A context constructivist account of contextual diversity

Shaorong Yan, Francis Mollica, Michael K. Tanenhaus
(syan13@ur.rochester.edu | mollicaf@gmail.com | mtanenha@ur.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627 USA

Abstract

Word frequency effects have long served as an empirical and theoretical test bed for theories of language processing. A number of recent studies have suggested that Contextual Diversity (CD) is a better metric of retrieval processes than word frequency. Motivated by these findings, we sketch an active account of lexical access during sentence processing: language users store statistics about contextualized lexical representations and use lexical-contextual relations to both *construct* context and *predict* words given the context. In line with our account, we provide evidence from a frequency judgment experiment suggesting that words are not stored independently of their contexts of use. To further examine CD effects in reading, we analyzed reading times in self-paced reading and eye-tracking corpora. We demonstrate that as context is constructed, the role of CD in lexical retrieval is attenuated, reflecting a trade-off between context construction and contextualized word prediction.

Keywords: Frequency; Contextual Diversity; Predictability

Introduction

How words are acquired, stored and retrieved are fundamental questions in psycholinguistics. To probe one's mental representation of word knowledge, i.e., the mental lexicon, researchers have hypothesized and investigated many lexical properties that might influence word reading/retrieval times. Among these, the word frequency (WF) effect—more frequently encountered words are processed faster than less frequent words—is perhaps the most established finding (for review, see Adelman & Brown, 2008). Recently, a body of research has shown that contextual diversity (CD), measured as the number of unique documents in which a word appears, predicts both reaction times in metalinguistic tasks (Adelman, Brown, & Quesada, 2006) and reading times for select words embedded in sentences (Plummer, Perea, & Rayner, 2014) over and above WF, raising a challenge to existing models of the mental lexicon. Inspired by these new findings, we propose a context constructive account of CD effects: Language users store fine-grained, contextualized statistical information about word distributions; this information is used to construct a discourse context and inform expectations about what words should be expected in the current context—i.e., a predictability effect.

Current accounts for CD effects are often revised versions of WF effects. At Marr (1982)'s computational level, WF reflects the probability that a word will be needed (i.e., *need probability*) (Anderson & Schooler, 1991). Retrieval processes should be optimized so that words that are more often needed are retrieved faster. At the algorithmic level, frequency effects are explained either via acquisition mechanisms, where more frequent words build stronger retrieval cues (e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996), or via models of lexical retrieval, where search is performed over frequency ordered representations (e.g., Murray & Forster, 2004). These algorithmic accounts of WF effects can be thought of as passive, meaning that frequency effects are a by-product of how the lexicon is learned/organized.

CD effects are given similar computational-level treatment. CD is argued to better reflect need probability than WF. After all, they are highly correlated quantities ($\rho = 0.98$ or $\tau = 0.91$ for SUBTLEX values). Similar to WF, CD has been incorporated into models of word learning (for review, see Jones, Dye, & Johns, 2017). To account for CD effect, it was proposed that the predictive power between a context and a word affects word learning (Jones, Johns, & Recchia, 2012). When encountering words in new contexts, if the word is not predicted by a context, it is more strongly encoded in memory. As a result, words that appear in more diverse contexts are more strongly encoded and more easily retrieved.

Our account is motivated by one important limitation to past research: Existing accounts treat CD effects as a *passive* by-product of how words are acquired or indexed. In other words, CD is a property of word storage that is independent of the current context and task. To be fair, this limitation reflects judicious restriction of theoretical conclusions given most accounts of CD were motivated by data from metalinguistic tasks—e.g., lexical decision times, naming latency for words presented in isolation. The few studies with sentence contexts (Chen, Huang, et al., 2017; Plummer et al., 2014) have used carefully controlled, experimenter constructed contexts. While such tasks are vital to uncovering how words are represented, they often do not reflect how words, or other linguistic information, are naturally acquired and used. Our account is motivated by the hypothesis that the mental representation of words should be adapted for naturalistic language processing where rich contexts of use are the norm.

Our Account

We propose an *active* account of lexical retrieval. Our account is motivated by considerations for how distributions of words are generated (e.g., Goldwater, Griffiths, & Johnson, 2011). People do not usually walk around the world saying words at random. Instead, people use language with purpose in specific contexts. Given that context is predictive of word use, we argue that the need probability of a word should be dependent on context[1] and context should be incorporated

[1] We do not confine our notion of context to local co-occurrence statistics, e.g., reflected by cloze scores (e.g., Rayner & Well, 1996) or N-gram probabilities (e.g., Smith & Levy, 2013) but also include more global and abstract notions of contexts like entities in one's surroundings (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Hanna & Tanenhaus, 2004) or the current topic/question under discussion (e.g., Piantadosi, 2014; Roberts, 1996).For a more richly

into lexical representations. With contextualized word representations, people can form expectations about what words are likely to be encountered given the current task/context. When encountering words within a specific context, words that are more frequent within the current context should be more expected, i.e., a predictability effect. On the other hand, when probing linguistic knowledge without any context, e.g., when words are presented in isolation in metalinguistic tasks, or without identifiable discourse context, words that appear in a larger number of (distinctive) contexts should be more expected, leading to faster reaction time and higher recognition accuracy for words with higher CD, as observed by Adelman et al. (2006).

The predictive power between words and contexts is mutual and bi-directional. Contexts are informative of what words are to be expected and words are indicative of the current topic/context as in a generative model of words and contexts (e.g., Griffiths, Steyvers, & Tenenbaum, 2007) and distributional models of semantics (e.g. Landauer & Dumais, 1997). To be specific, we predict simultaneous access to lexical identity and context of use. Through iterations of interactions between words and contexts, information about what message is being conveyed is updated and comprehension is achieved (for discussion, see Kuperberg & Jaeger, 2016). Based on what contexts a word is associated with, language users can form expectations about what is the current context and gradually construct a context as each word comes in; the constructed context allows them to make predictions about what words are to come and adjust the constructed context accordingly if such predictions are not met.

Our proposed account differs from traditional accounts of CD/WF effects in two aspects. First, we propose that WF/CD/predictability effects all result from an *active* word retrieval process triggered by the current task and context, instead of through a *passive* generic lexical access process. To be specific, WF, CD, and predictability effects reflect *active* use of stored knowledge to inform expectations about to-be-encountered words. Second, our account is "context-centric" rather than "word-centric", i.e., we argue that WF, CD, and predictability effects all result from the interaction between the current context and the contexts where a word has been encountered, rather than solely coming from the properties of the word itself. In other words, it is not the need probability of a word but the need probability of a word in context that language users must store.

The probability of a word is then a marginalization:

$$P(w) = \sum^{C} P(w|c) \cdot P(c) \tag{1}$$

Under this computational account, retrieval models intimately link words and contexts. Processing level accounts can now explain CD effects and predictability effects as the same thing. If CD is really a more accurate read on need probability than WF, we have now unified these three independent effects into one explanatory framework.

articulated notion of context, see (Clark, 1996).

In the remainder of the paper, we test our account by examining two predictions derived from it. In Experiment 1, we examine whether language users possess and use fine-grained, contextualized statistics of word distributions. In Experiments 2 & 3, we use corpora of reading time data to examine whether the CD effect decreases as context is gradually constructed. Taken together, these experiments constitute the first step towards building an active model of lexical retrieval.

Experiment 1[2]

To provide evidence that lexical representations are dependent on contexts, we adopt a binary 2AFC frequency judgment task (Landauer, 1986). As shown in Equation 1, word frequency is a function, i.e., marginalization, of contextualized word representations. Under our account, context should mediate the fidelity with which people retrieve word frequencies. The marginalizing in Equation 1 is costly if there are many contexts. As a result, people should approximate word frequency when the context is unknown. In this case, the search for contexts should serve as an anchoring bias resulting in less accurate frequency comparison judgments (Lieder, Griffiths, Huys, & Goodman, 2018). At the same time, if the words occur in the same contexts, direct comparison between their frequencies are possible, as there will be reduced bias due to search. Therefore, we expect judgments for words that are likely to occur in similar contexts to be more accurate than for words that are likely to occur in dissimilar contexts. On the other hand, if contexts are independently stored from words, there is no reason to expect any difference in accuracy when judging word frequencies across different contexts versus judging word frequencies within the same contexts. In Experiment 1, we manipulate the likelihood of words occurring in similar contexts by comparing words that come from the same or different semantic category.

Materials[3]

We sampled words from the lexical database SUBTLEX (Brysbaert & New, 2009) in 10 bins of varying log frequency. We removed words below the bottom 30'th percentile (frequency count of 1) and words above the upper 99'th percentile in word frequency in order to study the intermediate-frequency majority of the lexicon. As a proxy for context, we selected three semantic categories to group our materials: animals, clothing and food[4]. For each category, we chose two sets of words spanning the 10 frequency bins, resulting in six sets of words. We chose to use three semantic categories with two sets of words from each category in an attempt to reduce category or word effects. Twelve lists of words were constructed by pairwise combining the six word sets with the constraint that each list must span two semantic categories.

[2]Pre-registered: https://aspredicted.org/blind.php?x=ds7c7j

[3]All code and data are available at: github.com/mollicaf/ContextDiversity

[4]Food and animals were selected to be non-overlapping.

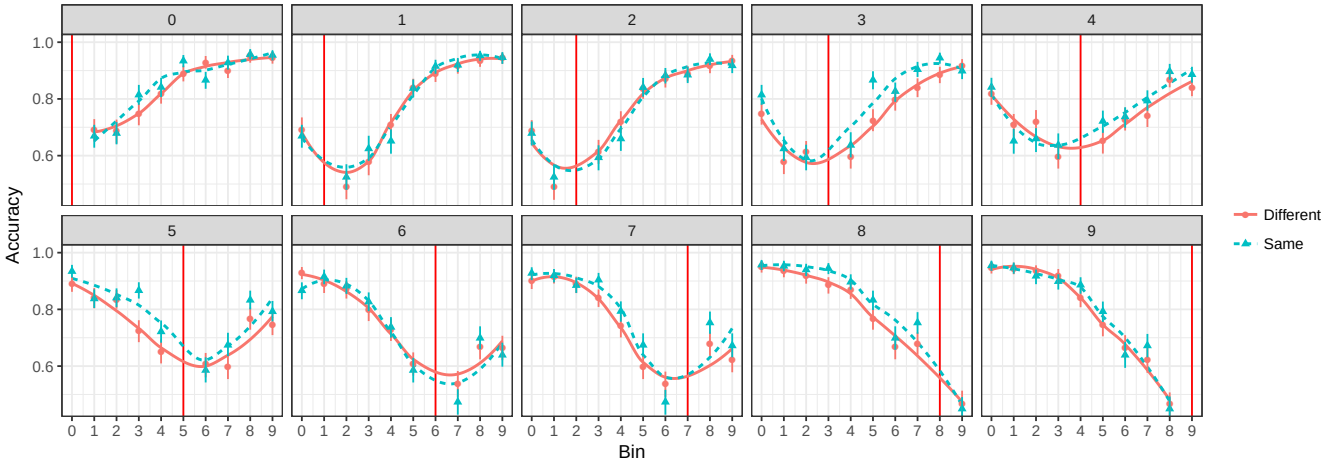

Figure 1: Average frequency discrimination accuracy as a function of log word frequency bin faceted by log reference word frequency bin. Vertical red lines denote within bin comparison. Line ranges reflect 95% bootstrapped confidence intervals.

Within each list, the words in the item set were pairwise enumerated to construct 190 frequency judgment trials. The experiment was conducted on Amazon Mechanical Turk using psiTurk (Gureckis et al., 2016).

Participants

Two hundred and forty two participants were recruited via Amazon Mechanical Turk and paid $2.50 for their participation which lasted approximately fifteen minutes. Nine participants were excluded from analyses for accuracy at chance.

Procedure

Participants were randomly assigned to one experimental list. Participants were asked to make a two-alternative forced choice to decide which of two words is more frequent. Participants were instructed to respond as quickly and accurately as possible. They were told that they should not look up word frequencies as we are interested in their first impression. On each trial, two words appeared on clickable buttons located on opposite sides of the computer screen. Participants were visually prompted, "Which word is more frequent?" When they clicked on their response, the next trial began. The trial order and the presentation order of words were randomized for each participant. Participants each completed 190 trials.

Results & Discussion

If lexical representations are contextualized, accuracy for judgments for words that are likely to occur within the same context should be greater than accuracy for judgments across contexts. We analyzed our data using mixed effect logistic regression with baseline fixed effects—the difference in frequency bin between the two words, max frequency bin of the comparison and an interaction between max bin of the comparison and difference in frequency bin, and random effects for subject and list–i.e., the maximally converging random effect structure. The baseline model controls for the intuitions that large differences in frequency are easier to discriminate, and the influence of the frequency gap may vary over the range of frequencies. For example, discriminating two low frequency words may require a larger gap than is necessary for discriminating two high frequency words. Consistent with our preregistration, accuracy for same category comparisons was greater than accuracy for cross-category comparisons ($\beta = 0.14$, $z = 5.53$, $p < 0.05$).

Participants' accuracy in answering is shown in Figure 1. The i'th subplot shows participants' accuracy (y-axis) in distinguishing the i'th bin from each other j'th bin, with the vertical line corresponding to $i = j$. This shows, for instance, that people are poor at distinguishing very close i and j (near the vertical line), as should be expected. The resolution with which participants store statistical information about word usage is also reflected in the shape of these accuracy curves. If participants store fine grained statistical information, accuracy should decreased sharply around the vertical lines and remain relatively high for all other bins. On the other hand, shallow dips around the vertical line spanning many adjacent frequency bins is indicative of low resolution representations of word statistics. The sharper increase in accuracy for contrasts within the same context as opposed to across contexts is consistent with our account and a good indication that our effect, albeit small, is signal driven.

There are two limitations from this experiment that we recommend be addressed by future work. Given the relatively high accuracy for comparisons distal to the vertical lines in Figure 1, future work should specifically target the frequency comparisons nearest the vertical lines rather than span the full range of frequency. This design consideration would permit more trials where the effect is largest, providing greater statistical power and better generalizability across items. Second, the simplification assumption that words from the same semantic category are more likely to occur in similar contexts needs to be validated, e.g., using distributional representation of word semantics (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013).

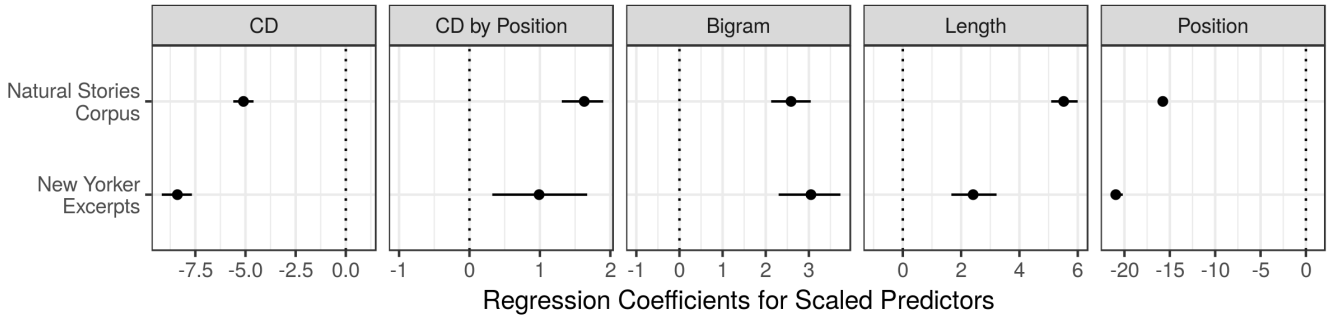

Figure 2: Fixed effect coefficients from linear mixed effect models with random intercepts for participant and story. Error bars reflect 95% bootstrapped confidence intervals. Qualitative results hold throughout five-fold cross validation.

Experiment 2

Context identification is a dynamic process; context is constructed, evolves rapidly and is even forgotten (see Futrell & Levy, 2017). If context is intimately linked to our distribution of words, predicting what context we are in is vital to language understanding. Knowing the context updates our expectations of future utterances. When the context is unknown, the CD of words, or the predictability of the word's context, provides updates for the context, much like the predictability of words provides updates for the sentence.

In non-specific contexts, the best guesses about what words to expect are the words that are more likely to appear in different contexts; as context is constructed, one can use the statistical information about the distribution of words within that context to form expectations for future words. Taking a global approach in Experiment 2, we argue that context is constructed over the course of reading a coherent discourse or story. Therefore, we predict that CD effects will be initially observed in reading times of natural text, and that the CD effect will attenuate as the context gets constructed.

Materials

We analyzed two corpora of self-paced reading (SPR) data: Mollica and Piantadosi (2017) and Futrell et al. (2017). Mollica and Piantadosi (2017) had participants ($N = 64$) read four $\sim$ 1000 word excerpts from *New Yorker* articles in different presentation conditions. For our analysis, we focus on the single-word, centrally-presented self-paced reading and masked self-paced reading conditions. Each participant read two stories each (one masked, one un-masked) in the lab. Futrell et al. (2017) had ($N = 181$) participants read ten 1000 word excerpts from natural texts slightly altered to include low frequency words and syntactic constructions. Participants each read between one and ten stories online via Amazon Mech Turk.

For our baseline model for reading time, we collected word length, position in text, and bigram surprisal for every word in our stories[5]. We used the bigram surprisal provided with the datasets, originally calculated from Google N-grams. For our main predictors, CD and WF, we used the statistics from SUBTLEX (Brysbaert & New, 2009). Following convention in the field, we log transformed both WF and CD for our analyses.

[5]The baseline model for Mollica and Piantadosi (2017) also contained a fixed effect for Mask type.

Table 1: Model Comparison for Natural Texts

	Add CD (χ^2)	*Add WF* (χ^2)
Natural Stories Corpus	461.18	131.84
New Yorker Excerpts	70.53	3.49

Results & Discussion

Before we analyze our main predictions, we first confirmed that CD is a better predictor than WF for reading times of continuous text. Following the methods in the literature (e.g., Adelman et al., 2006), we fit for each corpus, three linear mixed effect models predicting RT with baseline fixed effects for word length, bigram surprisal, and position in story and random intercepts for subject and story. One model includes WF as a fixed effect; one includes CD as fixed effect and the other includes both WF and CD. We individually compare the model including CD and WF to the model including both and replicate the findings of meta-linguistic tasks: adding CD to a model with only WF better predicts reading time data than adding WF to a model with CD for both corpora (Table 1).

To explore whether CD effects attenuate with the construction of discourse context, we analyzed the interaction between CD and word position in two self-paced reading datasets of natural stories. We fit a mixed effect linear regression with CD and CD by position fixed effects in addition to our baseline model. As predicted by our hypothesis (see Figure 2), we found a negative regression coefficient for CD, suggesting faster reading times for more contextually diverse words, and a positive regression coefficient for the CD by position interaction, suggesting that the CD effect attenuates as one reads further into a story—i.e., as context is constructed.

There is one main limitation in our analyses of these SPR corpora. Over time, participants read words faster. As a result, the variance in reading times and our ability to detect an effect "shrinks" over time. In this case, the negative interaction term that we observe between CD and position might be influenced or driven by shrinkage. We have attempted to remedy this by including by position interactions with other baseline parameters; however, we quickly run into collinear-

ity and convergence issues which render the model uninterpretable. We adopt an alternate measure of contextual constraint in Experiment 3 to partially address these concerns.

Experiment 3

We further test our account by looking at how CD effects are influenced by more local notions of contextual constraint. As context identification is a dynamic process, local properties of a text can influence a reader's certainty about the discourse context. We analyzed a corpus of eye-tracking data (the Provo corpus Luke & Christianson, 2017), which arguably is closer to naturalistic reading than SPR. The other benefit of this corpus is that it includes cloze test data so that we can measure how specific/constraining the context is without using word order as a coarse approximation, avoiding the potential shrinkage problem in Experiment 2.

Materials & Methods

The eye-tracking data used in the analysis are from the Provo corpus (Luke & Christianson, 2017) where 84 subjects read 55 short passages with an average length of 50 words (range: $39-62$). The passages were taken from a variety of sources, including "online news articles, popular science magazines, and public-domain works of fiction" (Luke & Christianson, 2017).

We used the cloze test data in the Provo corpus to calculate the contextual constraint upon reading each word. To be specific, we calculated the entropy (**H**) for each word position from the cloze probability (p_i) of each word completion:

$$\mathbf{H} = \sum_i -p_i * \log_2 p_i \tag{2}$$

This measures the uncertainty about what is the next word given the context for each word position and reflects the extent to which the context constrains future words. High entropy reflects greater uncertainty about what words are likely to come next.

We focused on two dependent measures argued to correlated with lexical retrieval (Rayner, 1998): first fixation durations (FFD)—i.e., the duration of the first fixation on a word, and gaze durations (GD)—i.e., the sum of all fixation durations when encountering a word during first-pass reading.

We fit linear mixed effect model to both eye-tracking measures. The key predictor of interest is the interaction between log CD (based on SUBTLX) and entropy, i.e., whether the magnitude of CD effect changes as a function of contextual constraint. We include baseline fixed effects for word order (a word's position within the paragraph), sentence order (a sentence's position within the paragraph), word position (a word's position within the sentence) and word length. Random intercepts for item and subject were included for each model.

Results & Discussion

The full model results can be find in Table 2.

	FFD		GD	
	Coef.	t	Coef.	t
Intercept	200.81	*67.65	186.83	*36.43
Word Order	0.18	*3.21	0.35	*3.69
Sentence Order	0.51	0.52	−1.43	−0.83
Word Position	0.06	0.95	0.06	0.59
Word Length	1.06	*6.75	10.20	*39.07
CD	−3.19	*−7.27	−9.59	*−13.14
Entropy	1.23	*4.40	1.57	*3.39
CD * Entropy	0.51	1.62	−2.39	*−4.57

Table 2: Coefficients and t-values from linear mixed-effects models for First Fixation Durations (FFD) and Gaze Durations (GD). (*: $p < 0.05$)

First Fixation Durations. Consistent with the literature, we find effects of CD and entropy. FFDs on words with larger CD are shorter than on words with smaller CD ($t = -7.27, p < 0.05$). FFDs on words that with higher entropy are longer than words that follow a context with lower entropy ($t = -4.40, p < 0.05$). In contrast to our account, we do not find a CD by entropy interaction.

Gaze Durations. Again, we find effects of CD and entropy. GDs on words with larger CD are shorter than on words with smaller CD ($t = -13.14, p < 0.05$). GDs on words that follow a context with higher entropy are longer than words that follow a context with lower entropy ($t = 3.39, p < 0.05$). Most importantly, we do find the expected interaction between CD and entropy ($t = 4.57, p < 0.05$). The CD effect attenuates when the context is more constraining, i.e., when entropy is smaller(for consistent findings, see also Chen, Wang, Xu, & Tanenhaus, in prep). Importantly this effect is not subject to the shrinkage problem in Experiment 2.

Discussion

In this paper, we propose an *active* account of lexical retrieval in language processing. Language users store statistics about contextualized lexical representations and use lexical-contextual relations to both construct context and predict words given the context. Our account unifies WF, CD and predictability effects, framing CD and WF as proxies to the probability a word will be needed and highlighting the role of context as an important component of the generative process of word distributions. In three experiments, we provided evidence for two predictions of our account: 1) Word representation in the lexicon is context-dependent; 2) The effect of CD attenuates as context is constructed. Our work accords with the large body of literature showcasing that language users store rich, contextualized knowledge about the distribution of linguistic information, and flexibly use such information to best accommodate the current task and context.

In support of active, predictive language processing, we demonstrated that the CD effects are smaller when the context is more constraining. This can be viewed as striking a balance between CD and predictability to best approximate the need probability in the current context. To further test

our account, future work is needed to calculate the mutual information between a words and its contexts and quantify the process of context construction. We can then directly quantify how such a trade-off is reached and whether it is sensitive to the task context. Recent approaches have modeled this trade-off using mixture models (Delaney-Busch, Morgan, Lau, & Kuperberg, 2017)

Although the focus on this paper is on lexical retrieval, we note that information at different levels of the linguistic hierarchy is also highly context-sensitive. For example, language users have been shown to exhibit sensitivity to what possible syntactic structures will likely follow a word (for review, see MacDonald, Pearlmutter, & Seidenberg, 1994) or even distributions of phonological information given the syntactic context (Farmer, Christiansen, & Monaghan, 2006). This explains why the CD effect has also been found at sub-lexical level (Chen, Zhao, Huang, Yang, & Tanenhaus, 2017). Taken together, our account is not specific to lexical access but reflects a general principle of how linguistic information is represented and used.

References

Adelman, J. S., & Brown, G. D. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review*, *115*(1), 214–227.

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological science*, *2*(6), 396–408.

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*(4), 977–990.

Chen, Q., Huang, X., Bai, L., Xu, X., Yang, Y., & Tanenhaus, M. K. (2017). The effect of contextual diversity on eye movements in chinese sentence reading. *Psychonomic bulletin & review*, *24*(2), 510–518.

Chen, Q., Wang, S., Xu, Y., & Tanenhaus, M. (in prep). Topic-based contextual diversity and frequency effects in visual word recognition.

Chen, Q., Zhao, G., Huang, X., Yang, Y., & Tanenhaus, M. K. (2017). The effect of character contextual diversity on eye movements in chinese sentence reading. *Psychonomic Bulletin & Review*, *24*(6), 1971–79. doi: 10.3758/s13423-017-1278-8

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. (2017). Comprehenders rationally adapt semantic predictions to the statistics of the local environment: a bayesian model of trial-by-trial n400 amplitudes. In *Proceedings of the 39th cognitive science society* (pp. 283–288). Cognitive Science society.

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, *103*(32), 12203–08.

Futrell, R., Gibson, E., Tily, H., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2017). The natural stories corpus. *arXiv preprint arXiv:1708.05763*.

Futrell, R., & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 1, long papers* (Vol. 1, pp. 688–698).

Goldwater, S., Griffiths, T. L., & Johnson, M. (2011). Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, *12*, 2335–82.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211–244.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, *28*(1), 105–115.

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an organizing principle of the lexicon. In *Psychology of learning and motivation* (Vol. 67, pp. 239–283). Elsevier.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *66*(2), 115–124.

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

Landauer, T. K. (1986). How much do people remember? some estimates of the quantity of learned information in long-term memory. *Cognitive Science*, *10*(4), 477–493.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211-240.

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic bulletin & review*, *25*(1), 322–349.

Luke, S. G., & Christianson, K. (2017). The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 1–8.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review*, *101*(4), 676–703.

Marr, D. (1982). *Vision: A computational investigation into*. WH Freeman.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mollica, F., & Piantadosi, S. T. (2017). An incremental information theoretic buffer supports sentence processing. In *Proceedings of the 39th annual meeting of the cognitive science society* (pp. 805–810).

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, *111*(3), 721–756.

Piantadosi, S. T. (2014). Zipfs word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, *21*(5), 1112–30.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological review*, *103*(1), 56–115.

Plummer, P., Perea, M., & Rayner, K. (2014). The influence of contextual diversity on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 275–283.

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

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

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, 91–136.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 1632–34.