

## Comparing Social and Geographical Explanations of Variation

Both geographical and social explanations of linguistic variation have been successful, but there have not been attempts to compare these quantitatively, which is what we attempt here. There are two difficulties that any such attempt at comparing putative explanations must face. First, we need representative material identifying social *and* geographical correlates, and second, we need a means of comparing potentially incommensurable explicanda. We overcome the first difficulty in this talk by analyzing LAMSAS pronunciation data (Kretzschmar 1994), which was collected with the goal of cataloging both social and geographic variation. Other data collections could be analyzed in the ways we suggest here, provided geographical and social coordinates are available.

The second difficulty is apparent in the many studies of social and geographical conditioning of linguistic variation in which the usual procedure is to identify some linguistic variable, and then to quantify the degree of influence of an extralinguistic variable, normally through logistic regression (VARBRULE). The difficulty in comparison arises because different linguistic variables are chosen for different studies. We overcome this difficulty by employing an aggregate measure of pronunciation difference over the entire data in the collection. The numeric measure is derived from the Levenshtein string distance metric, but modified for linguistic purposes (Heeringa 2004), and software for calculating the distance is available at [www.let.rug.nl/kleiweg/L04/](http://www.let.rug.nl/kleiweg/L04/).

The Linguistic Atlas of the Middle and South Atlantic States (LAMSAS) was designed by Kurath, who was well aware that social factors played a role in variation, and who included a question about educational level as a means of understanding social effects. We focus here on pronunciation variation in LAMSAS, in which the pronunciation of 151 words was elicited. Because field worker effects are present in LAMSAS, we compare analyses for the two most important field workers separately as well as including an analysis aggregating the samples they were (separately) responsible for. We focus on the material Lowman and McDavid collected (responsible for 71% and 25%, respectively, of the LAMSAS interviews), ignoring the 4% of the LAMSAS material collected by others.

We operationalize geography via the geographic distance between sites and social differences via the difference in educational level of the respondents. We then submit the aggregate pronunciation difference to a multiple regression analysis in R using geographic and social differences as (independent) explanatory variables. The result is that geography accounts for approximately 35% of the variance (the further two sites are from one another, the less similar their pronunciations), and that social factors are negligible.

There are many potential explanations for these findings, including a likely bias in LAMSAS toward geographic factors and the fact that the spread of geographic values is so much larger. We should also wish to examine social variation not only independently, but also separately for different areas (hierarchical design). It would therefore be foolish to attribute generality to these results, but they point the way to a research line which can answer questions about the relative importance of social and geographic influences.

## References

- Heeringa, Wilbert (2004), *Measuring Dialect Pronunciation Differences using Levenshtein Distance*, PhD thesis, Rijksuniversiteit Groningen.
- Kretzschmar, William A., ed. (1994), *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*, The University of Chicago Press, Chicago.

**Keywords** dialect geography, social dialects, quantitative dialectology, dialectometry