

Inter-dependency in linguistic data from one speaker: Assessing Competing Statistical Models

Joseph Roy (*University of Ottawa*)

Tokens drawn from the same speaker are not independent (in a probabilistic sense) from one another and this statistical inter-dependence must be accounted for when assessing the statistical significance and magnitude of the effects of linguistic factors on linguistic variables. The standard logistic regression based approaches (e.g. Cedergren and Sankoff (1974) *inter alia*), however, do not account for this inter-dependence. In both instances, the estimation of the effects and significance relies on all tokens being probabilistically independent from one another. There has, however, been a debate as to whether or not this inter-dependency is a problem at all (e.g. Young and Yandell's, 1999 response to Saito, 1999). If it *is* a statistical issue, how should we control for it in our model, as several alternatives have been proposed (e.g. Saito, 1999, and Johnson, 2009). While speaker-specific models have been proposed (Saito, 1999) and employed for linguistic data (Johnson, 2009), population-averaged models have been overlooked as a solution to the inter-dependency problem. Further, random-effect models have been misapplied to draw inference on the structure of grammar for a community without properly integrating out the random-effect for each speaker (Molensbergs and Verbeke, 2005:298-301)

This study first demonstrates the potential problem with GOLDVARB¹: A case study of the Future Temporal Reference sector in Canadian English reveals that the statistical significance and magnitude of two linguistic factor groups change when we employ a model that accounts for the inter-dependency of tokens from the same speaker. Then, several types of models are compared under data simulating a number of conditions (e.g. changes in speaker sample size (10, 30, 100) and tokens per speaker (10, 30, 100)) to assess alternatives to a logistic regression approach. All of our data indicate that inter-dependency is a problem when assessing the statistical significance and magnitude of factor groups and that population-average models perform more reliably under a variety of conditions than any of the alternatives.

References

- Cedergren, Henrietta and David Sankoff. 1974. Variable Rules: Performance as a Statistical Reflection of Competence. *Language*, 50(2):333-355.
- Johnson, Daniel Ezra. 2009. Getting off the Goldvarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Language and Linguistics Compass* 3(1): 359–383.
- Molensberghs, Geert and Geert Verbeke. 2005. *Models for Discrete Longitudinal Data*. New York: Springer.
- Saito, Hidetoshi. 1999. Dependence and interaction in frequency data analysis in SLA research. *Studies in Second Language Acquisition*, 21:453-75.
- Young, Richard, and Brian Yandell. 1999. Top-down versus bottom up analysis. *Studies in Second Language Acquisition*, 21:477-88.

¹ I use GOLDVARB as a stand-in for all logistic regression based approaches. The issue I am discussing applies regardless of which programming package we use (GOLDVARB/SAS/R/SPSS, if we're implementing any form of a standard logistic regression, the inter-dependency remains an issue).