PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2010) **365**, 3821–3828
doi:10.1098/rstb.2010.0048

Measuring the diffusion of linguistic change

John Nerbonne*

Alfa-informatica, University of Groningen, PO Box 716, 9700 AS Groningen, The Netherlands

We examine situations in which linguistic changes have probably been propagated via normal contact as opposed to via conquest, recent settlement and large-scale migration. We proceed then from two simplifying assumptions: first, that all linguistic variation is the result of either diffusion or independent innovation, and, second, that we may operationalize social contact as geographical distance. It is clear that both of these assumptions are imperfect, but they allow us to examine diffusion via the distribution of linguistic variation as a function of geographical distance. Several studies in quantitative linguistics have examined this relation, starting with Séguy (Séguy 1971 *Rev. Linguist. Romane* **35**, 335–357), and virtually all report a sublinear growth in aggregate linguistic variation as a function of geographical distance. The literature from dialectology and historical linguistics has mostly traced the diffusion of individual features, however, so that it is sensible to ask what sort of dynamic in the diffusion of individual features is compatible with Séguy's curve. We examine some simulations of diffusion in an effort to shed light on this question.

Keywords: linguistics; dialects; diffusion

1. INTRODUCTION

We summarize our key contributions in this introductory section, and provide a guide for the rest of the paper.

(a) *Key contributions*

There are two core contributions of the present paper. First, we extend arguments made by Nerbonne & Heeringa (2007) that dialectometric models provide a means for measuring linguistic variation in the aggregate and thence a means for measuring the influence of geography (and other factors) on linguistic variation. We extend these arguments by recalling Séguy's early demonstration that there was a sublinear relation between geographical distance and lexical variation, and then by examining six novel datasets, all of which confirm the relationship, which we propose dubbing SÉGUY'S CURVE.

But we also ask how we might engage the sociolinguistic literature, which has reflected profoundly on the mechanisms of diffusion, and the current exercise in measuring linguistic variation. We therefore turn secondly to a novel simulation of linguistic diffusion in which we can manipulate the strength of accommodation owing to geography. The results of the simulation suggest that the attractive force owing to gravity decreases linearly with distance, and not quadratically as the gravity model proposes.

(b) *Structure*

Section 2 reviews some of the linguistic literature on the geographical diffusion of language change, in particular, Trudgill's GRAVITY MODEL. Our point in this review is to note the need for a way of measuring diffusion and the influence geography has on it. Section 3 provides a very brief introduction to dialectometric techniques for measuring linguistic differences, and introduces SÉGUY'S CURVE of linguistic variation as a function of geography. Section 4 then reviews two recent papers exploring the gravity model using dialectometric techniques and extends their empirical base by examining six other datasets, reporting on the percentage of linguistic variation which can be explained by geography, even those which may not represent relatively stable settlements in which we can be sure that diffusion has worked 'normally'. Section 5 then introduces simulations as a tool to explore the relation between the diffusion patterns of individual lexical items and those of large aggregates, drawing the conclusion that the attractive power of geography is more probably a linear force than an inverse square relation of the sort proposed by the gravity model. Finally, §6 wraps things up a little, and also suggests why the ideas discussed here may be of more general interest for researchers interested in the determinants of linguistic diffusion.

2. THE SOCIOLINGUISTICS OF DIFFUSION

There is a substantial linguistic literature on diffusion which has documented and explained a large number of cases where individual features have spread, and most of the recent work has come from sociolinguists. We review this before turning to the aggregate analyses that have arisen in dialectometry.

(a) *The wave theory*

The *locus classicus* for linguists' discussion of diffusion is Schmidt's (1872) demonstration that there are important features that cut across the hierarchical classification of the Indo-European languages (Bloomfield 1933, pp. 312–319). Bloomfield uses Schmidt's demonstration to argue that in addition to

*j.nerbonne@rug.nl

One contribution of 14 to a Theme issue 'Cultural and linguistic diversity: evolutionary approaches'.

cases of sharp divisions between languages demonstrated by regular correspondences and modelled by family trees of relatedness, there must also be regular processes of diffusion even between the branches of the family trees, i.e. even between differentiated varieties (Bloomfield 1933, pp. 318). Bloomfield believed that speech habits were modified throughout life each time an individual entered into communication with another (Bloomfield 1933, pp. 46). He, therefore, predicted that processes of diffusion would follow the lines of communication DENSITY (Bloomfield 1933, pp. 46, 326) so that lines of dialect differentiation, reflecting processes of diffusion, should ultimately be explained by the density of communication. The idea is that diffusion is enabled and promoted by more frequent communication.

(b) *The gravity model*

Peter Trudgill's GRAVITY MODEL effectively recast Bloomfield's notion of density, focusing on distance and population sizes as predictors of the chance of communication (Trudgill 1974). In these models, inspired by social geography, the spread of linguistic innovation is always via social contact which is naturally promoted by proximity and population size. The gravity model foresees linguistic innovations not simply radiating from a centre, as they might in a pure version of the wave theory, but rather affecting larger centres first, and from there spreading to smaller ones, and so on.

In the special case of landscapes with a few larger-size cities, an innovation may spread from one large population centre directly to another intermediately sized one, often by-passing smaller, geographically intermediate sites. This is owing to the role of population size. Innovations are no longer seen as rolling over the landscape uniformly as waves, but rather as passing over immediate small neighbours in favour of larger, potentially more distant settlements. For this reason it is also referred to as a CASCADE model (Labov 2001, p. 285): linguistic innovations proceed as water falling from larger pools to smaller ones, and thence to smallest. Each population centre may be seen as having a sphere of influence in which further diffusion proceeds locally.

The connection to physical gravity may be appreciated if one considers the solar system, i.e. the sun, the nine planets and their moons. In understanding the movements of a given heavenly body, it is best to concentrate on the nearest very massive body. For example, even though the moon is affected by the sun's mass, its rotation is determined almost entirely by the much closer Earth. The physical theory of gravity accounts for this by postulating a force owing to gravity which is inversely proportional to the square of the distance between bodies. In this way very distant bodies are predicted to have much less influence than nearby ones.

There have been many reactions to the gravity model which we cannot elaborate on here for reason of space. We refer to our earlier paper (Nerbonne & Heeringa 2007) for discussion. In summary, research has been mixed in its reception of Trudgill's postulation of a gravity-like effect in linguistic diffusion. There have been voices of affirmation, but also of dissent.

This essay concentrates on the effect of geography, rather than population size, as both Nerbonne & Heeringa (2007) and Heeringa *et al.* (2007) show that geography is by far the more important factor. In contrast to most of the literature on this topic, this essay aims to *measure* the influence of geography on language variation in order to contribute to the discussion. Other contributions attempt no quantitative assessment of the strength of the influence, while this is possible and worthwhile. They also all require methodologically that the researchers identify one or more ongoing linguistic changes and find a way to track them, which is likewise non-trivial. We shall instead examine the residue of a large range of changes in a number of different language areas.

3. AGGREGATE (DIALECTOMETRIC) VARIATION

The remainder of this paper explores an alternative approach to studying the influence of geography on diffusion. We proceed from techniques for measuring linguistic variation, immediately obtaining the advantage of then being able to measure the influence of geography using standard (regression) designs. We shall aggregate the differences in many linguistic variables in order to strengthen their signals. We also assume that all the variation we encounter is the result of diffusion—even if we cannot identify its source. This makes it easier to apply the techniques without first studying where changes are occurring and in what direction.

But before presenting dialectometric approaches to diffusion, it is worthwhile reviewing how and why linguistic distances are measured. We do not have the time or space to review all of the background or range of techniques here, so the presentation will be sketchy. Fortunately, there are good introductions available (Goebl 1984; Heeringa 2004; Goebl 2006; Nerbonne 2009; Nerbonne & Heeringa 2009).

Roughly, dialectometry attempts to distil the aggregate relations from among a set of sites by systematically comparing a large set of corresponding linguistic items (Nerbonne 2009) and measuring differences. By aggregating over a large set of corresponding items, the dialectometric procedure attempts to immunize its work against the dangers of fortuitous, or biased selection of material.

The simplest dialectometric procedures analyse linguistic variation at a nominal or categorical level (Goebl 1984), at which linguistic items are either identical or not. Non-identical items contribute to the linguistic distance between sites, while identical elements do not. Various weighting schemes may be employed, as well (Nerbonne & Kleiweg 2007). Dialect similarity (s) is assayed as the fraction of overlap in the sample, and dialect dissimilarity (distance) is simply the inverse, $d = 1 - s$.

Our own developments in dialectometry have emphasized the advantage of applying LEVENSHTEIN DISTANCE, also known as EDIT DISTANCE, to phonetically transcribed data. Heeringa (2004) and Nerbonne (2009) present these techniques in more detail, so that we may summarize here that the techniques enable us to measure differences in pronunciation at

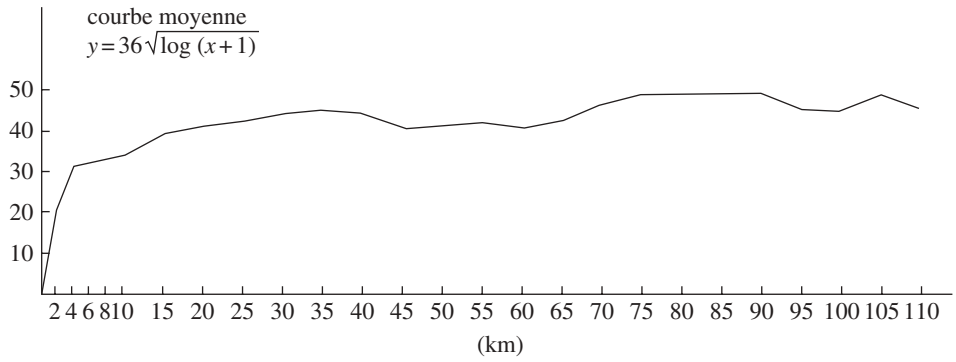

Figure 1. Séguy's (1971) plot of lexical distance, measured categorically, as a sublinear function of geography.

a finer level than merely 'same' or 'different'. Instead the technique ALIGNS the phonetic segments in a pair of word pronunciation transcriptions optimally and sums the differences between the segments in an optimal transcription. We illustrate the procedure via two aligned strings, one a transcription of the pronunciation of the word 'afternoon' in the American south, and the other a transcription of its pronunciation in the American north (or Midland):

æ ə f t ə Ø n ʉ n

æ Ø f t ə r n u n

The optimal alignment reveals three points of mismatch, one substitution of one version of [u] for another, one insertion of [ə] creating an initial diphthong and one deletion of a syllable-final [r]. Each of these points of mismatch is associated with a cost, and the total cost is regarded as the phonetic distance between the two pronunciations.

As Nerbonne (2009) shows, the procedure normally results in a consistent measure of pronunciation difference (Cronbach's $\alpha \geq 0.80$ with at least 30 words in the sample), and the procedure has been validated with respected to dialect speakers' judgements of dialect distance with the result that measurements correlated well with judgements $r \approx 0.7$ (Gooskens & Heeringa 2004).

We shall examine various dialectological situations below, all of which were examined using either the nominal level of analysis pioneered by Séguy (1973) or Goebl (1984) or the numeric level as realized by Levenshtein distance.

(a) *Séguy's curve*

The founder of dialectometry, Séguy (1971) examined the distribution of lexical distance, measured categorically (same or different variants) in the very first publication in this direction 'La relation entre la distance spatiale et la distance lexicale', and he compared the resulting curve with the one in which lexical distance varied with the square root of the logarithm of geographical distance. The result, as one might imagine, is a curve that shows an initial rise and then becomes quite flat. We show Séguy's distribution in figure 1. In view of Séguy's very early work in this direction, I propose that the sublinear curve of linguistic distance versus geographical distance be called SÉGUY'S CURVE.

Since Séguy (1971) appeared a little before Trudgill (1974), there was apparently never any early attempt to confront the two views of how geography influences linguistic variation until recently. The task of the following sections will be to show what these two views have to do with one another.

Séguy's insight is comparable to that of population geneticists, who had earlier found the same sublinear distribution of genetic diversity when viewed as a function of geography, a phenomenon they have come to call 'isolation by distance' (Jobling *et al.* 2004), tracing the idea back to work in mathematical biology of the 1940s and 1950s (Wright 1943; Malécot 1955). More recently, Holman *et al.* (2007) have examined the relation between geographical distance and typological distance as assayed by a set of structural features, referring to their results as establishing 'spatial autocorrelation'.

4. A DIALECTOMETRIC VIEW OF GRAVITY

Nerbonne & Heeringa (2007) and Heeringa *et al.* (2007) applied dialectometric designs to questions of diffusion in an attempt to add an aggregate quantitative perspective to the discussion, claiming two advantages for dialectometric approaches in approaching this question. First, some researchers may have relied on fortuitously chosen features which corroborate or contradict the lasting influence of geography and the chance of social contact, but which might be atypical. Dialectometry proceeds from the measurement of a large number of linguistic variables, and thus affords the opportunity to examine Trudgill's ideas from a more general perspective. Second, dialectometry enables the research to quantify the strength of attractive forces at least somewhat, and thus move beyond cataloguing examples which appear to obey or contradict the predictions of the theory.

Nerbonne & Heeringa derived linguistic distances from 52 towns in the Lower Saxon area of the Netherlands using a technique explained above (§3); they then attempted to explain the linguistic distances on the basis of geographical distance and the chance of social contact as reified in population size. Using a multiple regression model, they proceed from Trudgill's formulation of the gravity model:

$$I_{ij} = s \cdot \frac{P_i P_j}{(d_{ij})^2},$$

where I_{ij} represents the mutual influence of centres i and j, P_i is the population of center i, etc. and d_{ij} is the distance between i and j. s is a constant needed to allow for simple transformations, but it may be viewed as 'variable expressing linguistic similarity'. It will of necessity be ignored in what follows. See Nerbonne & Heeringa (2007) for discussion.

The model predicts that the 'attractive' (accommodating) force should correlate inversely with (the square of) geographical distance, and directly with the product of the population sizes. Reasoning that linguistic distance should reflect this attractive force, but inversely, Nerbonne & Heeringa (2007) examined whether linguistic distance therefore directly correlates with (the square of) geographical distance and inversely with the populations' product, and found that there appears to be no effect of population size on linguistic distance, but also that linguistic distance indeed correlates directly with geographical distance. But Nerbonne & Heeringa also noted that the correlation between geographical and linguistic distance appeared not to be quadratic, as the gravity model predicts, but rather sublinear, i.e. in the same family of relations that Séguy noted in 1971. The best predictor of aggregate linguistic distance was not the square of the geography, but rather the logarithm.

Heeringa *et al.* (2007) criticized the choice of sites in the Nerbonne & Heeringa study, replacing these with a set of sites from the entire Dutch area (Nerbonne & Heeringa had worked exclusively with Lower Saxon) which included rather more settlements of large population size. This study replicated the fact that aggregate linguistic distance depends in a sublinear fashion on geographical distance, but it vindicated the gravity model in showing that population size indeed played the role predicted. In the later study, population product size accounted for six per cent of the variance in linguistic distance.

(a) *Séguy's law*

Given Séguy's early demonstration that French lexical variation depends sublinearly on geographical distance, and Nerbonne & Heeringa's (2007) replication of this result for Dutch pronunciation, it is worth examining a range of other studies to see that they may contribute to the discussion.

Alewijnse *et al.* (2007) obtained pronunciation data from Bantu data collected in Gabon by researchers from the *Dynamique du Langage* (http://www.ddl.ish-lyon.cnrs.fr/) in Lyon. Let us note that since the Gabon Bantu population consisted of migratory farmers until recently, it might not be the right sort of population for this study as the relative mobility of the population might disturb the traces of 'normal diffusion'. The data involve broad phonetic transcriptions of 160 concepts taken from 53 sampling sites. Tone was not analysed as the Bantu experts were skeptical about how reliably it had been recorded and transcribed. The geographical locations recorded were those provided by native speaker respondents, but they should be regarded in some cases as 'best guesses' considering that the population has been fairly mobile (over long periods of time). The pronunciation differences were analysed using the procedure sketched in §3, and these correlate strongly with logarithmic geographical distances ($r = 0.469$).

Prokić (2007) obtained data on Bulgarian dialectology from Prof. Vladimir Zhobov's group at St Clement of Ohrid's University of Sofia. Prokić worked on broad phonetic transcriptions of 156 words from 197 sampling sites in Bulgaria. Palatalized consonants, which are phonemically distinct in Bulgarian, were represented in the data, but stress is not. The pronunciation difference measurement of §3 was applied, where alignments were constrained to respect syllabicity so that vowels only aligned with vowels and consonants only with consonants. Since Bulgaria was occupied by Turkey for several centuries (until 1872), its linguistic variation may display less reliable patterns vis-à-vis geography. The correlation of pronunciation and logarithmic geographical distance was measured at $r = 0.488$.

Nerbonne & Siedle (2005) obtained data from the *Deutscher Sprachatlas* in Marburg (http://www.uni-marburg.de/fb09/dsa/). The pronunciations of 186 words had been collected at 201 sampling sites for the project *Kleiner Deutscher Lautatlas*. A team of phoneticians transcribed the data narrowly; each word was transcribed twice independently and disagreements were settled in consultation so that there was consensus about the results. The pronunciation difference measurement of §3 was applied, where alignments were constrained to respect syllabicity so that vowels aligned only with vowels and consonants only with consonants. Logarithmic geographical distance correlates strongly with pronunciation in this dataset ($r = 0.566$).

Kretzschmar (1994) reports on the LAMSAS project (http://hyde.park.uga.edu/lamsas/), conceived and carried out mainly by Hans Kurath, Guy Lowman and Raven McDavid in the 1930s and again in the 1950s and 1960s. The data are publicly available at http://hyde.park.uga.edu/lamsas/. Due to differences in fieldworker/transcriber practices, we analyse only the 826 interviews which Guy Lowman conducted in the 1930s involving 151 different response items. LAMSAS used its own transcription system, which we converted automatically to X-SAMPA for the purpose of analysis, which was conducted using the measurements described in §3. Nerbonne (in press) describes some aspects of the analysis in more detail, in particular the degree to which phonological structure is present. Since the area of the present USA has only been English speaking for the last several centuries, it may retain traces of migration disturbance in the geographical distribution of linguistic variation. We nonetheless measured a strong correlation between pronunciation and geographical distance after applying a logarithmic correction to the latter ($r = 0.511$).

Wieling *et al.* (2007) analyses the data of the projects *Morphologische Atlas van Nederlandse Dialecten* (MAND) and *Fonologische Atlas van Nederlandse Dialecten* (FAND; Goeman & Taeldeman 1996). In order to eschew potential confounds owing to transcription differences Wieling *et al.* (2007) analyse only the data from the Netherlands, and not that of Flanders. The former included 562 linguistic items from 424 varieties. Since the Netherlands comprises only 40 km^2, the MAND/FAND is one of the densest

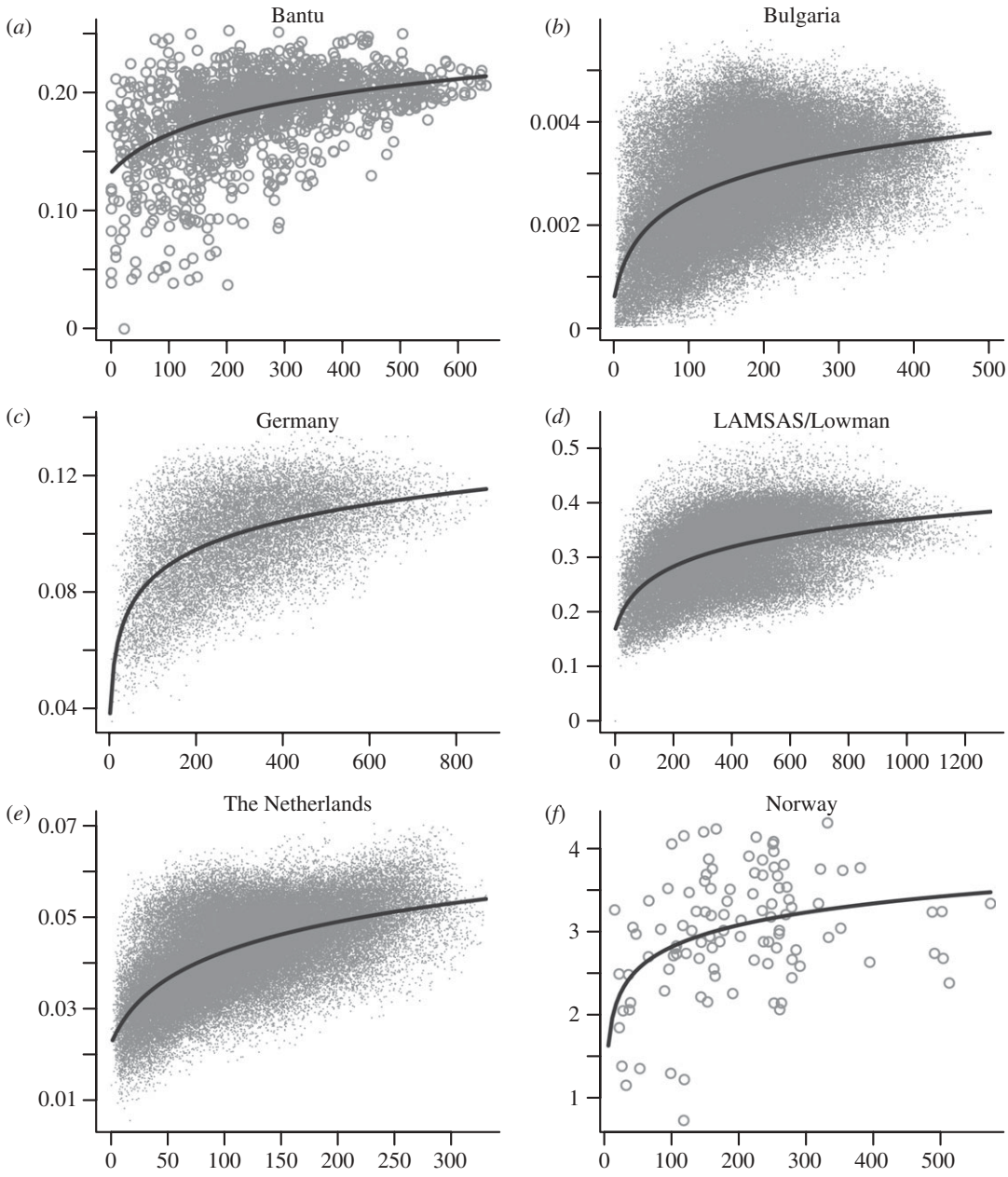

Figure 2. Six examinations of the influence of geography on linguistic variation; a logarithmic curve is drawn in every case. The *y*-axes vary owing to details of measurements, but all are linear scales. See text for details.

dialect samplings ever. The pronunciation differences were assayed using the technique described in §3, where alignments were constrained to respected syllabicity. Pronunciation distance correlates strongly with the logarithm of geographical distance ($r = 0.622$).

Gooskens & Heeringa (2004) analyse the variation in 15 Norwegian versions of the fable of the International Phonetic Association, 'The North Wind and the Sun', making use of material from http://www.ling.hf.ntnu.no/nos/. The material was again analysed using the pronunciation difference measurements in §3. Norwegian distinguishes pronunciations using lexical tone, and Gooskens & Heeringa experimented with measurements which incorporated this, with little distinction in the overall (aggregate) results. Interestingly from a geographical point of view (Britain 2002), Gooskens (2004) compares two geographical explanations of the linguistic differences, one based on 'as the crow flies' distances, and another based on the (logarithmic) travel time estimates of the late nineteenth century, showing an improvement in correlation (from $r = 0.41$ to $r = 0.54$). The motivation for examining the two operationalizations was naturally that Gooskens expected travel time to be the better reflection of the chance of social contact.[1]

We conclude from this section that there is a simple, measurable and normally sublinear influence which geography exerts on aggregate linguistic differences (figure 2). It is an empirical finding, not a theoretical prediction, that geography accounts for 16 per cent to about 37 per cent of the linguistic variation in these datasets ($100 \times r^2$). We note that the potential disturbances caused by migration, occupation, and recent settlement appear insubstantial enough in the cases examined so as not to disturb the overall tendency first noted by Séguy, namely that variation increases as a sublinear function of geography. We

should also note that Spruit (2006) obtains a slightly better analysis using a linear rather than a sublinear geographical model to explain SYNTACTIC distance, but we shall not pursue the issues this suggests here.

5. INDIVIDUAL VERSUS AGGREGATE DIFFERENCES

Taking stock a little, we note that the substantial sociolinguistic literature on diffusion has on the one hand recognized a major role for social contact and therefore geography, but has concentrated on identifying additional, what we might call 'extra-geographical' factors. Its data collection and analysis have exclusively concerned the patterns of diffusion found in individual linguistic items such as individual words or sounds. The dialectometric view on the other hand enables the measurement of the influence of geography on aggregate variation. Is there any way to bring these two perspectives to a more rewarding engagement?

We can think of two ways of exploring the relation between the individual variation discussed in the scholarly literature reviewed in §2 and the aggregate dialectal variation presented in §3, one empirical and the other simulation-based. The empirical path is conceptually simple, and involves examining the distributions of a large number of individual items to explore how their distributions are related to the aggregate distance curves of §4. As conceptually simple as that strategy is, still it requires identifying a large range of words, sounds, etc. about which there would be agreement that they constitute units of diffusion. This could be quite difficult.

But we can also simulate the diffusion of individual linguistic items to obtain insight about the relation between the diffusion of individual items and the aggregate diffusion curves examined in §4*a*. We turn now to a description of a simulation.

(a) *Simulating diffusion*

We wish to examine the effect of the 'gravity'-like, attractive force which influences diffusion, and we shall restrict our attention to the influence owing to geography, continuing to ignore the influence of population density. Like Holman *et al.* (2007) we created simulations in order to focus on the contribution of individual factors, in our case the relation between aggregate linguistic distance and the effect of attractive forces of varying strengths on individual linguistic features.

To investigate this process via simulation, we create several thousand sites, each of which is represented by a 100-dimensional binary vector. The sites are at regular distances from a single reference site so that the most distant site is several thousand times more distant from the reference site than the closest is. The 100 dimensions may be thought of as 100 linguistic items, e.g. 100 words or perhaps 100 pronunciation features, such as the pronunciation of the vowel in a words such as 'night' (i.e. as [nat] in the American south versus [nait] in standard American). The value '0' indicates that the site is the same as the reference site with respect to a given dimension, and '1' indicates that it is different. We intend to be deliberately vague about the units of diffusion. We proceed from the assumption that we are observing the differentiation of an initially homogeneous community, but we add some noise in the form of 100 random chances at change at every site. Unchanged linguistic items are assumed to be identical to those at the reference site. Since we finally compare each value in a given dimension only with other values in the same dimensions, we make no special assumptions about what these values are.

In reality, each settlement in a sample may potentially be influenced by any other sample, as Holman *et al.* (2007) note, but we wish to keep the simulation simple, so we shall examine the situation in which all the influence is exerted by a single reference site. The simulation will vary the strength with which that influence is exerted. We wish to contrast two possibilities concerning the strength with which the reference site influences others. In the first *linear* view, the distance of the simulation site predicts directly the chance with which the value of the reference site is adopted. In that case a site that is d distant from the reference site has twice the chance of being like it (with respect to a given linguistic dimension) as a site that is $2d$ distant. In the second *quadratic* view, a site that is d distant from the reference site has four times the chance of being like the reference site when compared with another that is $2d$ distant. The latter is the view advanced by the gravity model.

To simulate the diffusion of linguistic change we iterate once through the set of sites. At each site, we repeat the process of random change n times, where n depends on the distance of the site from the reference site. In the linear model n depends directly on the distance to the reference site, and in the quadratic model of influence n depends on the square of the distance. The random change itself is quite simple. We randomly select one dimension i in the 100-element vector, then generate a second random number, this time between 0 and 1. If the number is greater than 0.5, then we set the ith position to 1, indicating that the site differs from the reference site at dimension i. If the number is 0.5 or less, then the value of the ith dimension at the site is set to 0, indicating that it is linguistically the same as the reference site. (We note that it is distinctly possible that the *same* dimension is randomly chosen more than once when there is a large number of repetitions of the random change. In this case changes may cancel each other out.) In all cases the aggregate distance of the site from the reference site is simply the sum of the vector over all positions.

So the overall effect is that sites near the reference site have few chances of changing—the influence of the reference site is too strong. Sites twice as distant have twice as many chances to change, and, in the case of the 'gravity'-inspired simulated, four times as many. In both cases changes are introduced randomly, but while the chance of a change being attempted depends linearly on the geographical distance from the reference point in the linear model, it rises quadratically with the geographical distance in the quadratic model. Furthermore, since the stochastic events of change are competing for the same limited number of linguistic dimensions, the more distant sites are also more liable to change and also change back.

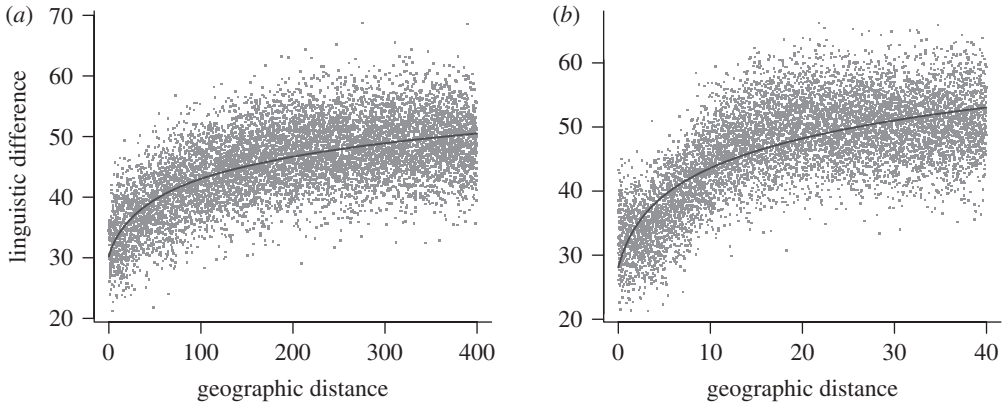

Figure 3. Two simulations of linguistic diffusion, with an attractive influence that (*a*) diminishes linearly and (*b*) diminishes quadratically. In both, we appear to obtain the characteristic sublinear Séguy curve of aggregate linguistic distance.

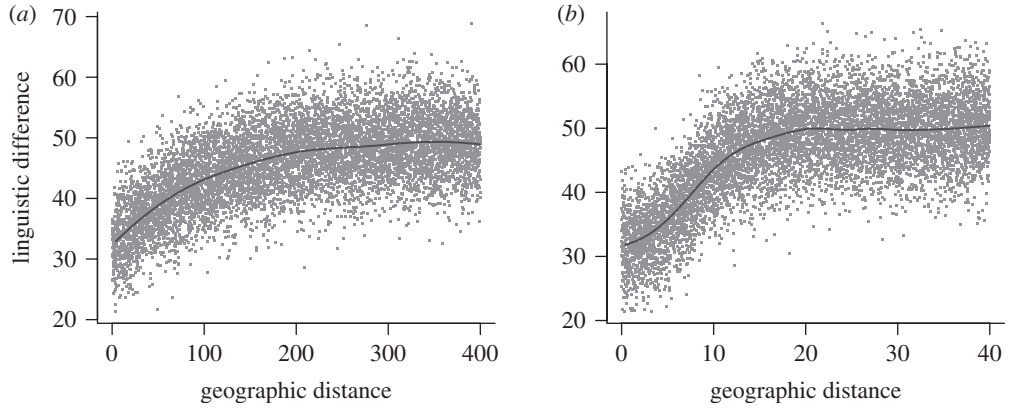

Figure 4. Local regression lines have been added to the scatterplots of the graphs in figure 3 revealing, in the case of the graph (*b*) representing (aggregate) inverse quadratic influences, a sigmoidal shape which is otherwise missing. The graph (*a*) shows the local regression line for inverse linear influence. This suggests that the inverse quadratic strength influence attributed to geography in the 'gravity' is bequeathed to the aggregate distribution as well, contrary to facts adduced in §4 (see figure 2). The local regressions were carried out using R with $\alpha = 0.4$ (each of 8000 steps considered 40% of the data using an inverse tricubic weighting).

(b) *Results of simulations*

Figure 3 compares the results of two single runs of the simulation, one in which the chance of change rises linearly in the distance to the reference point, and a second in which the same chance of change rises quadratically with respect to distance. In both cases we have drawn the logarithmic regression line, for which we obtain $r = 0.66$ (linear attraction) and 0.71 (inverse quadratic attraction), and in both cases we appear to obtain the characteristically sublinear Séguy curve of aggregate linguistic distance.

The curve in figure 3*b* appears somewhat sigmoidal, however, a suspicion we examine more closely by applying local regression to the same dataset. The result of the local regression is shown in figure 4, and, indeed, it appears that the quadratic influence results in a different curve in this respect. Local regression lines do not differ significantly from logarithmic lines in the other scatterplots (the left plot of cumulative linear influence in figure 3 or in any of the plots in figure 2).

Although the results clearly point to a linear effect of geography on the likelihood of an individual linguistic item differing from that of another site, we acknowledge that further simulations would be useful to be certain of the influence of some parameters, including the relatively great distance used at initialization, the effective ceiling of 50 per cent on average differences caused by restricting the model to binary choices, and the relatively constant variance in the simulations. Finally, it would be useful to view simulations in which locations interacted with each other and not merely with a single reference point. Holman *et al.* (2007) have analysed simulations with lattice structures with respect to other research questions.

6. CONCLUSIONS AND FUTURE WORK

Our foremost conclusion is that we can effectively test models of diffusion quantitatively, and, in particular, that these may be tested on the basis of large aggregates of linguistic material. This avoids the danger of picking material fortuitously, and it obviates the need to find material in the process of change.

We likewise conclude on the basis of several empirical studies that the chance of social contact, operationalized through geography, can account for

about one-quarter of the aggregate linguistic variation we find in large collections such as dialect atlases. Our experiment in simulation suggests that the attractive force which tends to resist linguistic change decreases linearly with geographical displacement while the gravity model suggests an attractive force that would decrease quadratically with geography. It would clearly be valuable to seek empirical data on individual linguistic variables with which the diffusion model might be tested.

If this approach to analysing diffusion is sound, and assuming that the relevant variables can be operationalized and that suitable data can be found, then this approach should likewise open the door to studies on the influence on non-geographical factors. These might be compared with geography. As the brief remarks on travel time (§4*a*) might suggest, it is also possible to examine alternative conceptions of geography, taking a step in the direction urged by Britain (2002).

Peter Kleiweg was responsible for all the programs used in this paper, in particular, the dialectometric package L04 (http://www.let.rug.nl/kleiweg/L04/). Two anonymous referees were generous in their comments.

ENDNOTE

[1]Van Gemert (2002) also examined the use of travel time in predicting Dutch dialect distances, but it turned out that travel time correlated nearly perfectly with geographical distance in the Netherlands, which lack the fjords and mountains that impede direct lines of travel in Norway.

REFERENCES

Alewijnse, B., Nerbonne, J., van der Veen, L. & Manni, F. 2007 A computational analysis of Gabon varieties. In *Proc. of the RANLP Workshop on Computational Phonology Workshop at Recent Advances in Natural Language Processing* (eds P. Osenova *et al.*), pp. 3–12. Borovetz, Bulgaria: RANLP.

Bloomfield, L. 1933 *Language*. New York, NY: Holt, Rhinehart and Winston.

Britain, D. 2002 Space and spatial diffusion. In *The handbook of language variation and change* (eds J. Chambers, P. Trudgill & N. Schilling-Estes), pp. 603–637. Oxford, UK: Blackwell.

Goebl, H. 1984 *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*, vol. 3. Tübingen, Germany: Max Niemeyer.

Goebl, H. 2006 Recent advances in Salzburg dialectometry. *Lit. Linguist. Computing* **21**, 411–435. (doi:10.1093/llc/fql042)

Goeman, A. & Taeldeman, J. 1996 Fonologie en morfologie van de nederlandse dialecten. Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten. *Taal en Tongval* **48**, 38–59.

Gooskens, C. 2004 Norwegian dialect distances geographically explained. In *Language variation in Europe: papers from ICLaVE 2* (eds B.-L. Gunnarson, L. Bergström, G. Eklund, S. Fridella, L. H. Hansen, A. Karstadt, B. Nordberg, E. Sundgren & M. Thelander), pp. 195–206. Uppsala, Sweden: Uppsala University.

Gooskens, C. & Heeringa, W. 2004 Perceptual evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Lang. Variation Change* **16**, 189–207.

Heeringa, W. 2004 Measuring dialect pronunciation differences using Levenshtein distance. PhD thesis, Rijksuniversiteit Groningen, The Netherlands.

Heeringa, W., Nerbonne, J., van Bezooijen, R. & Spruit, M. R. 2007 Geografie en inwoneraantallen als verklarende factoren voor variatie in het nederlandse dialectgebied. *Tijdschrift voor Nederlandse Taal- en Letterkunde* **123**, 70–82.

Holman, E. W., Schulze, C., Stauffer, D. & Wichmann, S. 2007 On the relation between structural diversity and geographical distance among languages: observations and computer simulations. *Linguist. Typology* **11**, 393–421. (doi:10.1515/LINGTY.2007.027)

Jobling, M. A., Hurles, M. E. & Tyler-Smith, C. 2004 *Human evolutionary genetics: origins, peoples and diseases*. New York, NY: Garland.

Kretzschmar, W. A. (ed.) 1994 *Handbook of the linguistic atlas of the Middle and South Atlantic States*. Chicago, IL: The University of Chicago Press.

Labov, W. 2001 *Principles of linguistic change: social factors*, vol. 2. Malden, MA: Blackwell.

Malécot, G. 1955 The decrease of relationship with distance. *Cold Spring Harbor Symp. Quant. Biol.* **20**, 52–53.

Nerbonne, J. 2009 Data-driven dialectology. *Lang. Linguist. Compass* **3**, 175–198. (doi:10.1111/j.1749-818X.2008.00114.x)

Nerbonne, J. In press. Various variation aggregates in the LAMSAS south. In *Language variety in the South III* (eds C. Davis & M. Picone). Tuscaloosa, AL: University of Alabama.

Nerbonne, J. & Heeringa, W. 2007 Geographic distributions of linguistic variation reflect dynamics of differentiation. In *Roots: linguistics in search of its evidential base* (eds S. Featherston & W. Sternefeld), pp. 267–297. Berlin, Germany: Mouton De Gruyter.

Nerbonne, J. & Heeringa, W. 2009 Measuring dialect differences. In *Theories and methods, language and space* (eds J. E. Schmidt & P. Auer). Berlin, Germany: Mouton De Gruyter.

Nerbonne, J. & Kleiweg, P. 2007 Toward a dialectological yardstick. *Quant. Linguist.* **14**, 148–167. (doi:10.1080/09296170701379260)

Nerbonne, J. & Siedle, C. 2005 Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. *Z. Dialektol. Linguist.* **72**, 129–147.

Prokić, J. 2007 Identifying linguistic structure in a quantitative analysis of dialect pronunciation. In *Proc. of the ACL 2007 Student Research Workshop*, pp. 61–66. Prague: Association for Computational Linguistics.

Schmidt, J. 1872 *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. Weimar, Germany: Böhlau.

Séguy, J. 1971 La relation entre la distance spatiale et la distance lexicale. *Rev. Linguist. Romane* **35**, 335–357.

Séguy, J. 1973 La dialectométrie dans l'atlas linguistique de Gascogne. *Rev. Linguist. Romane* **37**, 1–24.

Spruit, M. R. 2006 Measuring syntactic variation in dutch dialects. *Lit. Linguist. Computing* **21**, 493–506. (doi:10.1093/llc/fql043)

Trudgill, P. 1974 Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Lang. Soc.* **2**, 215–246. (doi:10.1017/S0047404500004358)

van Gemert, I. 2002 *Het geografisch verklaren van dialectafstanden met een geografisch informatiesysteem* (*gis*). Master's thesis, Rijksuniversiteit Groningen, The Netherlands. www.let.rug.nl/alfa/scripties.html.

Wieling, M., Heeringa, W. & Nerbonne, J. 2007 An aggregate analysis of pronunciation in the Goeman-Taeldeman-van Reenen-project data. *Taal en Tongval* **59**, 84–116.

Wright, S. 1943 Isolation by distance. *Genetics* **28**, 114–138.