9.19: Computational Psycholinguistics, Pset 7
due 8 December 2023

29 November 2023

Note: there will be a third problem added to this pset soon; please keep an eye out for a class announcement when the problem is ready.

Uniform Information Density and Availability-Based Production

The theory of Uniform Information Density (Levy & Jaeger, 2007) proposes that:

(I) An utterance is communicatively optimal if each of its parts is equally surprising given what precedes it. If we apply this idea at a level where what is meant by a "part" of an utterance is a *word* w_i, for example, then an utterance of a sentence $w_1 \dots w_n$ in context C is communicatively optimal if the surprisals $\log \frac{1}{P(w_i|w_{1\dots i-1},C)}$ are the same for all $i \in 1, \dots, n$.

(II) Speakers use the optionality afforded to them by their language to optimize their utterances according to principle (I) wherever possible.

Let's look at how this applies to a sentence like the following:

A. How big is the family you cook for?

Here, the phrase *you cook for* is a relative clause that modifies *family*. From a left-to-right incremental-processing point of view, we can break down the new information that its first word, *you*, conveys into two pieces:

1. The fact that a relative clause (RC) has started
2. Part of the contents of the relative clause (namely, its subject).

Because in this context, it is pretty much obligatory that what follows *you* is the continuation of a relative clause, not some other kind of grammatical object, we can write $P(\text{you}, \text{RC} \mid \text{How big is the family…}) \approx P(\text{you} \mid \text{How big is the family…})$, and we can use the chain rule decomposition to rewrite

$$\begin{aligned} P(\text{you}|\text{How big is the family…}) &\approx P(\text{you}, \text{RC}|\text{How big is the family…}) \\ &= P(\text{you}|\text{How big is the family…}, \text{RC}) P(\text{RC}|\text{How big is the family…}) \end{aligned}$$

so:

$$\log \frac{1}{P(\text{you}|\text{How big is the family…})} \approx \log \frac{1}{P(\text{you}|\text{How big is the family…}, \text{RC})} + \log \frac{1}{P(\text{RC}|\text{How big is the family…})}$$

The second term on the right-hand side of this last equation corresponds to the surprisal associated with piece of information (1); the first term on the right-hand side corresponds to the surprisal associated with piece of information (2).

Now, one feature of sentences like A is that the word *that* can optionally be inserted at the start of the relative clause:

B. How big is the family *that* you cook for?

Question: intuitively, what effect does inserting *that* have on where these two pieces of information are conveyed in the sentence? Are they both still conveyed at the word *you*?
Question: Suppose that a speaker follows principle (II). What effects would the following situations have on the speaker's preference for using vs. omitting *that* at the start of an RC, and why?

- The preceding context makes it highly likely that an RC will come next—e.g., *My manager directed me to do only the things…*
- The first word of the RC is very rare and unpredictable, e.g., *The ideas (that) zealots espouse are usually poorly conceived.* [A *zealot* is a person who is fanatically committed to something.]

Now, let's look at how this applies to Mandarin classifier choice. Mandarin is one of many languages with NUMERAL CLASSIFIERS. Numeral classifiers are a special grammatical category that, in order to modify COUNT NOUNS (nouns like *person* or *chair* that pick out individuated objects, as opposed to MASS NOUNS like *water* or *mud* that pick out substances) with numerals, must be used together with the numeral. In Mandarin, the classifier occurs immediately after the numeral, and the whole numeral+classifier combination occurs before the noun. There is a GENERAL classifier, 个, which can be used in combination with nearly any noun, and there are about 100 SPECIFIC classifiers, which are consistent with narrower sets of possible nouns, generally corresponding to coherent semantic categories. The examples below in (1), for example, are synonymous but use different classifiers: (1-a) uses the general classifier, whereas the (1-b) uses the specific classifier 台, which more or less requires that the modified noun refer to some kind of machine.

(1) a. 我买了 三 个 电脑
I bought three CL.general computer
"I bought three computers."
b. 我买了 三 台 电脑
I bought three CL.machine computer
"I bought three computers."

Question: in cases where either the general classifier or a specific classifier can be used for a particular numeral-modified noun, what kind of predictions are made by Uniform Information Density regarding speaker choice of classifier, and why? How do they compare with the predictions for *that*-omission?

Finally, we'll move on to the predictions of Availability-Based Production for the cases of English *that*-omission and Mandarin classifier choice. **Question:** Zhan and Levy (2018) argued that the patterns of the predictions of Availability-Based Production versus Uniform Information Density the same for *that*-omission, but may be different for Mandarin classifiers. What are those predictions, and what are the key assumptions required to make those predictions?

Unsupervised Word Segmentation

The Goldwater et al. (2006, 2007, 2009) unigram model involves the following parameters (collectively termed θ):

h	The parameter defining the geometric probability distribution over utterance length $P(L)$ (so that $P(L) = (1-h)^{L-1}h$)[1]
α	The concentration parameter defining the probability of the next word being novel (see GGJ 2007, page 5)
$p_\#$	The parameter defining the geometric probability distribution over word length
V	The number of phonemes in the language

The utterance `ba.di.ba` has the likelihood

$$P(\texttt{ba.di.ba}|\theta) = \overbrace{(1-h)^2 h}^{\text{P(L=3)}} \overbrace{\frac{\alpha}{\alpha}\frac{\alpha}{1+\alpha}\frac{2}{2+\alpha}}^{P(\text{new,new,old|L})} \overbrace{(1-p_\#)p_\#\frac{1}{V^2}}^{P(w_1|\text{new})} \overbrace{(1-p_\#)p_\#\frac{1}{V^2}}^{P(w_2|\text{new})} \overbrace{\frac{1}{2}}^{P(w_3|\text{old})}$$

(Note that this ignores the possibility of having multiple distinct lexical entries with the same phonemic form `ba`—this oversimplification is OK for the purposes of this homework.)

1. Calculate the likelihood of the utterance `ba.diba` and use it to calculate the likelihood ratio

[1] Actually, the GGJ model also puts a probability distribution over h, but we will ignore this detail here.

$$\frac{P(\texttt{ba.diba})}{P(\texttt{ba.di.ba})}$$

(we did this in class). You can think of this likelihood ratio as a posterior belief ratio for two alternative lexicons—`{ba,di}` and `{ba,diba}`.

2. Imagine that the (unsegmented) utterance were extended to become `badibaba`. What are the likelihoods of the segmented utterances `ba.di.ba.ba` and `ba.diba.ba`? What is the effect of adding this additional `ba` on the likelihood ratio for the two lexicons described in problem 1?

References

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation, In *Proceedings of coling/acl*.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2007). Distributional cues to word segmentation: Context is important, In *Proceedings of the 31st Boston University conference on language development*.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction, In *Proceedings of the 20th conference on Neural Information Processing Systems (NIPS)*.

Zhan, M., & Levy, R. (2018). Comparing theories of speaker choice using a model of classifier production in Mandarin Chinese [**Marr Prize for Best Student Paper**].