word in 55 paragraphs taken from various sources. Second, these same paragraphs were presented to a different set of participants, who read them while their eye movements were being tracked.

2.1. Participants

2.1.1. Survey

Four hundred seventy-eight participants from Brigham Young University completed an online survey for course credit through the Psychology Department subject pool. Responses from eight participants were discarded because they were not native speakers of English or did not complete the survey. In total, data from 470 people (267 females) were included. Participants' ages ranged from 18 to 50 years (*M*: 21). All were high school graduates with at least some college experience, and approximately 10% had received some degree beyond a high school diploma.

2.1.2. Eye tracking

Eighty-four participants from Brigham Young University completed the eye-tracking portion of the study. All participants were native English speakers with 20/20 corrected or uncorrected vision. They received course credit through the Psychology Department subject pool. None had participated in the survey.

2.2. Materials

Fifty-five short passages were taken from a variety of sources, including online news articles, popular science magazines, and works of fiction. These passages were an average of 50 words long (range: 39–62) and contained 2.5 sentences on average (range: 1–5). Sentences were on average 13.3 words long (range: 3–52). Across all texts, there were 2689 words total, including 1197 unique word forms.

The words were tagged for part of speech using the Constituent Likelihood Automatic Word-tagging System (CLAWS; Garside & Smith, 1997). Using the tags provided by CLAWS, words were then divided into 9 separate classes. In total, the passages contained 227 adjectives, 169 adverbs, 196 conjunctions, 364 determiners, 682 nouns, 287 prepositions, 109 pronouns, 502 verbs, and 153 other words and symbols. In addition, inflectional information was also coded for the words within each class where appropriate. Nouns were coded for number and verbs were coded for tense.

Target words ranged from 1 to 15 letters long (*M*: 4.76). The frequency of each target word was obtained from the Corpus of Contemporary American English (COCA; Davies, 2009). Word frequencies ranged from 0 to 55,701 (*M*: 7934) words per million. Frequencies were log transformed for analysis. Transitional probabilities, which have been implicated as a possible source of contextual information in reading (McDonald & Shillcock, 2003a, 2003b; but cf. Frisson, Rayner, & Pickering, 2005), were computed from the COCA corpus by dividing the frequency of the collocation of the target word and the previous word (e.g., the frequency of "I agree") by the frequency of the previous word alone (e.g., the frequency of "I" in all contexts). This provides an estimate of the predictability of the upcoming word, given that the reader has just encountered the previous word. Transitional probabilities in the corpus ranged from 0 to 0.91 (*M*: 0.039), and the distribution was highly skewed, with only about 10% of words having a transitional probability of greater than 0.1. A measure of the semantic association between the target word and the entire preceding passage context was obtained using Latent Semantic Analysis (Landauer & Dumais, 1997). This LSA context score was obtained using the General Reading – Up to First Year of College topic space with 300 factors. LSA cosines typically range from 0 to 1, with larger values indicating greater meaning overlap between two terms. LSA context scores ranged from 0.03 to 0.97 (*M*: 0.53). Target word position within the passage (sentence number) and within the sentence (word-in-sentence number) were also obtained.

2.3. Apparatus

For the eye-tracking portion of the study, eye movements were recorded via an SR Research Eyelink 1000 plus eye tracker (spatial resolution of 0.01°) sampling at 1000 Hz. Subjects were seated 60 cm away from a monitor with a display resolution of 1600 × 900, so that approximately 3 characters subtended 1° of visual angle. Head movements were minimized with a chin and forehead rest. Although viewing was binocular, eye movements were recorded from the right eye. The experiment was controlled with SR Research Experiment Builder software.

2.4. Procedure

2.4.1. Survey

Participants completed an online survey administered through Qualtrics Research Suite software (Qualtrics, Provo, UT). Participants first answered a few demographic questions (gender, age, education level, language history), then proceeded to complete the main body of the survey. For each question, participants were instructed to "Please type the word that you think will come next." Beneath this instruction was a portion of one of the texts, with a response box for the participant to type in a word. For the first question about a text, only the first word in the text was visible, then the first two words for the second question, the first three words for the third, and so on, until for the last question about a text all words in the text but the final word were visible. Thus participants provided responses for all but the first word in each text. Participants were required to give a response before proceeding to the next question, and within a text all questions were presented in a fixed order, so that participants were never given a preview of the upcoming words in a text.

Each participant was randomly assigned to complete five texts, giving responses for an average of 227 different words. For each word in each text, an average of 40 participants provided a response (range: 19–43).

2.4.2. Eye-tracking

Participants were told that they would be reading short texts on a computer screen while their eye movements were recorded. These texts were the same 55 texts that were used in the survey. Each trial involved the following sequence. The trial began with a gaze trigger, a black circle presented in the position of the first character in the text. Once a stable fixation was detected on the gaze trigger, the text was presented. The participant read the text and pressed a button when finished. Then a new gaze trigger appeared and the next trial began. Texts were presented in a random order for each participant. Participants had no task other than to read for comprehension.

3. Results

Responses were edited for spelling. When a response contained contractions or multiple words, the first word was coded. Each survey response was then tagged for part of speech using CLAWS, and the responses were then divided into word classes and coded for inflection as described previously for the target words. Responses and targets (the word that actually appeared in that position in the text) were compared to see if they matched in three different ways: orthographically (cloze score), by word class, and (for nouns and verbs) by inflection. Responses and target were considered to match orthographically if the two full word forms were orthographically identical. For the purposes of this comparison all letters were in lower case. A word class match was coded if the response and target belonged to the same word class, and an inflectional match was coded if the words belonged to the same word class and carried the same inflectional suffix. Latent semantic analysis (Landauer & Dumais, 1997) was also used to provide an estimate of the relatedness of the responses and targets for all content word targets. The Latent Semantic Analysis (LSA) cosine between each response and target was obtained using the General Reading topic space via the web-based LSA interface (lsa.colorado.edu). Note that this procedure, which compared the response and target words, is different from the LSA procedure previously described, in which the target words were compared to the entire preceding passage. Comparing two words together provides an estimate of the semantic relatedness of these two words, while comparing the target word with its context estimates the contextual fit of the target word. The latter (word in context) was used in the analyses of survey responses, while the former (comparing two words) was used in the eye-tracking analyses.

Prior to the analysis of eye-tracking data, the data were cleaned, with fixations shorter than 80 and longer than 800 ms removed (about 4% of the data). Different reading measures were computed for pre-defined interest areas around each word in each passage, comprising the letters of each word and half of the white space surrounding each word, both vertically and horizontally. These included measures of reading time: first fixation duration, the length of the first fixation on a word; gaze duration, the sum of all fixations on a word before leaving it for the first time; total time, the sum of all fixations, including rereading; and go-past time, the sum of all fixations, including fixations on previous words, from the time a word is first fixated until the eyes exit the word and move past it to the right. Four binary measures were also included: word skipping probability, whether a word was passed over on first reading; word refixation probability, whether a word was fixated more than once before moving on; regression in probability, whether the reader's eyes returned to a word after having moved on to the right; and regression out probability, whether the reader's eyes left the word to re-read earlier material before moving past it to the right.

3.1. Survey results

3.1.1. Influences on cloze scores

Staub et al. (2015) showed that participants in a cloze task respond more rapidly when contextual constraint is higher. Smith and Levy (2011) showed that lexical properties of a word such as frequency, length, familiarity, and age of acquisition can influence the likelihood that a given word will be produced in a cloze task. These results underscore the fact that the cloze task is a production task, and so responses on the cloze task should be influenced by lexical variables, such as word length or frequency. In order to explore this, and to investigate how strong the effects of lexical, semantic, and positional variables are on cloze task performance, we performed a series of analyses on the survey results to investigate the variables that influence orthographic match probability (cloze scores). Binary match probabilities were analyzed using logit mixed models, with the lme4 package (Bates, Mächler, Bolker, & Walker, 2016) in R 3.2.3 (R Core Team, 2015) with random by-participant intercepts. Random slopes for the different predictors were included if they both contributed to model fit, as indicated by likelihood ratio tests, and had an influence on the significance of the results. This was done because including multiple random slopes prevented the logistic models from converging, so each slope was tested individually to avoid Type I errors that can be associated with incomplete random effects structure (Barr, Levy, Scheepers, & Tily, 2013). Only the random slopes for word length in the analysis of function words had any significant influence on the model output. Content words and function words were analyzed separately. Predictors included the following characteristics of the target word: Word Length, Word Frequency (log transformed), Transitional Probability, and LSA context score (the semantic relationship between the target word and the previous context, as measured by latent semantic analysis (Landauer & Dumais, 1997)), Sentence Number, and Word-In-Sentence Number (position of the word in the sentence). All predictors were scaled using the scale( ) function in R and then centered so that all means were 0.

In the analysis of content words, all predictors were highly significant (see Table 1). The probability of an orthographic match increased significantly as sentence number and word-in-sentence number increased, indicating that content words became more predictable as the reader progressed through each sentence and the passage as a whole. Orthographic match probability also increased as Word Frequency increased. Word Length was also a significant predictor, with a decreased probability of an orthographic match as length increased. Transitional Probability was a significant predictor of orthographic match probability as well. LSA Context Score was also a significant predictor, with increasing match probability as LSA Context Score increased. In sum, shorter, more frequent content words that occurred later in the sentence and later in the passage, that were more likely to follow the preceding word, and that were more semantically related to the content of the previous passage were more predictable from context.

The results for the analysis of function words was similar for Word Frequency, Transitional Probability, LSA Context Score, Sentence Number, and Word-In-Sentence Number (see Table 2). However, when controlling for these other factors, the length of the word had no influence on the probability of an orthographic match. This is likely because function words are more restricted in the range of their lengths than content words.

Table 1
Model output of the analysis of orthographic match probability for content word targets.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.95	0.02	−96.81	<.0001
Word length	−0.21	0.018	−11.61	<.0001
Word frequency	0.27	0.019	14.15	<.0001
Transitional probability	0.11	0.012	9.29	<.0001
LSA context score	0.24	0.014	16.47	<.0001
Sentence number	0.16	0.012	13.4	<.0001
Word-in-sentence number	0.28	0.011	24.27	<.0001

Table 2
Model output of the analysis of orthographic match probability for function word targets.

	b	SE	z value	p-value
(Intercept)	−0.91	0.016	−58.26	<.0001
Word length	0.024	0.016	1.56	.12
Word frequency	0.52	0.017	30.28	<.0001
Transitional probability	0.46	0.012	38.67	<.0001
LSA context score	0.0052	0.015	3.42	.00062
Sentence number	0.049	0.012	3.95	<.0001
Word-in-sentence number	0.23	0.012	19.69	<.0001

These results serve as a useful reminder that the cloze procedure is a production task (Smith & Levy, 2011; Staub et al., 2015) and is subject to the same cognitive constraints as other production tasks: words that are longer and/or less frequent can be harder to produce (and thus to generate in a cloze procedure). These findings also indicate that the predictability of a word increases both within a given sentence and within a passage, as the number of sentences accumulates. Additionally, it is interesting to note that the size of the Transitional Probability effect was small, at least for content words. At the same time, the sizes of all these effects taken together were relatively small. The conditional R^2 (from the sem.model.fits function from the R package piecewiseSEM (Lefcheck, 2015)) for the content words model was 0.14, and for the function words model was 0.18. These low R^2 values indicate that while the fixed effects have significant predictive value, they still leave a great deal of variance in the cloze scores unaccounted for. In other words, cloze scores capture information about the predictability of upcoming words that is much greater than the sum of these lexical, positional and semantic predictor variables, and should thus be preferred as an index of predictability (Staub, 2015; Staub et al., 2015).

3.1.2. How frequent are highly constraining contexts?

As noted in the introduction, researchers investigating the influence of predictability on reading and language processing typically use highly predictable or unpredictable words (e.g., words with very high or very low cloze probabilities) as stimuli. Studies comparing high and low constraint contexts have shown that prediction appears to occur when context is highly constraining (e.g. Ehrlich & Rayner, 1981). However, because of the unbounded nature of language, such highly-constraining contexts appear to be quite rare in natural language (Gough, 1983; Gough et al., 1981; Jackendoff, 2002; Rubenstein & Aborn, 1958). We used the results of our survey data to quantify how rare such contexts actually are, by examining the distribution of cloze scores.

Table 3 shows mean cloze scores, as well as the proportion of highly predictable words (>0.67 cloze probability) by word class. Mean cloze scores were observed to be generally low, consistent with previous studies (Gough, 1983; Gough et al., 1981). These numbers further indicate that highly-predictable words are not common, comprising about 5% of content words and about 19% of function words in the texts examined here. Fig. 1 shows the distribution of cloze scores for both content words and function words. Both distributions are skewed to the right, with many more words falling at the low end of the cloze score continuum (i.e. low predictability) than the high end. This skewness is more pronounced for content words than for function words. It should therefore be apparent from Fig. 1 that highly predictable words are the exception rather than the rule; the majority of content words are not even moderately predictable from context.

Fig. 2 and Table 3 show that the predictability of content words varies by word class. Nouns and verbs were rarely highly predictable, but more often than adverbs and especially than adjectives; most adjectives were unpredictable from context. This makes sense when one considers that verbs and nouns are required elements in a sentence, while adverbs and adjectives, which modify verbs and nouns, respectively, are optional elements and are therefore more difficult to predict.

Finally, Table 3 reports the proportion of target words, by word class, that were the modal response, the most frequent response given by survey participants. Of all target words, only about 21% of content words and 40% of function words were the modal response. This means that for the

Table 3
Summary of survey results for each of the different major word classes. Cloze probability represents the mean proportions of responses that matched the target orthographically. Highly predictable targets were defined as target words with probability scores >0.67.

	Content words					Function words					Grand mean
	Nouns	Verbs	Adjectives	Adverbs	Mean	Prepositions	Pronouns	Conjunctions	Determiners	Mean	
Mean target cloze probability	0.17	0.16	0.07	0.13	0.13	0.38	0.28	0.26	0.28	0.3	0.22
Proportion of highly predictable targets	0.08	0.05	0.02	0.04	0.05	0.28	0.13	0.11	0.13	0.19	0.11
Proportion of targets that were modal response	0.26	0.28	0.09	0.19	0.21	0.46	0.35	0.4	0.4	0.4	0.3

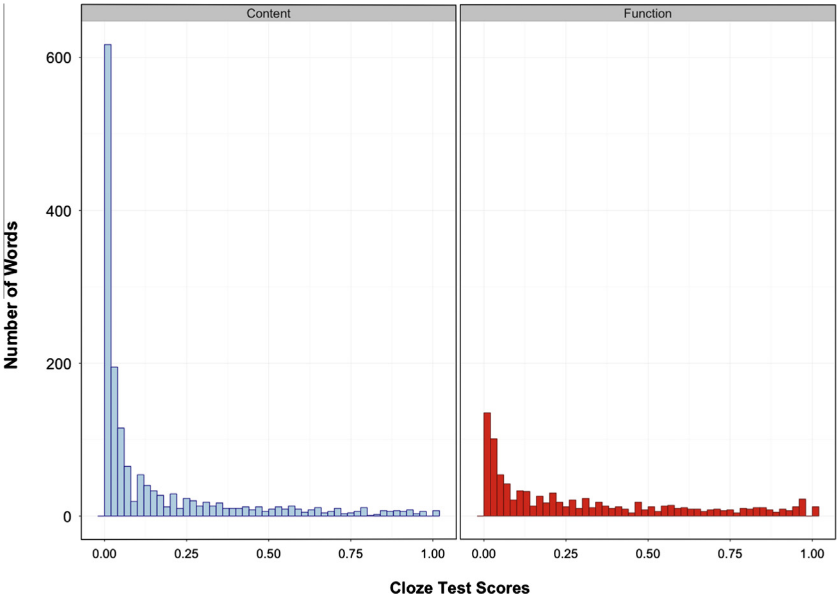

Fig. 1. Histogram of cloze scores for content (left) and function words (right). Bin width is 0.02.

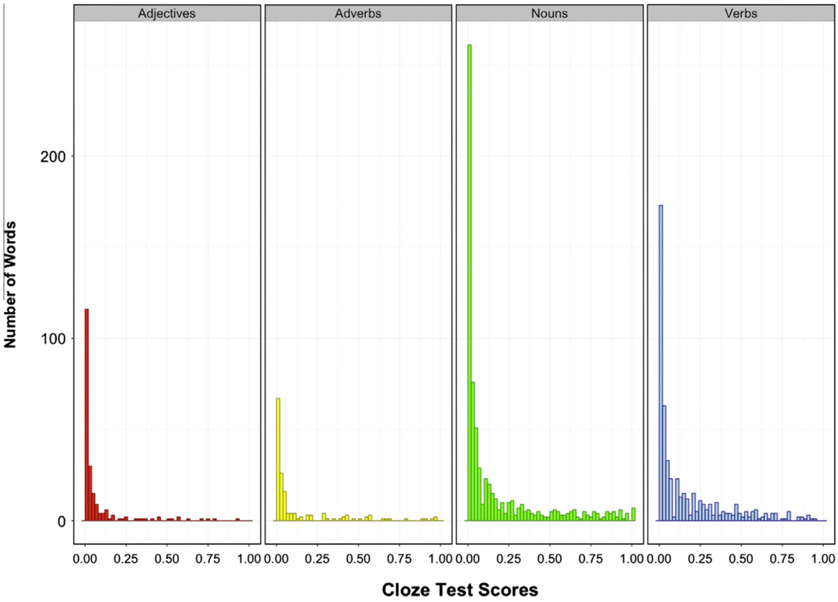

Fig. 2. Histogram of cloze scores for different classes of content words. Bin width is 0.02.

majority of both content and function words, some word other than the target word was more expected by participants.

In sum, it appears that highly constraining contexts that permit specific lexical predictions are quite rare in passages of connected text. Further, for most words in a given text readers will expect

Table 4
Summary of syntactic relationship between target and responses for each of the different major word classes.

	Content words					Function words					Grand mean
	Nouns	Verbs	Adjectives	Adverbs	Mean	Prepositions	Pronouns	Conjunctions	Determiners	Mean	
Mean part of speech match probability	0.72	0.72	0.30	0.25	0.5	0.42	0.34	0.51	0.42	0.42	0.46
Proportion of targets with highly predictable word class	0.69	0.67	0.15	0.08	0.4	0.2	0.13	0.35	0.28	0.24	0.32

Table 5
Results of LSA comparison.

	Nouns	Verbs	Adjectives	Adverbs
Mean latent semantic analysis score	0.34	0.47	0.27	0.55
Mean LSA score from random target-response pairings	0.17	0.34	0.25	0.42
t-test by target word	$t(638) = 15.85$, $p < .0001$	$t(512) = 13.53$, $p < .0001$	$t(214) = 2.3$, $p = .031$	$t(165) = 6.5$, $p < .0001$

a word other than the one that is actually present. While there is considerable evidence that prediction can and does occur when contextual constraint is strong, if predictive mechanisms in language processing require a high level of constraint, then such mechanisms can be expected to be inoperable for a majority of words. If such a mechanism were employed, most lexical predictions would turn out to be actively misleading.

3.1.3. What other information is predictable from context?

Typically, the only result reported for cloze tasks is cloze probability, the likelihood of a particular response being given. These scores reflect the likelihood of full orthographic matches between responses, and as we have seen there are relatively few cases where most participants produced a response that was an orthographic match to the actual target word. However, responses may match the target word in other ways: responses may have the same part of speech as the target word, or may be semantically similar. A response that matches the target word syntactically and/or semantically but not orthographically indicates that the participant was able to use context to predict something about the target word even if the actual word itself was not fully predictable.

Table 4 indicates that word class can be highly predictable from context. Note that while mean cloze scores are quite low (Table 3), especially for content words, responses matched targets' part of speech at much higher rates, sometimes as high as 0.7 (for nouns and verbs). Further, about 2/3 of the nouns and verbs were classified as highly predictable (>67% POS match). Thus, the part of speech of an unpredictable word can still be highly predictable.

In addition to word class, morphosyntactic information should logically be predictable from context as well. In English, nouns carry information about number, and verbs are morphosyntactically marked for number and for tense. In the present data, when the target word was a noun, responses and targets carried the same inflection (e.g., both were plural nouns) about 72% of the time (70% for plural nouns, 75% for singular). For verbs, the response was also a verb with the correct inflection (tense) 78% of the time (range for the different tenses: 74–87%). This suggests that morphosyntactic information may be readily available to the predictive system.

Word class and inflectional information are associated with syntactic processing, and so can be predicted independently of semantic information. In other words, knowing that an upcoming word will be a verb does not necessarily entail knowing anything about the meaning of the upcoming verb. Semantic information might be predictable from context as well (Wlotko & Federmeier, 2015); predicting semantic content is clearly a prerequisite for predicting the full word form. To investigate the prediction of semantic information, mean LSA Match Scores (a measure of the semantic relationship between the target and response; see Materials, above) were computed for content words. These means are reported in Table 5. As a control, target-response pairings were randomized and the LSA scores recomputed. For all content word classes, the mean LSA score was significantly lower when the target-response pairs were randomized. These means indicate that responses were, on average, semantically related to the target words. We note that the meaning of a given LSA score can be difficult to evaluate in isolation. To facilitate this, here are some example target-response pairs with the mean score for each word class: Nouns – process-procedure, Verbs – reduce-decrease, Adjectives – highest-top, Adverbs – nearly-not.

In sum, the present results are consistent with previous research (Dikker et al., 2009, 2010; Luke & Christianson, 2015; Roland et al., 2012; Szewczyk & Schriefers, 2013) in providing evidence that the language processing system could make less-detailed, or graded, predictions that do not specify the

Table 6
Means (and standard deviations) for all eye movement measures.

	Content	Function
Word skipping probability	0.32 (0.47)	0.62 (0.49)
First fixation duration	214 (84)	201 (79)
Refixation probability	0.21 (0.41)	0.09 (0.28)
Gaze duration	261 (145)	216 (103)
Regression in probability	0.18 (0.38)	0.23 (0.42)
Total reading time	314 (201)	245 (144)
Regression out probability	0.17 (0.38)	0.16 (0.36)
Go-past time	363 (396)	302 (362)

full word form. Such predictions could include information about part of speech, morphosyntax, and semantics, and may occur even if the actual identity of the word is not predictable from context.

3.2. Eye movements in reading

In order to explore the questions outlined in the introduction – when does prediction occur and what information is predicted – a series of analyses on the eye-tracking data were conducted. Primary dependent variables included reading time measures (first fixation, gaze duration, total time, go-past time), word skipping probability, word refixation probability, and the probability of regressions in and regressions out. Summary statistics for all dependent measures analyzed are available in Table 6. These dependent measures were analyzed using linear or logit mixed models, with the lme4 package (Bates et al., 2016) in R (R Core Development Team, 2015). Response times were log transformed. Linear models included random by-participant and by-word intercepts and all possible random slopes, except where non-convergence necessitated the removal of a slope, in which case the slope for the effect with the larger *t* value was removed, as removal of this slope was least likely to allow a non-significant effect to reach threshold and thus lead to a Type 1 error. When the linear model investigated an interaction, a random effect for this interaction was modeled only when the fixed effect interaction was first observed to be significant in a simpler model that included no random slope interactions. *P*-values for linear models were obtained using Satterthwaite's approximation, as implemented in the lmerTest package in R (Kuznetsova, Brockhoff, & Christensen, 2014). Logit models (for skipping, refixations, and regressions) included random by-participant intercepts, and any random slopes that were found to contribute to model fit, as indicated by likelihood ratio tests, and to influence the significance of the results. This was done because these models did not converge when the random effects structure was too complex or when random by-word intercepts were included. Content words and function words were analyzed separately. Overall patterns of results are reported in the text, and full model outputs for all models that included more than one fixed effect are presented in Appendix A.

Predictability (cloze score) was log transformed in all analyses. Some of the analyses reported below are specifically comparing slopes of different functions, so we felt it was important to select the appropriate transformation. Other researchers doing similar large-scale analyses of eye-tracking data transformed predictability using a logit function (Kliegl et al., 2004). By contrast, Smith and Levy (2013) suggest that the relationship between predictability and response time is logarithmic. Finally, for simplicity it may be preferable to not transform the data at all. To evaluate the desirability of these different transformations, we compared the goodness of fit of several simple linear/logistic regression models (12 in total, one for each dependent variable except go-past time and regressions out in the content words data and again in the function words data) to see which transformation fit the data best. The marginal R^2 (from the sem.model.fits function from the R package piecewiseSEM (Lefcheck, 2015)) was lowest for the raw cloze scores in all but 1 case, indicating that these models generally explained the least variance in the data. The R^2 values were comparable for the logit and log transformations, but because the log transformation is somewhat more empirically well-founded (Smith & Levy, 2013), we chose to employ the log transformation.

Values of 0, which were very common in our cloze scores, cannot be log transformed, and so in order to perform this transformation it was necessary to replace the raw cloze scores with the fitted values generated by models similar to those reported in Tables 1 and 2 which included only random by-word intercepts. The fitted values obtained from these models ranged from 0.005 to 0.977, never including 0 or 1. The correlation between the raw cloze scores and the fitted cloze scores variables was 0.999.

3.2.1. When does word predictability influence eye movements?

As noted above (see Table 3), only about 30% of the words in the presented paragraphs were the most common response from participants in the cloze task. That means that if readers generated a prediction for each of the remaining 70% of the words, it was most probably not the target word. We analyzed the effect of cloze probability on reading behavior when the target word was either the modal response or not.

There were two purposes to these analyses. The first was to see if the predictability of a given word influences reading even when the word is unexpected (i.e., not the modal response). That is, when readers encounter a word that is likely not the word they would predict, does contextual constraint still operate to facilitate processing of that word? Further, is this influence similar to that observed for modal words (i.e., is the slope of the function the same when a word was the expected continuation or not)?

The second purpose was to see if there is any evidence of inhibition in the eye movement data for words that were unexpected (non-modal). One might expect that generating an incorrect prediction could slow down processing of the target word. If incorrect predictions disrupt processing, then reading might be slowed when the target word is not the word that should be predicted from context.

3.2.1.1. Does predictability influence the processing of unexpected words?. In order to compare the slopes of the cloze probability functions, the Predictability (cloze probability) for each group (modal vs. non-modal) were centered on their respective means. This was done because modal and non-modal responses by definition have different mean cloze probabilities. The question here is not whether the cloze probabilities are different (they are) but whether the slopes of the Predictability function (the decrease in reading times as predictability increases) are similar for the unexpected continuations compared to the expected ones. Models were fitted to the reading data that contained the expectation group variable (modal vs. non-modal) and these centered Predictability values. If the effect of Predictability is more pronounced when the target word was also the modal response, which would be expected if full lexical prediction provides a boost to processing, then these two variables should interact. Predictability was significant (and facilitative) in all analyses. For function words the effect of predictability was statistically indistinguishable for modal and non-modal words in all analyses (i.e. no interactions were significant). For content words, this was also true for all duration analyses, as well as refixation probability (see Fig. 3 and Appendix A). For content word skipping, the influence of Predictability was greater for modal words, suggesting that modal, highly predictable words are especially likely to be skipped. For regression in probability, the influence of Predictability that was negative for non-modal words did not appear to be present for modal words, suggesting a floor effect. For regression out probability, the effect of Predictability was statistically significant for modal words but not for non-modal words; increasing Predictability led to *more* regressions out for modal words only. These results suggest that, overall, the effect of Predictability is not restricted to high-probability words or to the most predictable words, but is generally statistically equivalent for expected and unexpected word continuations. The only indications that modal words are processed differently come from the analyses of content word skipping and regressions out: the slope of the predictability function was greater for modal words in both cases.

Because the choice of transformation could potentially influence the outcome of these analyses, we note that when raw or logit-transformed cloze scores were used instead of log-transformed scores, a somewhat different pattern of results was obtained. In many of the analyses where a log transformation resulted in insignificant interactions, the interaction was significant when either no transformation or a logit transformation was used. However, these interactions indicated that the influence of cloze scores was observed to be *weaker* for modal words than for non-modal ones, indicating that

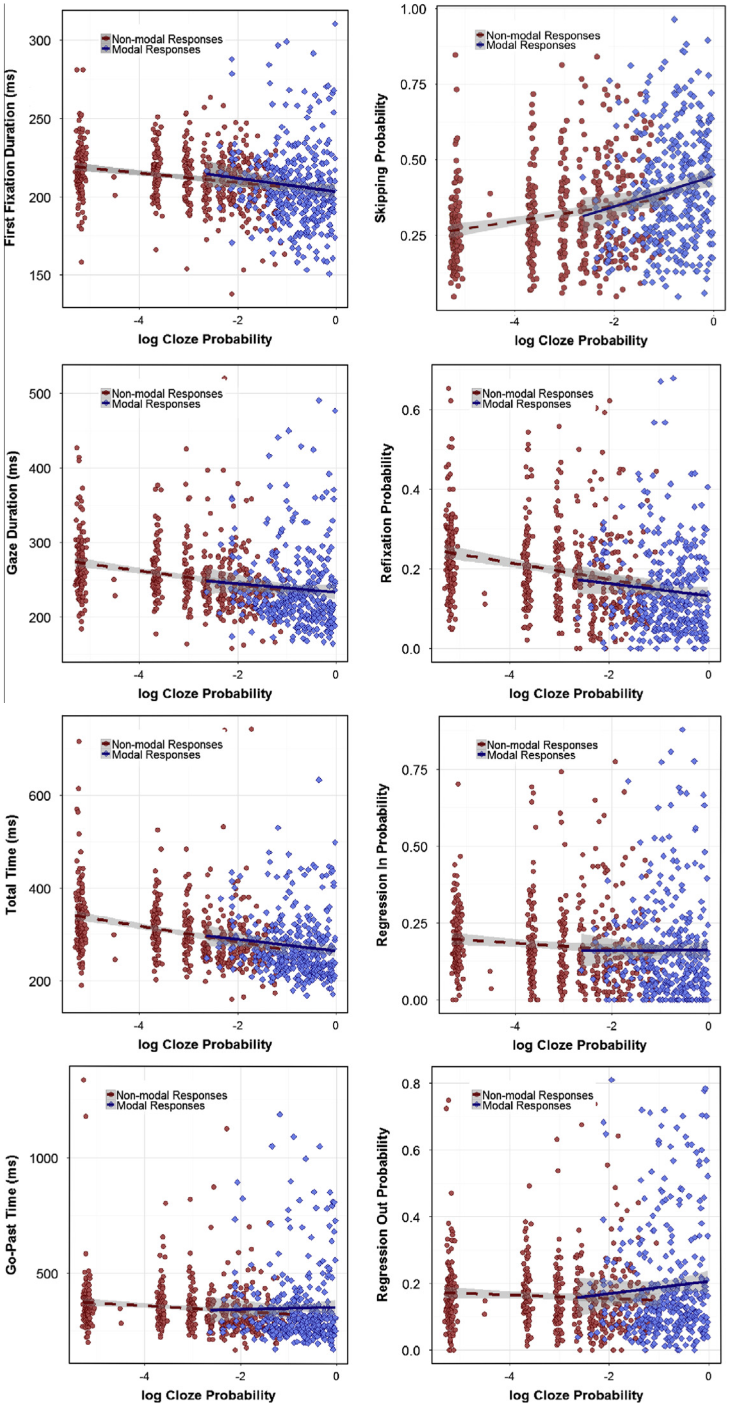

Fig. 3. Slope of the function of cloze probability on all dependent measures for content words. The dashed line represents the function for targets that were the non-modal responses on the cloze task, while the solid line represents targets that were the modal responses. Error bands represent 95% confidence intervals.

the effect of predictability flattened out for higher cloze score values. While this pattern of results is different from those reported above, it too is inconsistent with the idea that facilitation requires the generation of full lexical predictions. It does suggest that the influence of predictability is greater at the lower range of the continuum, which is consistent with the suggestion by Smith and Levy (2013) that the effect of predictability on reading time is logarithmic.

3.2.1.2. Is there a penalty for mis-prediction?. As noted in the introduction, if prediction is an all-or-nothing process involving pre-activation of full lexical forms, then it might be that an incorrect prediction would lead to a processing slow-down. In order to investigate whether mis-predictions incur a processing penalty, reading measures on target words that were not the modal response were analyzed separately. For these non-modal target words, cloze scores by definition could not exceed 0.5. For these words, the cloze probability of the modal response was also computed and used as a measure of entropy. Entropy is an informational measure of how uncertain an upcoming bit of information, in this case a word, is; the more entropy, the more uncertainty. In the present case greater entropy reflects a greater diversity of responses for a particular item on the cloze task. We computed entropy for each word using the entropy package in R (Hausser, Strimmer, & Strimmer, 2014). However, like Staub et al. (2015), we found that entropy is strongly correlated with the cloze probability of the modal response ($r = -0.95$). In other words, as participants became more unanimous in their response on the cloze task, the number of different responses decreased. We therefore, following Staub et al. and others (Schwanenflugel & LaCount, 1988), use the cloze probability of the modal response as our measure of uncertainty, as it is easier to interpret and is on the same scale as the cloze probability of the target word. Because the word *entropy* refers to Shannon entropy, a particular information-theoretic measure, we will use the term *certainty* to refer to the modal cloze probability, as a higher modal cloze probability is associated with less entropy and more certainty about the identity of the upcoming word. We will refer to the actual cloze probability of the target word as *predictability*. *Certainty* was weakly and non-significantly correlated with *predictability* ($r = 0.1$).

The logic behind this analysis is that as *certainty* increases, participants have a stronger expectation of what the upcoming word will be. However, by analyzing only those target words that were not the modal response, we restrict the data to cases where *certainty* reflects the strength of a *false* expectation; as *certainty* increases, participants have a stronger expectation of the modal response, but the target word is not the modal response, so it should be unexpected. Thus, if incorrect predictions lead to a processing slow-down, then reading times on these words should increase as *certainty* increases.

Analyses were conducted to see if *certainty* did influence reading times. Target *predictability* was significantly predictive of all reading time measures, including go-past time, with shorter reading times for higher cloze probabilities. In addition, higher target cloze scores were associated with more word skipping and fewer refixations. *Predictability* was not predictive of regressions out. *Certainty* was also predictive in some of the analyses, and was, surprisingly, mostly facilitative when it was significant (see Fig. 4 and Appendix A). For content words, higher *certainty* was associated with more word skipping and fewer refixations. The effect of *certainty* was not significant for any content word reading time measure (although the effect for total time was marginal) or for regressions in. Higher *certainty* was associated with more regressions out of content words. For function words, *certainty* was a significant predictor in all reading time analyses except go-past time, and in the analyses of refixations and word skipping, and was again also facilitative when significant. As with content words, higher *certainty* was associated with more regressions out, but was not associated with regressions in. In sum, these analyses provide little evidence that mis-predictions slow down reading. Instead, it appears that the opposite may be true; when a more predictable competitor word is available, processing of the actual, unexpected word is facilitated, even when the effect of target *predictability* is controlled for. This was especially apparent for function words. The one exception to this pattern is in regressions out. It appears that a higher *certainty* is associated with a greater likelihood of re-reading previous text, although these effects were small (see Fig. 4, bottom). The amount of time spent when re-reading (go-past time) was associated with the *predictability* of the target only. This finding is consistent with the result from the previous set of analyses that *predictability* had a stronger effect on regressions out for modal words. For these words, the target *predictability* and *certainty* were the same values, suggesting that it is the *certainty*, not the *predictability*, of a given word that increases the prob-

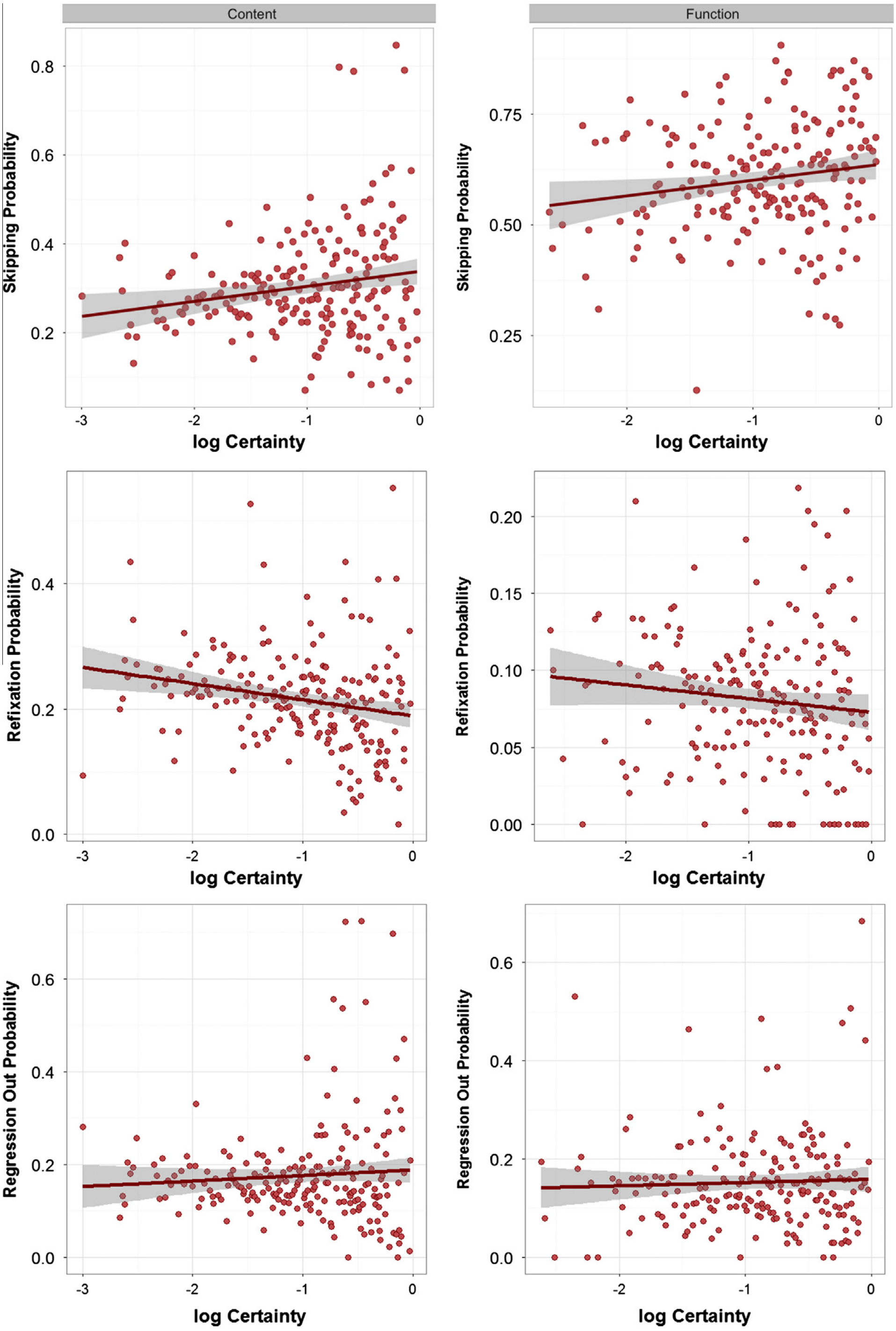

Fig. 4. Slope of the function of *certainty* on word skipping (top), refixations (middle), and regressions out (bottom) for content words (left) and function words (right). Only target words that were not the modal response in the cloze task were included in these analyses. *Certainty* represents the cloze score for the modal response. Error bands represent 95% confidence intervals.

ability of regressing from that word to previous portions of the text. A follow-up analysis revealed that the influence of *certainty* on regressions out was the same for modal and non-modal responses (i.e. the interaction was not significant; see Appendix A, Table 39).

3.2.2. Beyond cloze scores

As noted above, participants in the cloze procedure were able to reliably predict the part of speech and semantic characteristics of upcoming words as well as information about the inflectional status of nouns and verbs. These results indicate that it is possible that readers generate partial predictions about upcoming words that do not include word identity. To see if such incomplete predictions can facilitate processing, several analyses were conducted, which included as predictors the *predictability* (cloze probability) of the target word, the Part of Speech Match score (proportion of responses that had the same word class as the target), and (for content words only) the LSA Match score (the mean semantic relationship between the responses and the target word, as measured by latent semantic analysis (Landauer & Dumais, 1997)). We performed separate analyses on the nouns and verbs alone, including the aforementioned predictors as well as Inflection Match probability. All proportion data were log transformed prior to analysis, for reasons described previously. The purpose of these analyses was to see if the predictors other than predictability of the full word form (i.e. cloze probability), such as POS Match, LSA Score, and Inflection Match, were predictive of reading behaviors when *predictability* was controlled for. This was accomplished by entering all of these predictors into the same model. These variables are all correlated, by definition; in order to produce a response in the cloze task that matched the target word, a participant had to have predicted the part of speech, semantic content, and inflectional status of the word as well. When correlated predictors are entered into a single lme model, the resulting model output reflects the strength of each predictor's effect when controlling for the other predictors (Wurm & Fisicaro, 2014). We note here that we chose to include only those variables that are related by definition with cloze scores, and not others that would almost certainly be correlated, such as word length or frequency. The purpose of these analyses is to explore whether the predictability of partial information about an upcoming word influences eye movements over and above the predictability of the full word form. Including frequency or length or other lexical variables would address a different question, the question of how such partial predictions are formed and what information helps us generate these predictions, which is beyond the scope of these analyses.

3.2.2.1. Content words. LSA Match score was significant in almost all analyses, even when *predictability* was controlled for, indicating that when the responses generated by the survey participants were more semantically related on average to the target word, reading times, refixations, and regressions were reduced and word skipping probability increased. Thus, the overall pattern was facilitative, although participants did make more regressions in to words that had a higher LSA score.

Part of Speech Match probability had a significant influence on the reading of content words in all analyses except word skipping. Thus, being able to predict a word's part of speech has a generally facilitative effect over and above that of *predictability*; for two words with similar cloze probability, the one with higher Part of Speech Match probability was read faster and regressed to less often. However, higher Part of Speech Match did lead to more refixations.

In order to investigate the influence of inflectional predictability, separate analyses were conducted on the nouns and on the verbs. These analyses included Inflection Match Score in addition to *predictability*, Part of Speech Match score and LSA Match score. Models were fitted as described previously. For both nouns and verbs, Inflection Match Score had a significant effect on all eye movement measures except first fixation duration. These findings suggest that the predictability of the inflection of a noun or verb had a significant effect on eye movements in reading above and beyond other measures. The results further suggest that this effect can be seen early, in word-skipping behavior, but that if a word is fixated, the effect does not arise until gaze duration or later, suggesting that refixations and rereading are often influenced by uncertainty about a word's morphosyntax.

3.2.2.2. Function words. Part of Speech Match probability was significant in all reading time analyses and in the analysis of word skipping. It is possible that refixations were too infrequent to show any effect (see Table 6). This result indicates that the predictability of the part of speech of a function word has a consistent and significant influence on the processing of that word, and that this influence is independent of the influence of the cloze probability of the word. That is, being able to predict the word class has a benefit on reading times above and beyond being able to predict the word itself.

4. Discussion

Prediction is thought by many to be central to language comprehension (for reviews, see DeLong et al., 2014; Huettig, 2015; Kuperberg & Jaeger, 2016; Kutas et al., 2011; Van Petten & Luka, 2012). Indeed, the evidence that prediction can occur is robust (Altmann & Kamide, 1999, 2007; Brothers et al., 2015; DeLong et al., 2005; Federmeier & Kutas, 1999; Federmeier et al., 2002; Kamide et al., 2003; Kleinman et al., 2015; Laszlo & Federmeier, 2009; Luke & Christianson, 2012; Otten et al., 2007; Rommers et al., 2013; Staub et al., 2012; Van Berkum et al., 2005; Wicha et al., 2003), as is the evidence that predictable words are processed more easily (Ashby et al., 2005; Balota et al., 1985; Ehrlich & Rayner, 1981; Fischler & Bloom, 1979; Kliegl et al., 2004; Kutas & Hillyard, 1984; Rayner & Fischer, 1996; Rayner et al., 2011; Schwanenflugel & LaCount, 1988; Schwanenflugel & Shoben, 1985; Zola, 1984). Clark (2013) argues that prediction is how the brain operates, not just in language but in all domains. However, to say that prediction occurs is insufficient; the next step is to systematically identify the timing and content of predictions as well as the mechanisms that support prediction (Farmer, Brown, & Tanenhaus, 2013; Huettig, 2015).

The purpose of the present study was to contrast two different theoretical approaches to prediction in reading. *Lexical prediction* is the all-or-nothing activation of a specific lexical item before it is encountered in the linguistic input. *Graded prediction* is passive and partial. These different definitions give rise to different hypotheses about how prediction should be manifest in eye movement behavior during reading. Lexical prediction should be all-or-nothing, while graded prediction should lead to the formation of partial predictions that do not necessarily include word identity. As a result, lexical prediction should be most facilitative when predictability is highest, with little or no effect at lower levels of contextual constraint. Graded prediction should be facilitative at all levels of contextual constraint. Finally, lexical prediction might be expected to involve processing costs when the prediction is incorrect, while graded prediction should be relatively cost-free.

The results of the large-scale cloze procedure reported here show that highly predictable words are rare in connected text, comprising only about 5% of content words and 19% of function words. While these percentages can be expected to vary from text to text, it seems clear that accurate full lexical predictions are unlikely to be formed for the vast majority of words during normal reading. Further, for words that are not highly predictable, any prediction that might be generated would most likely be misleading: the actual word in the text was the most expected word only about 21% of the time for content words and 40% of the time for function words, so that for an average of about 7 out of 10 words the word encountered was not the word most likely to be predicted. Thus, a comprehension strategy that relied heavily on lexical prediction would prove both inefficient and costly; accurate predictions would be generated too infrequently to be of consistent help and incorrect predictions would be frequent. Taken together, these results suggest that the language processor is unlikely to rely on lexical prediction during reading.

In the eye-tracking data there was no evidence that mis-predictions come at a cost. Instead, the influence of cloze scores on reading was in most cases statistically identical whether the word was the expected continuation or not, and the presence of a more expected competitor word was found to be facilitative, rather than inhibitory. These results provide further evidence that the language processor does not rely on lexical prediction, as defined above. At the same time, robust effects of cloze probability were observed across the full range of cloze scores in multiple reading measures, from early measures such as skipping and first fixation to late measures such as total time and go-past time. This ubiquity of the facilitative effects of context on reading is most consistent with graded prediction.

In a recent review of prediction in language comprehension, Kuperberg and Jaeger (2016) note that:

> Most empirical work has focused on the effects of lexical constraint, as operationalised using cloze procedures . . . Contexts that are lexically constraining, by definition, constrain strongly for multiple

> types of representation (semantic, phonological, and syntactic). It is important to recognise, however, that a context can constrain strongly for just one type of upcoming representation, leading just to facilitation of incoming information at this representational level, independently of any other (pg. 11).

Consistent with this idea, the survey results indicate that features of upcoming words other than their full identity, such as word class, morphosyntax, and semantic content, were more reliably available to the processor, and the eye-tracking results indicate that the processor makes use of this information. The data thus are more consistent with graded prediction.

4.1. What is the function linking predictability and word reading time?

It appears, based on our data and the work of others (Lavigne et al., 2000; Rayner & Well, 1996; Rayner et al., 2005, 2006; Smith & Levy, 2013), that predictability has a continuous and graded effect on reading times. However, the shape of this function is not entirely clear. Knowing what shape this function takes could help to adjudicate between different theoretical proposals (Smith & Levy, 2013). In our analyses, we employed a log transformation of predictability, following Smith and Levy (2013). There is some reason to believe that this choice was appropriate: the log transformation provided an overall better fit to the data than did raw cloze scores, and the function was more linear when a log transformation was employed. On the other hand, there is some evidence that the relationship is not truly logarithmic (Staub, 2015). Also, Smith and Levy used trigram frequency as their operationalization of predictability, which is similar to transitional probability, and, as we have seen here, cloze scores from human participants appear to capture much more information than such corpus-based measures do. Clearly, more work is needed to address this issue.

4.2. Late-arising effects of predictability

In his review of predictability effects in reading, Staub (2015) notes that while early predictability effects appear to be observed consistently across reading studies, later eye-movement measures such as regressions and go-past time are not reported often enough to draw firm conclusions about whether these effects are observed late as well as early. Our analyses replicated the observed facilitative influence of predictability in early measures. Facilitative predictability effects were also observed for several later measures, including total time, go-past time, and regressions in to the word. This may indicate that predictability is involved in post-lexical integration processes as well as pre-lexical and/or lexical processes. However, the relationship between predictability and regressions made out of a word to re-read previous material was less straightforward. No effect of predictability on regressions out was observed for function words. For content words, an effect was observed for modal targets (i.e. targets that were the most-produced response in the cloze task), and this effect was in an unexpected direction: higher predictability led to *more* regressions out. Further investigation revealed that it was not predictability, the cloze score of the target word, but rather certainty, the cloze score of the modal response, that was influencing regressions out. This unexpected finding may suggest that regressions from a target word to previous portions of a text are often initiated before significant processing of the word itself has occurred. The direction of the effect, with more regressions out when certainty was greater, may suggest that when contextual constraint is greater readers have a tendency to go back and review previous context before fully processing the critical word. We note, however, that this effect was small (see Fig. 4, bottom panels), and that, overall, predictability has a significant and facilitative effect on reading in late as well as early measures.

4.3. Prediction versus integration

The present study explored the predictability of words in connected text, and the effect this predictability has on eye movements during reading. Strictly speaking, predictability effects are not direct evidence for prediction. While it does seem clear from our eye-tracking data that contextual facilitation occurs across the full range of predictability, this may either indicate that some form of anticipatory processing is occurring, or that the contextual facilitation observed here reflects ease of integration rather than pre-activation of lexical information. "Ease-of-integration" has been put forth many times as a counter-argument to prediction. However, we echo the argument by Staub (2015) that an integration account of predictability effects is inconsistent with the early onset of such effects; predictability influences the earliest eye-tracking measures, including word skipping and first fixation. These early measures are generally thought to reflect word identification processes, not integration. However, as noted above, our results suggest that predictability may also influence integration processes.

4.4. Entropy, predictability, and reading

Entropy is a measure of informational uncertainty. In our data, we operationalized entropy as *certainty*, the cloze probability of the most common response on the cloze task. This provides an inverse function of entropy: when the cloze score for the modal response is higher, it means that there is less uncertainty about what word will come next. Consider sentence fragment (1), below.

(1) Her primary achievement

For sentence fragment (1), 87% of participants in the cloze task continued the sentence with the word "was". For this sentence, there was high certainty (and thus low entropy). As another example, consider sentence fragment (2).

(2) When early Europeans discovered Easter Island, its somewhat

For this fragment, entropy was high, with the most common response, "barren", being produced only 10% of the time. Perhaps the most surprising finding from our study is that certainty was observed to be often facilitative and never inhibitory, even when the expected word was not actually encountered. The next words for these example sentences were actually "lay" and "isolated", respectively, with each word produced as a response by only one participant in the cloze task. Nevertheless, our results indicate that higher *certainty* (lower entropy) was facilitative in several analyses (see Fig. 4), even when the effect of the predictability of the actually-encountered target word was controlled for.

Staub et al. (2015) found that when participants are producing responses in a cloze task, onset of vocalization was faster when contextual constraint was higher, even for items with the same cloze probability. That is, lower entropy sped up responses, even when predictability was controlled for. This finding from a production task is highly consistent with the results from our comprehension task. Staub et al. (2015) demonstrated that this independent influence of entropy on cloze task response times can be replicated using a simple activation-based race model, in which the first word that reaches a particular threshold of activation is the one that is produced. This account makes the prediction that in high entropy contexts where no particular word is expected, it should be possible for a wide range of weakly activated words to reach threshold. However, in a low entropy context where one word is highly expected, any other word that is produced even once in the cloze task must also be highly activated. Thus, even though the word "was" is most expected in (1), "lay" must have received a great deal of contextual activation to be produced even once as a possible continuation. By contrast, for sentence fragment (2) there was a great deal of entropy, so that the actual continuation "isolated" did not need as much activation to be produced once as a response. This means that "lay" likely received more pre-activation from context than "isolated" did, even though they had similar cloze scores. This account is highly consistent with graded prediction as defined above.

4.5. Prediction costs

The strongest version of lexical prediction assumes that the processor commits in some way to a particular lexical item, and that violations of this prediction should be expected to lead to processing costs (see Kuperberg & Jaeger, 2016 for a discussion of this issue). Our data provide no evidence for prediction costs in reading. It might be argued that signs of cost exist elsewhere, however, and that eye movements do not reflect the costs. Some researchers have reported evidence for such costs when comparing words with the same cloze scores in highly constraining or non-constraining contexts, mostly using ERPs (see Kuperberg, 2013; Van Petten & Luka, 2012).

The N400 effect in ERP studies has been used as a measure of prediction in language processing, specifically linking decreases in the amplitude of the N400 waveform with facilitated processing when a highly predictable word is encountered. As Van Petten and Luka (2012) discuss at length, however, the opposite is apparently not the case: an inflated N400 does not signal a cost for a failed prediction. Federmeier, Wlotko, De Ochoa-Dewald, and Kutas (2007) found no differences in N400 amplitudes for unexpected words in highly-constraining vs. weakly-constraining contexts. Thus, whatever the relationship between N400 modulation and predictability is proposed to be, it cannot be described as signaling a cost for incorrect predictions.

One aspect of the ERP N400 literature that Van Petten and Luka (2012) do not discuss is the tracking of N400 amplitude for each word of a sentence, rather than just the final word. Since the 1990s (e.g. Van Petten & Kutas, 1990) it has been demonstrated that all words in a sentence elicit spikes in the N400 waveform, and that these spikes generally decrease as the sentence progresses and as contextual constraint accrues. For example, Payne, Lee, and Federmeier (2015) compared congruent (predictable) sentences as in (3a) to syntactic prose (3b) and scrambled (3c) sentences, tracking N400 peak amplitudes across all the words in the sentence.

(3) a. She kept checking the oven because the cake seemed to be taking an awfully long time to *bake*.
b. She went missing the spring because the court began to be making an awfully poor art to *bake*.
c. The court the she spring making missing awfully art poor to because an to be went began *bake*.

Payne et al. (2015) observed decreases in the mean amplitude of the N400 across sentence position in all conditions in both open-class and closed-class words (the latter result contrary to Van Petten & Kutas, 1991), and, notably, a much steeper decline in N400 amplitude in congruent (contextually constrained) sentences such as (3a) compared to the conditions in (3b-c). The N400 in analyses of open-class words, however, all bottomed out at the final word (*bake*). These N400 effects, then, mirror the cloze measures we report here: N400 peaks decrease as a sentence unfolds, and cloze scores increase, both of which suggest accumulating effects of contextual constraint. Given that our cloze data here (and older data, e.g., Gough et al., 1981) show that 95% of open-class words are *not* anywhere near as contextually constrained as in typical ERP materials, it is reasonable to conclude that the normal state of affairs for the N400 is a spike at every word, and it is only in *abnormal* or atypical instances in which the N400 amplitude drops so dramatically, beyond a more generalized relatively monotonic decrease in amplitude across a sentence (Payne et al., 2015). Although this description of the N400 has been recognized by ERP researchers for over 25 years, it is often overlooked outside the ERP literature, and more work remains. For example, no ERP experiments have to our knowledge have examined N400 amplitudes across sentences such as (4), where (4) contains few semantic cues and, consequently, little contextual constraint. We might predict that N400 amplitudes would decrease across (4) compared to (3b), but not as much as (3a). In any case, the idea that regularly spiking N400 waveforms represents the normal state of affairs, rather than an unexpected failure in prediction, makes it unlikely that the N400 indexes the cost of a failed prediction, as Van Petten and Luka (2012) also argue.

(4) She kept checking because it seemed to be taking an awfully long time to bake.

Van Petten and Luka (2012) discuss at length the possibility that other segments of the ERP waveform might index a cost associated with incorrect lexical predictions. They speculate that post-N400 components – specifically a late frontal positivity (post-N400 positivity, or PNP), typically peaking between 600 ms and 900 ms post stimulus onset – might be a good candidate. Notably, studies in Spanish (Wicha, Moreno, & Kutas, 2004) and Dutch (Van Berkum et al., 2005), which sought to replicate the effects of DeLong et al. (2005), did find effects of mismatching grammatical gender marking on adjectives prior to highly predictable nouns. But these effects were manifested in late positivities, not N400s. Thornhill and Van Petten (2012) observed inflated PNPs for low-cloze completions of both strongly and weakly constrained contexts, suggesting that the PNP is insensitive to more graded predictions of the sort we describe here, and might instead be an index of purely lexical prediction. If so, examining the PNP in tandem with eye movements might reveal costs for failed predictions, if indeed any are to be found in the eye movement record.

In the previous section we discussed the evidence that highly constraining contexts have facilitative effects above and beyond the cloze scores of the individual words (see also Staub et al., 2015). Thus, neurophysiological results that are interpreted as reflecting prediction costs might actually be reflecting this additional influence of contextual constraint. Given the limited evidence for prediction costs overall, we suggest that it is unlikely that the processor commits to a particular lexical item under most circumstances.

4.6. Implications for previous findings of specific lexical prediction

Contrary to our findings here, there is a body of work that is marshaled as support for lexical predictions (see above). Notable among these is DeLong et al. (2005). In this study, sentences such as (5a) were used, and completed with high-cloze probability noun phrases such as (5b) or low-cloze probability noun phrases such as (5c). Importantly, Delong et al. analyzed the N400 for both the articles (”a” vs. “an”) and the nouns (“kite” vs. “airplane”). They found inflated N400s on the articles associated with the low-cloze nouns as well as the nouns themselves. Importantly, the cloze probability of these articles varied in a continuous fashion, as did the N400 response, and DeLong et al. interpreted this graded effect of article cloze probability as evidence for highly specific prediction of the phonology of the upcoming noun.

(5) a. The day was breezy so the boy went outside to fly
b. a kite.
c. an airplane.

The effects reported by DeLong et al. (2005) are sometimes interpreted as supporting lexical prediction. That is, the language processor is predicting particular nouns, which in turn allows the processor to predict the phonological content of the preceding article. However, the continuous nature of the effect reported by DeLong et al. suggests graded pre-activation of phonological and other information, and thus is most consistent with graded prediction as defined here.

ERP studies have proven extremely useful for investigating prediction effects, but they are not without limitations. DeLong et al. (2005) and other such ERP studies employ word-by-word presentation, with 350–500 ms devoted to each word. This mode of presentation may enable and encourage a greater reliance on predictions than natural reading. Given that mean reading times here (and in the vast body of previous eye-tracking research) are significantly less than those single-word presentation times, it might be expected that participants would use the extra time to try to predict the next word, which they might not do in less luxurious circumstances. Wlotko and Federmeier (2015) explicitly tested the effect of SOA on the amplitude of the N400 as an index of failed prediction, and found that when presentation of words was set to 250 ms, within-category “violations” (e.g., when the context created an expectation for “palms” but the target turned out to be another tree, such as “pines”) failed to elicit a spike in the N400. Between-category violations (e.g., seeing “tulips” instead of “palms”) still elicited an inflated N400, however. Interestingly, with a 500 ms presentation, both within-category and between-category violations triggered equivalent spikes in the N400s. Furthermore, whether participants experienced the 250 ms or 500 ms block first influenced N400 amplitudes. Participants who experienced the 500 ms block first had larger N400 amplitudes in response to within-category violations even in the later 250 ms block. Wlotko and Federmeier interpreted their results as suggestive of flexible processing, in which, depending on various factors related to task, presentation, etc., the language processor can switch into or out of "predictive processing mode" as appropriate given the current processing circumstances. The block order effect on N400 amplitudes reported by Wlotko and Federmeier is an example of this flexibility: once participants had been conditioned to make predictions at longer SOAs, they continued to do so at shorter SOAs. Thus, the way that ERP studies are traditionally designed and conducted may encourage more predicting and/or more detailed predictions (Huettig & Mani, 2016), so their results should be interpreted with caution.

4.7. Implications for models of language comprehension

The majority of published studies investigating prediction focus exclusively on highly predictable content words for which conscious prediction is possible and likely does occur. Huettig and Mani (2016) argue that this fact, along with other "prediction-encouraging" aspects of most prediction experiments (such as word-by-word, experimenter-timed presentation), leads to over-estimates of the importance of prediction in language comprehension. The selection of such stimuli can therefore be interpreted as an implicit endorsement of all-or-nothing lexical prediction. Different theoretical accounts based on these studies, therefore, may also reflect this bias.

For example, Pickering and Garrod (2013; see also Pickering & Garrod, 2007) propose that the comprehender's production system is employed to generate predictions. They note explicitly that:

> prediction is very powerful, because it is often the case that language is highly predictable at one linguistic level at least. An upcoming content word is sometimes predictable. Often, a syntactic category can be predicted when the word itself cannot. On other occasions, the upcoming phoneme is predictable. We propose that comprehenders make whatever linguistic predictions they can (pg. 341).

This proposal thus appears on the surface to be consistent with the idea of graded prediction. Pickering and Garrod (2013) further suggest that prediction error (i.e. the difference between the expected continuation and the actual one) is an important factor. Dell and Chang (2014) similarly argue that prediction is ubiquitous, that it employs the production system, and that prediction error is important for language comprehension and learning. These accounts acknowledge that prediction will occur at multiple levels. However, our data suggest that most predictions would be restricted to the earliest stages of production (syntax and semantics), while phonology and orthography will be predictable only rarely. Because these features are so rarely predictable, one might expect that early processing measures that are more sensitive to them, such as word skipping and first fixation duration, would be less affected by predictability, at least for low- and moderate-constraint contexts, but the data suggest otherwise. Furthermore, over-active prediction in speech production leads to a wide variety of speech errors, so if predictive processes in comprehension do rely of the production system it is surprising that mis-prediction costs have not been observed (see above). Adapting existing computational models of language production to model predictive processes would be useful in testing the extent to which these production-based accounts of prediction are consistent with graded prediction.

Christiansen and Chater (2016) argue that, because of the rapidity of linguistic input and the fragility of memory representations, language must be processed as rapidly and efficiently as possible. As part of this, they also suggest that prediction must be ubiquitous, as it gives the language processor a head-start on resolving ambiguities. Our data suggest that in many cases any predictions generated would provide little help in the current input and little guidance for making future predictions, and that over-active reliance on prediction would be disruptive and misleading instead of facilitative. In Christiansen and Chater's account, prediction error is thought to serve as an important source of learning, so that language acquisition can in a sense be understood as learning to predict (Dell & Chang, 2014, also argue that prediction has an important role in language acquisition). This emphasis on prediction error across theories raises an important concern: are predictions specific enough often enough to be useful, or will prediction error be too great on average to be informative? In other words, it may be that researchers have overestimated the accuracy of predictions for the reasons outlined above, and so have also overestimated the informativeness of prediction error. Christiansen and Chater (2016, p. 22) state, "If language processing involves prediction, in order to make the encoding of material sufficiently rapid, then a critical aspect of language acquisition is *learning* to make such predictions successfully" (see also Altmann & Mirković, 2009). Based on our results, it seems that, given the nature of language input, it is practically impossible to learn to make predictions "successfully," if success is defined as making accurate lexical predictions. Under these circumstances, learning to predict could be defined in two different but not mutually exclusive ways. First, it could mean an increased ability to make graded predictions of the sort described in this paper. Learning to predict could also mean learning *when* to make lexical predictions. If learning to predict is more accurately characterized as "learning when to predict," then error-driven learning mechanisms are not leading to better predictions of upcoming material, but rather better recognition of the very infrequent contexts in which lexical predictions might be more likely to be correct. Such a learning process could plausibly lead to the sort of flexible predictive processor proposed by Wlotko and Federmeier (2015). Under this view the extreme rarity of highly constraining contexts might actually result in a small processing penalty when one is encountered, such as the small yet significant rise in regression out probabilities that we observed here as log certainty increased.

Finally, it seems clear that prediction error will vary considerably from one level of linguistic representation to the next. This concern is especially potent in the context of language learning, given that the ability to generate useful predictions may require language knowledge as a prerequisite (cf. Mani & Huettig, 2012, 2014). On the other hand, Kuperberg and Jaeger (2016) suggest that even a minimal amount of (accurate) pre-activation would reduce Bayesian surprise relative to a situation where no pre-activation occurred, meaning that even a little bit of prediction is better than nothing. Further research and further theoretical refinement are needed to address these concerns.

4.8. Does lexical prediction ever occur, and if so when?

As noted in the introduction, a number of studies (including some of our own) have demonstrated that, under certain circumscribed circumstances, with tightly controlled materials, the language processor can and does generate fine-grained details about upcoming material, including prediction of specific lexical items. Our results indicate that, for most words and sentences, normal silent reading of connected text is not one of those circumscribed circumstances. However, other linguistic contexts, such as speaker-listener interactions, may provide more cues to the listener and thus make prediction more likely (Bögels, Magyari, & Levinson, 2015). Yet even in the more prediction-friendly environment of speaker-listener interactions, it may be that predictions are made about the gist of the speaker's meaning or the timing of the end of an utterance, rather than about specific lexical items (Bögels et al., 2015; Magyari, Bastiaansen, de Ruiter, & Levinson, 2014). Further research is needed to explore the limits and nature of predictive processes across different linguistic tasks.

5. Conclusion

This study collected a large corpus of cloze probabilities in naturally occurring extended texts. Analyses of these cloze probabilities revealed that highly predictable words of the sort used in most prediction experiments occur rarely, comprising only approximately 5% of content words and 20% of function words. Part of speech, inflectional properties and semantic features were somewhat more consistently predictable from context. Analyses of eye movement data from a different set of participants who read the same texts revealed facilitation for reading times in proportion to increased cloze probability across the full range of those probabilities, equally for expected targets and for non-modal targets for which there was a more expected competitor. Further, no evidence of a penalty for inaccurate prediction was observed in any eye movement measures. Instead, the presence of a more expected competitor had a facilitative effect in several measures. Taken together, the results strongly suggest that (1) all-or-nothing *lexical prediction* does not occur often; (2) *graded prediction*, a more continuous pre-activation process, is a better characterization of the facilitative effects of increased cloze probability on reading time; and (3) most predictions will be partial and even sparse, but semantic and morphosyntactic features of upcoming words are more likely to be predicted than the actual words themselves, and the language processor appears to make use of this information.

Appendix A. Statistical results

A.1. When does word predictability influence eye movements?

In the analyses reported in this section, we investigated whether cloze scores have a similar influence on eye movements in reading across the full range of predictability, and whether any evidence could be found of inhibition or disruption arising from incorrect predictions.

A.1.1. Does predictability influence the processing of unexpected words?

In these analyses, we compared the slopes of the *Predictability* (cloze scores, log transformed) function for modal and non-modal target words (i.e. target words that were or were not the most expected continuation of the sentence; these groups were represented by the variable *IsModal*). A significant interaction of IsModal and Predictability indicates that the slope for modal targets were different than for non-modal targets (the simple effect of *Predictability* indicates the slope for non-modal targets). Dependent variables included first fixation duration, gaze duration, total reading time, word skipping probability, refixation probability, regression in probability, regression out probability, and go-past time.

Content words. Tables 7–14 show the analyses for content words only.

Table 7
Skipping probability – content words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−0.94	.058	−16.2	<.0001
IsModal = modal	0.54	0.022	24.88	<.0001
Predictability	0.13	0.0072	17.84	<.0001
IsModal × predictability	0.096	0.023	4.08	<.0001

Table 8
First fixation duration – content words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.3	0.011	474.22	<.0001
IsModal = modal	−0.044	0.0063	−6.96	<.0001
Predictability	−0.011	0.0023	−4.72	<.0001
IsModal × predictability	−0.014	0.0089	−1.56	=.11

Table 9
Refixation probability – content words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.32	0.055	−23.99	<.0001
IsModal = modal	−0.46	0.02	−22.75	<.0001
Predictability	−0.12	0.0069	−17.71	<.0001
IsModal × predictability	0.032	0.033	0.99	=.32

Table 10
Gaze duration – content words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.45	0.015	354.86	<.0001
IsModal = modal	−0.096	0.01	−9.29	<.0001
Predictability	−0.027	0.0037	−7.22	<.0001
IsModal × predictability	−0.0019	0.015	−0.13	=.9

Table 11
Regression in probability – content words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.59	0.051	−31.08	<.0001
IsModal = modal	−0.15	0.024	−6.1	<.0001
Predictability	−0.064	0.0081	−7.89	<.0001
IsModal × predictability	0.13	0.036	3.64	=.00027

Table 12
Total time – content words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.6	0.019	295.66	<.0001
IsModal = modal	−0.14	0.013	−10.58	<.0001
Predictability	−0.043	0.0047	−9.27	<.0001
IsModal × predictability	−0.0017	0.018	−0.092	=.93

Table 13
Regression out probability – content words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.68	0.048	−34.71	<.0001
IsModal = modal	0.18	0.02	8.94	<.0001
Predictability	−0.0077	0.0082	−0.95	=.34
IsModal × predictability	0.19	0.031	6.02	<.0001

Table 14
Go-past time – content words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.65	0.019	289.89	<.0001
IsModal = modal	−0.07	0.016	−4.45	<.0001
Predictability	−0.031	0.0057	−5.46	<.0001
IsModal × predictability	0.029	0.023	1.26	=.21

Function words. Tables 15–22 show the analyses for function words only.

Table 15
Skipping probability – function words.

	b	*SE*	*z* value	*p*-value
(Intercept)	0.39	0.051	7.69	<.0001
IsModal = modal	0.29	0.018	15.63	<.0001
Predictability	0.13	0.0093	13.66	<.0001
IsModal × predictability	0.027	0.025	1.09	=.28

Table 16
First fixation duration – function words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.24	0.013	413.27	<.0001
IsModal = modal	−0.021	0.0061	−3.4	=.00071
Predictability	−0.0081	0.0037	−2.26	=0.024
IsModal × predictability	−0.014	0.001	−0.9	=.37

Table 17
Refixation probability – function words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−2.46	0.061	−40.57	<.0001
IsModal = modal	−0.27	0.037	−7.46	<.0001
Predictability	−0.18	0.019	−9.55	<.0001
IsModal × predictability	0.061	0.064	0.95	=.34

Table 18
Gaze duration – function words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.29	0.014	377.72	<.0001
IsModal = modal	−0.036	0.0072	−4.85	<.0001
Predictability	−0.017	0.0043	−3.92	<.0001
IsModal × predictability	−0.011	0.012	−0.9	=.37

Table 19
Regression in probability – function words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.27	0.066	−19.4	<.0001
IsModal = modal	−0.15	0.024	−6.23	<.0001
Predictability	−0.051	0.014	−3.71	<.0001
IsModal × predictability	−0.08	0.042	−1.92	=.055

Table 20
Total time – function words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.41	0.015	351.78	<.0001
IsModal = modal	−0.072	0.0051	−14.17	<.0001
Predictability	−0.035	0.0029	−12.17	<.0001
IsModal × predictability	−0.014	0.0075	−1.88	=.061

Table 21
Regression out probability – function words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.74	0.041	−42.34	<.0001
IsModal = modal	−0.0039	0.027	−0.14	=.89
Predictability	−0.0044	0.016	−0.28	=.78
IsModal × predictability	0.09	0.05	1.89	=.059

Table 22
Go-past time – function words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.47	0.017	321.3	<.0001
IsModal = modal	−0.038	0.011	−3.32	=.00094
Predictability	−0.021	0.0065	−3.2	=.0014
IsModal × predictability	−0.0032	0.019	−0.17	=.87

A.1.2. Is there a penalty for mis-prediction?

In these analyses, only data from non-modal targets (i.e. targets that were not the most common response on the cloze task) were included. Variables included *Predictability* (cloze scores, log transformed) and *Certainty*, which was the cloze score of the modal response for each target, also log transformed. Dependent variables included first fixation duration, gaze duration, total reading time, word skipping probability, refixation probability, and the probability of regression back to the word. To look for possible later inhibitory effects of *Certainty*, regressions out and go-past time were also included.

Content words. Tables 23–30 show the analyses for content words only.

Table 23
Skipping probability – content words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−0.25	0.063	−4	<.0001
Certainty	0.15	0.013	11.56	<.0001
Predictability	0.12	0.0054	21.76	<.0001

Table 24
First fixation duration – content words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.26	0.015	351.6	<.0001
Certainty	0.0031	0.0047	0.65	=0.52
Predictability	−0.011	0.002	−5.31	<.0001

Table 25
Refixation probability – content words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.98	0.065	−30.55	<.0001
Certainty	−0.12	0.016	−7.71	<.0001
Predictability	−0.12	0.0069	−16.8	<.0001

Table 26
Gaze duration – content words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.32	0.021	248.66	<.0001
Certainty	−0.013	0.0083	−1.62	=.11
Predictability	−0.026	0.0037	−7.19	<.0001

Table 27
Regression in probability – content words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.82	0.063	−29.04	<.0001
Certainty	0.019	0.016	1.13	=.26
Predictability	−0.062	0.0074	−8.37	<.0001

Table 28
Total time – content words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.4	0.026	210.03	<.0001
Certainty	−0.021	0.011	−1.96	=0.051
Predictability	−0.042	0.0046	−9.16	<.0001

Table 29
Regression out probability – content words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.66	0.056	−29.47	<.0001
Certainty	0.05	0.019	2.68	=.0073
Predictability	−0.0081	0.0085	−0.95	=.34

Table 30
Go-past time – content words.

	b	*SE*	*z* value	*p*-value
(Intercept)	5.51	0.029	187.61	<.0001
Certainty	−0.014	0.012	−1.14	=.25
Predictability	−0.031	0.0053	−5.77	<.0001

Function words. Tables 31–38 show the results for function words only.

Table 31
Skipping probability – function words.

	b	*SE*	*z* value	*p*-value
(Intercept)	0.87	0.059	14.74	<.0001
Certainty	0.097	0.018	5.42	<.0001
Predictability	0.12	0.008	15.54	<.0001

Table 32
First fixation duration – function words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.2	0.018	296.32	<.0001
Certainty	−0.016	0.0077	−2.04	=.043
Predictability	−0.0077	0.0034	−2.29	=0.022

Table 33
Refixation probability – function words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−3.16	0.11	−29.13	<.0001
Certainty	−0.13	0.05	−2.7	=.007
Predictability	−0.18	0.019	−9.41	<.0001

Table 34
Gaze duration – function words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.21	0.021	253.44	<.0001
Certainty	−0.02	0.0095	−2.056	=.04
Predictability	−0.017	0.0041	−4.17	<.0001

Table 35
Regression in probability – function words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.44	0.082	−17.63	<.0001
Certainty	−0.032	0.029	−1.13	=.26
Predictability	−0.042	0.013	−3.32	=.00089

Table 36
Total time – function words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.25	0.026	203.92	<.0001
Certainty	−0.027	0.012	−2.26	=.024
Predictability	−0.031	0.0052	−5.98	<.0001

Table 37
Regression out probability – function words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.66	0.074	−22.48	<.0001
Certainty	0.091	0.033	2.75	=.006
Predictability	−0.006	0.016	−0.39	=.7

Table 38
Go-past time – function words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.41	0.029	184.05	<.0001
Certainty	−0.0027	0.014	−0.18	=.85
Predictability	−0.021	0.0065	−3.19	=.0015

Follow-up analysis. Table 39 show the results for a follow-up analysis on regressions out for content words. Compare these results to those reported in Table 13.

Table 39
Regression out probability follow-up analysis – content words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.65	0.057	−28.79	<.0001
IsModal = modal	−0.051	0.052	−0.98	=.33
Certainty	0.087	0.036	2.45	=.015
IsModal × certainty	−0.0053	0.056	−0.094	=.92

A.2. Beyond cloze scores

These analyses investigated whether partial information available to the language processor had any influence on reading above and beyond the influence of the predictability of the full word form. Content and function words were analyzed separately, and nouns and verbs were separated out from the content word data for further analysis. Dependent variables included first fixation duration, gaze duration, total reading time, word skipping probability, refixation probability, and the probability of regression back to the word.

A.2.1. Content words

Results of the analyses on all content words are reported in Tables 40–45. These analyses included the variables *Predictability* (cloze scores, log transformed), *Latent Semantic Analysis (LSA) Score*, and *Part-Of-Speech (POS) Match Probability* (log transformed).

Table 40
Skipping – content words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−0.89	0.059	−15.08	<.0001
Predictability	−0.021	0.0052	−4.13	<.0001
LSA score	1.72	0.036	47.94	<.0001
POS match probability	0.0031	0.0061	0.51	=.61

Table 41
First fixation duration – content words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.27	0.012	431.52	<.0001
Predictability	−0.0045	0.0022	−2.065	=.039
LSA score	−0.07	0.015	−4.66	<.0001
POS match probability	−0.0057	0.0024	−2.41	=.016

Table 42
Refixation – content words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.49	0.06	−24.88	<.0001
Predictability	−0.0036	0.0067	−0.54	=.59
LSA score	−1.3	0.05	−26.3	<.0001
POS match probability	0.048	0.0072	−6.59	<.0001

Table 43
Gaze duration – content words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.41	0.018	301.41	<.0001
Predictability	−0.0042	0.0036	−1.18	=.24
LSA score	−0.22	0.026	−8.42	<.0001
POS match probability	−0.012	0.0039	−3.05	=.0024

Table 44
Regressions in – content words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−2.01	0.056	−36.29	<.0001
Predictability	−0.1	0.007	−14.34	<.0001
LSA score	0.66	0.05	13.17	<.0001
POS match probability	−0.051	0.0078	−6.53	<.0001

Table 45
Total time – content words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.52	0.021	258.77	<.0001
Predictability	−0.013	0.0043	−3.05	=.0023
LSA score	−0.27	0.032	−8.37	<.0001
POS match probability	−0.013	0.0048	−2.78	=.0056

Nouns. Results of the analyses on the nouns alone are reported in Tables 46–51. These analyses included the same variables as the content words analyses, with the additional variable of *Inflection Match Probability* (log transformed).

Table 46
Skipping – nouns.

	b	*SE*	*z* value	*p*-value
(Intercept)	−0.78	0.066	−11.7	<.0001
Predictability	0.058	0.0097	5.98	<.0001
LSA score	0.46	0.068	6.77	<.0001
POS match probability	−0.21	0.02	−10.42	<.0001
Inflection match probability	0.18	0.017	10.69	<.0001

Table 47
First fixation – nouns.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.27	0.016	337.89	<.0001
Predictability	−0.008	0.0039	−2.07	=.039
LSA score	−0.025	0.028	−0.91	=.36
POS match probability	−0.0018	0.0076	−0.24	=.81
Inflection match probability	−0.000008	0.0062	−0.001	0.999

Table 48
Refixations – nouns.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.63	0.067	−24.13	<.0001
Predictability	−0.069	0.011	−6.08	<.0001
LSA score	−0.24	0.086	−2.85	=.0043
POS match probability	0.99	0.021	4.83	<.0001
Inflection match probability	−0.14	0.017	−8.35	<.0001

Table 49
Gaze duration – nouns.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.37	0.025	212.85	<.0001
Predictability	−0.019	0.0066	−2.89	=.004
LSA score	−0.03	0.048	−0.63	=.53
POS match probability	0.012	0.013	0.96	=.34
Inflection match probability	−0.021	0.011	−2.017	=0.044

Table 50
Regressions in – nouns.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.97	0.062	−31.89	<.0001
Predictability	−0.072	0.013	−5.68	<.0001
LSA score	0.073	0.096	0.76	=.45
POS match probability	0.2	0.024	−8.19	<.0001
Inflection match probability	−0.086	0.021	4.17	<.0001

Table 51
Total time – nouns.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.46	0.03	467.7	<.0001
Predictability	−0.035	0.0079	−4.37	<.0001
LSA score	−0.0072	0.057	−0.13	=.9
POS match probability	0.013	0.015	0.82	=.41
Inflection match probability	−0.03	0.013	−2.32	=.021

Verbs. Results of the analyses on the verbs alone are reported in Tables 52–57. These analyses included the included the same variables as the content words analyses, with the additional variable of *Inflection Match Probability* (log transformed).

Table 52
Skipping – verbs.

	b	*SE*	*z* value	*p*-value
(Intercept)	−0.57	0.063	−9.04	<.0001
Predictability	−0.011	0.0093	−1.22	=0.22
LSA score	2.23	0.07	31.75	<.0001
POS match probability	0.1	0.016	6.64	<.0001
Inflection match probability	−0.031	0.013	−2.35	=0.019

Table 53
First fixation – verbs.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.27	0.016	334.94	<.0001
Predictability	−0.0058	0.0038	−1.5	=.13
LSA score	−0.089	0.029	−3.12	=.002
POS match probability	−0.0022	−0.006	−0.38	=.71
Inflection match probability	−0.00041	0.005	−0.082	0.93

Table 54
Refixations – verbs.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.58	0.07	−22.48	<.0001
Predictability	−0.05	0.013	3.82	=0.00014
LSA score	−1.87	0.1	−18.65	<.0001
POS match probability	−0.076	0.02	−3.83	=.00013
Inflection match probability	−0.073	0.018	−4.15	<.0001

Table 55
Gaze duration – verbs.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.39	0.022	245.25	<.0001
Predictability	0.00088	0.0055	0.16	=.87
LSA score	−0.28	0.042	−6.65	<.0001
POS match probability	−0.011	0.0087	−1.31	=.19
Inflection match probability	−0.0098	0.0075	−1.32	=.19

Table 56
Regressions in – verbs.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.73	0.07	−24.81	<.0001
Predictability	−0.036	0.013	−2.81	=0.005
LSA score	0.84	0.096	8.72	<.0001
POS match probability	0.11	0.02	5.59	<.0001
Inflection match probability	−0.15	0.017	−8.69	<.0001

Table 57
Total time – verbs.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.51	0.026	213.81	<.0001
Predictability	−0.00046	0.0067	−0.069	=.95
LSA score	−0.38	0.051	−7.45	<.0001
POS match probability	−0.01	0.011	−0.96	=.34
Inflection match probability	−0.02	0.0092	−2.151	=.032

A.2.2. Function words

Results of the analyses on function words are reported in Tables 58–63. These analyses included only *Predictability* (cloze score, log transformed) and *Part-Of-Speech (POS) Match Probability* (log transformed).

Table 58
Skipping – function words.

	b	*SE*	*z* value	*p*-value
(Intercept)	0.77	0.051	15.18	<.0001
Predictability	0.11	0.0049	23.4	<.0001
POS match probability	0.021	0.0055	3.72	=.0002

Table 59
First fixation –function words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.2	0.012	435.28	<.0001
Predictability	−0.0069	0.0022	−3.15	=.0017
POS match probability	−0.0059	0.0023	−2.55	=.011

Table 60
Refixations – function words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−2.86	0.066	−43.33	<.0001
Predictability	−0.13	0.012	−11.22	=.0017
POS match probability	−0.0056	0.014	−0.41	=.68

Table 61
Gaze duration – function words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.23	0.013	396.63	<.0001
Predictability	−0.014	0.0026	−5.22	<.0001
POS match probability	−0.0057	0.0028	−2.07	=.039

Table 62
Regressions in – function words.

	b	*SE*	*z* value	*p*-value
(Intercept)	−1.45	0.067	−21.49	<.0001
Predictability	−0.056	0.0081	−6.95	<.0001
POS match probability	0.00086	0.0092	0.094	=.93

Table 63
Total time – function words.

	b	*SE*	*t* value	*p*-value
(Intercept)	5.29	0.015	350.96	<.0001
Predictability	−0.026	0.0033	−7.83	<.0001
POS match probability	−0.0074	0.0035	−2.13	=.034

References

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247–264.

Altmann, G. T., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57*(4), 502–518.

Altmann, G., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33*(4), 583–609.

Ashby, J., Rayner, K., & Clifton, C. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *Quarterly Journal of Experimental Psychology, A: Human Experimental Psychology, 58*(6), 1065–1086. http://dx.doi.org/10.1080/02724980443000476.

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology, 17*(3), 364–390.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2016). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-11 Retrieved from<http://CRAN.R-project.org/package=lme4>.

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports, 5*.

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition, 136*, 135–149.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 1–72.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(03), 181–204.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*(2), 159–190.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1634), 20120394.

DeLong, K. A., Troyer, M., & Kutas, M. (2014). Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. *Language and Linguistics Compass, 8*(12), 631–645.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117–1121.

Dikker, S., Rabagliati, H., Farmer, T. A., & Pylkkänen, L. (2010). Early occipital sensitivity to syntactic category is based on form typicality. *Psychological Science, 21*(5), 629–634.

Dikker, S., Rabagliati, H., & Pylkkänen, L. (2009). Sensitivity to syntax in visual cortex. *Cognition, 110*(3), 293–321.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*(6), 641–655.

Farmer, T. A., Brown, M., & Tanenhaus, M. K. (2013). Prediction, explanation, and the role of generative models in language processing. *Behavioral and Brain Sciences, 36*(03), 211–212.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41*(4), 469–495.

Federmeier, K. D., McLennan, D. B., Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology, 39*(2), 133–146.

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research, 1146*, 75–84.

Finn, P. J. (1977). Word frequency, information theory, and cloze performance: A transfer feature theory of processing in reading. *Reading Research Quarterly*, 508–537.

Fischler, I., & Bloom, P. A. (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. *Journal of Verbal Learning and Verbal Behavior, 18*(1), 1–20.

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

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In *Corpus annotation: Linguistic information from computer text corpora* (pp. 102–121).

Gough, P. B. (1983). Context, form, and interaction. *Eye Movements in Reading, 331*, 358.

Gough, P. B., Alford, J., & Holley-Wilcox, P. (1981). Words and contexts. In *Perception of print: Reading research in experimental psychology* (pp. 85–102).

Hausser, J., Strimmer, K., & Strimmer, M. K. (2014). Estimation of entropy, mutual information and related quantities. R package, version 1.2.1 Retrieved from<http://strimmerlab.org/software/entropy/>.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*.

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience, 31*(1), 19–31.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution* : . Oxford University Press.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*(1), 133–156.

Kennedy, A., Hill, R., & Pynte, J. (2003). The Dundee Corpus. *Paper presented at the proceedings of the 12th European conference on eye movement*.

Kennedy, A., Pynte, J., Murray, W. S., & Paul, S.-A. (2013). Frequency and predictability effects in the Dundee Corpus: An eye movement analysis. *The Quarterly Journal of Experimental Psychology, 66*(3), 601–618.

Kleinman, D., Runnqvist, E., & Ferreira, V. S. (2015). Single-word predictions of upcoming language during comprehension: Evidence from the cumulative semantic interference task. *Cognitive Psychology, 79*, 68–101.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16*(1–2), 262–284.

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*(1), 12–35.

Kuperberg, G. (2013). The proactive comprehender: What event-related potentials tell us about the dynamics of reading comprehension. In *Unraveling the behavioral, neurobiological, and genetic components of reading comprehension* (pp. 176–192).

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32–59.

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In *Predictions in the brain: Using our past to generate a future* (pp. 190–207).

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature* (307), 161–163.

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2014). *LmerTest: Tests for random and fixed effects for linear mixed effect models. R package, version 2.0-3* : . .

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211.

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language, 61*(3), 326–338.

Lavigne, F., Vitu, F., & d'Ydewalle, G. (2000). The influence of semantic context on initial eye landing sites in words. *Acta Psychologica, 104*(2), 191–214.

Lefcheck, J. S. (2015). piecewiseSEM: Piecewise structural equation modelling in r for ecology, evolution, and systematics. *Methods in Ecology and Evolution*.

Luke, S. G., & Christianson, K. (2012). Semantic predictability eliminates the transposed-letter effect. *Memory & Cognition, 40*(4), 628–641.

Luke, S. G., & Christianson, K. (2015). Predicting inflectional morphology from context. *Language, Cognition and Neuroscience, 30*(6), 735–748.

Magyari, L., Bastiaansen, M. C., de Ruiter, J. P., & Levinson, S. C. (2014). Early anticipation lies behind the speed of response in conversation. *Journal of Cognitive Neuroscience*.

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance, 38*(4), 843.

Mani, N., & Huettig, F. (2014). Word reading skill predicts anticipation of upcoming spoken language input: A study of children developing proficiency in reading. *Journal of Experimental Child Psychology, 126*, 264–279.

McDonald, S. A., & Shillcock, R. C. (2003a). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science, 14*(6), 648–652.

McDonald, S. A., & Shillcock, R. C. (2003b). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research, 43*(16), 1735–1751.

Nicholson, T., & Hill, D. (1985). Good readers don't guess-taking another look at the issue of whether children read words better in context or in isolation. *Reading Psychology: An International Quarterly, 6*(3–4), 181–198.

Otten, M., Nieuwland, M. S., & Van Berkum, J. J. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience, 8*(1), 89.

Payne, B. R., Lee, C. L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology, 52*(11), 1456–1469.

Perfetti, C. A., Goldman, S. R., & Hogaboam, T. W. (1979). Reading skill and the identification of words in discourse context. *Memory & Cognition, 7*(4), 273–282.

Piai, V., Roelofs, A., & Maris, E. (2014). Oscillatory brain responses in spoken word production reflect lexical frequency and sentential constraint. *Neuropsychologia, 53*, 146–156.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11*(3), 105–110.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(04), 329–347.

Posner, M. I., & Snyder, C. R. R. (1975). Facilitation and inhibition in the processing of signals. In *Attention and performance V* (pp. 669–682).

R Core Team (2015). *R: A language and environment for statistical computing (version 3.2.2)*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from<http://www.R-project.org/>.

Rayner, K., & Fischer, M. (1996). Mindless reading revisited: Eye movements during reading and scanning are different. *Perception & Psychophysics, 58*(5), 734–747. http://dx.doi.org/10.3758/BF03213106.

Rayner, K., Li, X., Juhasz, B. J., & Yan, G. (2005). The effect of word predictability on the eye movements of Chinese readers. *Psychonomic Bulletin & Review, 12*(6), 1089–1093.

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

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance, 37*(2), 514.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review, 3*(4), 504–509.

Roland, D., Yun, H., Koenig, J.-P., & Mauner, G. (2012). Semantic similarity, predictability, and models of sentence processing. *Cognition, 122*(3), 267–279.

Rommers, J., Meyer, A. S., Praamstra, P., & Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia, 51*(3), 437–447.

Rubenstein, H., & Aborn, M. (1958). Learning, prediction, and readability. *Journal of Applied Psychology, 42*(1), 28.

Schatz, E. K., & Baldwin, R. S. (1986). Context clues are unreliable predictors of word meanings. *Reading Research Quarterly*, 439–453.

Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(2), 344.

Schwanenflugel, P. J., & Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language, 24*(2), 232–252.

Smith, N. J., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. *Paper presented at the proceedings of the 33rd annual conference of the Cognitive Science Society*.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302–319. http://dx.doi.org/10.1016/j.cognition.2013.02.013.

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass, 9*(8), 311–327.

Staub, A., Abbott, M., & Bogartz, R. S. (2012). Linguistically guided anticipatory eye movements in scene viewing. *Visual Cognition, 20*(8), 922–946.

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language, 82*, 1–17.

Szewczyk, J. M., & Schriefers, H. (2013). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and Language, 68*(4), 297–314.

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

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology, 83*(3), 382–392.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(3), 443.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition, 18*(4), 380–393.

Van Petten, C., & Kutas, M. (1991). Electrophysiological evidence for the flexibility of lexical processing. In G. Simpson (Ed.), *Understanding Word and Sentence* (pp. 129–174). Amsterdam: North-Holland Press.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*(2), 176–190.

Wicha, N. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003). Potato not Pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters, 346*(3), 165–168.

Wicha, N. Y., Moreno, E., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience, 16*(7), 1272–1288.

Wlotko, E. W., & Federmeier, K. D. (2013). Two sides of meaning: The scalp-recorded N400 reflects distinct contributions from the cerebral hemispheres. *Frontiers in Psychology, 4*.

Wlotko, E. W., & Federmeier, K. D. (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex, 68*, 20–32.

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language, 72*, 37–48.

Zola, D. (1984). Redundancy and word perception during reading. *Perception & Psychophysics, 36*(3), 277–284.