Published in final edited form as:
Cognition. 2013 September ; 128(3): 302–319. doi:10.1016/j.cognition.2013.02.013.

The effect of word predictability on reading time is logarithmic

Nathaniel J. Smith and **Roger Levy**
University of California, San Diego

Abstract

It is well known that real-time human language processing is highly incremental and context-driven, and that the strength of a comprehender's expectation for each word encountered is a key determinant of the difficulty of integrating that word into the preceding context. In reading, this differential difficulty is largely manifested in the amount of time taken to read each word. While numerous studies over the past thirty years have shown expectation-based effects on reading times driven by lexical, syntactic, semantic, pragmatic, and other information sources, there has been little progress in establishing the quantitative relationship between expectation (or prediction) and reading times. Here, by combining a state-of-the-art computational language model, two large behavioral data-sets, and non-parametric statistical techniques, we establish for the first time the quantitative form of this relationship, finding that it is logarithmic over six orders of magnitude in estimated predictability. This result is problematic for a number of established models of eye movement control in reading, but lends partial support to an optimal perceptual discrimination account of word recognition. We also present a novel model in which language processing is highly incremental well below the level of the individual word, and show that it predicts both the shape and time-course of this effect. At a more general level, this result provides challenges for both anticipatory processing and semantic integration accounts of lexical predictability effects. And finally, this result provides evidence that comprehenders are highly sensitive to relative differences in predictability – even for differences between highly unpredictable words – and thus helps bring theoretical unity to our understanding of the role of prediction at multiple levels of linguistic structure in real-time language comprehension.

Keywords

Psycholinguistics; Expectation; Probabilistic models of cognition; Information theory; Reading

Making probabilistic predictions about the future is a necessary component of essentially every task that the brain performs, to the point that it has been proposed as a fundamental principle underlying its operation (Bar, 2009). One example of this is in language comprehension: As you read this text, you are unconsciously anticipating upcoming words based on the constantly-evolving context. For example, the sentence

(1) My brother came inside to…

Correspondence concerning this article should be addressed to Nathaniel J. Smith, School of Informatics, University of Edinburgh, Informatics Forum 3.29, 10 Crichton Street, Edinburgh, EH8 9AB. nathaniel.smith@ed.ac.uk.

Author Note: Nathaniel J. Smith, Department of Cognitive Science, University of California, San Diego; Roger Levy, Department of Linguistics, University of California, San Diego. Nathaniel J. Smith is now at School of Informatics, University of Edinburgh.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

may well continue any number of ways, but native English speakers are in general agreement—and you will likely immediately recognize—that the sentence

(2) The children went outside to…

is almost certain to continue with the word *play*. Although *play* is perfectly reasonable as a continuation of (1), in (2) it is read more quickly on average (Ehrlich & Rayner, 1981; Rayner & Well, 1996; McDonald & Shillcock, 2003a; Kliegl, Nuthmann, & Engbert, 2006). A wide range of studies have shown that such effects of predictability or expectation for specific words affect not only reading times but also neural responses (DeLong, Urbach, & Kutas, 2005; Kutas & Hillyard, 1984; Kutas & Federmeier, 2011; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Bates, Moreno, & Kutas, 2003) and interpretation of temporarily ambiguous input (Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003; Dahan & Tanenhaus, 2004; Knoeferle, Crocker, Scheepers, & Pickering, 2005; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995).

A second, related strand of research has shown that incremental processing difficulty is also affected by expectations for more abstract levels of linguistic content, including the predictability of different syntactic (Ferreira & Clifton, 1986; McRae, Spivey-Knowlton, & Tanenhaus, 1998; Demberg & Keller, 2008), semantic (Federmeier & Kutas, 1999), and pragmatic (Ni, Crain, & Shankweiler, 1996) structures. However, the relationship between the effects of expectations for specific-words and expectations for more abstract structures remains poorly understood. The most widespread method for assessing expectations for specific words is the cloze task (Taylor, 1953), in which native speakers are asked to write continuations of an incomplete sentence; in the examples above, *play* is the first word in over 90% of continuations of (2) but almost never appears as the first word of continuations of (1). However, the cloze task makes it quite difficult to precisely measure predictabilities <5–10%, and it is commonly assumed that differences in lexical expectation between items in this range do not produce behavioral effects. This contrasts with studies involving more abstract levels of linguistic structure, where expectation-based effects are observed even though the specific word instantiating the structure may rarely or never be produced in a cloze task. To take one recent example, Levy, Fedorenko, Breen, and Gibson (2012) showed that the word *who* and the immediately following region of the sentence were read more quickly in sentences like (3) than in sentences like (4):

(3) After the show, a performer who had really impressed the audience bowed.

(4) After the show, a performer bowed who had really impressed the audience.

The word *who* never occurs in practice as a cloze continuation in either contexts (unpublished data), and so this result would conventionally be interpreted as arising from syntactic expectations (Levy et al., 2012; Hale, 2001; Levy, 2008; Lau, Stroud, Plesch, & Phillips, 2006; Staub & Clifton, 2006; Ilkin & Sturt, 2011): In (4) *who* is introduced by a grammatical construction (relative-clause extraposition) that corpus data indicate is both lower frequency and less likely given the grammatical context than the construction in (3) (ordinary postmodification by a relative clause), even though both are infrequent and unlikely in absolute terms.

Probability theory, however, tells us that such differences in syntactic expectation should also produce differences in lexical expectation, even if these latter differences are too small to measure via the cloze task. We can quantify the predictability of a word w in context C as its conditional probability of occurrence in that context, $P(w|C)$. Similarly, we write the predictability of a syntactic construction S as $P(S|C)$. In these particular contexts, w = *who* can only occur if S = relative clause. The laws of conditional probability then let us decompose the lexical predictability as the product of two terms (Demberg & Keller, 2008; Fossum & Levy, 2012; Roark, Bachrach, Cardenas, & Pallier, 2009):

$$P\left(who|C\right)=P\left(\text{rel. clause}|C\right)\times P\left(who|\text{rel. clause},C\right)$$

The first term is the syntactic predictability, and the second measures the likelihood that this relative clause will begin with the word *who* (as opposed to, say, *that*). The latter is presumably roughly constant between these two contexts, which means that while the precise lexical predictabilities in (3) and (4) are too small to measure directly, the ratio between them should be similar to the ratio between their syntactic predictabilities.

Motivated by such considerations, Hale has suggested that syntactic and other types of abstract expectations may affect processing difficulty purely by modulating lexical predictability, which under the *surprisal theory* of incremental language processing is measured as log-probability (Hale, 2001; Levy, 2008). Within surprisal theory, lexical predictability forms a “causal bottleneck” through which the many different kinds of more abstract expectation discussed above must act. But as the above example shows, an essential requirement for this theory is that small absolute differences in expectation for low-predictability words must be capable of producing relatively large effects on processing difficulty, and it is not known whether this is the case. In fact, almost nothing is known about the quantitative form of the relationship between word predictability and the measurable correlates of processing difficulty such as reading time. This is in striking contrast to the study of isolated word recognition, where it has been known since the 1950s that recognition time varies almost exactly as a logarithmic function of frequency[1] (Howes & Solomon, 1951), and the need to explain this pattern has motivated a wide range of theories (Adelman, Brown, & Quesada, 2006; Murray & Forster, 2004; Morrison, Hirsh, & Duggan, 2003; Norris, 2006; Baayen, 2010a). But the few extant published studies (Rayner & Well, 1996; Kliegl et al., 2006) that have investigated the quantitive relationship between word predictability and processing time have yielded only limited insights, particularly regarding the shape of this relationship for highly unpredictable words, partly because of the cloze method’s limitations at measuring small differences in absolute predictability.

Here we overcome these limitations by combining two large behavioral datasets of word-by-word reading times with probabilistic language models from computational linguistics (Kneser & Ney, 1995; Chen & Goodman, 1998; Manning & Schütze, 1999) and nonparametric statistical analysis. These methods allow us to establish for the first time the functional form of the predictability/reading-time relationship, and we do so over six orders of magnitude in probability, from near-obligatory to one-in-a-million events. (Due to the Zipfian distribution of language we encounter instances of the latter class of events relatively often, even though each individual such event is extremely rare.) We first describe a number of potential functional forms which have been hypothesized in the literature (see also Fig. 1)—as well as giving a new theoretical motivation for a previously-hypothesized functional form—and then proceed to our empirical analysis.

Theories relating word predictability and reading time

Simple guessing (prediction: linear)

The simplest possible curve that might relate predictability and processing time is a straight line; indeed, several modern models of eye movement control in reading apply a logarithmic

[1] Note that in psycholinguistics, the term *frequency* refers specifically to a word’s unconditional probability of occurrence without regard to context, $P(w)$, making it quite distinct from context-dependent predictability.

transformation to frequency, but enter predictability linearly (Reichle, Pollatsek, Fisher, & Rayner, 1998; Engbert, Nuthmann, Richter, & Kliegl, 2005). While we are not aware of any published justification for this practice, it does arise naturally from a simple and intuitive theory: Suppose that before reading each word, comprehenders make a guess at its identity by sampling from the distribution $P(w|C)$ (a probability matching strategy). If their guess is correct, then they continue undisturbed and can read the word in some baseline amount of time t_{baseline}. Otherwise, they must spend some fixed additional amount of time recovering from their error — call this time $t_{\text{incorrect}}$. If comprehenders' estimates of this probability are accurate, then they will guess correctly on $P(w|C)$ proportion of trials and incorrectly on $(1 - P(w|C))$ proportion. Thus the *average* reading time will be $(t_{\text{baseline}}) + (1 - P(w|C))t_{\text{incorrect}}$ — i.e., reported reading time will vary linearly with the word's probability.

Other authors have proposed a reciprocal (Narayanan & Jurafsky, 2004) or logarithmic (Hale, 2001; Levy, 2008) relationship, but based more on principles of elegance than any particular mechanism. There are, however, several more detailed reasons we might expect some specific non-linear relationship.

Analogy with frequency (prediction: none)

At first glance, we might expect a logarithmic relationship for predictability because of an argument by analogy: Predictability is conceptually similar to frequency, and frequency has a logarithmic effect. However, the predictability and frequency of any particular word vary on very different time-scales: a word's predictability may be radically different every single time it is encountered, because it is encountered in different contexts, but that same word's context-independent frequency will remain effectively constant over a timescale of months at least. This presents an obstacle to wholesale importing of theories designed to explain the shape of the frequency effect. To illustrate this, consider Forster's serial-search model (Forster, 1976; Murray & Forster, 2004), perhaps the most explicit extant proposal for why frequency has a logarithmic effect. In essence, this model assumes a frequency-sorted lexicon accessed by serial search; therefore accessing the 100th most frequent word takes twice as long as accessing the 50th most frequent word. Then, by Zipf's law, these rank frequencies turn out to be approximately equal to the log numerical frequencies. The fact that frequencies are relatively stable over time makes a frequency-sorted lexicon plausible. Predictabilities, though, are dependent on context, and so change radically from one word to the next. A serial search model of predictability effects would thus require us to accept a lexicon that is completely reordered before processing each word. This is difficult, given that such a reordering would necessarily involve examining every lexical item, which would seem to remove the need for a second search step.

Similarly, connectionist models of the word frequency effect explain it as arising from connection weights which are determined solely by training, and do not vary between contexts during the testing phase (Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989; Zevin & Seidenberg, 2002; Zorzi, Houghton, & Butterworth, 1998). Other theories have attributed this effect to, for example, the age at which the word was first learned (Morrison & Ellis, 1995; Morrison et al., 2003), or differences in the ease of memory retrieval as determined by the number of previous exposures to the word, modulated by recency (Morton, 1969) or the diversity of previous contexts (Adelman et al., 2006). What all of these theories have in common is that they attribute the classic word frequency effect to some property of the word itself, while predictability is intrinsically a property of the interaction between a word token and its current context. We are currently unaware of any theoretical mechanism which would link the effects of long-term frequencies and local predictabilities, or cause them to behave in similar quantitative fashions — and the absence of such a mechanism makes it difficult to accord any weight to the argument by analogy.

Optimal visual discrimination (prediction: logarithmic)

Norris, in the tradition of previous work using the Sequential Probability Ratio Test as a model of human choice reaction time (Stone, 1960; Laming, 1968; Carpenter & Williams, 1995; Gold & Shadlen, 2007), has proposed a model of word recognition as optimal visual discrimination: The Bayesian Reader (Norris, 2006, 2009). In this theory, the comprehender receives samples of noisy visual information at a fixed rate from their perceptual system, and their goal is to identify a word to some fixed degree of certainty as quickly as possible. According to Bayesian principles, the proper way to do this is to initially set one's belief in the word's identity to match its prior probability of occurrence — i.e., its predictability. Then, as each input sample arrives, this belief is updated by multiplying in the likelihood of the new sample and renormalizing. When the belief reaches some threshold, the word is declared to be identified. When transformed to log-probability space, this multiplication becomes addition, and thus the belief update process becomes a random walk in which each sample on average causes the log posterior probability of the correct word to increase with a near-constant step size. The expected number of steps—and thus the expected time—before reaching threshold for the word's true identity is thus linear in the log of the word's prior probability, so that in this model, the amount of time required to identify a word is proportional to its log-predictability (Carpenter & Williams, 1995).

Highly incremental processing (prediction: logarithmic)

Here we introduce a second theory which predicts a logarithmic relationship between word probability and comprehension time. In fact, we show that if processing is "incremental enough", then such a relationship will arise almost inevitably.

We start from the observation that, so far as we can determine, stimuli which have higher predictability are processed more efficiently in every task and species where this has been studied (e.g., Carpenter & Williams, 1995; Froehlich, Herbranson, Loper, Wood, & Shimp, 2004; Janssen & Shadlen, 2005; Pang, Merkel, Egeth, & Olton, 1992). We take this as evidence that there are domain-independent mechanisms by which predictability modulates the efficiency of cognitive processing, and assume that the relationship between word predictability and reading time arises because of these mechanisms affecting linguistic processing, rather than some mechanism specific to language comprehension (Smith & Levy, 2008). (Of course, the processes which produce expectancies for particular words are themselves highly sensitive to linguistic structure and usage; here we're speaking only of the mechanisms that link expectancies to reading time.) This assumption alone does not suggest any specific quantitative relationship between predictability and reading time. But if this is an instance of a more general predictability effect in which processing time for any stimulus is sensitive to that stimulus's predictability, P(stimulus|context), then we need to ask what constitutes a linguistic 'stimulus'. Words are a privileged unit of linguistic representation, and psycholinguistic theories therefore tend to assume that cognitive mechanisms for manipulating linguistic representations will be sensitive to the properties of individual words. But are words the units by which real-time language comprehension proceeds?

Our theory's second assumption is that they are not: that instead, language comprehension proceeds incrementally, by which we mean that processing a word involves processing a sequence of sub-word fragments (Tanenhaus et al., 1995). In auditory comprehension, this is well established; speech unfolds over time, and if a speaker says *candy* then the listener's language processor will be quite happy to start working on the initial /kæn-/ before they hear /-di/ (Tanenhaus et al., 1995). In reading, incrementality at the sub-word level is less well established, but might involve this same process after phonological recoding (Frost, 1998), or alternatively might involve multiple visual features (Morton, 1969) which arrive with different latencies and are processed in sequence. (Note that this does not require that the order in which features arrive matches the left-to-right order of letters within the word.)

Putting these two assumptions together, we have that the processing time for each fragment depends on the predictability *of that fragment*. Furthermore, we assume that prediction for each fragment takes into account the previous fragments, and that the total time required to process two sequences of fragments is the sum of the time required to process the first sequence plus the time required to process the second. These conditions are trivially satisfied if fragments are processed in a strictly serial manner, but this is not a requirement; they are also satisfied by, e.g., models in which the processing for adjacent fragments overlaps in time, but uses a limited pool of shared computational resources so that higher degrees of parallelism result in slower overall processing. This is analogous to the observation that while there may be overlap in the processing of adjacent words (i.e., spillover effects), nonetheless reading multiple words takes longer than reading a single word.

In this model, the predictability of words *per se* has no direct effect on their reading time; unpredictable words take longer to process only because they contain unpredictable fragments. This would seem to make it difficult to test the theory, because experimentally we can only measure word predictability and word reading time, not fragment predictability and fragment reading time. But fortunately, effects at the fragment level turn out to produce characteristic patterns at the word level. This follows from another difference between predictability and frequency. With frequency, there is no necessary relationship between fragment frequency and word frequency; e.g., a word can be rare without containing any rare syllables. True word frequency effects must therefore arise from whole-word processing. By contrast, the axioms of probability dictate that the context-conditional predictability of a word is the product of the context-conditional predictability of its parts,

$$P\,(/\text{kændi}/|\text{context}) = P\,(/\text{kæ n}-/|\text{context})$$

$$\times P\,(/-\text{di}|\text{context}, /\text{kæn}-/)\,.$$

More formally, take a word with conditional probability p_{word} that is composed of k fragments with conditional probabilities $p_1, \ldots, p_k$ respectively. (E.g., if *candy* is processed as /kæn-/, /-di/ then $k = 2$; processing it as /k-/, /-æ-/, /-n-/, /-d-/, /-i/ gives $k = 5$. Processing it truly continuously gives $k = \infty$.) And, let $f(x)$ be the function that gives the processing time for a fragment that has probability x (we assume that there is a single f for all fragment types). Then looking only at the portion of processing time which is dependent on predictability, we have two equalities:

$$\begin{aligned} p_{\text{word}} &= p_1 \times \cdots \times p_k \\ \text{total processing time} &= f\,(p_1) + \cdots + f\,(p_k)\,. \end{aligned}$$

These equations let us simulate the total processing time that would result from different choices of p_{word}, k, $f(x)$, and $p_1, \ldots, p_k$. Fig. 2 shows the result of such simulations, and reveals a regularity: As k increases, the total processing time becomes a better and better approximation to a logarithmic function of word predictability p_{word}:

$$\text{total processing time} \approx -\theta \log p_{\text{word}}$$

(Here θ is an arbitrary scaling parameter). Ultimately, this pattern is caused by the fact that in the equations above, probabilities multiply, while times add. If we set $f(x) = -\theta \log x$, then the above approximation becomes exact for all k, because logarithms convert products into sums. This makes the logarithm the unique fixed point of this process, which other choices of $f(x)$ converge to as k increases. (A proof is given in Appendix A.)

We do not, of course, suggest that the brain is actually literally calculating any limits; presumably some specific $f(x)$ and k apply for each word that is read. What the above analysis means, though, is that so long as k is large — that is, processing is 'highly incremental' — then we will be near the limit, and the details of choice of $f(x)$ or the distribution of probability within the word will have only minimal effect on the observed whole-word reading time. A logarithmic reading time curve arises inevitably — and, perhaps, epiphenomenally — from using a coarse whole-word measure to examine a collection of fine-grained sub-word processes, each of which are sensitive to predictability.

Uniform information density (prediction: super-logarithmic)

The uniform information density (UID) effect is that speakers seem to use various strategies to lengthen or shorten parts of their utterances so that the average predictability (as measured in bits) per unit time ends up being roughly constant (Aylett & Turk, 2004; Genzel & Charniak, 2002; Jaeger, 2010; Piantadosi, Tily, & Gibson, 2011). Levy and Jaeger (2007) proposed that one possible source of this pattern is as an audience design strategy. If comprehension difficulty induced by low-predictability words grows more quickly than the logarithm, then out of all ways of distributing a fixed amount of information across an utterance, the one which adheres most closely to the UID principle is also the one which will produce the lowest total comprehension difficulty. Intuitively, this occurs because for a super-logarithmic difficulty curve, Fig. 1, peaks in unpredictability produce a disproportionate amount of difficulty, which cannot be balanced out by an adjacent trough of similar size. So the logic of the proposed mechanism is: if producers attempt to make their utterances easy to comprehend, and if predictability has a super-logarithmic effect on comprehension time, then producers should adhere to the UID principle. The logic of empirical inference then inverts this: producers in many circumstances do adhere to the UID principle, and there must be some reason for this, so we should be predisposed to expect a super-logarithmic relationship between predictability and reading time; and, if one is found, that would provide further support for this account of UID effects.

Materials and methods

Accurately assessing the shape of the word predictability effect requires a large number of data points distributed evenly over a wide range of predictability values. The availability of such data has previously been restricted by the difficulty and expense of gathering cloze data, and its analysis limited by the use of factorial designs. Three aspects of our approach allow us to overcome these challenges. First, instead of relying on cloze, we estimate word probabilities using a state-of-the-art computational language model trained on a large corpus. While undoubtedly more errorful than good cloze norming, this allows us to estimate predictability for relatively unexpected words and over very large stimulus sets, which compensates for the increase in noise. Other psycholinguistic studies have used such computational methods (Demberg & Keller, 2008; Roark et al., 2009; Boston, Hale, Kliegl, Patil, & Vasishth, 2008); the primary difference is that our language model is chosen to give best-effort broad-coverage word probability estimates, not to proxy for any particular psycholinguistic theory (Frank & Bod, 2011) or to give high-quality estimates for specific grammatical structures (Levy, 2008). Second, we avoid the factorial approach in favor of a spline-based regression technique designed for measuring non-linear curve shapes (Wood, 2006). (For a previous application of this technique to psycholinguistic data, see Baayen, 2010b.) This also enables us to control for confounds (e.g., word frequency) post hoc, which allows us to analyze large stimuli sets using relatively natural texts rather than carefully normed sentences. Finally, the use of regression allows us to directly ask how the probability of word n affects reading time for word $n + 1$, after controlling for the probability of word $n + 1$. This reduces confounding by controlling for word-to-word correlations in frequency, predictability, etc. More importantly, it gives us a new and powerful way to measure spill-over effects (Mitchell, 1984; Rayner, 1998), letting us better capture predictability's full effect while additionally giving insight into its time-course.

Eye-tracking

First pass gaze durations (Rayner, 1998) were extracted from the English portion of the Dundee corpus (Kennedy, Hill, & Pynte, 2003), which records eye movements of 10 native speakers each reading 51,502 words of British newspaper text. Previous work (Demberg & Keller, 2008; Frank & Bod, 2011; Kennedy, Pynte, Murray, & Paul, in press) has reported predictability effects in this corpus, but did not examine curve shape.

Self-paced reading

Moving-window self-paced reading times (Just, Carpenter, & Woolley, 1982) were measured for 35 UCSD undergraduate native speakers each reading short (292–902 word) passages drawn from the Brown corpus of American English (2860–4999 total words per participant, mean 3912). In this paradigm, the participant must press a button to reveal each word in turn, and the time elapsed between button presses is recorded. 3 participants with comprehension-question performance at chance were excluded.

Probability estimation

Interpolated modified Kneser-Ney trigram word probabilities (Kneser & Ney, 1995; Chen & Goodman, 1998) were estimated from the British National Corpus (BNC Consortium, 2001) using SRILM v1.5.7 (Stolcke, 2002), and combined with a conditional bigram cache (Goodman, 2001). Self-paced reading analyses were adjusted for British/American spelling differences using VARCON (Atkinson, 2004). Our primary consideration in selecting this model was to maximize what Frank and Bod (2011) term 'linguistic accuracy', i.e., the model's ability to accurately predict words in corpora (perplexity), without regard to behavioural data. We certainly do not claim that this model is an appropriate theory of *how* the human comprehension system goes about making predictions. But, to the extent that our model and the brain are both attempting to achieve linguistic accuracy, they should arrive at numerically similar estimates (see also Fossum & Levy, 2012), and the analyses we present here depend only on our estimated probabilities acting as an accurate statistical proxy for the true subjective probabilities.

One possible source of inaccuracy is that in practice, our model relies primarily on local context for estimating predictabilities; in this respect it is similar to the transitional probabilities used in previous research (Demberg & Keller, 2008; Frisson, Rayner, & Pickering, 2005; McDonald & Shillcock, 2003a, 2003b), though we use a larger local context and a substantially more sophisticated estimation procedure. In addition, the bigram cache portion of our model reaches beyond local context to create increased expectancies for repeated mentions of words and short phrases across the entirety of each stimulus text. Nonetheless, there remain a variety of long-distance linguistic dependencies induced by syntax, semantics, etc., which this model captures only imperfectly. This is in contrast to humans, who are generally sensitive to such long-distance dependencies. While this sensitivity is of great importance to theories about human expectancy generation, it does not affect our analyses here unless such dependencies have a large and systematic effect on the numerical magnitude of expectation for a large proportion of the words in our stimuli, which seems unlikely. Extensive experience in computational linguistics confirms that local-context models empirically outperform syntax-based models when it comes to achieving high linguistic accuracy in unrestricted-domain texts like ours. This suggests that while distant context can have large effects on the predictability of some words, in practice it usually does not; on average, local context is the most reliable cue to word predictability. Our estimates, therefore, while sometimes noisy, should serve as an accurate statistical proxy overall. Most importantly, there is no clear reason why this choice of language model would bias our results regarding curve shape in any particular direction.

Curve estimation

We used mgcv v1.6-2 (Wood, 2004, 2006) to predict reading times using penalized cubic spline functions (20 d.f.) of word log-probability. As controls, we entered a spline function of position in text, a two-dimensional tensor spline interaction between orthographic word length and log-frequency, and factors indicating participant identity and (eye-tracking only) whether the previous word had been fixated. (This was motivated by preliminary analyses which revealed weak or non-existent interactions between predictability and either frequency or word length, but a substantial interaction between frequency and word length, which is consistent with previous findings (Kliegl et al., 2006; Pollatsek, Juhasz, Reichle, Machacek, & Rayner, 2008).) To capture spillover (Mitchell, 1984; Rayner, 1998), log-frequency/word-length interaction and probability terms were included for each word in an M-word window up to and including the current word, where M was chosen empirically to capture the effect present in each data set (eye-tracking: $M = 2$; self-paced reading: $M = 4$). All words were analyzed except those at the beginning or end of a line, or for which some word in the window did not appear in the British National Corpus, occurred adjacent to punctuation, or contained numerals. Eye-tracking analyses excluded unfixated words; self-paced reading analyses excluded outliers (reading times <80 ms, >1500 ms, or >4 sd above participant-specific means). Eye-tracking: $N = 166522$; self-paced reading: $N = 51552$. Fitting was by (penalized) least squares; confidence intervals were estimated by bootstrapping both participants and cases within participants, using the mgcv fitter's weights parameter to avoid replicating data across folds in its internal cross-validation routine. (This method takes subject random effects into account; for a discussion of item random effects in these analyses, see Appendix B). All reported results were robust to choice of spline basis, use of least squares estimation versus maximum likelihood estimation with the assumption of heavy-tailed (gamma-distributed) error, and the use of larger spillover windows (increased M); see Appendix C for further validation of penalized spline regression in this setting.

Results

Fig. 3 shows how the probability of a word w affects the reading time for w and the words immediately succeeding it (the spillover region). Our two data-sets show marked differences in time-course. For eye-tracking, the effect begins immediately and extends onto the next word, but is not seen on words further downstream. For self-paced reading, the effect does not begin until the succeeding word, and lasts through the third succeeding word. Nonetheless, if we sum these curves to find the total slowdown due to a particular unpredictable word (Fig. 4), then we find nearly identical effect sizes. This suggests that these tasks involve similar processing, though this processing is differently distributed through time with respect to saccades and button-presses respectively.

Crucially, Fig. 4 shows clearly that the relationship between word predictability and reading time is, in fact, logarithmic across at least six orders of magnitude in probability. (Lower probability items occur, but not often enough to reliably estimate curve shape without a larger data set; see online supplementary information for graphs with the full x-axis.)

Fig. 4 contains little visual evidence for super-logarithmicity. To check this more formally, we re-ran the above model fits, but now entering linear and quadratic functions of log-probability instead of an arbitrary spline (but keeping the same controls). A positive β coefficient on the quadratic term would indicate a super-logarithmic curve. We found no support for any quadratic component, positive or otherwise (eye-tracking: total $\beta = -0.05$, 95% CI = (−0.45,0.37), one-tailed $p = 0.59$; self-paced reading: total $\beta = 0.04$, 95% CI = (−0.90,1.07), one-tailed $p = 0.46$; statistics via bootstrap).

For the Dundee corpus, there were sufficient data to fit participant-specific models; results from these analyses are shown in Fig. 5. Nine out of ten participants showed clear effects of log-probability, all of which are overall linear in shape. The individual-participant data for the Brown self-paced reading corpus were not plentiful enough to conduct participant-specific analyses.

Discussion

The predictability effect on word comprehension in context takes a regular logarithmic form over at least six orders of magnitude in estimated predictability. This finding has both practical and theoretical consequences.

Practically speaking, predictability is potentially affected by nearly any manipulation one can make to linguistic structure. It is therefore a potential confound in most psycholinguistic studies, and knowing the quantitative form of this confound allows us to better control it. This non-linearity is very severe — Fig. 6. When word predictability is included as a covariate in regression analyses it should be log transformed; in factorial designs where average predictability is matched between conditions, it should be log-predictability rather than raw predictability that is matched. Since the uncertainty in the estimate of a word’s log-predictability for any given context will grow as the word’s predictability decreases, this also implies that in practice it is very difficult to assert with confidence from cloze norms that two different sets of word/context pairs are truly “equally” unpredictable in the sense that matters for real-time comprehension behavior. For example, a word whose true probability is 10^{-2} will act more like a word whose true probability is 1 than like one whose true probability is 10^{-6} — yet these will most likely be measured as having 0%, 100%, and 0% cloze, respectively. Our results suggest that there is no such thing as an unexpected word; there are only words which are more or less expected.

Anticipation versus integration

Our results bear on the theoretical debate about whether predictability effects in general arise from anticipatory pre-activation of specific words, or from post-hoc effects that arise while integrating the word into some kind of larger semantic context. The integration difficulty account (Brown & Hagoort, 1993; Foss, 1982; Hagoort, Baggio, & Willems, 2009; Hess, Foss, & Carroll, 1995; Traxler & Foss, 2000) holds that predictability itself does not affect comprehension difficulty, but rather that words which have high predictability scores are also those which are somehow more related to the prior context, and words which are more related to the prior context are also easier to integrate semantically. For example, processing the word *play* in examples (1) and (2) presumably requires us to construct some representation of two different scenarios: one involving my brother playing inside, and another involving children playing outside. If the latter scenario is easier to construct, then we expect *play* to be read more quickly in (2) than in (1). Crucially, under this account, predictability effects do not arise until after the comprehension system encounters the actual word; there may appear to be effects of predict*ability*, but they do not result from any cognitive process of predict*ion*. On the other hand, the anticipatory processing account holds that predictability effects do arise from some kind of processing which is predictive in the sense that it is dependent on the identity of the upcoming word, but occurs before this word identity is known (DeLong et al., 2005; Van Berkum et al., 2005). Our results provide challenges for both of these accounts.

It seems plausible that predictability will, in general, be correlated with semantic integration difficulty, so perhaps the apparent effects of predictability in empirical studies are actually a result of this confounding. But, is this correlation tight enough to explain our results? Intuitively, we expect these measures to be similar in some cases, but to diverge in others. For instance, producers avoid saying things which are too obvious from context, and so statements of obvious facts presumably have a simultaneously low integration difficulty and a low predictability; similarly, syntactic alternatives with similar semantic content presumably produce similar degrees of integration difficulty, but may have wildly different predictabilities. Our results do not rule out an integration difficulty account, but given the precise and law-like relationship we found, the challenge for such accounts becomes to explain why integration difficulty should vary in a quantitatively exact way with the logarithm of predictability.

The anticipatory processing account avoids this difficulty, because it is obvious why predictive processing would be sensitive to predictability *per se*; if you want to start processing words in some manner before you actually encounter them, then a word’s probability of occurrence given the available information, $P(w|C)$, may be a useful guide to decide which words should receive such processing, and to what degree. (Compare this to the situation after you have encountered the word, at which point it would seem mostly irrelevant how predictable it used to be when you had less information available.) And it has other appealing properties. It is independently motivated: there is ample independent evidence that the comprehension system anticipates upcoming material in at least some situations (Altmann & Kamide, 1999; DeLong et al., 2005; Kamide, Altmann, & Haywood, 2003; Kamide, Scheepers, & Altmann, 2003; Knoeferle et al., 2005; Van Berkum et al., 2005; Wicha et al., 2003). And, it provides an obvious reason why predictability differences would produce differences in reading time (as higher predictability words will receive more anticipatory processing, and thus require less post-hoc processing).

However, the most straightforward instantiation of anticipatory processing idea is the ‘simple guessing’ model we formalized above, which predicted a linear effect of predictability. Our results clearly rule this out. More generally, and for the same reason, these results are incompatible with any theory which assumes both that (a) predictability effects on reading time arise from processing which precedes the actual appearance of the word, and (b) the comprehension system can only apply this processing to a small number of words at any given moment (relative to the size of the lexicon). When such a model encounters a word with probability $<10^{-5}$, it will almost never have formed any expectation regarding it — yet the observed effect is just as strong in this region of word log-probability as anywhere else. The reading time difference between words with probability 10^{-6} and words with probability 10^{-5} is just as large as the difference between words with probability 10^{-2} and those with probability 10^{-1}. Thus, we must reject either (a) or (b). Integration difficulty accounts reject (a). If we wish to preserve an anticipatory processing account of these data, we must instead reject (b), and build theories in which expectancies do not take the form of simple guesses; instead, the comprehension system must be able to simultaneously pre-activate large portions of its lexicon in a quantitatively graded fashion (Smith & Levy, 2008).

Yet another possibility would be for anticipatory processing to be directed not at words, but at word fragments, which would make this account consistent with the incremental processing theory we propose here (which takes as granted that there is *some* mechanism linking predictability and processing time, and focuses on explaining the resulting curve shape), while potentially reducing the degree to which parallel pre-activation is necessary (as there are e.g. far fewer potential upcoming phonemes than there are potential upcoming words).

Consequences for UID

Our results lend no support to the audience-design account of uniform information density effects proposed by Levy and Jaeger (2007), which required a super-linear relationship between log probability (surprisal) and processing difficulty. We find no evidence for deviation from a pure logarithmic curve, which under their analysis would suggest that overall audience interpretation time is entirely unaffected by the uniformity or non-uniformity of information density. However, this need not be taken to rule out the possibility that other forms of audience design might motivate a UID principle. For example, if speech is consistently produced more quickly than it can be comprehended then it will eventually become incomprehensible, which gives producers an incentive to slow down on difficult content and let comprehenders catch up. In the case of predictability-related difficulty, producers who follow this strategy will end up following the UID principle, though under this revised account more local variation in information density would be acceptable. The original theory predicts that information density should be optimized on the time scale of individual processing fragments; here, what would matter is uniformity on a time-scale only fine-grained enough to avoid overloading comprehenders' working memory.

The Bayesian Reader versus the incremental processing account

There are two theories which predict the precise logarithmic effect we found: The Bayesian Reader (Norris, 2006, 2009) and the incremental processing account. Both find support in Fig. 4, but they make different predictions about the time-course of these effects, Fig. 3. In the Bayesian Reader model as originally formulated, predictability affects how much visual information the eye needs to gather from each word. This makes a clear prediction in the case of self-paced reading, where only one word is displayed at a time: Since you can't gather perceptual input from a word that has disappeared, the model expects a word's predictability to affect viewing time for that word only, with no spillover effect. Fig. 3b, however, shows the exact opposite pattern: There is little or no effect on the word itself, with a large spillover effect. We can perhaps overcome this difficulty at the cost of some theoretical elegance, by postulating that the noise bottleneck occurs not in visual perception per se, but at some later moment where word identity must be communicated between two internal processing stages connected by a noisy channel. Further analysis would be needed to determine whether such a mechanism could produce slowdowns distributed over such a wide temporal span (2-3 words).

The incremental processing account, by contrast, is based on the assumption that predictability affects not perception, but the speed of cognitive processing generally. Therefore, under this account we expect to see predictability effects at every moment that lexically associated processing occurs. Since spillover effects occur consistently in the psycholinguistic literature, this model makes the opposite prediction from the Bayesian Reader — that predictability effects should *not* be restricted to the period when the word is actually visible, which is what we find. While this prediction is not particularly surprising, it does make this the only extant model which can directly explain our full pattern of results. And this model, if correct, raises a number of new questions. Most obviously, what is the form of the true underlying function $f(x)$ relating predictability and processing time? Is it an arbitrary and idiosyncratic function that, say, varies between individuals, or is there some regularity to it, and if so, what? Answering this would require some other methodology, as per-word reading times are too coarse-grained a measurement to yield much insight. Of even greater theoretical interest is the question of the value of k, the grain-size of incremental processing; larger k would correspond to the processor operating on finer-grained or perhaps even truly continuous chunks of input (McClelland & Elman, 1986; Spivey, 2007). Although these new results do not give direct knowledge of k, consider that most possible functions are not logarithms, and that Fig. 2 indicates that for some possible functions a rather large k ($\geqslant$10) is required to produce a near-logarithmic curve shape like the ones we observe. Such a high degree of incrementality goes beyond what has already been established in the visual world paradigm (Tanenhaus et al., 1995) for the incremental processing of the speech signal, and certainly beyond what has otherwise been observed in reading. These results and model together, then, may provide an initial, tantalizing glimpse of a more fine-grained linguistic processor than has so far been exposed to experimental view. Other methods which allow more detailed measures of the time-course of processing, such as EEG/MEG, mouse-tracking (Spivey, Grosjean, & Knoblich, 2005), or hazard function analysis of eye movements (Feng, 2009) may yield further insights in this regard.

Surprisal as a causal bottleneck

Finally, these results confirm that it is plausible that all reading time predictability effects are mediated by lexical predictability, in accordance with the the causal bottleneck hypothesis of surprisal theory. Since the seminal work of Shannon on quantifying the bit rate of English (Shannon, 1951), information-theoretically informed work on language has recognized that all types of hierarchical predictive information present in language—syntactic, semantic, pragmatic, and so forth—must inevitably bottom out in predictions about what specific word will occur in a given context, and that when measured in bits, expectations at each successive level combine naturally in a simple additive fashion. This is illustrated by our example from the introduction, where the total bits carried by the word *who* is the sum of the bits associated with the fact of a relative clause's appearance in context C and the bits associated with the fact that the particular word introducing the relative clause is *who*:

$$\log P(who|C) = \log P(\text{rel. clause}|C) + \log P(who|\text{rel. clause}, C)$$

Our present results reveal that the bit is also the correct unit for measuring the processing time needed in general during incremental language comprehension by a native speaker; a logarithmic effect of lexical predictability both implies and subsumes logarithmic effects of transitional probability, syntactic predictability, semantic predictability, etc., allowing us to explain these apparently disparate effects as arising via a single unified mechanism. With contemporary probabilistic models of language structure we can measure the bits carried by a wide variety of abstract linguistic structures; the way is thus paved for their contributions to the time required for incremental language comprehension to be investigated and quantified using this common currency.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by NSF grant 0953870 and NIH grant HD065829 to R.L., NIH grant T32-DC000041 to the Center for Research in Language, NIH grant T32-MH20002 to the Institute for Neural Computation, funding from the Army Research Laboratories Cognition & Neuroergonomics Collaborative Technology Alliance, and the European Commission through the EU Cognitive Systems project Xperience (FP7-ICT-270273). We thank: H. Lu and N. Katz for assistance in collecting self-paced reading data, and V. Ferreira, T. F. Jaeger, R. Kliegl, M. Kutas, E. H. Levy, E. Mukamel, and K. Rayner for comments on the manuscript.

Appendix

Appendix A:

In the main body, we claimed that if these two equalities hold:

$$\begin{aligned} p_{\text{word}} &= p_1 \times \cdots \times p_k \\ \text{total processing time} &= f(p_1) + \cdots + f(p_k) \end{aligned}$$

then as k goes to infinity, the total processing time will become a closer and closer approximation to a logarithm of p_{word}. Here we formalize and prove this claim.

First, we need a way to talk about the p_i values as k increases. Let's assume we have a sequence of increasingly long vectors $(p_1,\ldots, p_k)_k$, for $k = 1,2,\ldots$. We'll write the ith entry in the kth vector as $p_{i,k}$, with $1 \le i \le k$. As per above, we require that for all k, $p_{1,k} \times\ldots\times p_{k,k} = p_{\text{word}}$. We do not, though, require that there be any necessary relationship between $p_{i,k}$ and $p_{i+1,k}$; our result holds even if each new vector is totally different than the one before. What we do need is a condition to avoid degenerate behavior. What we want to avoid is a sequence like this one, where all the uncertainty remains restricted to a single fragment:

$$p_{i,k} = \begin{cases} p_{word} & i=1 \\ 1 & i \neq 1 \end{cases}$$

Therefore, we require that as k increases, the probability of the most-surprising fragment should go to 1:

$$\lim_{k\to\infty} \min_i p_{i,k} = 1$$

This forces the word's overall uncertainty to be distributed over an increasing number of fragments.

Next we consider $f(x)$. For convenience, we assume that $f(1) = 0$. If it doesn't, then we can just subtract off a constant c, and this won't alter the shape of $f(x)$. This will also affect the total processing time graph by shifting it up or down by $k \times c$ units, but this only affects the baseline reading time, not the shape of the curve. To avoid degenerate behavior, we also assume that the derivative of $f(x)$ is defined at $x = 1$. This derivative will play an important role in our story, so we give it a name: $f'(1) = -\theta$.

Now given these conditions, we wish to show that

$$\lim_{k\to\infty}\left(\sum_{i=1}^{k} f(p_{i,k})\right) = -\theta \log p_{\text{word}}$$

It will be easier to work with this after a change of variables. Let $\tilde{p}_{\text{word}} = \log p_{\text{word}}$, and $\tilde{p}_{i,k} = \log p_{i,k}$. Plugging these transformations into our assumptions above, we have as given that

$$\begin{gathered}\tilde{p}_{\text{word}}=\sum_{i=1}^{k}\tilde{p}_{i,k}\\ \lim_{k\to\infty}\left(\min_{i}\tilde{p}_{i,k}\right)=0\\ g\left(0\right)=f\left(1\right)=0\\ g^{'}\left(0\right)=f^{'}\left(1\right)\times e^{0}=f^{'}\left(1\right)=-\theta\end{gathered}$$

and wish to show that

$$\lim_{k\to\infty}\left(\sum_{i=1}^{k}g\left(\tilde{p}_{i,k}\right)\right)=-\theta\times\tilde{p}_{\text{word}}.$$

Note that since $\sum\tilde{p}_{i,k}=\tilde{p}_{\text{word}}$, our result would hold if $g\left(\tilde{x}\right)$ were the linear function $-\theta\times\tilde{x}$. The idea of our proof is to use the derivative to approximate $g\left(\tilde{x}\right)$ as a linear function near 0. This turns out to be enough since as k increases, the $\tilde{p}_{i,k}$ values approach 0.

More formally, recall from the definition of the derivative that

$$\begin{aligned}-\theta=g^{'}\left(0\right)= & \lim_{h\to0^{-}}\frac{g(0+h)-g(0)}{h}\\ = & \lim_{h\to0^{-}}\frac{g(h)}{h}\end{aligned}$$

Therefore, for any $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $h \in (-\delta,0)$,

$$\left|\frac{g\left(h\right)}{h}-\left(-\theta\right)\right|<\varepsilon$$

which implies that

$$\begin{gathered}-\varepsilon<\frac{g(h)}{h}-\left(-\theta\right)<\varepsilon\\ -\theta-\varepsilon<\frac{g(h)}{h}<-\theta+\varepsilon\\ \left(-\theta-\varepsilon\right)h<g\left(h\right)<\left(-\theta+\varepsilon\right)h\end{gathered}$$

That is, for $h \in (-\delta,0)$, $g(h)$ is approximately linear with slope $-\theta \pm \varepsilon$.

Now, for k sufficiently large, we are guaranteed that $p_{i,k} \in (-\delta,0]$. And so for such a k, we can infer that $\sum_{i=1}^{k}g\left(\tilde{p}_{i,k}\right)$ is bounded by

$$\begin{aligned}\left(-\theta\pm\varepsilon\right)\sum_{i=1}^{k}\tilde{p}_{i,k}= & \left(-\theta\pm\varepsilon\right)\times\tilde{p}_{\text{word}}\\ = & \left(-\theta\times\tilde{p}_{\text{word}}\right)\pm\left(\varepsilon\times\tilde{p}_{\text{word}}\right).\end{aligned}$$

That is, we can always pick k so as to guarantee that $\sum_{i=1}^{k}g\left(\tilde{p}_{i,k}\right)$ is as close as we like to $-\theta\times\tilde{p}_{\text{word}}$, which is what we wanted to prove.

Appendix B:

A discussion of item effects in our analyses: In null-hypothesis significance testing within psycholinguistics, it is widely recognized that it is essential to take into account idiosyncratic differences among individuals and experimental items relevant to the dependent measure—both overall proclivities and sensitivities to psychological variables—because they break the conditional independence assumptions implicit in non-hierarchical ("flat") regression models. This issue is what motivates the use of by-participant and by-item "random effects" in repeated-measures ANOVA and mixed-effects regression models (Clark, 1973; Baayen, Davidson, & Bates, 2008a; see also Barr, Levy, Scheepers, and Tily, 2013 for specific discussion on this issue). The analyses above use a hierarchical bootstrap procedure (first on participant clusters, then on observations within each participant) that takes into account the by-subjects clustering structure in our data, but it does not take into account the clustering structure deriving from the fact that our data involve multiple measurements taken from a given [word,context] pairing—what would be called *item* random effects. Upon this view, the theoretically critical effect of word probability is "between items" rather than "within items", since a given [word,context] pair always has the same conditional probability in our language model, so that a reasonable way to model item random effects would be to assume that the underlying hypothetical "average" reading time for a given [word,context] pair across the potential participant population is offset from that predicted by the other components by a factor b_w drawn from some distribution with zero mean. This is known as an ITEM RANDOM INTERCEPT in the mixed-effects models literature (Baayen et al., 2008a). It is worth carefully considering how the results we obtained here might be affected by our omission of item random intercepts if they are present with non-negligible variance in the underlying generative process:

1. Omitting them could induce overconfidence in the parametric form of the probability/time relationship, artificially narrowing our confidence intervals;
2. Omitting them could induce the non-parametric model to overfit arbitrary, small deviations from the true underlying probability/reading-time function, since closely matching observed mean reading times for specific [word,context] pairs would produce spuriously high cross-validation scores and lead to the selection of too small a penalization term.

The former could lead to unjustified overconfidence in the inferred effect size and shape; the effect of the latter would, if anything, make us *less* likely to obtain a cleanly linear effect shape. Since we nevertheless did obtain a nearly perfectly linear effect shape in both datasets, is the first of these two possibilities, (1), that is of primary concern.

We first demonstrate that when the effect of word log-probability is assumed to be linear, its effect is highly significant even when crossed subject and item random effects are taken into account. For both the Brown and Dundee datasets we fit parametric linear mixed-effects models (Baayen, Davidson, & Bates, 2008b) to reading times. As fixed effects, we entered all predictors used in the main analysis except for participant identity, and with linear effects substituting for all splines. Our random effects structure included (i) a random intercept for word token, and (ii) random subject slopes for all word probability measures entered as fixed effects, with all correlations allowed (a "maximal" random-effects structure in the sense of (Barr et al., 2013)). Models were fit with (unrestricted) maximum likelihood estimation using the lme4 package (Bates, Maechler, & Bolker, 2011). Consistent with the findings of our non-parametric analyses, these analyses found that the linear effects of word log-probability were highly significant in all cases except for that of current-word log-probability in the Brown self-paced reading dataset (Dundee: current-word $|t|$ = 2.88, one-back $|t|$ = 5.59; Brown: current-word $|t|$ = 1.35, one-back $|t|$ = 4.83, two-back $|t|$ = 3.68, three-back $|t|$ = 3.13).

In addition, we conducted a “by-items” analysis of each dataset, computing mean reading time for each word token (aggregating across subjects) and then fitting our non-parametric model to each dataset. (For the eye-tracking dataset this meant discarding the predictor of whether the previous word was fixated.) Results are shown in Fig. B1 and B2; once again we recovered effects on reading time that were linear in word log-probability. The main difference from our earlier results is that in Fig. B2, the self-paced reading effect appears somewhat stronger than the eye-tracking effect. If true, this may be an artifact of our excluding unfixated words from the eye-tracking analysis.

Appendix c:

Penalized spline regression as implemented by mgcv (Wood, 2004, 2006) is a powerful and principled technique for estimating unknown non-linear relations. In order to fit nearly-arbitrary smooth curves, it uses a high-dimensional spline basis; in order to avoid the over-fitting that otherwise plagues such high-dimensional models, it combines the standard maximum likelihood criterion with a curvature penalty term that biases the regression towards less ‘wiggly’ curves. Critically, the relative weight placed on the likelihood term (which attempts to follow the data) versus the penalty term (which attempts to make the line smoother and closer to a straight line) is determined by cross-validation. In theory, therefore, this method’s fitted curves should be biased towards smoothness only to the extent that this helps it better match the true curve describing the underlying phenomenon.

But since our key empirical finding is that such a fit produces a straight line, it seems prudent to verify that this is not an artifact introduced by penalization. We therefore repeated our analysis, but using two different methods to remove this potential bias.

First, we ran the same model fits, but entering raw probability in place of log-probability; in this case we predict that the splines should attempt to form a steep logarithmic curve (since we believe that is the true underlying relationship), while the penalization pushes towards a linear relationship (Fig. 1). As expected, mgcv’s algorithm chose to apply very small penalization weights (ranging from 65 to 64000 times smaller than the corresponding weights chosen in the original analyses), which in turn allowed the resulting spline fit to form an highly-nonlinear, approximately logarithmic curve with substantial local variation around this underlying trend (Fig. C1b; note that while the fit was performed using *raw* probability, we plot the result against *log* probability to facilitate comparison with other fits). Second, we fit our original model, but with penalization simply disabled (i.e., using standard least-squares); this produced similar results (Fig. C1c). The three models thus agree that the underlying relationship is approximately logarithmic.

Finally, we would like to confirm that the local non-linear deviations from this trend that we see in models (b) and (c) are the result of over-fitting rather than a true effect. We verified this by performing 1000-fold cross-validation on all three models, and found that in both data sets, the original penalized model (Fig. C1a) achieved the highest log-likelihood on held-out data. (Similar results, not shown, were obtained when performing cross-validation of the penalized model versus the other models on the “by item” data set described above.) Thus we conclude that, to the limits of our data, the underlying relationship between word probability and processing time is in fact logarithmic.

References

Adelman JS, Brown GDA, Quesada JF. Contextual diversity, not word frequency, determines word-naming and lexical decision times. Psychological Science. 2006; 17(9):814–823. doi: 10.1111/j.1467-9280.2006.01787.x. [PubMed: 16984300]

Altmann GTM, Kamide Y. Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition. 1999; 73(3):247–264. doi: 10.1016/S0010-0277(99)00059-1. [PubMed: 10585516]

Atkinson, K. The VARCON database, version 4.1. 2004. Retrieved from http://wordlist.sourceforge.net/

Aylett M, Turk A. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Language & Speech. 2004; 47(1):31–56. [PubMed: 15298329]

Baayen RH. Demythologizing the word frequency effect: A discriminative learning perspective. The mental lexicon. 2010a; 5(3):436–461. doi: 10.1075/ml.5.3.10baa.

Baayen, RH. The directed compound graph of English: an exploration of lexical connectivity and its processing consequences. In: Olsen, S., editor. New impulses in word-formation. Vol. Vol. 17. Buske; Hamburg: 2010b. p. 383-402.

Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language. 2008a; 59(4):390–412.

Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language. 2008b; 59(4):390–412. doi: 10.1016/j.jml.2007.12.005.

Bar M. Predictions: a universal principle in the operation of the human brain. Philosophical Transactions of the Royal Society B: Biological Sciences. 2009; 364(1521):1181–1182. doi: 10.1098/rstb.2008.0321.

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

Bates D, Maechler M, Bolker B. lme4: Linear mixed-effects models using S4 classes [Computer software]. 2011 (R package version 0.999375-42).

BNC Consortium.. (version 2, BNC World) [Text corpus]. The British National Corpus. 2001. Retrieved from http://www.natcorp.ox.ac.uk/

Boston MF, Hale J, Kliegl R, Patil U, Vasishth S. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. Journal of Eye Movement Research. 2008; 2:1–12.

Brown C, Hagoort P. The processing nature of the n400: Evidence from masked priming. Journal of Cognitive Neuroscience. 1993; 5(1):34–44.

Carpenter RHS, Williams MLL. Neural computation of log likelihood in control of saccadic eye movements. Nature. 1995; 377(6544):59–62. doi: 10.1038/377059a0. [PubMed: 7659161]

Chen, SF.; Goodman, J. An empirical study of smoothing techniques for language modeling. Computer Science Group, Harvard University; Cambridge, MA: 1998. (Technical Report No. TR-10-98)Retrieved from ftp://ftp.deas.harvard.edu/techreports/tr-10-98.ps.gz

Clark HH. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior. 1973; 12:335–359.

Dahan D, Tanenhaus MK. Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. Journal of Experimental Psychology: Learning, Memory, & Cognition. 2004; 30(2):498–513.

DeLong KA, Urbach TP, Kutas M. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nature Neuroscience. 2005; 8:1117–1121.

Demberg V, Keller F. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition. 2008; 109(2):193–210. [PubMed: 18930455]

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

Engbert R, Nuthmann A, Richter EM, Kliegl R. SWIFT: A dynamical model of saccade generation in reading. Psychological Review. 2005; 112(4):777–813. [PubMed: 16262468]

Federmeier KD, Kutas M. A rose by any other name: Long-term memory structure and sentence processing. Journal of Memory and Language. 1999; 41:469–495.

Feng G. Time course and hazard function: A distributional analysis of fixation duration in reading. Journal of eye movement research. 2009; 3(2):1–22.

Ferreira F, Clifton C Jr. The independence of syntactic processing. Journal of Memory and Language. 1986; 25:348–368.

Forster, K. Accessing the mental lexicon. In: Wales, RJ.; Walker, ECT., editors. New approaches to language mechanisms. North-Holland; Amsterdam: 1976. p. 257-287.

Foss DJ. A discourse on semantic priming. Cognitive Psychology. 1982; 14(4):590–607. doi: 10.1016/0010-0285(82)90020-2. [PubMed: 7140212]

Fossum, V.; Levy, R. Sequential vs. hierarchical syntactic models of human incremental sentence processing. Proceedings of the 3rd workshop on cognitive modeling and computational linguistics (CMCL 2012); Montréal, Canada. Association for Computational Linguistics; 2012. p. 61-69.Retrieved from http://www.aclweb.org/anthology/W12-1706

Frank SL, Bod R. Insensitivity of the human sentence-processing system to hierarchical structure. Psychological Science. 2011; 22(6):829–834. doi: 10.1177/0956797611409589. [PubMed: 21586764]

Frisson S, Rayner K, Pickering MJ. Effects of contextual predictability and transitional probability on eye movements during reading. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2005; 31(5):862–877.

Froehlich AL, Herbranson WT, Loper JD, Wood DM, Shimp CP. Anticipating by pigeons depends on local statistical information in a serial response time task. Journal of Experimental Psychology: General. 2004; 133(1):31–45. [PubMed: 14979750]

Frost R. Toward a strong phonological theory of visual word recognition: True issues and false trails. Psychological Bulletin. 1998; 123(1):71–99. (WOS:000071350100006) doi: 10.1037//0033-2909.123.1.71. [PubMed: 9461854]

Genzel D, Charniak E. Entropy rate constancy in text. Proceedings of ACL. 2002

Gold JI, Shadlen MN. The neural basis of decision making. Annual Reviews of Neuroscience. 2007; 30(1):535–574. doi: 10.1146/annurev.neuro.29.051605.113038.

Goodman, JT. A bit of progress in language modeling, extended version. Microsoft Research; Redmond, WA: 2001. (Technical Report No. MSR-TR-2001-72)

Hagoort, P.; Baggio, G.; Willems, RM. Semantic unification. In: Gazzaniga, MS., editor. The new cognitive neurosciences. 4th ed. Vol. Vol. 4. MIT Press; Boston, MA: 2009. p. 819-836.

Hale, J. A probabilistic Earley parser as a psycholinguistic model. Proceedings of NAACL-2001; Stroudsburg, PA. Association for Computational Linguistics; 2001. p. 159-166.

Hess DJ, Foss DJ, Carroll P. Effects of global and local context on lexical processing during language comprehension. Journal of Experimental Psychology: General. 1995; 124(1):62–82. doi: 10.1037/0096-3445.124.1.62.

Howes DH, Solomon RL. Visual duration threshold as a function of word-probability. Journal of Experimental Psychology. 1951; 41:401–410. [PubMed: 14873866]

Ilkin Z, Sturt P. Active prediction of syntactic information during sentence processing. Dialogue & Discourse. 2011; 2(1):35–58.

Jaeger TF. Redundancy and reduction: Speakers manage syntactic information density. Cognitive Psychology. 2010; 61(1):23–62. doi: 10.1016/j.cogpsych.2010.02.002. [PubMed: 20434141]

Janssen P, Shadlen MN. A representation of the hazard rate of elapsed time in macaque area LIP. Nature Neuroscience. 2005; 8(2):234–241. doi: 10.1038/nn1386.

Just MA, Carpenter PA, Woolley JD. Paradigms and processes in reading comprehension. Journal of Experimental Psychology: General. 1982; 111(2):228–238. [PubMed: 6213735]

Kamide Y, Altmann GTM, Haywood SL. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and Language. 2003; 49(1):133–156. doi: 10.1016/S0749-596X(03)00023-8.

Kamide Y, Scheepers C, Altmann GTM. Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from german and english. Journal of

Psycholinguistic Research. 2003; 32(1):37–55. doi: 10.1023/A:1021933015362. [PubMed: 12647562]

Kennedy A, Hill R, Pynte J. The Dundee corpus. Proceedings of the 12th European conference on eye movement. 2003

Kennedy A, Pynte J, Murray WS, Paul S-A. Frequency and predictability effects in the Dundee corpus: An eye movement analysis. The Quarterly Journal of Experimental Psychology. (in press). doi: 10.1080/17470218.2012.676054.

Kliegl R, Nuthmann A, Engbert R. Tracking the mind during reading: The influence of past, present, and future words on fixation durations. Journal of Experimental Psychology: General. 2006; 135(1):12–35. [PubMed: 16478314]

Kneser R, Ney H. Improved backing-off for M-gram language modeling. Proc. ICASSP. 1995:181–184.

Knoeferle P, Crocker MW, Scheepers C, Pickering MJ. The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. Cognition. 2005; 95(1):95–127. doi: 10.1016/j.cognition.2004.03.002. [PubMed: 15629475]

Kutas M, Federmeier KD. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology. 2011; 62(1):621–647. doi: 10.1146/annurev.psych.093008.131123.

Kutas M, Hillyard SA. Brain potentials reflect word expectancy and semantic association during reading. Nature. 1984; 307:161–163. [PubMed: 6690995]

Laming, DRJ. Information theory of choice-reaction times. Academic Press; London: 1968.

Lau E, Stroud C, Plesch S, Phillips C. The role of structural prediction in rapid syntactic analysis. Brain & Language. 2006; 98:74–88. [PubMed: 16620944]

Levy R. Expectation-based syntactic comprehension. Cognition. 2008; 106:1126–1177. [PubMed: 17662975]

Levy R, Fedorenko E, Breen M, Gibson T. The processing of extraposed structures in English. Cognition. 2012; 122(1):12–36. doi: 10.1016/j.cognition.2011.07.012. [PubMed: 22035959]

Levy R, Jaeger TF. Speakers optimize information density through syntactic reduction. Advances in neural information processing systems. 2007; 19:849.

Manning, CD.; Schütze, H. Foundations of statistical natural language processing. MIT Press; 1999.

McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive Psychology. 1986; 18(1):1–86. doi: 10.1016/0010-0285(86)90015-0. [PubMed: 3753912]

McDonald SA, Shillcock RC. Eye movements reveal the online computation of lexical probabilities during reading. Psychological Science. 2003a; 14(6):648–652. [PubMed: 14629701]

McDonald SA, Shillcock RC. Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. Vision Research. 2003b; 43:1735–1751. [PubMed: 12818344]

McRae K, Spivey-Knowlton MJ, Tanenhaus MK. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. Journal of Memory and Language. 1998; 38(3): 283–312.

Mitchell, DC. An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. In: Kieras, DE.; Just, MA., editors. New methods in reading comprehension research. Lawrence Erlbaum Associates; Hillsdale, New Jersey: 1984.

Morrison CM, Ellis AW. Roles of word frequency and age of acquisition in word naming and lexical decision. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1995; 21(1): 116–133. doi: http://dx.doi.org/10.1037/0278-7393.21.1.116.

Morrison CM, Hirsh KW, Duggan GB. Age of acquisition, ageing, and verb production: Normative and experimental data. The Quarterly journal of experimental psychology. A, Human experimental psychology. 2003; 56(4):705–730. doi: 10.1080/02724980244000594.

Morton J. Interaction of information in word recognition. Psychological Review. 1969; 76(2):165–178.

Murray WS, Forster KI. Serial mechanisms in lexical access: The rank hypothesis. Psychological Science. 2004; 111(3):721–756.

Narayanan, S.; Jurafsky, D. A Bayesian model of human sentence processing. Nov. 2004 Retrieved from http://www.icsi.berkeley.edu/~snarayan/newcog.pdf

Ni W, Crain S, Shankweiler D. Sidestepping garden paths: Assessing the contributions of syntax, semantics and plausibility in resolving ambiguities. Language & Cognitive Processes. 1996; 11(3): 283–334.

Norris D. The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. Psychological Review. 2006; 113(2):327–357. [PubMed: 16637764]

Norris D. Putting it all together: A unified account of word recognition and reaction-time distributions. Psychological Review. 2009; 116(1):207–219. [PubMed: 19159154]

Pang K, Merkel F, Egeth H, Olton DS. Expectancy and stimulus frequency: a comparative analysis in rats and humans. Perception & psychophysics. 1992; 51(6):607–615. [PubMed: 1620572]

Piantadosi ST, Tily H, Gibson E. Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences. 2011; 108(8):3526–3529.

Plaut DC, McClelland JL, Seidenberg MS, Patterson K. Understanding normal and impaired word reading: Computational principles in quasi-regular domains. Psychological Review. 1996; 103(1): 56–115. doi: 10.1037/0033-295X.103.1.56. [PubMed: 8650300]

Pollatsek A, Juhasz BJ, Reichle ED, Machacek D, Rayner K. Immediate and delayed effects of word frequency and word length on eye movements in reading: A reversed delayed effect of word length. Journal of Experimental Psychology: Human Perception & Performance. 2008; 34(3):726–750. [PubMed: 18505334]

Rayner K. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin. 1998; 124(3):372–422. [PubMed: 9849112]

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

Reichle ED, Pollatsek A, Fisher DL, Rayner K. Toward a model of eye movement control in reading. Psychological Review. 1998; 105(1):125–157. [PubMed: 9450374]

Roark B, Bachrach A, Cardenas C, Pallier C. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. Proceedings of the 2009 conference on empirical methods in natural language processing. 2009:324–333.

Seidenberg MS, McClelland JL. A distributed, developmental model of word recognition and naming. Psychological Review; Psychological Review. 1989; 96(4):523–568.

Shannon CE. Prediction and entropy of printed English. Bell Systems Technical Journal. 1951; 30:50–64.

Smith, NJ.; Levy, R. Optimal processing times in reading: a formal model and empirical investigation. In: Love, BC.; McRae, K.; Sloutsky, VM., editors. Proceedings of the thirtieth annual conference of the Cognitive Science Society; Austin, TX. Cognitive Science Society; 2008. p. 595-600.

Spivey, MJ. The continuity of mind. Oxford University Press; New York: 2007.

Spivey MJ, Grosjean M, Knoblich G. Continuous attraction towards phonological competitors. Proceedings of the National Academy of Sciences. 2005; 102:10393–10398.

Staub A, Clifton C. Syntactic prediction in language comprehension: Evidence from either…or. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2006; 32(2):425–436. doi: 10.1037/0278-7393.32.2.425.

Stolcke, A. Proc. intl. conf. on spoken language processing. Vol. Vol. 2. Denver; 2002. SRILM — an extensible language modeling toolkit; p. 901-904.

Stone M. Models for choice-reaction time. Psychometrika. 1960; 25:251–260.

Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC. Integration of visual and linguistic information in spoken language comprehension. Science. 1995; 268:1632–1634. [PubMed: 7777863]

Taylor WL. “Cloze procedure”: A new tool for measuring readability. Journalism Quarterly. 1953; 30:415–433.

Traxler MJ, Foss DJ. Effects of sentence constraint on priming in natural language comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2000; 26(5):1266–1282. doi: 10.1037/0278-7393.26.5.1266.

Van Berkum JJA, Brown CM, Zwitserlood P, Kooijman V, Hagoort P. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. Learning and Memory. 2005; 31(3):443–467.

Wicha NYY, Bates EA, Moreno EM, Kutas M. Potato not pope: human brain potentials to gender expectation and agreement in spanish spoken sentences. Neuroscience Letters. 2003; 346(3):165–168. doi: 10.1016/S0304-3940(03)00599-8. [PubMed: 12853110]

Wood SN. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association. 2004; 99:673–686.

Wood, SN. Generalized additive models: An introduction with R. Chapman and Hall/CRC; Boca Raton: 2006.

Zevin JD, Seidenberg MS. Age of acquisition effects in word reading and other tasks. Journal of Memory and Language. 2002; 47(1):1–29. doi: 10.1006/jmla.2001.2834.

Zorzi M, Houghton G, Butterworth B. Two routes or one in reading aloud? a connectionist dual-process model. Journal of Experimental Psychology: Human Perception and Performance. 1998; 24(4):1131–1161. doi: http://dx.doi.org/10.1037/0096-1523.24.4.1131.

* Human reading time for words varies logarithmically with word probability.
* This is predicted by a novel incremental processing model.
* It is also partially predicted an existing optimal perceptual discrimination model.

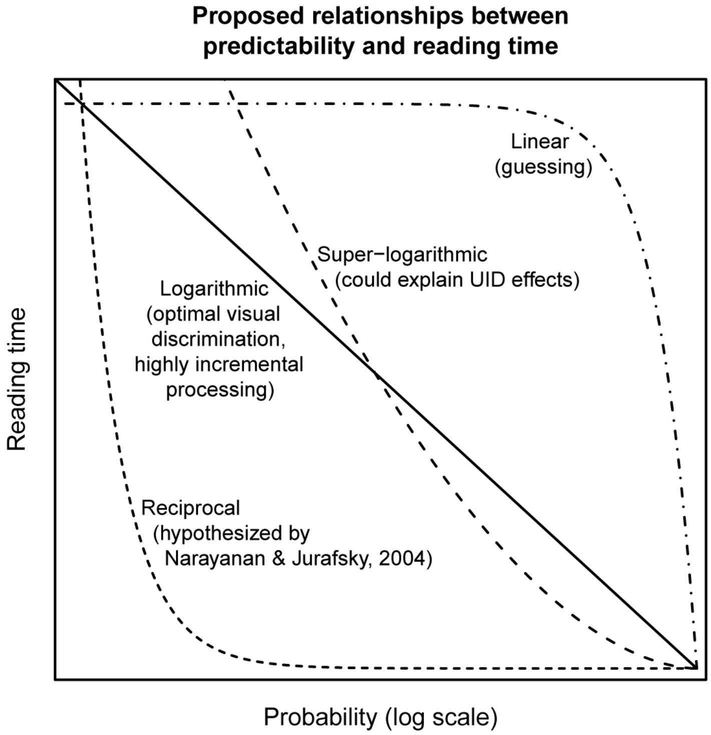

Figure 1.

Several hypothesized forms for the predictability effect, plotted in log space. (Of course many other forms are also possible *a priori*; here we show only those previously mentioned in the literature.) A linear effect is predicted by a simple guessing model. A logarithmic effect is predicted by both an optimal visual discrimination account (Norris, 2006) and an incremental processing account (see text). A super-logarithmic effect is predicted by the audience design theory of uniform information density effects (Levy & Jaeger, 2007).

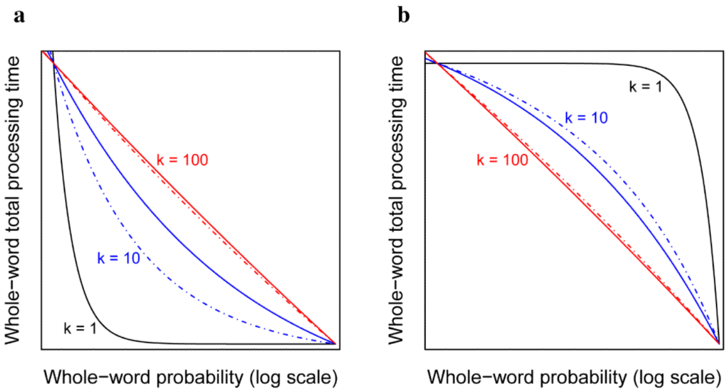

Figure 2.

These graphs show the whole-word processing times resulting from different variants of the incremental processing model. We consider: A linear effect at the fragment level (**a**, $f(x) = -x$) versus a reciprocal effect (**b**, $f(x) = 1/x$), for different values of k. For $k > 1$, we also consider two different possibilities for how probability is distributed among the fragments: Either uniformly ($p_i = \sqrt[k]{p_{\text{word}}}$, solid lines) or with later fragments more predictable than earlier fragments ($p_i = p_k^{(k+1-i)^2}$ with p_k chosen so that $p_1 \times \ldots \times p_k = p_{\text{word}}$, dashed lines). In all cases, more highly incremental processing (larger k) produces a logarithmic effect at the word level ($f(p_1) + \ldots + f(p_k) \approx \log p_{\text{word}}$).

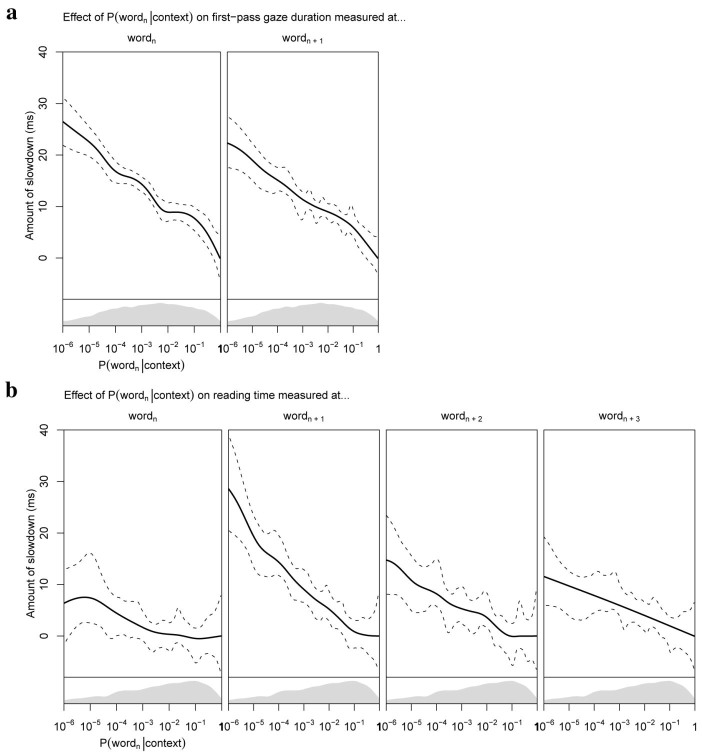

Figure 3.

The effect of the probability of word *n* on reading time measured at word *n* and on successive words (the spill-over region). Curves are penalized splines with point-wise 95% confidence intervals. To correct for inter-subject variability, we measure the effect of probability against the notional baseline of a perfectly predictable word; zero on this graph does not indicate an instantaneous overall reading time. Confidence intervals do not include the uncertainty induced by measurement error in probability estimation. Lower panels show the proportion of data available at each level of probability. (**a**) First-pass gaze durations. (**b**) Self-paced reading times.

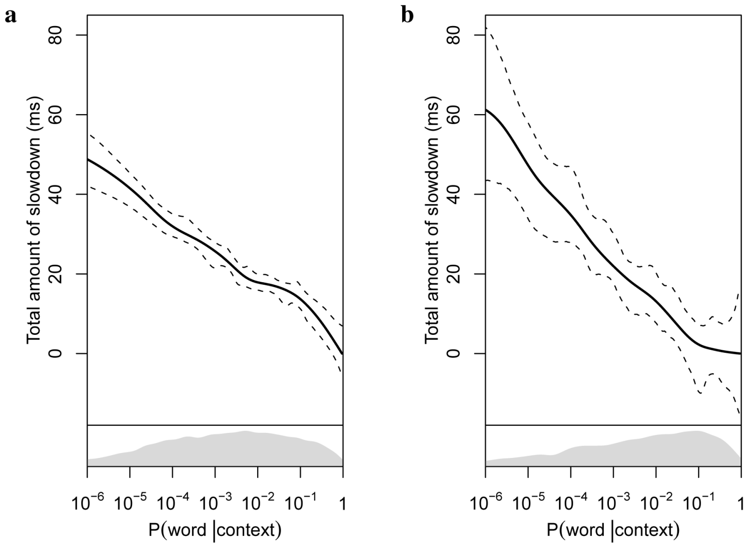

Figure 4.

By summing the curves in Fig. 3, we can estimate the total slowdown caused by an unpredictable word, regardless of where in the spillover region this slowdown occurs. (**a**) First-pass gaze durations. (**b**) Self-paced reading times. Lower panels show the proportion of data available at each level of probability.

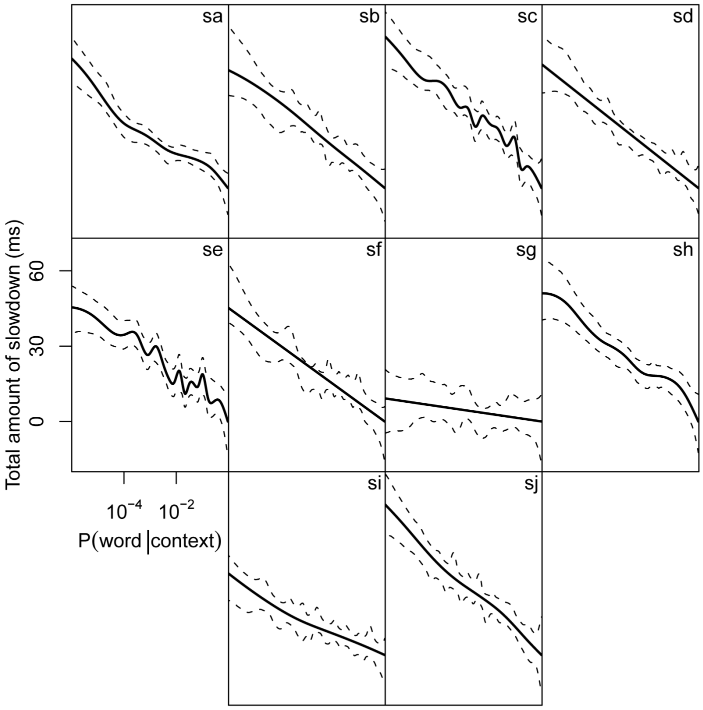

Figure 5.

To visualize inter-individual variation, we break down the Dundee corpus summed slowdown data (Fig. 4a), analyzing each participant separately. Participant codes from the corpus are shown in the upper right of each panel. Dashed lines represent bootstrapped point-wise 95% confidence intervals. The variation in 'wiggliness' of the main curves results in part from noise and numerical instability in mgcv's GCV-based penalization selection (Wood, 2004) allowing over/under-fitting in some cases. Even so, 9 out of 10 participants show effects of log-probability with an overall linear trend, while no effect was found for participant 'sg'.

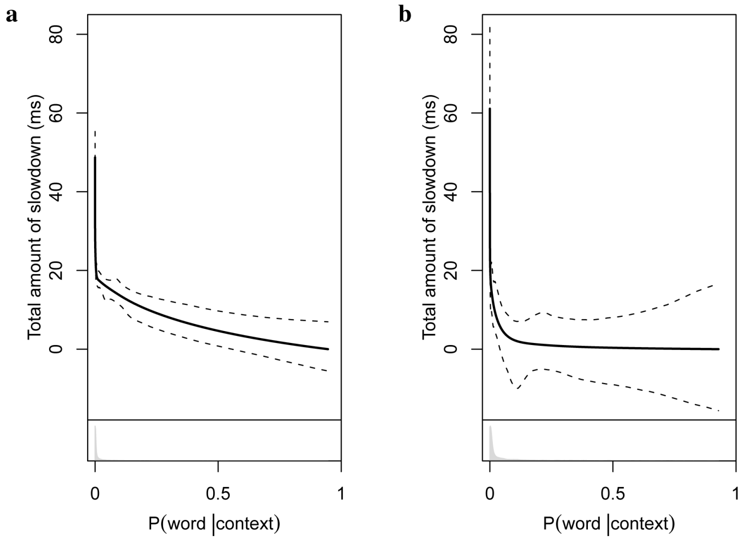

Figure 6.

The same curves shown in Fig. 4, but here plotted against raw predictability to better show the severity of the non-linearity. (**a**) First-pass gaze durations. (**b**) Self-paced reading times. Lower panels show the proportion of data available at each level of probability. (While these accurately indicate that the majority of our data is concentrated in the <0.1 range, the scale here is somewhat misleading; both analyses contain >10 000 data points with conditional probability >0.1.)

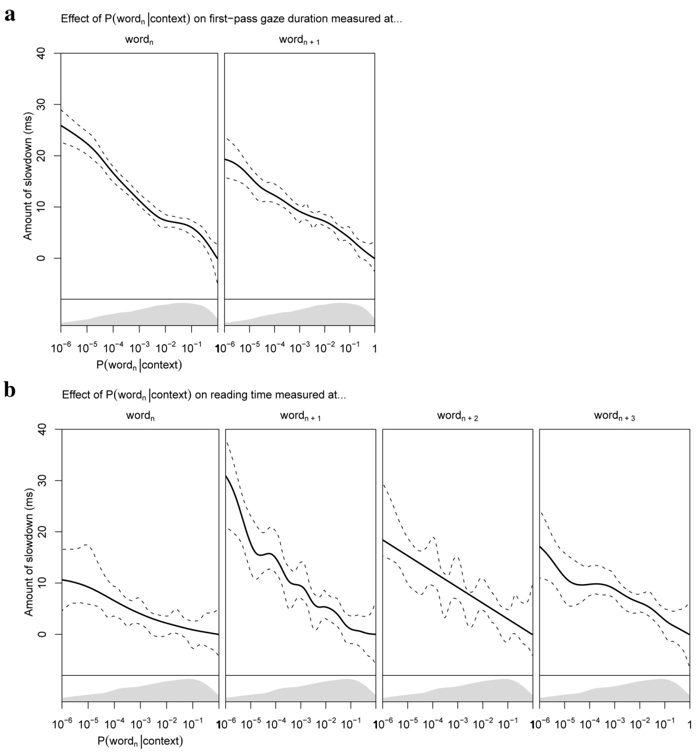

Figure B1.

“By item” analysis of per-token mean reading times aggregated across participants, showing the effect of the predictability of word$_n$ on word$_n$ and succeeding words. 95% confidence intervals calculated by bootstrapping over cases. (**a**) Eye-tracking. (**b**) Self-paced reading. Lower panels show the proportion of data available at each level of probability.

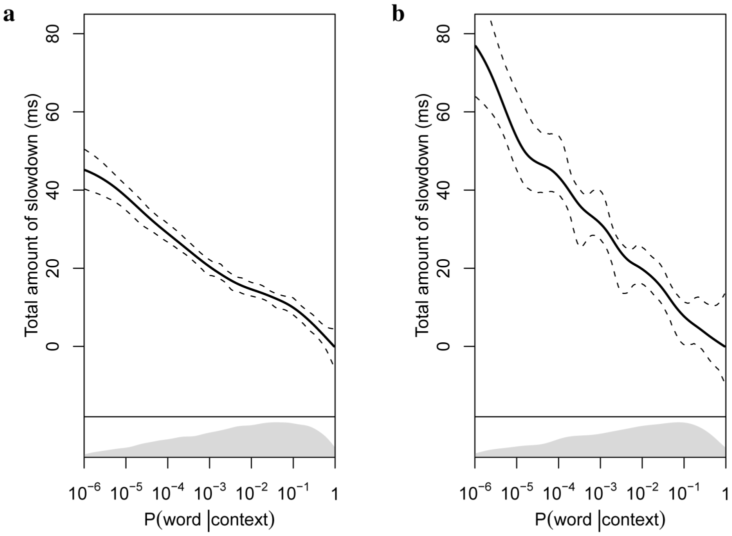

Figure B2.

"By item" analysis of per-token mean reading times aggregated across participants, showing the total reading time slowdown attributable to word predictability. (The sum of the curves in Fig. B1). 95% confidence intervals calculated by bootstrapping over cases. (**a**) Eye-tracking. (**b**) Self-paced reading. Lower panels show the proportion of data available at each level of probability.

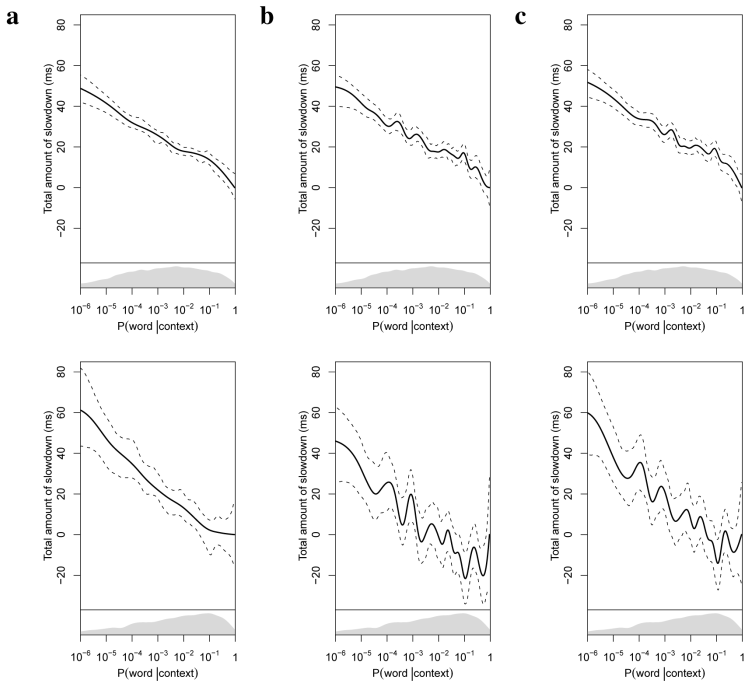

Figure C1.

The effect of penalization in controlling over-fitting. (**a**) Our original, penalized model (a repeat of Fig. 4). (**b**) The same model as in a, but fit with raw probability entered instead of log probability, then plotted in log-space. (**c**) The same model as in a, but fit without penalization. Upper panels show first-pass gaze durations; lower panels show self-paced reading times. That the lower panels show more wiggliness than the upper ones is presumably due to the relative sizes of the two data sets; in the absence of penalization, the smaller data set allows more overfitting than the larger. Dashed lines denote point-wise 95% confidence intervals. Lower panels show the proportion of data available at each level of probability.