

New results from hierarchical models of the community grammar

Josef Fruehwald and Laurel MacKenzie (*University of Pennsylvania*)

The most powerful theoretical construction of modern sociolinguistics is the notion of a “community grammar.” That is, the linguistic behavior of individuals is un-interpretable unless examined vis-à-vis larger norms of language usage. It is these larger norms, and the way in which individuals situate themselves in relation to them, which are the object of sociolinguistic investigation. In this paper, we statistically model this paradigm utilizing hierarchical regression, and uncover two novel results. First, speakers deviate from the communal norm less for phonological variation than for morphological variation. Second, an abstract variable displays the same amount of speaker-level deviation from the communal norm in all environments.

Though logistic regression (Cedergren and Sankoff, 1974) has long been the statistical analysis tool of choice among sociolinguists, recent work has argued that sociolinguistic data violates the assumption of independent, identically distributed observations that logistic regression is based on, and mixed-effects, hierarchical regression has been proposed as an alternative (Johnson, 2009). Guy (2009), however, expressed concern that these hierarchical models abandon the sociolinguistic project by including random speaker effects, which, he argued, de-emphasizes community-level variation.

We argue that in fact, hierarchical regression is ideal for modeling the sociolinguistic paradigm of speaker-level deviation from a community grammar. Utilizing a variant of hierarchical regression (Markov chain Monte Carlo estimation), we analyze several morphological and phonological variables: York English was/were variation (Tagliamonte and Baayen, forthcoming), English auxiliary contraction (MacKenzie, forthcoming), English semi-weak past tense variation (Fruehwald, 2011), /l/-vocalization (Dodsworth, 2005), t/d-deletion (Fruehwald, 2008). We model variation at both the community level and the speaker level, to produce two parameters of interest: 1) the community rate of variation, 2) the degree of inter-speaker deviation from the communal norm. By more closely modeling the sociolinguistic paradigm, we uncover a novel result: speakers deviate from the communal rate more in cases of morphological variation than phonological variation. Table 1 displays the estimated “community cohesion coefficients” for these five variables, with larger coefficients indicating greater conformity of speakers to the communal pattern. These cohesion coefficients are not correlated with the rate of community-level variation, nor to the model’s confidence in its estimate of the community-level rate.

| Morphological | | | Phonological | |
|---------------|----------------|-------------|------------------|--------------|
| York was/were | Semi-weak Past | Contraction | /l/-Vocalization | t/d-Deletion |
| 4.18 | 4.48 | 5.21 | 7.09 | 23.63 |

Table 1. Community cohesion coefficient for five variables.

Our second novel result pertains to the cohesion coefficients for the contraction of had, has, and have. Separate cohesion coefficients for each auxiliary would indicate that speakers are more similar to each other in their contraction of some auxiliaries than others. However, a model goodness-of-fit comparison finds that a model with only one cohesion coefficient is preferred over a model with three (DIC difference = -3.83). We take this as evidence that speakers are situated relative to the community rate of a single abstract process of contraction, rather than multiple auxiliary-specific processes, in keeping with the proposal of MacKenzie (forthcoming). In summary, hierarchical regression allows us to investigate the relationship between the community and the individual, something not possible with ordinary logistic regression. Its use in the current paper has uncovered novel results and raised further questions concerning the situation of individuals to communal norms.