Fixation durations in first-pass reading reflect uncertainty about word identity

Nathaniel J. Smith
njsmith@cogsci.ucsd.edu
UC San Diego Department of Cognitive Science
9500 Gilman Drive #515, La Jolla, CA 92093-0515 USA

Roger Levy
rlevy@ling.ucsd.edu
UC San Diego Department of Linguistics
9500 Gilman Drive #108, La Jolla, CA 92093-0108 USA

Abstract

In reading, it is often assumed that words are recognized sufficiently quickly, accurately, and unambiguously that downstream processes may proceed with perfect information about word identity. For example, word predictability is believed to affect early reading time measures, yet a word's predictability cannot be calculated without knowledge of the word's identity. We argue that such information is not, in general, available to the language processing system, and that it proceeds with only probabilistic information about word identity. We predict therefore that what have been analyzed previously as predictability effects must instead be based on noisy estimates of word predictability that are influenced by the predictability of visually similar words (neighbors). We test this prediction by building a Bayesian model of visual word recognition, using it to compute the 'average neighborhood surprisal' of words in a corpus, and testing the ability of this novel measure to explain human reading time data.

Keywords: Psychology, Cognitive Science, Perception, Language Understanding, Bayesian Modeling, Neighborhood Effects, Visual Uncertainty, Reading

Performance in isolated word recognition tasks is often affected by the existence or properties of words that are not presented, but that are visually similar to the words which are presented. For instance, a word with many neighbors — especially high frequency neighbors — tends to produce a faster response in the lexical decision task, and a slower response in naming or reading tasks (Perea & Rosa, 2000). Norris (2006) has argued that these divergent results can be best explained by uncertainty in the processing system. That is, noise is an inevitable component of all biological computation, and if the processor receives only noisy information about a word's shape, then it must consider all similar looking words as candidates for identification. When there are many such candidates, identifying the single correct candidate (as in the naming task) becomes more difficult, because there are many incorrect distractors and only one correct target; resolving this difficulty requires the acquisition of more sensory information, which requires more time. In the lexical decision task, however, it is not necessary to determine *which* word is seen, only *whether* a word is seen, and therefore increasing the number of candidates only makes it easier to give a correct response (even if for the 'wrong' reason).

However, in reading — that is, processing connected language, rather than isolated words — another consideration arises. In a naming task, no response can be given until the word is identified, but when reading, the ultimate outcome is not the name of a single word, but an understanding of the text as a whole. Here, we ask: can the linguistic processing associated with a word proceed before that word is uniquely identified? And if so, what are the consequences for processing? It's possible that neighborhood effects are limited to some early, serial, word identification process, in which any uncertainty is resolved before higher-level linguistic processes begin. Alternatively, this uncertainty may be propagated through the linguistic processing system itself.

Most current models of high-level language processing fall into the former category; for instance, they take as input words, rather than probability distributions over words. However, there is some reason to suspect that the latter possibility is more plausible. Spoken language, in particular, is a very noisy signal, in which word identification is generally impossible without reference to high-level linguistic constraints. Furthermore, listeners are willing to revise their identification of perceptually ambiguous phonetic material in light of disambiguating material that follows within a short period (Connine, Blasko, & Hall, 1991). In reading, the availability of a stable visual record makes it possible in principle to acquire substantially more detailed perceptual information — but in practice the average fixation length in reading is 200 ms, comparable to the time required to plan a motor saccade. This suggests that the next saccade must be initiated almost as soon as the fixation begins, and that decisions about its timing — and thus the fixation time for the current word — must be made before the current word is fully processed. In addition, Levy, Bicknell, Slattery, and Rayner (2009) have recently used evidence from a reading task to argue that certain syntactic constructions associated with garden-path-like processing difficulty may arise from uncertainty about the identity of critical words earlier in the sentence. Therefore it seems plausible that the language processing system not only has the capacity to handle uncertain input, but that this ability is used in natural reading.

Here, we examine this question via the well-known effect of word predictability on reading time (predictable words are read more quickly, Ehrlich & Rayner, 1981). This is a useful tool, because *(i)* the effect is very early; it affects the duration of initial fixations on a word, in the 200–300 ms range, when we would most expect some uncertainty to remain, and *(ii)*

as word predictability depends on the fit between the present word and its context, it implicates higher-level linguistic processing in a way that word frequency, for instance, might not, and yet *(iii)* it cannot affect processing until the word is fully identified, because different words are differently predictable. All theories which invoke word predictability to explain early reading time measures therefore implicitly assume that word identification occurs early and fully.

We hypothesize that this effect does not arise from predictability *per se*, but from the processing system's 'best guess' at the word's predictability, given the uncertain information available to it. To test this hypothesis, we build a simple Bayesian model of visual word recognition, use it to estimate 'best guess' predictabilities on a corpus, and test whether this improves our ability to predict human reading-time measures.

Word recognition model

We begin with a standard Bayesian model of word recognition in sentence context, in which beliefs about the identity of the word on which the eyes are currently fixated are formed by integrating top-down prior expectations from language knowledge and context with bottom-up perceptual input:

$$P(\text{word}|\text{context, input}) = \frac{P(\text{word}|\text{context})P(\text{input}|\text{word, context})}{P(\text{input})} \quad (1)$$

The first term in the numerator, $P(\text{word}|\text{context})$, corresponds to top-down prior expectations and can be estimated from any of a variety of language-modeling techniques standard in computational linguistics (Manning & Schütze, 1999). The second term in the numerator, $P(\text{input}|\text{word, context})$, corresponds to bottom-up perceptual evidence and is the present focus: we are investigating the possibility that this evidence is imperfect and that this imperfection may be reflected in rapid eye-movement decisions in reading.

We introduce three simplifying assumptions to make our model of perceptual evidence more tractable. First, we assume conditional independence between input and context given word identity, which is natural since it is the word being identified rather than the preceding context that generates the relevant perceptual input. Second, we assume that readers are aware of how many letters exist in the word that they are looking at, and only their identity is in doubt. (A more detailed model would certainly relax this assumption, but we believe that the high visual salience of inter-word spaces makes it a reasonable initial approximation.) Third, we assume that the subjective evidence for a given letter depends only on the noisy input we receive describing that letter (and this noisy input, of course, depends in turn on the letter that is actually present in the world). In particular, we assume that our bottom-up perceptual evidence for each letter in a word is probabilistically independent of that for the other letters. Therefore, we can write the perceptual evidence for a word as simply the product of the evidence for each of the n letters which comprise it. If E is the complete perceptual input derived from a word and E_i is the component of that perceptual input arising from the i-th letter, then normative Bayesian inference for the word's identity looks as follows:

$$\begin{aligned} P(\text{letters}|\text{input}) &= \frac{P(E|\text{letters})P(\text{letters})}{P(E)} \\ &\propto P(E_1,\ldots,E_n|\text{letters})P(\text{letters}) \\ &= P(\text{letters})\prod_{i=1}^{n} P(E_i|\text{letter}_i) \end{aligned}$$

The term $P(\text{letters})$ is simply the prior probability of the word in question; the perceptual evidence for the word is represented by the term $\prod_{i=1}^{n} P(E_i|\text{letter}_i)$.

To estimate the perceptual evidence $P(E_i|\text{letter}_i)$ obtained from each position in the word, we made use of letter-confusion matrices derived from previous norming experiments with the lowercase English alphabet (Engel, Dougherty, & Jones, 1973; Geyer, 1977). In each of these experiments, participants were presented with isolated letters for durations brief enough to induce considerable identification error, and the frequency with some presented letter α was identified as some letter β was tabulated as $f_{\alpha\beta}$. Here $\alpha = \beta$ implies correct identification and $\alpha \neq \beta$ implies misidentification. Finally we used these frequency tables to obtain a matrix M, in which each entry $M_{\alpha\beta}$ denotes the estimated probability of identifying letter α as β. For example, M_{ti} is relatively high, presumably reflecting the visual similarity of the letters t and i, whereas M_{tn} is relatively low.

Since these norming studies used viewing conditions rather unlike those that occur in natural reading, we assume that the matrix entries $M_{\alpha\beta}$ specify only the *relative* perceptual evidence provided by each letter of the word, rather than the *absolute* evidence. We therefore introduce a single free parameter q which scales the matrix as a whole, so that for the i-th letter of a word in a sentence,

$$P(E_i = \alpha|\text{letter}_i = \beta) \propto (M_{\alpha\beta})^q. \quad (2)$$

This allows us to estimate the overall level of noise in the model when analyzing human reading-time data.[1] The parameter q can be interpreted as the overall quantity of information acquired by the reader and used to inform downstream decisions; each entry in the confusability matrix is raised to the power q, and then rows are renormalized. Thus, $q = 0$ creates a uniform posterior distribution over letters, or perfect ignorance, while in the limit as q goes to infinity, the matrix becomes diagonal — representing perfect in-

[1] Note that we are making a simplifying assumption by equating the perceptual evidence from the i-th letter with the letter actually in the word, rather than with noisy perceptual input generated from the actual letter, as is done in models such as (Norris, 2006). This simplifying assumption can be interpreted roughly as marginalizing over the perceptual input itself; see (Smith, Chan, & Levy, 2010) for discussion of the justification for and implications of this simplifying assumption.

formation about letter identity. Varying q between these extremes smoothly varies the overall accuracy of letter information available, while preserving relative differences in letter similarity and recognizability. Figure 1 depicts the resulting letter-confusion matrices for $q = 1$ and $q = 2$.

This idea of rescaling was also used in producing our perceptual confusion matrix M from the raw norming data. We assumed that the two experiments had different overall levels of perceptual noise, and we used maximum likelihood to find the single matrix M that — when rescaled for each experiment — best explained the data from both. However, simply averaging the two norming matrices would produce similar results.

In aggregate, these assumptions give us the following final estimate of the subjective probability that we are observing a particular word given both context and visual input:

$$P(\text{word}|\text{context, visual input}) \propto P(\text{word}|\text{context}) \prod_i P(\text{letter}_i|\text{visual input}). \quad (3)$$

Average neighborhood surprisal

Now that we have a model of the uncertainty affecting the language processing system, we can model its consequences for the predictability effect. Word predictability itself is well-described computationally by surprisal — the negative log-probability of a word in context (Hale, 2001; Levy, 2008). For clarity, in this paper we will refer to this as the *raw surprisal* (RS). We now define the *average neighborhood surprisal* (ANS) of a word in some context to be the average of the surprisal of every word that might occur in that context, weighted by that word's similarity to the visible word, $P(\text{word}|\text{context}, \text{visual input})$. More formally,

$$ANS(\text{word}_k|\text{context}) = \sum_i P(\text{word}_i|\text{context, word}_k) RS(\text{word}_i|\text{context}). \quad (4)$$

Our fundamental prediction is that ANS will better predict reading times than RS.

The intuition here is that the processing system would prefer to spend an amount of time on a word proportional to its RS, but since visual noise makes the RS unavailable, the ANS is the best available approximation. The visual system is accurate enough that in most cases $P(\text{word}_k|\text{context, word}_k)$, the subjective probability that one is looking at word k given that one is, in fact, looking at word k, will be close to one; therefore ANS will generally be close to the RS for any given word. However, if a word has visually similar neighbors with higher surprisals, then this will pull up the ANS, and the reader will spend more time on that word 'just in case' it turns out to be one of those high-surprisal neighbors that require more time to process. Contrariwise, if a word has visually similar neighbors with lower surprisals, then this will pull down the average, and our reader will hurry onward faster than they otherwise might. Note especially that in this model, a word with a dense neighborhood may be read either faster *or* slower than a word with a sparse neighborhood. It's not the size of your neighborhood that matters, it's who your neighbors are.

It should also be noted that other models of neighborhood effects generally predict that the presence of higher-frequency neighbors will produce an inhibitory effect on word identification, as these neighbors interfere with recognition of the true word (e.g., Perea & Rosa, 2000). Our prediction is nearly the opposite — that in reading, the presence of high probability neighbors should lead to shorter initial fixations (though it is possible that later, as more information about the word's true identity becomes available, the eyes may slow or regress in compensation).

Methods

We compared average neighborhood surprisal to raw surprisal as predictors of human reading times in the Dundee eye-movement corpus (Kennedy, Hill, & Pynte, 2003), which consists of all eye-movements made by 10 subjects while reading a collection of newspaper articles totaling approximately 50,000 words. Several previous studies have already demonstrated surprisal effects on reading times in the Dundee corpus (Demberg & Keller, 2008; Frank, 2009; Smith & Levy, 2008). We analyzed both first fixation times — defined as the duration of the first fixation to land on each fixated word in a text — and second fixation times, defined as the duration of the second fixation to land on each word that was fixated a second time. We eliminated all fixations on words that occurred at the beginning or end of a line, which preceded or followed punctuation, that did not occur in the BNC (i.e., unknown words), or that occurred in the BNC but in segmented form (e.g., the BNC codes *don't* as two words, *do* followed by *n't*). Finally, we eliminated any remaining words containing uppercase letters, since our confusion norms only cover the lowercase alphabet. This left 182,169 first fixations and 42,024 second fixations for further analysis.

To obtain conditional word probabilities for both raw surprisal estimates and noisy conditional word-probability estimates (Equation 3) we used a trigram language model trained on the 100 million word British National Corpus (BNC), using the SRI Language Modeling Toolkit (Stolcke, 2002); the trigram model was smoothed using modified Kneser-Ney (Kneser & Ney, 1995), a standard technique for broad-coverage language modeling. Average neighborhood surprisal was estimated for each fixated word by plugging in raw surprisal estimates to Equation (4), and repeating this process at each value of q required by the fitting process.

As Smith and Levy (2008) have previously demonstrated that the relationship between surprisal and first fixation times in this corpus is linear, we simply regressed fixation time on RS and ANS simultaneously, with frequency (estimated from the BNC) and word length as controls. The noise parameter q was fit simultaneously with the regression coefficients by maximum likelihood. Gamma distributed error was assumed,

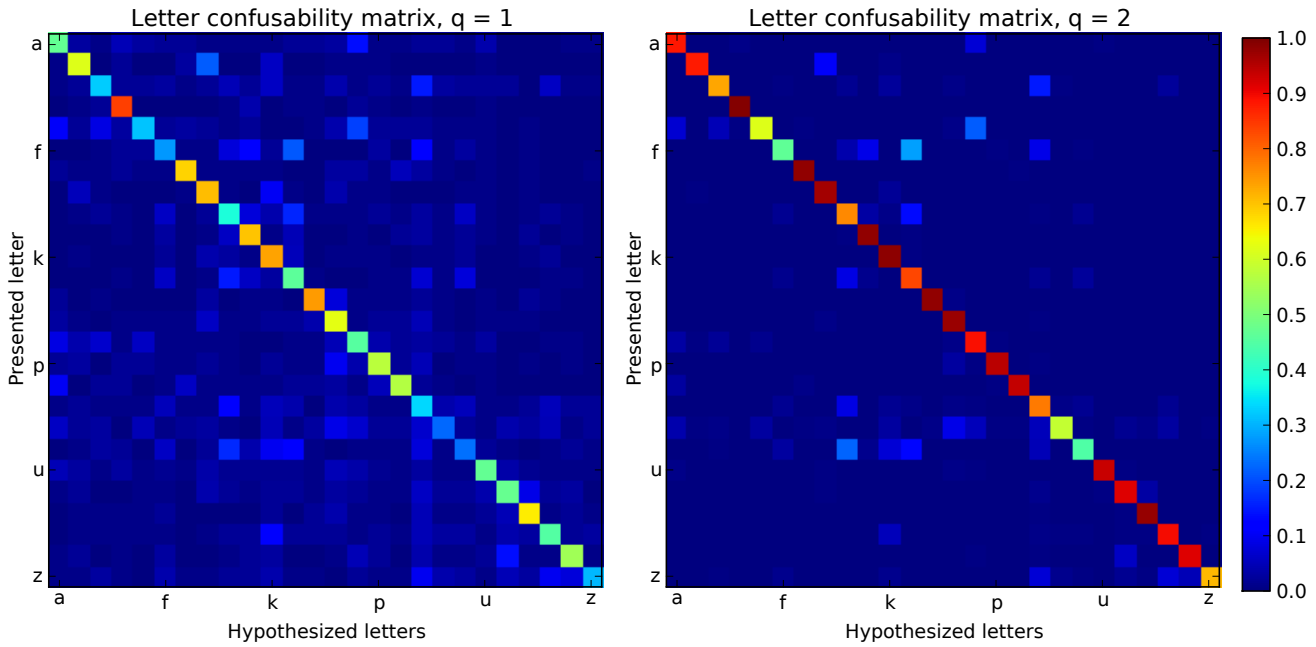

Figure 1: The letter confusability matrix, for different values of the scaling parameter q. For instance, presentation of the letter a to the noisy input system eventually gives rise to a particular posterior distribution over letters that is represented by the top row in each matrix. The diagonal represents the probability of veridical perception; we can see that the letter d is the least confusable in the lowercase English alphabet. As q increases (right), more information becomes available, causing the posterior distribution to cluster around the diagonal.

in order to properly account for the long right-ward tail in fixation durations.

Results

First fixations

The best fitting model had a moderate level of noise ($q = 1.306$), corresponding to a mean naming accuracy for individual letters of 66%. While this may seem low, most words contain enough letters that, combined with the constraint of linguistic context, this allows for substantial information about word identity. As a result, ANS and RS are highly correlated ($R^2 = 0.96$) — suggesting that while the models differ greatly in terms of the cognitive processes they postulate, they may be difficult to disentangle experimentally.

Even so, our data set turned out to be large enough for the regression model to give an unambiguous result: ANS better predicts human behavior than RS. That is, ANS is highly significant after controlling for RS ($t(182155) = 4.164, p \ll 0.001$), while RS has no significant effect after controlling for ANS ($t(182155) = 0.489, n.s.$). This result also remains after controlling for neighborhood size (N).

Second fixations

The same analysis on second fixations produces analogous results; ANS is highly significant ($t(42010) = 4.209, p \ll 0.001$), while RS is marginally significant in the wrong direction ($t(42010) = -1.847, p = 0.06$). More interesting, however, is examination of the q value; we predicted that a second fixation would provide more visual information about word identity, and thus result in a higher q. In fact, for second fixations, we found $q = 2.939$, suggesting that by the end of the second fixation, the eye movement control system has access to somewhat more than twice the information it has at the end of the first fixation.

Frequency prior

Equation (4) suggests that to compute the estimated, average neighborhood surprisal, the processing system must be able to, in some sense, compute the probability of *all possible* words in the current context, and sum over all of them, in time to affect the first fixation. This is a strong claim, and so to test it we calculated a simplified version of ANS in which we modified Equation 3 to replace the context-sensitive prior over words, $P(\text{word}|\text{context})$, with a simple, context-insensitive word frequency prior, $P(\text{word})$. This modified ANS was then added to our regression as an additional control. Our original context-sensitive ANS remained highly significant ($t(182154) = 4.413, p \ll 0.001$), suggesting that in the neighborhood effects we describe, the definition of 'neighborhood' is indeed sensitive to linguistic context.

Other determiners of reading time

While in this preliminary work we have focused on surprisal as a model reading time predictor, the essential argument applies to any word property which is believed to affect reading time, and one could define *average neighborhood X* for any interesting property *X* that was believed to affect reading time (or language processing behavior more generally). Generally, we would predict that to the extent the brain processes sensitive to property *X* must work from noisy representations of linguistic input, *average neighborhood X* would also be a better predictor of human behavior than *X* alone.

We have begun to examine this more general prediction, and in the process discovered a mystery. Using the above model to define average neighborhood word frequency, we find our regression against reading times gives just as unambiguous results as for surprisal — but the other way. That is, raw frequency is significant, and average neighborhood frequency is not. This suggests that whatever process produces word frequency effects in reading times appears to have exact information about the frequency (and therefore identity) of the word being processed, while the process which produces predictability effects has only noisy and imperfect information. Furthermore, this is true even on first fixations, so it cannot be a simple matter of the frequency effect arising later in the processing stream, when more information is available. (Evidence for frequency as a later effect than predictability would also, it seems safe to say, surprise most experts in the field.)

Discussion

Our fundamental prediction — that early predictability effects in reading are modulated by the predictability of visually similar (but unseen) words — was confirmed. Furthermore, the reduction of this effect on second fixations gives insight into the time course for resolution of uncertainty about word identity, and the failure of the word frequency prior to adequately explain the data argues for the ability of high-level linguistic constraint to quickly and robustly modulate the resolution of visual uncertainty. All our results — with the possible exception of the mysterious frequency non-effect — are compatible with a model of reading in which uncertainty about the input is propagated forward into the linguistic processing system itself.

Going forward, a major question is whether the noise we observe is truly visual noise, or whether it has another source. After all, biological computation necessarily involves noise and uncertainty at every level. When reading, for example, visual information must be gathered at the retina, transmitted and analyzed by the visual system, and converted to some higher level representation of word identity; then, this representation must be maintained in memory for semantic processing and integration. None of these processes can be perfectly veridical or reliable; all must introduce some amount of noise and uncertainty. Here, we built a specifically visual noise model, relying on a visual confusability matrix and a letter-based word representation, but presumably all models of word similarity/confusability are similar to the first order, and we did not compare against any other noise model; therefore, while our results suggest that average neighborhood surprisal drives reading time, it may be premature to conclude that the visual system is the source of uncertainty being averaged over.

In future work, we hope to make a sharper test of this part of the model in two ways. First, we can fit a different noise parameter q for letters at different degrees of eccentricity from visual fixation; if this reproduces the classic curve of acuity falling off with increasing eccentricity, then that would be stronger evidence that our noise arises from visual processing limitations. Second, looking the other direction, we plan to build a simple phonological/auditory noise model, and use it to estimate ANS for written words. If this model outperforms the visual noise model, then that would be strong evidence that the noise is in fact noise in some post-recoding internal representation. Finding auditory noise in a visual paradigm would be quite curious, but there is some precedent; for instance, it has been argued that the true determiner of neighborhood size for purposes of word naming effects is the number of words which are simultaneous visual and phonological neighbors (Adelman & Brown, 2007).

Finally, we hope that further investigation may shed light on the lack of a neighborhood effect on word frequency. One possibility is that further study of the noise, as described above, will provide a clue — perhaps visual information is highly accurate, the frequency effect is a relatively early and low-level effect acting on this low-level, accurate visual representation, and the predictability-sensitive process is working with a later representation more subject to internal noise. However, this remains mere speculation, and we welcome any suggestions on this matter. In another way, though, this dissociation of frequency and predictability is quite exciting, as it suggests a possible avenue for understanding the relationship between these highly similar linguistic properties. (Indeed, as they are inherently confounded in any study using isolated words stripped of context, and quite difficult to accurately measure and deconfound in more naturalistic stimuli, it has long been unclear whether they represented distinct effects at all.) This is, to our knowledge, the first study to find qualitatively different effects of each, and we hold high hopes that our current confusion may lead to a deeper future understanding.

Acknowledgments

We are grateful to Michael Tanenhaus for the initial suggestion of averaging surprisal over the visual neighborhood, and to Shane T. Mueller for maintaining the invaluable Letter Similarity Data Set Archive (`http://obereed.net/lettersim/`). This research was partially supported by NIH Training Grant T32-DC000041 to the Center for Research in Language at UC San Diego to NJS, and by NSF grant 0953870 to RL.

References

Adelman, J. S., & Brown, G. D. A. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review*, *14*, 455–459.

The British National Corpus, version 3 (BNC XML edition). (2007). (Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: `http://www.natcorp.ox.ac.uk/`)

Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints. *Journal of Memory and Language*, *30*(2), 234–250.

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

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.

Engel, G. R., Dougherty, W. C., & Jones, G. B. (1973). Correlation and letter recognition. *Canadian Journal of Psychology*, *27*(3), 317–326.

Frank, S. L. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 1139–1144).

Geyer, L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception and Psychophysics*, *22*, 487–490.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL-2001* (pp. 159–166).

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

Kneser, R., & Ney, H. (1995). Improved backing-off for M-gram language modeling. In *Proc. ICASSP* (pp. 181–184).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1093–1582.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357.

Perea, M., & Rosa, E. (2000). The effects of orthographic neighborhood in reading and laboratory word identification tasks: A review. *Psicológica*, *21*(3), 327–340.

Smith, N. J., Chan, W.-H., & Levy, R. (2010). Is perceptual acuity asymmetric in isolated word recognition? evidence from an ideal-observer reverse-engineering approach. In *Proceedings of the 32nd annual meeting of the cognitive science society.*

Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the thirtieth annual conference of the Cognitive Science Society.*

Stolcke, A. (2002). SRILM — an extensible language modeling toolkit. In *Proc. intl. conf. on spoken language processing* (Vol. 2, pp. 901–904). Denver.