Is it the size, or how you use it? Comparing the effects of subject length and predictability on contraction

English auxiliary contraction (e.g., *John is* ~ *John's here*) is a frequent and conspicuous instance of linguistic variation. Recent work has demonstrated that significant predictors of contraction include the length of an auxiliary's subject in words (MacKenzie, 2013) and information content (predictability) of the contractible form (Frank & Jaeger, 2008). In this paper, we probe these sources of variation in contraction, comparing several measures of subject size and predictors capturing information content. We find that no measure of information content predicts the observed variation better than structural measures of subject length.

To model contraction in conversational speech, we analyzed 1995 tokens of the auxiliaries *has, is*, and *will* after non-pronoun subjects from the Fisher (Cieri et al., 2004), Switchboard (Godfrey et al., 1992), and Philadelphia Neighborhood (Labov & Rosenfelder, 2011) corpora. Tokens of *has* and *is* were coded as contracted if they surfaced as [z] or [s]; tokens of *will* were coded as contracted if they surfaced as [əl] (MacKenzie, 2013). Each auxiliary's subject was coded for several size-related predictors—speaking rate, orthographic word count, prosodic word count, and syllable count—and for predictability-related predictors—probability of auxiliary given preceding word, probability given following word, and frequencies of preceding and following words.

A mixed-effects logistic regression analysis including fixed effects for speaker demographics and random effects for speaker and surrounding words finds that subject length—whether measured as orthographic or phonological words or number of syllables—is significant ($p < 0.001$), but information-related and word frequency predictors are not (all $p > .4$). The observed subject length effect thus cannot be reduced to information content or predictability, but must instead be structural in nature.

We subsequently investigate the cause of the subject length effect in more detail by using residualization to examine several multicollinear predictors of length. We find that structural predictors such as number of words predict variation better than a gross phonological size measure such as number of syllables. In fact, the best predictor of subject length incorporates separate measures of content and function word counts. We conclude by noting the connections between this latter finding and the units governing production planning (e.g., Ferreira, 1991).

References

Cieri, Christopher, Miller, David, and Walker, Kevin. 2004. The Fisher corpus: A resource for the next generations of speech-to-text. In Lino, M. T., Xavier, M. F., Ferreira, F., and Silva, R., editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation.*

Ferreira, Fernanda. 1991. Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, 30:210–233.

Frank, Austin, and T. Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *The 30th Annual Meeting of the Cognitive Science Society (CogSci08)*, 939–944.

Godfrey, John J., Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 1, 517–520.

Labov, William and Rosenfelder, Ingrid. 2011. The Philadelphia Neighborhood Corpus.

MacKenzie, Laurel. 2013. Variation in English auxiliary realization: A new take on contraction. *Language Variation and Change*, 251:17–41.