Title
Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 33(33)

ISSN
1069-7977

Authors
Smith, Nathaniel
Levy, Roger

Publication Date
2011

Peer reviewed

Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing

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

Abstract

When performing online language comprehension, comprehenders probabilistically anticipate upcoming words. Psycholinguistic studies thus often depend on accurately estimating stimulus predictability, either to control it or to study it, and this estimation is conventionally accomplished via the cloze task. But we do not know how effectively — or even, strictly speaking, whether — cloze probabilities reflect comprehender predictions. This is both methodologically worrisome and an obstacle to detailed understanding of online predictive mechanisms. Here, we demonstrate first that cloze probabilities vary substantially and systematically from normative corpus statistics, and secondly that some portion of these deviations are also reflected in online comprehension measures. Therefore, while there is some reason to be concerned that cloze norming may be distorting the results of psycholinguistic studies, these apparent distortions may instead reflect genuine errors in native speakers' probabilistic models of their language.

Keywords: Psychology; Linguistics; Prediction; Language Understanding; Reading; Rationality

There's currently a great deal of interest in how the brain makes and uses predictions (Bar, 2009). Within psycholinguistics, this interest dates back 30 years, to the discovery that the predictability of a word — its probability of occurrence given preceding context — has large and robust effects on both reading times (Ehrlich & Rayner, 1981) and event-related brain potentials (Kutas & Hillyard, 1984). These early studies, and innumerable others since, rely on the cloze task (Taylor, 1953) to measure the predictability of their stimuli. Many more studies use cloze to control for predictability in order to isolate some other variable of interest. Yet despite its ubiquitous use as an estimate of predictability, we know almost nothing about what this task is actually measuring.

The cloze task consists of presenting a large group of participants with sentence stems like *In the winter and* _____, and asking each to fill in the blank with some plausible continuation — some might write *spring*, others *summer*, and so on. We then count up what proportion of participants responded with each word; this proportion is called the cloze probability of that word in that context. Our goal is to get some estimate of the subjective probability distribution over continuations which skilled comprehenders compute implicitly during online comprehension; Fig. 1 summarizes the logical relationship between these subjective probability distributions, cloze probability distributions, and alternative corpus based measurements.

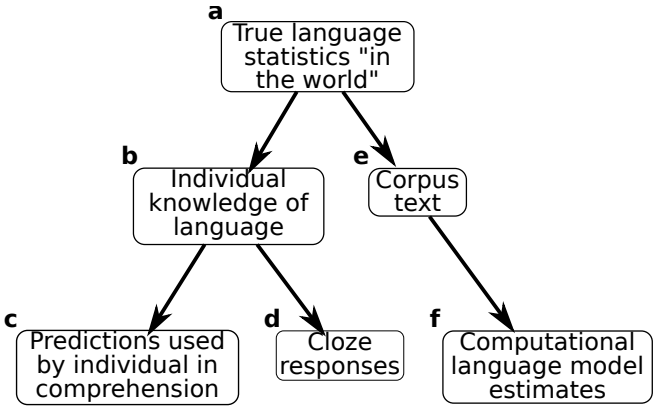

Figure 1: An informal illustration of the situation faced by those who wish to study linguistic prediction. Language is actually used in some particular ways in the real world (**a**); some subset of these uses are recorded in corpora (**e**), and may be used to train computational language models (**f**). A different subset is experienced by human language users, who use these experiences to create some internal model of the statistics of their language (**b**). They then draw on this internal model to make predictions during online linguistic comprehension (**c**) and also, presumably, when responding in the cloze task (**d**). But the actual relationship between the items on the left side of the diagram remains obscure — do cloze completions match online predictions? Do online predictions match real-world statistics?

We know that the participants in a cloze task have some knowledge of their language (Fig. 1**b**), which they presumably draw on when producing continuations. But isn't clear how they use this knowledge. If they generated their cloze responses by sampling from their subjective probability distribution ('probability matching'), then cloze probabilities would be identical to subjective probabilities.[1] But cloze norming is an offline, untimed, and rather unnatural task, which leaves ample room for conscious reflection and other strategic effects to distort this process — if participants are

[1] At least if we ignore inter-subject variation, as is conventional.

even probability matching in the first place. So our first question is: what distortions, if any, are introduced by the processes that produce **d** from **b**?

This question has important methodological implications, because if there are systematic biases in cloze estimates, then in the worst case attempts to measure or control for predictability might actually *introduce* confounds. For instance, if there were some measure M that affected cloze but did not affect reading times, then a reading time study that compared cloze-matched high-M and low-M items might find a spurious effect of M on reading times, because at a fixed level of cloze, variation in M would be confounded with variation in true predictability. This result would indicate not that M affected language processing, but only that it allowed us to better estimate predictability by correcting the biased cloze estimates.

The ideal solution to this problem would be to measure subjective probabilities and cloze on the same items for comparison. Sadly, this is impossible, since we have no reliable independent measure of subjective probabilities. Fortunately, several extremely large corpora have recently become available, which allow us in Experiment 1 to compare cloze distributions to true distributions of continuations in large corpora of real text (**e**).[2] Logically, the relationship between cloze and corpus distributions is determined by the arrows linking them. If none of the processes denoted by **a**→**e**, **a**→**b**, or **b**→**d** introduced distortions of any kind, then cloze and corpus distributions would be identical. Of course, it turns out that we instead find large and systematic differences. So the next question is where these distortions arise.

If they come from strategic task effects in the **b**→**d** link, then that has important methodological implications, as discussed above. But another possibility is that they arise from learning or processing biases in the **a**→**b** link — that is, biases in cloze responses might reflect actual errors or inefficiencies in language users' predictions about upcoming material. Such errors would be of great theoretical interest, but have not previously been possible to study, since you cannot measure biases when your measuring tool has a matched set of biases. And, of course, we must also consider the less interesting possibility that some portion of these differences are simply caused by biases in the sampling process (**a**→**e**) used to construct our corpus. In Experiment 2, we begin to distinguish these possibilities by directly comparing cloze and corpus probability in their ability to explain self-paced reading times. While more work remains to fully isolate these effects, we find preliminary evidence to at least rule out the possibility that strategic effects in the cloze task are the *sole* source of these biases.

[2]Note that this is quite different from comparing cloze probabilities to probabilities generated by the computational language models (**f**) that are also sometimes used in research (Hale, 2001; McDonald & Shillcock, 2003; Demberg & Keller, 2008; Levy, 2008; Smith & Levy, 2008). Those models can use sophisticated mathematics (represented by the **e**→**f** link) to estimate predictabilities, but for computational reasons are still limited in considering a small fraction of available context — usually just one to two words, or a parse tree stripped of word identity — which makes them unrealistic models of human performance. Here, we use materials constructed so that our critical word always appears after just four words of context, and then use a corpus so large that we can simply count how often that word appears given the *full* context.

Experiment 1

Methods

Materials We selected 300 four-word sentence initial stems from the Web 1T 5-gram corpus (Brants & Franz, 2006), which was compiled from one trillion words of English web text. (By 'stem' we mean nothing more or less than four words which begin a sentence.) The messy nature of this corpus required a complex selection procedure; we summarize the most important points: Our stems were required to have occurred often enough to allow reasonable probability estimates (median count 1906, minimum count 250), to meet a minimum perplexity threshold according to a separate trigram model trained on the British National Corpus (*The British National Corpus, version 3 (BNC XML Edition)*, 2007), to induce mostly open-class word continuations ($\geq 90\%$), and to vary substantially in the range of probability for their most-likely and second-most-likely continuations. They were then screened by hand to eliminate obvious spam (any phrase used in a spam web page is repeated in many locations, which causes it to be over-counted relative to its actual usage), high-frequency stereotyped phrases (*e.g., Designated trademarks and brands...*), excessively technical usages that we judged participants were unlikely to have had much experience with (*The study protocol was...*), or web-specific usages (on the web, *If you leave the...* is usually followed by *...field blank...*, because web pages are very concerned about explaining web forms). Finally, whenever two stems were judged 'too similar' to each other (e.g., because they differed from each other only in the gender of pronouns), one of them was eliminated.

Procedure Participants performed a computerized sentence continuation (cloze) task, in which they were given each stem and asked to type one or more words which naturally continued the sentence. Spelling was corrected by hand, with computer assistance.

Participants 140 students from UC San Diego participated for course credit. All were native English speakers. 114 participated via an online web form; of these, 6 were eliminated for admitting in a post-test questionnaire that they had used Google to find continuations. The remaining 26 participants performed the identical task in a lab environment.

Results and Discussion

As the online and in-person participant groups performed similarly in all analyses reported here, we present only pooled data ($N = 134$).

When we started this project, the Web 1T corpus was the only corpus available that was large enough for our purposes, and so it was used to optimize the design of both this and

He played a key...			*After a cup of...*		
role	94%	42%	coffee	39.6%	28%
part	2.3%	3.8%	tea	39.1%	61%
			hot	3.0%	—
When she began to...			*The time needed to...*		
speak	5.2%	9.0%	complete	41%	6.7%
cry	2.5%	19%	reply	3.2%	—
work	2.2%	1.5%	finish	0.3%	10%
It usually takes the...			*In the winter and...*		
form	34%	1.5%	spring	66%	40%
shape	23%	—	early	13%	—
following	2.7%	—	summer	4.2%	32%
cake	—	8.3%	fall	2.3%	19%

Table 1: Sample continuation distributions from Experiment 1. In each case, the left column is corpus probability, and the right column is measured cloze probability. '—' denotes continuations that were never observed.

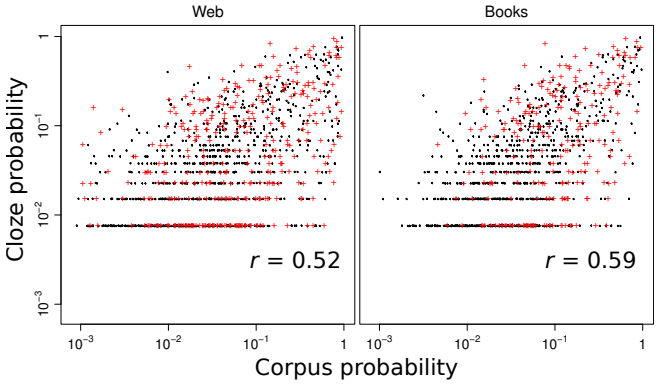

Figure 2: Cloze versus corpus probability. Each point represents a single stem/continuation pair that appeared in the corpus and was given by at least one cloze participant. (Regression analyses included responses that were given by zero participants, but they can't be shown on this log scale.) Correlations are computed in log space. Red crosses mark stem/continuation pairs that were selected for use as stimuli Experiment 2.

the next experiment. However, a second large corpus has recently become available, derived from scanned books rather than the raw web (Michel et al., 2011; we use the subset containing American English since 1960, consisting of ∼89 billion words). This corpus seems more representative of real usage (i.e., it has no spam), but is smaller and our experimental design is not optimized to take full advantage of it; therefore, we perform all analyses with both corpora.

Fig. 2 shows the overall relationship between cloze and corpus probability; our first result is that their correlation is only moderate. So the next question is, if corpus probability does not determine cloze, then what does?

To find out, we fit cloze responses to a log-linear model of

	Web	Books
Corpus probability	0.79	0.79
Corpus constraint ($\log_{10}$)	−0.40	−0.33
Stem frequency ($\log_{10}$)	−0.23	0.19
Familiarity	0.96	0.78
Concreteness	0.14	0.03
Imageability	−0.14	0.07
Age of acquisition	0.28	0.24
Frequency ($\log_{10}$)	0.20	0.16
Contextual diversity ($\log_{10}$)	0.72	0.31
Length	0.00	−0.11
Lexical prime probability	4.37	6.33

Table 2: Estimated coefficients from a log-linear model regressing cloze responses against corpus probability and other measures. Shaded cells are significant. Positive/green cells indicate a response preference, while negative/red cells indicate a response dispreference; e.g., cloze participants use familiar words more than would be expected given corpus probability but may avoid long words. Corpus constraint and stem frequency are entered as modulating the effect of corpus probability, rather than having an independent effect. Not much should be read into the absolute magnitude of coefficients, since different predictors are on different scales.

the form

$$P(\text{response}_{ij}|\text{stem}_i) = \frac{1}{Z_i} \times p_{ij}^{(\alpha_0+\alpha_1\text{center}(\text{StemProp1}_i)+\cdots)} \times \exp(\beta_1\text{WordProp1}_j+\cdots)$$

Here, p_{ij} indicates the corpus probability of continuation j given stem i, computed as the number of times we observed this continuation following the stem divided by the total number of times that we observed the stem. α_0 is a free parameter that measures the sensitivity of cloze to corpus probability. An α_0 of 0 would indicate no sensitivity, and a value of 1 would indicate that cloze matches corpus probability perfectly (at least, until the word-specific parameters come in to further influence matters). A value between 0 and 1 would indicate that cloze distributions are overall flatter (have higher entropy) than the corresponding corpus distribution, while a value greater than 1 would indicate that cloze distributions are more peaked (have lower entropy), as might happen if participants preferred to provide the most-probable continuation instead of probability matching. StemProp1,... are properties of the stem which might modulate the overall effect of α_0. WordProp1,... are word properties that might cause participants to give particular responses more or less often than predicted by corpus probability alone. And Z_i is a normalizing constant (not a free parameter).

Stem predictors included the corpus constraint, $\max_i P(\text{continuation}_i|\text{stem})$, the total number of times that the stem was observed in the corpus (a proxy for participants likely amount of experience with each particular

stem). Word predictors included familiarity, concreteness, imageability, and age of acquisition (from Wilson, 1988; Stadthagen-Gonzalez & Davis, 2006; Nelson, McEvoy, & Schreiber, 1998), word frequency and contextual diversity (from Brysbaert & New, 2009), word length, and a measure of interlexical priming from the stem to the target ('lexical prime probability'). This measure was computed by looking up the probability p_i that each word i in the stem would produce the continuation as a response in a free-association task (Nelson et al., 1998), and then combining these probabilities as $1-\prod_i(1-p_i)$.

We analyzed the subset of the data for which all of these norming values were available, for which the continuation was recorded in the corpus, and, for the book corpus analysis, for which the continuation was observed in the corpus with a stem frequency of >100. (This allowed the analysis of 5015 responses for the web data, and 4636 for the book data.) The model was fit by maximum-likelihood, with all predictors entered simultaneously, and significance computed with the likelihood ratio test and corrected for multiple comparisons by sequential Bonferroni. The results of this analysis are shown in Table 2.

Reassuringly, the exponent α_0 on corpus probability is significantly greater than 0 (web: $\chi^2(1)=992, p \ll 0.001$, books: $\chi^2(1)=784, p \ll 0.001$), indicating that cloze is sensitive to corpus probability, as expected. However, it is also significantly smaller than 1 (web: $\chi^2(1)=62, p \ll 0.001$, books: $\chi^2(1)=44, p \ll 0.001$), indicating that cloze distributions are systematically more variable (higher entropy, more flattened) than corpus distributions. If we interpret cloze task responses as reflecting participant predictions, then this might suggest that our participants are substantially more confused about upcoming linguistic material than would be expected of an optimal rational agent (compare Griffiths & Tenenbaum, 2006). This increase in entropy is more pronounced for contexts that are particularly constraining (web: $\chi^2(1)=36, p \ll 0.001$, books: $\chi^2(1)=29, p \ll 0.001$).

The effect of high frequency stems is more complicated. In the web corpus, these stems produce particularly high entropy cloze distributions ($\chi^2(1)=54, p \ll 0.001$); in the book corpus, they produce cloze distributions that are lower entropy and closer to the normative corpus values ($\chi^2(1)=8.6, p < 0.005$). This may indicate that on the web, high frequency phrases are ones that are contaminated by spam and other distributional oddities, but in print, high frequency phrases are ones that participants genuinely have more experience with, and that they are able to use this experience to make better predictions.

For word properties, there is evidence that participants prefer to respond to words which are familiar, concrete, have high contextual diversity, and are primed by words in the stem (e.g., this may explain the *winter and fall* responses; in ordinary usage people would usually say *fall and winter*, which makes *fall* an unlikely continuation according to the corpus; but in any case *fall* is primed by *winter*). They may avoid words which are long — perhaps because they require more effort to type — and also avoid words which are imageable or are acquired early — perhaps because in the formal context of an experiment they attempt to use more formal language.

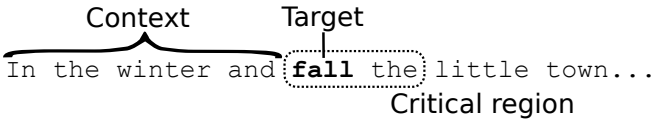

Figure 3: Sample stimulus for Experiment 2.

Experiment 2

Having established that cloze and corpus probability vary in substantial and systematic ways, our next question is whether to attribute these effects to biases in the corpus sampling (Fig. 1, **a**→**e**), to biases in language acquisition and processing (**a**→**b**), or to biases in cloze task performance (**b**→**d**). If these effects were caused by the cloze task alone, then we would expect corpus probabilities to be more closely correlated with online subjective probabilities than cloze probabilities are, and therefore corpus probabilities should outperform cloze probabilities in explaining performance in an online comprehension task. So in this experiment, we pit cloze probabilities against corpus probabilities in explaining self-paced reading times.

Methods

Materials Experiment 1 produced 2350 stem-plus-continuation pairs for which we had both cloze and corpus predictability measurements. From these we selected 179 four-word stems, then for each stem selected 2 target continuations, producing a total of 358 five-word sentence beginnings. We then completed each sentence and divided them into two 179-sentence lists, so that no participant saw any stem or any target more than once. These stems and target continuations were selected to maximize our ability to distinguish cloze and web corpus probability according to a power analysis. No fillers were used.

Procedure Participants read sentences in random order in a self-paced moving-window paradigm (Just, Carpenter, & Woolley, 1982) with a comprehension question presented after each sentence.

Participants 38 students students from UC San Diego participated for course credit. All were monolingual English speakers.

Results and Discussion

Comprehension questions All subjects performed significantly above chance on comprehension question accuracy (minimum 77%, median 92%).

Reading times We analyzed the total reading time for a region consisting of the target word plus the following word (to capture spillover; see Fig. 3). After removing sentences

with incorrect comprehension question answers or outlier reading times in the critical region (reading times $<$80 ms, $>$1500 ms, or $>$4 sd above participant-specific means, 1.9% of data removed), reading times were entered into a multilevel mixed-effects regression model using lme4 (Bates & Maechler, 2010). Our first question was whether cloze or corpus probabilities better explained reading times, so both were log-transformed (Smith & Levy, 2008 demonstrated that the empirical relationship between word predictability and reading time is in fact log-linear) and entered into the regression. In addition, for controls, we entered the log-frequency (from Brysbaert & New, 2009), word length, log-frequency/word-length interaction, and log-corpus probability for three different words: the word preceding the target, the target itself, and the word following the target. Because our stimuli were not optimized to produce corpus estimates of the probability of the word following the target, this probability was estimated using only a two-word context (i.e., an unsmoothed trigram model). As random effects, we allowed the intercept to vary by stem, and the intercept, slope of the cloze effect, and slope of the target word corpus probability effect to vary by subject (this structure selected by model comparison). Significance was assessed by assuming calculated t values were distributed as standard normal under the null hypothesis (Baayen, Davidson, & Bates, 2008).

We found that cloze was significant after controlling for corpus probability and the other factors described (web: $t = -2.87, p < 0.005$, books: $t = -2.32, p < 0.03$), but that after controlling for cloze and these other factors, corpus probability was not significant (web: $t = 1.83$, *n.s.*, books: $t = 1.13$, *n.s.*; both trends in the wrong direction). This would suggest that some of the the biases we observe in cloze probability are also present in readers' subjective probabilities — but before we conclude this, there is another possible interpretation we must consider. It might be that readers have accurate knowledge of word predictability, but that their reading times are *independently* sensitive to some of the properties listed in Table 2. In that case, cloze might outperform corpus probability simply because cloze is able to account for two factors that affect reading times, while corpus probability can account for only one.

To rule out this possibility, we re-ran the regression described above, this time adding the properties from Table 2 as additional controls; if cloze is only performing well because of its partial confounding with these other measures, then including them directly should cancel out its effect. On the contrary, however, our results were essentially unchanged in the web corpus — cloze remains significant ($t = -2.33, p < 0.02$), while corpus probability remains insignificant ($t = 1.75$, *n.s.*). For the book corpus, cloze drops to insignificance ($t = -1.6$, *n.s.*) while corpus probability remains insignificant ($t = 0.59$, *n.s.*). However, as the general trend remains the same, we suspect this is simply a consequence of our reduced statistical power when working with this corpus. We start out with corpus probability estimates for only a fraction of our data (roughly 75%), and adding these additional controls reduces our usable data still more, making this a worst case for statistical analysis. However, further data collection should resolve this issue in one way or the other.

Conclusion and Future Directions

Provisionally, at least, we can conclude that not only is cloze systematically biased relative to corpus probability, but that at least some portion of these biases are also reflected in comprehender's subjective probability estimates (i.e., arise along the **a**→**e** or **a**→**b** arrows in Fig. 1).

The next challenge for future work is to break down the different biases we have observed, and further identify their locus of effect. For instance, it could be the case that of the effects observed in Experiment 1, the age of acquisition bias arises from strategic effects in the cloze task (**b**→**c**), while simultaneously the familiarity effect is caused by biases in subjective probability estimation (**a**→**b**), and then in addition corpus sampling problems (**a**→**e**) are making our corpus estimates more noisy across the board.

Fortunately, testing such hypotheses is possible with the tools we have described. Our proposed strategy would be to use the log-linear modeling approach from Experiment 1 to estimate and then correct for different biases — that is, to estimate what cloze *would* look like if different biases didn't exist. In the situation described in the previous paragraph, we would predict that removing the age of acquisition bias should produce a measure that explains reading times even better than cloze itself does, while removing the familiarity bias should reduce our ability to explain reading times. Since these tests would be comparing cloze-based measures against each other, they reduce the possibility of artifacts caused by corpus sampling problems. However, if such problems do exist, we suspect that they should make the corpus estimates more noisy but without producing any *systematic* errors. Therefore, we can test for their presence by removing all of the known, systematic biases from cloze, and testing whether this 'unbiased' cloze continues to outperform corpus probability.

Thus, while the differences between cloze and corpus probability continue to raise worrisome questions about current methodology (it remains possible that many or most of the biases we found *are* artifactual), the way is clear to not only resolve this issue, but also shed new light onto the mechanisms underlying linguistic prediction in general.

Acknowledgments

We thank Erin Bennett, Tiffany Chiou, Megha Ram, Maria Sokolov, and Daphne Tan for assistance in collecting data. This research was partially supported by NIH Training Grant T32-DC000041 to the Center for Research in Language at UC San Diego to NJS, NSF grant 0953870 to RL, and funding from the Army Research Laboratory's Cognition & Neuroergonomics Collaborative Technology Alliance.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008, November). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Bar, M. (Ed.). (2009). Predictions in the brain: using our past to prepare for the future [theme issue]. *Phil. Trans. R. Soc. B*, *364*.

Bates, D., & Maechler, M. (2010). lme4: Linear mixed-effects models using s4 classes [Computer software manual]. Available from `http://CRAN.R-project.org/package=lme4` (R package version 0.999375-37)

Brants, T., & Franz, A. (2006). *Web 1T 5-gram version 1.1, LDC2006T13.* Philadelphia, PA: Linguistic Data Consortium.

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

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990.

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.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

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

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238.

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

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

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

Michel, J., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., et al. (2011, January). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms.* Available from `http://www.usf.edu/FreeAssociation/`

Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the thirtieth annual conference of the Cognitive Science Society* (pp. 595–600). Austin, TX: Cognitive Science Society.

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The bristol norms for age of acquisition, imageability and familiarity. *Behavior Research Methods*, *68*, 598–605.

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

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine Readable Dictionary, version 2. *Behavioural Research Methods, Instruments and Computers*, *20*(1), 6–11.