The Geographic Distribution of Linguistic Variation

John Nerbonne and Wilbert Heeringa

Alfa-informatica, University of Groningen

{j.nerbonne,w.j.heeringa}@rug.nl

1 Introduction

This paper builds on techniques for measuring pronunciation differences in order to analyze large quantities of dialect atlas data. We aggregate the pronunciation differences between lists of dialect atlas entries in order to assay the difference between two varieties. Given these techniques we can measure pronunciation differences for largish numbers of varieties. In this paper we take the further step of asking about the extent to which geographic distance predicts aggregate pronunciation differences, structuring our questions around the "gravity hypothesis," proposed by Trudgill (1974), which predicts a joint influence of distance and population size on the propagation of linguistic changes.

We are able to show that, while Trudgill ought to be vindicated vis-à-vis sociolinguistic critics who suspect that social factors are dominant, the influence of geography and population size take a rather different form than he predicted.

In the course of this argument, two novel uses of linguistic evidence are proposed and defended.

2 Previous Work

Trudgill (1974) suggested that the diffusion of dialect features might obey a "gravity"-like law (also known as a hierarchical model or a cascade model), where the influence of distance is inversely proportional to its square, and population plays the role of mass, so that settlements with large populations are particularly likely to adopt each other's changes. His hypothesis has been tested several times using individual features undergoing change. The results have varied, and researchers have tended to suspect that social and political factors are much more important than geographic ones (Bailey, Wikle, Tillery, and Sand, 1993; Boberg, 2000; Horvath and Horvath, 2001).

The present paper replaces the examination of individual features with a dialectometric measure of aggregate differences, thus eschewing the focus on individual features undergoing change for an examination of pronunciation differences in a large sample. We furthermore innovate in the sort of evidence we bring to bear on the problem.

2.1 Synchronic Differences Reflect Diachronic Dynamics

Earlier discussions attempted to track ongoing changes, at least implicitly. We innovate in the discussion by interpreting present variation as evidence of earlier dynamics in linguistic variation. In particular, we postulate that linguistic differences should—in large numbers, and on average—reflect the dynamics of linguistic change. If linguistic changes spread via Trudgill's inverse square law, then, as a result of the accumulation of linguistic changes, very close sites should linguistically be particularly similar. Since we measure differences, the prediction is that closer sites are very little different. If we then plot linguistic (pronunciation) difference as function of geographic distance, we should therefore see the (positive half of) the familiar x^2 parabola.

We apply the same reasoning to the second prediction of the gravity hypothesis, viz., that population size promotes the propagation of linguistic changes, in this case the prediction that larger population centers should propagate changes more readily than smaller ones. Again, we interpret this to mean that the number of accumulated similarities should rise as a function of population, and that differences should therefore decline when we compare larger centers.

3 Experiment

We measured the pronunciation difference between segments using spectrogram differences (curve distance between the two dimensional spectrogram surfaces), and then used the resulting segmental distances within a Levenshtein sequence distance algorithm. We used the smaller of adjacent-segment distances as costs for insertions and deletions, and we normalized distance for word length. Heeringa (2004, Ch. 7) demonstrated that this combination of techniques was most consistent and most valid vis-à-vis dialect speakers' judgment of dialect pronunciation. At the same time, this technique correlates highly with many other ways of measuring pronunciation difference—so that we may be confident that our results are unlikely to be overturned by advances in measurement technique.

To carry out this measurement we chose 52 settlements from the Lower Saxony part of the *Reeks Nederlandse Dialectatlassen*. Because of our interest in comparing our results to the predictions of Trudgill's (1974) GRAVITY HYPOTHESIS, we obtained both the distances between each pair of sites and also their respective populations (see

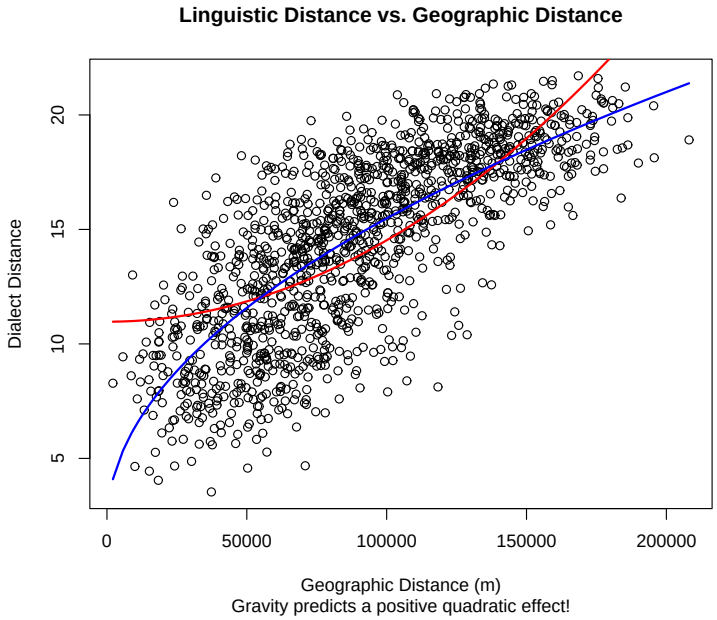

Figure 1: Geography predicts pronunciation difference in a sub-linear (or perhaps linear) fashion. Gravity (rising line) predicts a quadratic relation.

below). On the basis of this data, linguistic distances between settlements were calculated using Levenshtein distance (see above). 125 words, involving all the phonemes used in the area, were then chosen as the basis of the calculations (Heeringa, 2004, App.B).

Finally we analyzed the dependence of varietal distance on geographic distance and population size via a multiple regression analysis, examining linear, sub-linear and quadratic models for geography (the last on account of the gravity hypothesis), and only linear models of population size effects.

4 Results

The correlation of varietal distance with geographic distance turned out to be very substantial in the linear model ($r = 0.76$), and insignificantly smaller in the sub-linear model, while the quadratic model is distinctly less successful.

We find no dominant gravity-like (inverse-square) force evident in the residue of linguistic differences (see figure). The analysis furthermore indicates that the role of population, while very weak, is actually the opposite of that postulated by the gravity model (not shown here).

5 Conclusions

In order to assess the impact of geography on linguistic variation, we have emphasized an approach which utilizes *all* available data and subjects it to a rough, but provenly valid analysis. We characterize the difference between the pronunciation of two linguistic varieties via the sum of word pronunciation differences. We postulate then that the dynamics of linguistic change should be reflected in the patterns of linguistic similarities and differences, thus making a first contribution to broadening the scope of linguistic evidence.

Our most important conclusion is not surprising: geography exerts a very substantial influence on linguistic variation. But we also propose a measure of this influence, apparently for the first time, so that the degree of influence may be quantified.. Geography accounts for over 50% of the variance in the varietal data ($r^2 > 0.5$). When we compare this to non-quantitative work we note, first, that the non-quantitative work has had little to say about the degree of influence, and second, that it appears to be misled about the importance of social and political factors. We leave work contrasting the influence of these factors vis-à-vis geography to the future, however.

Since the aggregate analysis indicates that geography plays an overwhelming role, we suspect that the sociolinguistic critique of the gravity hypothesis is focusing on atypical cases, a conjecture which naturally requires further examination and further proof. The danger of focusing on atypical cases is ever present in a methodology built on examples, and with no means of assessing the relative contribution of various changes. This is our second contribution to the discussion of novel sorts of linguistic evidence.

Finally, it is interesting to speculate about the deeper import of this result. If we postulate that whatever force drives linguistic change must weaken over space (longer distances), then the gravity hypothesis reflects the view that linguistic accommodation is the primary dynamic in variation. The data demonstrate, however, that particularly close sites are especially different from one another, suggesting that the more dominant force is dialect differentiation.

References

Bailey, G., T. Wikle, J. Tillery, and L. Sand (1993). Some patterns of linguistic diffusion. Language Variation and Change, **3**(3):241–264.

Blancquaert, E. and W. Pée (1925-1982). Reeks Nederlandse Dialectatlassen. De Sikkel, Antwerpen.

Boberg, C. (2000). Geolinguistic diffusion and the U.S.-Canada border. Language Variation and Change, **12**(1):1–24.

Heeringa, W. (2004). Measuring Dialect Pronunciation Differences using Levenshtein Distance. Ph.D. thesis, Rijksuniversiteit Groningen.

Horvath, B. M. and R. J. Horvath (2001). A multilocality study of a sound change in progress: The case of /l/ vocalization in New Zealand and Australian English. Language Variation and Change, **13**(1):37–57.

Trudgill, P. (1974). Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. Language in Society, **2**:215–246.