A computational approach to detecting the envelope of variation

Isaac L. Bleaman and Rhea Kommerell (University of California, Berkeley)

Variationist sociolinguistic methodology is grounded in the principle of accountability (Labov 1972:72; Tagliamonte 2006:12–3), which compels the researcher to enumerate all the contexts in which a variable occurs or fails to occur (where one variant is used categorically or where the choice is neutralized). The process of defining the envelope of variation and determining which tokens "count" for analysis is notoriously time- and labor-intensive (Labov 1978:6). Moreover, although the variationist enterprise rejects the use of grammaticality/acceptability intuitions as data (Bayley 2013:89), researchers routinely rely on such intuitions when *selecting* data—especially in studies of morphosyntactic, lexical, and discourse variables.

In this paper, we demonstrate the usability of pre-trained computational language models to automatically identify tokens of sociolinguistic variables in raw text. We focus on two English-language variables from different linguistic domains: intensifier choice (lexical; e.g., *he is* {*very*, *really*, *so*} *cute*) and complementizer selection (grammatical; e.g., *they thought* {*that*, Ø} *she understood*). These variables exemplify different challenges for automatically detecting the envelope of variation: Intensifier variants are one-word strings, but basic search techniques cannot distinguish intensifier from non-intensifier usages (e.g., exclusions such as *she's the* {*very*, **really*, **so*} *person I had in mind*). Complementizer selection involves one variant that is overt and another that is phonetically null; the overt variant also appears in non-complementizer contexts (e.g., determiner or relativizer *that*), and the null variant necessarily eludes most search methods.

We employed BERT (Devlin et al. 2019) to train classifiers to predict whether sentences in raw text fall within or beyond the envelope of variation for each variable. The classifiers were trained and evaluated using manually annotated data. We adapted the dataset from Tagliamonte & Roberts's (2005) study of intensifiers in episodes of the American sitcom *Friends* to compile a list of sentences containing the words *very*, *really*, or *so* in both intensifier and non-intensifier contexts. We used the Penn Treebank to obtain sentences containing an overt complementizer, a null complementizer, or no complementizer. For each variable, classifier models were trained on random samples of different sizes in order to compare their performance; for complementizers, separate classifiers were trained for the overt and null variants (though these were combined during evaluation).

Our findings show that computational language models, like BERT, can dramatically reduce the burden of combing through raw language data in search of tokens of a variable—including when the surface forms are highly polysemous or phonetically null. Very little hand-annotated training data is required to achieve relatively high accuracy. Precision is somewhat lower than recall, but this is not crucial for our methodological purposes because it is much easier to remove false positives than it is to recover false negatives. Furthermore, by manually inspecting the sentences that receive high scores (indicating prototypical examples of the variable), low scores (likely exclusions), and intermediate scores around 0.5 (tricky edge cases), researchers can identify patterns that should be written into the description of the variable context for further study.

References

Bayley, Robert. 2013. The quantitative paradigm. In J. K. Chambers & Natalie Schilling (eds.), *The handbook of language variation and change*, chap. 4, 85–107. Malden, MA: Wiley-Blackwell.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL 2019*, 4171–4186. Minneapolis: Association for Computational Linguistics. doi:10.18653/v1/N19-1423.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. 1978. Where does the linguistic variable stop? A response to Beatriz Lavandera. In Richard Bauman & Joel Sherzer (eds.), *Working papers in sociolinguistics*, vol. 44, 1–17. Austin, TX: Southwest Educational Development Laboratory. https://eric.ed.gov/?id=ED157378.

Tagliamonte, Sali A. 2006. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.

Tagliamonte, Sali & Chris Roberts. 2005. So weird; so cool; so innovative: The use of intensifiers in the television series *Friends*. *American Speech* 80. 280–300. doi:10.1215/00031283-80-3-280.