Replication of Mahowald et al. (2013): Speakers choose shorter words in predictive contexts

Lucy Li
Symbolic Systems Program
Stanford University
lucy3@stanford.edu

1 Introduction

Synonymy offers language users flexibility in how we encode information. For example, the words *hippo* and *hippopotamus* denote the same animal, but have different lengths, so in some cases we may prefer using one over the other (Jaeger and Buz, 2017). Mahowald et al. (2013) examined these kinds of long-short word pairs through the perspective of information theory, where speakers aim to efficiently use words to encode "bits" of information. Information theory predicts that low-information words are shorter than high-information ones (Shannon, 2001; Piantadosi et al., 2011). Mahowald et al. (2013) investigated this idea in experiments that controlled for word meaning and syntactic category but varied word length. They hypothesized that participants would choose shorter forms of words more often in predictive or supportive contexts than in neutral ones.

Mahowald et al. (2013)'s paper consisted of two parts: a corpus study that replicated Piantadosi et al. (2011) and a behavioral study. Their corpus study measured the surprisal of long and short forms of words, which is a measure of unpredictability. Words with lower surprisal hold less information. Mahowald et al. (2013) calculated surprisal using an unsmoothed trigram model from the Google corpus, where the context of each long or short word were the two words preceding it. They found that the mean surprisal of long forms of words, such as *examination*, was significant higher than that of their short forms, such as *exam*. For their behavioral study, Mahowald et al. (2013) created supportive and neutral sentence completions for pairs of long and short words. The first half of this study consisted of free-form responses, where Amazon Mechanical Turk participants could supply any word of their choosing. The long or short form of the target word was chosen 52.4% of the time in supportive contexts and 1.6% of the time in neutral contexts, demonstrating a significant difference in their constructed context types (Mahowald et al., 2013).

The critical experiment of Mahowald et al. (2013)'s behavioral study involved forced-choice sentence completions, where the same contexts were presented to participants but they could only choose either the long or short form of each word. They found a significant effect of context type on participants' responses, and their results supported the hypothesis that supportive contexts result in a higher preference for the short form of words. Later work such as Seyfarth (2014)'s study on the shorter acoustic duration of low-informativity words have reinforced the conclusion that shorter words hold less information.

This present project replicates the second portion of Mahowald et al. (2013)'s behavioral experiment, using the same setup and stimuli[1]. I also looked at several additional predictors aside from context and order, which included differences in ambiguity, frequency, character length, and syllable length. The first additional predictor was motivated by the idea that shorter forms of words can be ambiguous: for example, *vet* can mean *veterinarian* or *veteran*, and *shake* can mean *milkshake* or a shaking movement. In the case of neutral contexts, I expected that words with more ambiguous word forms would have an even higher preference for the long form. I included frequency difference as well, because Piantadosi et al. (2011) and Mahowald et al. (2013) used it as a predictor for their corpus studies of the surprisal of long and short forms. Since short words are easier to produce, they tend to be more frequent than longer ones (Zipf, 2016), and I hypothesized that the more fre-

[1]The preregistration for this experiment can be found on the Open Science Framework.

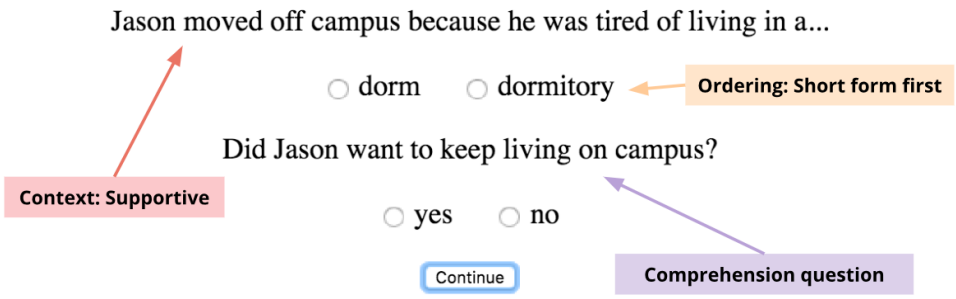

Figure 1: An example of what a participant may see for the long-short pair *dorm*/*dormitory*, if the condition was chosen for the context to be supportive and the ordering to be the short form first.

quent a form is, the more participants would prefer to choose it. Finally, I looked at word length because Mahowald et al. (2013)'s work is based on the claim that a word's length indicates its informativeness, and calculating the number of syllables or characters would transform the binary variable of short versus long into something more continuous. I would expect that smaller differences in word length between the long and short forms result in a lesser preference for the short form over the long form.

2 Experiment

2.1 Methods

Fifty Amazon Mechanical Turk workers participated in this experiment. The stimuli were the same as the original paper, which includes 40 interchangeable long-short pairs and 40 filler pairs, along with sentence completion prompts and comprehension questions[2]. Examples of filler pairs include *shoes*/*socks* and *explanation*/*clarification*, and examples of long-short pairs include *bike*/*bicycle* and *math*/*mathematics*. The comprehension questions were asked to make sure participants paid attention. One participant was excluded from analysis due to an accuracy on these questions below 75%, resulting in 49 participants. Mahowald et al. (2013) had 58 participants in their behavioral study, and this current project used a smaller number due to cost constraints.

Each participant saw all 80 sentence completions and comprehension questions, shuffled in a random order. For each long-short pair, they saw one of 4 different possible versions: long form first and neutral context, long form first and supportive context, short form first and neutral context, or short form first and supportive context. An example of one of these can be found in Figure 1. Since the condition for each word and participant was randomly decided, the split across all data points is not perfectly even but close: 26.17% were long first and neutral, 24.18% were long first and supportive, 24.08% were short first and neutral, and 25.56% were short first and supportive. The instructions asked participants to chose the word that sounded most natural[3].

Ambiguity of words was determined by their number of WordNet synsets, and frequency was based on Google Ngram English counts in the year 2008[4]. WordNet is a lexical database where synsets are the different senses for each word (Miller, 1995).

2.2 Results

I replicated Figure 2 in Mahowald et al. (2013)'s paper[5] using my data (Figure 3). In that plot, having a word above the horizontal line at 0 means its long form was much more preferred in neutral contexts than in supportive ones. If our main hypothesis about context and word choice were true, we would expect most words to be above that line.

Some words tend to have a higher preference for the long form than others, and this spread is wide (Figure 4). The variance of the random effect of word was 1.7578, and the variance of the random effect of participant was 0.8686.

In Mahowald et al. (2013), the long form was chosen in 33% of supportive contexts and 44% of neutral contexts. In my replication, this gap between context types was smaller: the long form was chosen in 41.6% of supportive contexts,

[2] All stimuli can be found here.

[3] The MTurk task, as a participant would see it.

[4] R package for querying frequencies.

[5] Original figure found here.

	Estimate	Standard Error	z value	p
Intercept	0.3118	0.2549	1.223	0.221
Context - supportive	0.3953	0.1127	3.506	4.55 x 10^{-4}
Order - short first	0.3606	0.1122	3.213	1.313 x 10^{-3}

Table 1: Results of a mixed logistic regression model, with random effects of participant and word pair and fixed effects of mean-centered context and order.

Added predictor	Estimate	Standard Error	z value	p
syllable	0.4603	0.2343	1.964	0.04949
character	0.03628	0.08653	0.419	0.674989
frequency	0.1486	0.1689	0.880	0.378813
ambiguity	0.11489	0.07007	1.640	0.101068

Table 2: The effect of each additional mean-centered predictor when they are each separately added to the main mixed-effect logistic model. These additional predictors are the value for the short form subtracted from the value for the long form.

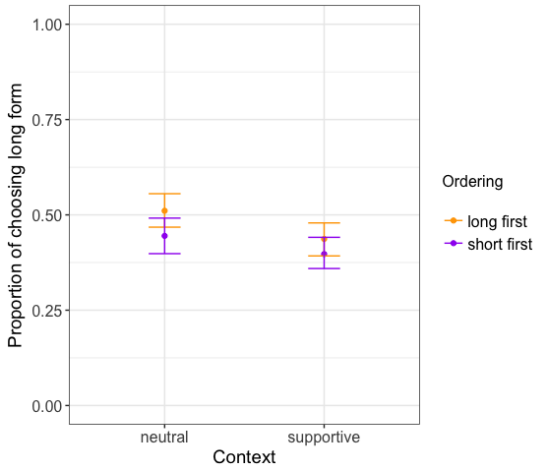

Figure 2: The proportion of times the long form of each word pair was chosen, depending on the context and answer choice ordering.

47.9% of neutral contexts, and 44.8% overall. Figure 2 shows the proportion of times the long form was chosen, given context and ordering. On average, having the long form first resulted in a slight tendency for the long form to be chosen, though the confidence intervals overlap.

The results of a mixed logistic regression model shown in Table 1 show that order and context are significant predictors of whether a participant chooses the long form or short form of a word, with random effects of word and participant. The positive slopes for both fixed effects show that a supportive context and having short first resulted in a preference for the short form of a word.

Syllable and character differences between long and short forms were highly correlated ($r = 0.61$).

When each additional predictor was individually added to the main mixed logistic regression model, only syllable difference had a slightly significant effect (Table 2, Figure 5). An ANOVA comparing the main model (AIC = 2190.5, BIC = 2218.4, deviance = 2180.5) and the model with syllable difference added (AIC = 2188.8, BIC = 2222.3, deviance = 2176.8) showed a negligible improvement with the additional predictor, with Df = 1, Pr(>Chisq) = 0.05432.

3 Discussion

These results confirm Mahowald et al. (2013)'s main hypothesis that participants would choose shorter forms of word more often in supportive contexts than in neutral ones. Their mixed-effect logistic regression showed that context had a significant effect on participants' word choice ($\beta = .75$, $z = 3.65$, $p < .001$), and there was a baseline preference for the short form, independent of context ($\beta = .77$, $z = 2.76$, $p < .01$). My slopes and intercepts were lower, but the effects of context and order were both significant. Since the p-value for the intercept was > 0.05 in Table 1, once the model took in account context, order, participant, and word, I cannot conclude that theres an overall preference for the short form of a word. The high p-value of the intercept remained even after removing context and order as predictors.

It is surprising that most of our additional predictors were not significant, and only syllable difference was slightly significant. In the case of ambiguity, this is especially surprising because *vet/veterinarian* was a high outlier in both Figure 3

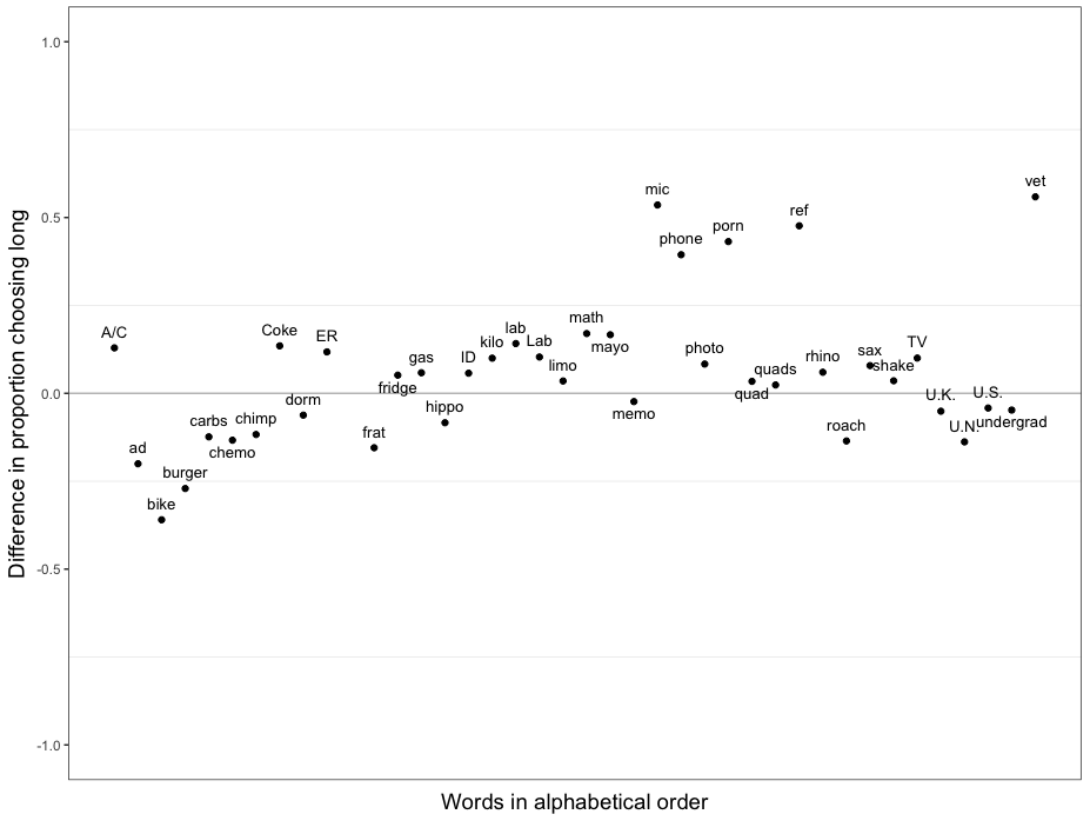

Figure 3: The words are sorted in alphabetical order and the y axis is the proportion of times the long form was chosen in neutral contexts minus the proportion of times it was chosen in supportive contexts.

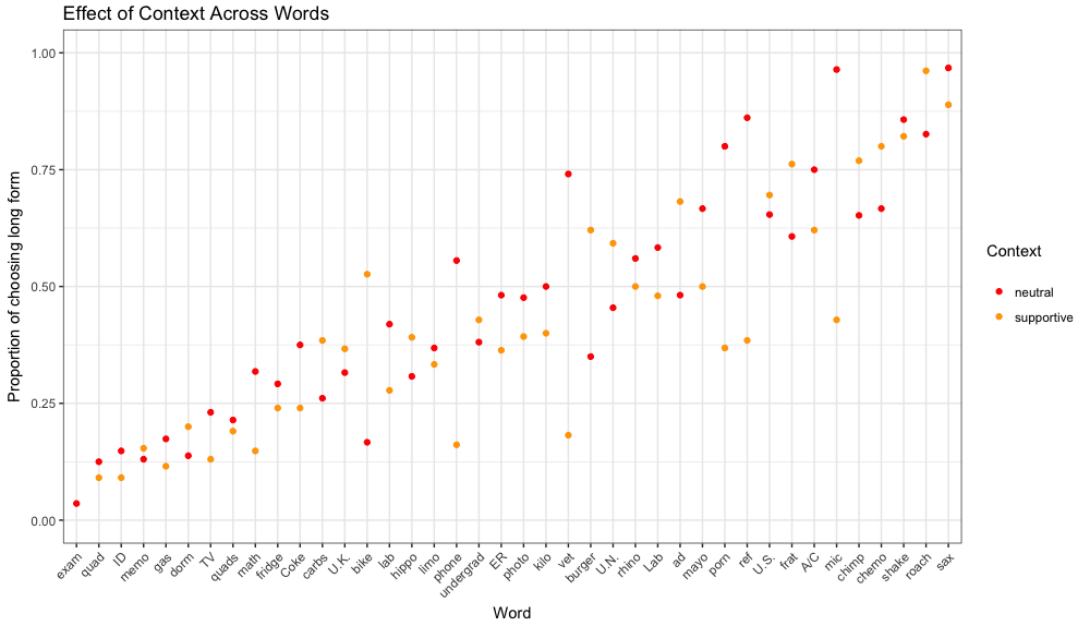

Figure 4: The proportion of times the long form was chosen across words and contexts. Note that there is no data point for *exam* in the supportive context because no participant chose the long form of that word in that context.

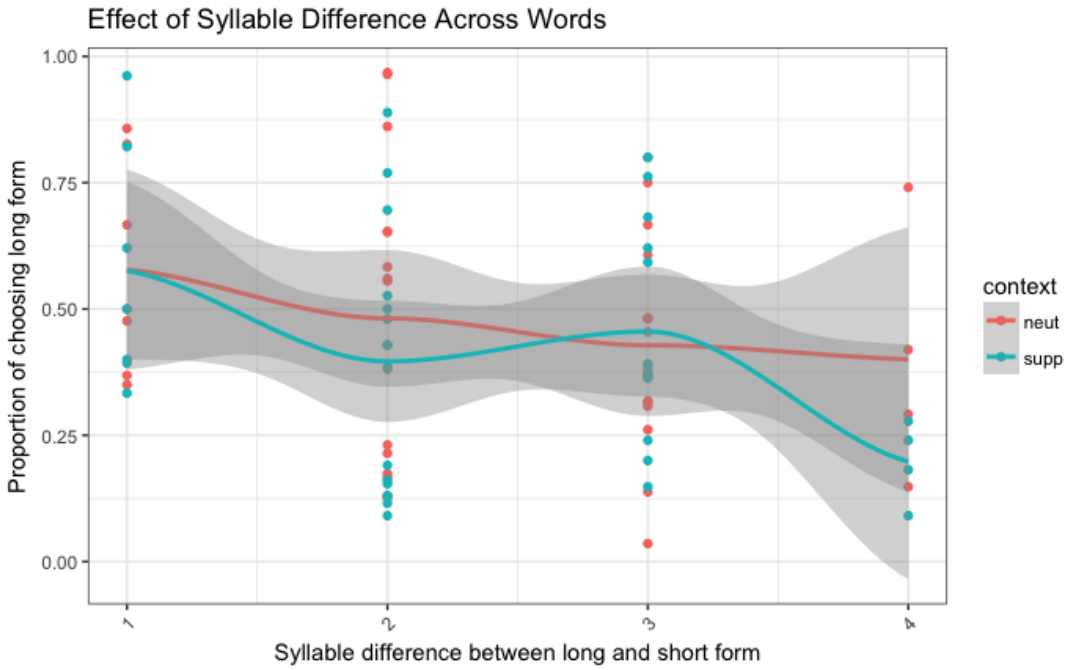

Figure 5: Syllable difference (long form - short form) versus the proportion of times participants chose the long form in their response, for different contexts.

and Mahowald et al. (2013)'s version of that plot, and I thought one possible explanation would be that a participant would chose the long form so that *vet* is not confused with *veteran*. In fact, *vet* has 6 WordNet synsets while *veterinarian* only has 1. However, it seems like there's no clear trend with the number of synsets and the effect of context. For example, *gasoline* has 1 synset while *gas* has 8, but *gas* is near the horizontal difference line at 0 in Figure 3. It could be that WordNet is noisy since it includes some very rare senses: for example, it claims that *hippo* could either mean hippopotamus or Hippo Regius, an ancient Numidian town in Africa.

Interestingly, the majority of frequency differences, with the short frequency subtracted from the long, were positive. This means long forms of words were usually more frequent in the Google Ngrams corpus than the short forms, which is the opposite of what Zipf (2016) would suggest. For example, using the `ngram` command in the Google Ngrams R package, *mayonnaise* has a frequency of 7.9 x 10^{-7}, while *mayo* has a frequency of 1.7 x 10^{-7}.

Thus, the question of how we should explain the variance across words in our results remains. What makes *vet* so different from *bike* and *burger* for them to experience almost opposite effects with context type? Perhaps this variance is just an effect of how the supportive and neutral contexts were written, and if we reran the experiment with new supportive and neutral contexts, the values would shift around. Still, we saw that some word pairs such as *exam*/*examination* and *quad*/*quadrangle* have a generally low preference for their long forms for both context types while *roach*/*cockroach* and *sax*/*saxophone* have a higher general preference (Figure 4).

One possible limitation to this work is that the long-short pairs were manually chosen by Mahowald et al. (2013), and the criteria for this filtering step was based on whether the two forms are used the same way by English speakers. It may be possible that some of these pairs are not actually semantically aligned. For example, participants rarely chose *quadrangle* in both neutral and supportive contexts, which could suggest that its meaning may be very different from *quad*. I also wondered about the inclusion of acronyms such as *U.K.*, *U.N.*, and *U.S.*, and whether there is something different with how speakers use acronyms versus shortened words. Finally, what kind of long-short word pairs would not satisfy the inclusion criteria set by the original paper? For example, would *CD* and *compact disc* not qualify since the latter is so rare? One possible way to

see how well the meanings of the long and short forms match would be to calculate the cosine similarity between their word embeddings (Mikolov et al., 2013).

In conclusion, more supportive contexts significantly increased the preference for the short form of a word, replicating Mahowald et al. (2013). Having a higher syllable difference between the long and short forms may also predict a higher short form preference as well, while character length, frequency, and ambiguity are not significant predictors.

References

Jaeger, T. F. and Buz, E. (2017). Signal reduction and linguistic encoding. *The handbook of psycholinguistics*, pages 38–81.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., and Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

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

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1):140–155.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

Zipf, G. K. (2016). *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.