Reconciling Fieldworkers' Reports: Lowman vs. McDavid

Large-scale variation studies normally employ several different fieldworkers, who may leave traces of their methods in data even when they are convinced of the need to collect data in a controlled and consistent fashion: this is the phenomenon of "fieldworker isoglosses". The present paper reports on an attempt to normalize conflicting fieldwork practices in the LAMSAS data (Kretzschmar 1994) which leads to a quantitative analysis without evidence of fieldworker isoglosses.

The Linguistic Atlas of the Middle and South Atlantic States (LAMSAS) consists of 1162 interviews conducted over a period of thirty years. We focus here on the lexical variation in LAMSAS, in which 151 lexicalizations were elicited. It is probably true that lexical data are less subject to fieldworker effects than pronunciation data since phonetic transcription is also involved, but the effects in lexical studies are also problematic. Guy Lowman and Raven McDavid were responsible for 71% and 25% of the LAMSAS interviews, and our analysis focuses on reconciling differences in their techniques which at first confound statistical analysis. The differences in technique emerge from profiles of the interviews: Lowman elicited 150.1 responses/interview (with a standard deviation of SD = 25.3 responses/interview) while McDavid recorded 197.3 (SD = 76.8). Lowman encountered 15 "no response"/interview and McDavid 22.5. We use overlap measures suggested by Séguy (1971) Goebl (1982, 1984) throughout this work, and analyze results using a variety of techniques. But as Nerbonne & Kleiweg (2003) suggested, fieldworker divisions dominate the resulting, unreconciled analysis.

The fundamental idea in reconciliation is to normalize the measurements, i.e., to express distances as z-scores, $z_i = (x_i - m)/\text{SD}$, where each score is normalized according to the mean (m) and standard deviation (SD) of the respective fieldworkers. Using normalized distances, many fieldworker effects disappear from the analysis.

On the other hand, McDavid conducted interviews over a larger area than Lowman, and this could distort normalization since linguistic distances are well known to increase with geographic distance. McDavid, in fact, was responsible for the most extreme parts of the LAMSAS areas, northern New York and areas in South Carolina, Georgia and Northern Florida. Lowman's area is "sandwiched" by McDavid's. This suggests that we distinguish three areas and normalize depending on which of the six pairs of areas is involved (Lowman-Lowman, Lowman-McDavid/North, ...).

We show (i) that the apparent fieldworker glosses disappear in this step; (ii) that the resulting division into dialect areas accords with traditional scholarship; and (iii) that one aspect of the distribution of distances involving McDavid's data is nonetheless puzzling. Finally, we consider whether the distinction we draw to counteract the large geographic distances that would otherwise infect distances derived from McDavid's data might motivate a generalized linear model in which linguistic distance were expressed as a function of root geographic distance. In this step, we examine $d_{\text{ling}}(\text{site}_1, \text{site}_2)/\sqrt{d_{\text{geo}}(\text{site}_1, \text{site}_2)}$, in which "true" linguistic distance is regarded as linguistic distance corrected by the square root of geographic distance, and normalize accordingly. The general correction is justified by the dependence of lexical distance on geography: a regression analysis shows that lexical distance (1 - proportional overlap) correlates with the root of geographic distance significantly ($r = 0.62$), and we expect such a dependence for theoretical reasons, knowing that linguistic variation is influenced by the chance of social contact, in turn dependent on geography. We note, however, (iv) that the normalization based on geographically corrected data is less satisfactory.

Finally, we note questions which these procedures raise, e.g., about the relative reliability of the fieldworkers, about linking this analysis to the geographical distribution of linguistic variation, and about the potential limitations of dialectometry in face of such (common) difficulties with data sources.

References

Goebl, Hans (1982), *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*, Österreichische Akademie der Wissenschaften, Wien.

Goebl, Hans (1984), *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. 3 Vol.*, Max Niemeyer, Tübingen.

Kretzschmar, William A., ed. (1994), *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*, The University of Chicago Press, Chicago.

Nerbonne, John & Peter Kleiweg (2003), 'Lexical variation in LAMSAS', *Computers and the Humanities* **37**(4). To appear.

Séguy, Jean (1971), 'La relation entre la distance spatiale et la distance lexicale', *Revue de Linguistique Romane* **35**, 335–357.

Keywords dialectometry, lexical variation, fieldworker isogloss, correction,