Grouping speakers and assessing speaker groups: A case study of Chinese Americans in New York City
Tyler Kendall, William Rivers and Robin Dodsworth
University of Oregon, JNCL-NCLIS, North Carolina State University

The grouping of speakers based on social characteristics has been a central part of the variationist sociolinguistic endeavor from its beginnings (Labov 1966). In a variable analysis (cf. Tagliamonte 2006), individual speakers each contribute a number of tokens and then some forms of speaker aggregation (grouping by, e.g., socioeconomic scale (SES), sex, ethnicity, community of practice orientation, etc.) are used to examine the linguistic patterns underlying the variable, and, simultaneously, to test whether these social factors stratify the use of the given variable in the community. The number of possible social factors that might be used for grouping speakers is vast. We typically are satisfied to find that some social categories we expect to be relevant for our community of interest emerge as significant in our statistical tests.

The question remains, however: How do we know when or if we have determined the most significant social factor(s) for a particular linguistic feature in our community? We typically do not ask what the best statistical results would look like, or further how likely some other, unexamined social factor(s) might outperform those of the current analysis. Horvath and Sankoff (1987) reversed the typical grouping process and used principle components analysis (PCA) to examine individuals based on their linguistic features. They then interpreted the most linguistically meaningful social differences from the linguistic groupings. Beyond this paper and a handful of others (e.g., Van de Velde et al. 1997), the use of PCA or other clustering techniques have rarely been pursued by sociolinguists, but they represent a useful opportunity for assessing the most linguistically meaningful social groups in a community (given a particular feature).

In the present paper, we examine consonant cluster reduction and plural –s absence in the English speech of 12 Chinese Americans in New York City. Asian Americans have been massively underrepresented in the sociolinguistic literature (cf. Bucholtz 2009) and thus our study represents a rare quantitative analysis of common sociolinguistic variables for this demographic. However, we focus our primary consideration on a range of grouping techniques for investigating these speakers, based on social/demographic factors (such as Chinese dialect/region-of-origin) and ideological/discourse-based factors (Emergent Theme Analysis [Strauss & Corbin 1990], and Attribute Network Analysis [ANA; Dodsworth 2005]). Basic social factors (such as Chinese background) do not arise as significant for the variables, but a number of the discourse-based, cultural factors do, primarily centered on acculturation psychology models of integrated identity (Berry 1990). We then implement a combinatorial cluster analysis, comparing statistical models of every possible arrangement of the speakers into groups, to ask which of our groupings is actually statistically best, and further, how possible it is that some other grouping would outperform those we tested. Our cluster analysis indicates that ANA most successfully differentiates the 12 individuals, and provides a specific measure of how optimal this grouping is. Unlike other studies using cluster analyses in sociolinguistics, we focus on the use of clustering as a validation technique rather than a primary means to identify the social factors.