Size matters: the effect of subject length on contraction

English auxiliary contraction (e.g., *John is ~ John's here*) is a frequent and conspicuous instance of linguistic variation. The fact that both grammatical (e.g., Kaisse, 1983; Labov, 1969) and extragrammatical (e.g., Frank & Jaeger, 2008) constraints on this process have been identified suggests that a speaker's use of contractions must depend on factors that span multiple levels of linguistic analysis. However, given the insufficient empirical work on this variable, little more is known about what conditions its occurrence.

In this paper, we present evidence that the "size" of an auxiliary's subject, as measured in multiple dimensions, is a strong predictor of whether contraction occurs. We compare syntactic, phonological, and phonetic measures of subject size against a measure of number of orthographic words to evaluate both structural and extragrammatical effects on contraction.

To model contraction in conversational speech, we analyzed 223 tokens of the auxiliaries *has, is*, and *will* after non-pronoun subjects from the Switchboard corpus (Godfrey et al., 1992). 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, to appear). Syntactic parses of each subject were extracted from those parsed conversations available in the Penn Treebank (Marcus et al., 1993). Each auxiliary's subject was coded for speaking rate, ratio of syntactic nodes to words, and number of orthographic words and syllables.

Separate mixed-effects regression models with a fixed effect of auxiliary identity and random effect of speaker were run using each factor in isolation, showing that all factors except speaking rate have a significant effect (all $p<.001$) on auxiliary realization. However, correlation between predictors obscures the true source of the effect. To address this issue, we used residualization to make predictors orthogonal with respect to each other, testing the unique contribution of each residualized predictor beyond number of words. None of the residualized predictors reached significance after number of words was partialed out. We also evaluated whether number of words has a unique contribution over all other predictors, finding it to still reach significance ($p=.01$).

Despite the fact that the number of orthographic words has little structural basis compared to the other measures, it is the most robust of the available measures and encodes information not present in any of the other predictors examined. We conclude by discussing whether number of words may be a proxy for some other measure.

References

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.

Kaisse, Ellen M. 1983. The syntax of auxiliary reduction in English. Language 59:93–122.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. Language 45:715–762.

MacKenzie, Laurel. To appear. English auxiliary contraction as a two-stage process: Evidence from corpus data. In Proceedings of WCCFL 29.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19:313–330.