In class exercises: binomials and logistic regression

9.19(0) Fall 2023, Instructor: Roger Levy

30 October 2023

Binomial ordering preferences: corpus frequency and subjective judgments

For studying binomial and other short multi-word expressions, n-gram datasets are an incredible resource. One of the best is derived from the Google Books project. There is a very useful web-based interface for this: https://books.google.com/ngrams/.

As a reminder, here are the main constraints influencing binomial ordering constraints that were covered in the prerecorded lecture (and keep in mind that these are not necessarily a perfectly correct inventory, but they constitute a hypothesis that seems to have fairly good empirical coverage):

- Iconic/scalar sequencing (e.g., *open and read*)
- Perceptual markedness (e.g., *good and bad*)
- Power (e.g., *clergymen and parishioners*)
- Formal markedness (e.g., *changing and improving*)
- No final stress (e.g., *abused and neglected*)
- Frequency (e.g., *smile and wink*)
- Length (e.g., *tense and irritable*)

Task: think of binomials that seem interesting to you from the point of view of these constraints, and introspect on what you think the ordering preference is, and how strong. Then check the empirical ordering preference using the n-grams search interface above. Iterate until you've come up with one or two binomials that you think are interesting. At the end of this exercise, we'll compare notes and solicit class-wide ordering-preference judgments to see how closely they match corpus frequencies.

Binomial ordering reversals

Goldberg and Lee (2021) noted that the ordering preference for *aunts and uncles*/*uncles and aunts* shows an interesting diachronic pattern: the ordering preference reversed late in the 20th century. Explore this pattern in related binomials. (Some other cases of known ordering reversals are *salt and pepper* and *nuts and bolts*; see Mollin, 2014).

Bayesian inference for logistic regression

The impact of semantic constraints for binomial ordering are, of course, potentially dependent on context. Imagine for example that you enter a department store, ask an employee where you can find jackets, and they point in a general direction and say "over there, just after the sweaters and shirts". You might draw an inference about the relative perceptual prominence of sweaters vs. shirts in that part of the department store. What is that inference? Why would you draw it? Formalize this using logistic regression and Bayesian inference. **Hint:** you can use dichotomous random variables to represent relative perceptual prominence (e.g., "sweaters more prominent than shirts" vs. "shirts more prominent than sweaters") and the formally-based ordering constraints. You will want to put a prior distribution on the relative perceptual prominence random variable; think about what that prior represents. Recall that the formula for logistic regression with predictors $\{X_i\}$ and weights $\{\beta_i\}$ is:

$$\eta = \sum_i X_i \beta_i$$

$$P(\text{"success"}) \propto \frac{e^\eta}{1 + e^\eta}$$

Follow-up question: Suppose that what the employee said is "over there, just after the sweaters and trenchcoats". Normatively speaking, would the inference about the relative perceptual prominence of sweaters vs. trenchcoats be the same as that for sweaters vs. shirts in the previous example, or different? Why?

References

Goldberg, A. E., & Lee, C. (2021). Accessibility and historical change: An emergent cluster led uncles and aunts to become aunts and uncles. *Frontiers in Psychology*, *12*, 1418.

Mollin, S. (2014). *The (ir)reversibility of English binomials: Corpus, constraints, developments* (Vol. 64). John Benjamins Publishing Company.