
Computers and the Humanities **37:** 245–255, 2003.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

Introducing Computational Techniques in Dialectometry

JOHN NERBONNE and WILLIAM KRETZSCHMAR
Humanities Computing, University of Groningen, the Netherlands and Atlas Project, University of Georgia, USA

Abstract. Dialectology is the study of dialects, and dialectometry is the measurement of dialect differences, i.e. linguistic differences whose distribution is determined primarily by geography. The earliest works in dialectology showed that language variation is complex both geographically and linguistically and cannot be reduced to simple characterizations. There has thus always been a perceived need for techniques which can deal with large amounts of data in a controlled means, i.e. computational techniques. This special issue of *Computers and the Humanities* presents a range of recent work on this topic.

Key words: dialect, dialectology, dialectometry

1. Introduction

DIALECTOLOGY is the study of dialects, and DIALECTOMETRY is the measurement of dialect differences, i.e. linguistic differences whose distribution is determined primarily by geography. Dialectology may be classified within the more general study of how languages vary – not only along geographical, but also social lines or along lines of age and gender. Dialectology is the oldest, and best understood branch of variationist linguistics, which includes, in addition to dialectology, the study of linguistic variation as it correlates with social class, age, sex, and occupation. We expect the more general study of variation to benefit from the techniques developed for dialectology.

The earliest works in dialectology showed that language variation is complex both geographically and linguistically and cannot be reduced to simple characterizations. There has thus always been a perceived need for techniques which can deal with large amounts of data in a controlled way, i.e. computational techniques. Dialectological data is available digitally and challenging. This special issue of *Computers and the Humanities* sketches some ways in which computational techniques can be put to use in the study of variation.

1.1. MOTIVATION

The study of language variation has always been an important aspect of linguistic research. It provides insights into historical, social and geographical factors of language use in society. Gilliéron, the father of French dialectology, was, for example, famous for showing that several linguistic divisions, running roughly East-West across French, corresponded closely with well established cultural divisions, in particular the ethnic split between slightly Romanized Celts in the North, and thoroughly Romanized non-Celts in the South, the legal division between the common law North and the Roman law South, and patterns of agriculture and architecture (see Chambers and Trudgill, 1998, pp. 95–103). Once we know that shared linguistic traits arise through interaction and shared history, we may then reverse the perspective and suggest, on the basis of shared linguistic traits, that the people speaking related varieties must have been in contact. In recent years theoreticians have also turned increasingly to the study of dialects as a means of demarcating the possible range of human language in more detail (Benincà, 1987).

In the nineteenth century historical linguists turned to dialectology when they found that irregularities discovered in the history of standard languages were sometimes illuminated by dialect facts (Bloomfield, 1933, p. 322). There was a short-lived hope that the historical record in the local dialects would prove better susceptible to historical analysis. Rather quickly they learned that dialects are likewise complex and that they show regularities which, however, are subject to exception. Bloomfield's (1933) authoritative discussion of the problems (p. 328) of determining dialect areas is a *locus classicus*: the vowels in Dutch *huis, muis* ('house', 'mouse') were the same historically, but they do not align with other linguistic distinctions, and thus do not determine dialect areas satisfactorily (Figure 1). In a sense this discussion set the stage for a central analytical question of twentieth-century dialectology: given that the geographic coherence of language variation is imperfect, how must it be analyzed?

Older dialectology focused on the identification of DIALECT AREAS, where a dialect area is an area distinguished from its neighbors by its relatively more limited range of linguistic variation. While older studies were able to reach a reasonable level of consensus on which areas those are, still the characterization resisted analysis. More than one report is accompanied by a sigh, and a remark that variation might better be understood as "a fairly unbroken chain of dialects [. . .] the furthest extremes of the continuum being unintelligible to one another" (Tait, 1994, p. 3).

Competition between the idea of dialect areas and the idea of dialect continua can be characterized historically as a contrast between the German (NeoGrammarian) model and the French model that emerged from the work of Gaston Paris and his student Gilliéron (see Kretzschmar, 1995). The root of the puzzle generated by this contrast – that dialectologists cannot demonstrate in detail the existence of dialect areas that we perceive to exist – may lie in two different senses of "dialect" which Kretzschmar (1998) has dubbed ATTRIBUTIVE DIALECTS and

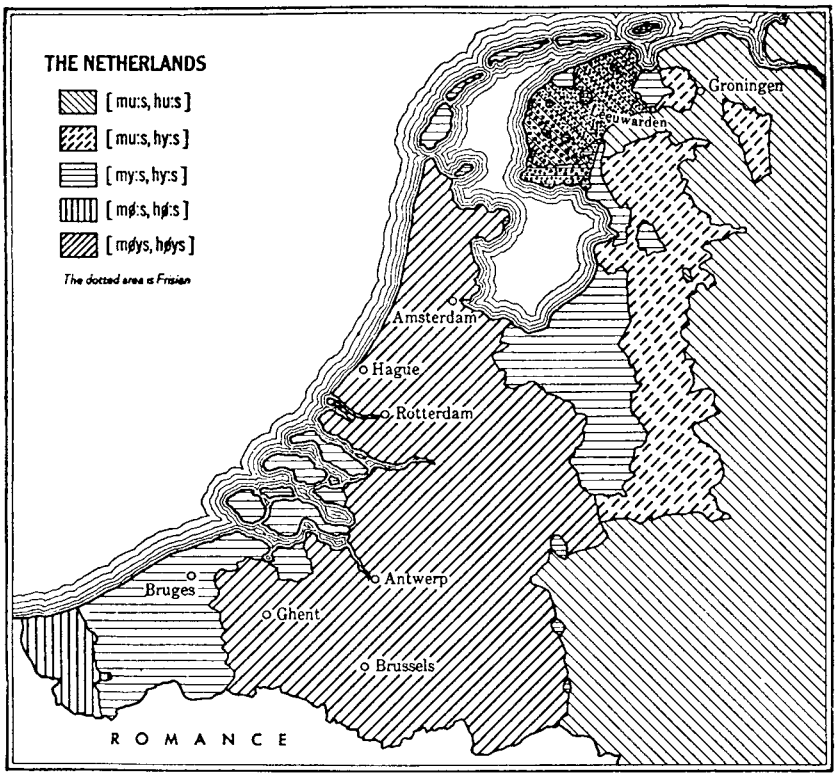

FIGURE 6. Distribution of syllabic sounds in the words *mouse* and *house* in the Netherlands. — After Kloeke.

Figure 1. Bloomfield's (1933, p. 328) classical discussion of the problems of determining dialect areas. The vowels in Dutch *huis, muis* ("house", "mouse") were the same historically, but they do not align with the distribution of other linguistic variables, and thus do not determine dialect areas satisfactorily.

BLIND DIALECTS. Sometimes we refer e.g. to "the dialect Smith speaks" or the "dialect of South Boston" without reflecting on whether it is distinctive in any way. In such a case we "attribute" a dialect to a location by noting the linguistic features in use there. An attributive dialect is simply the linguistic variety used in a particular place. Note that a field linguist will generally succeed in the task of cataloguing the linguistic features in a given place, i.e. in specifying the attributive dialect, but this does not guarantee success in determining how the local (attributive) dialect compares to the speech of other places. It does not guarantee that the field linguist has noted anything linguistically distinctive about the variety. Dialectologists working in the French traditon most often focus on the careful cataloguing of individual linguistic features (as did Gilliéron himself, especially on their etymologies) instead of the question of what is distinctive to some groups of varieties.

The task is different for an analyst who examines a range of varieties and seeks to abstract the features distinctive for one or more varieties while turning a "blind" eye to extralinguistic properties associated with the linguistic data. This analyst seeks subareas distinguished by the linguistic features commonly used there, but he works only on the basis of linguistic features and without reference, e.g. to geographical or cultural factors. This is a challenging task. Because linguistic variation is gradual (showing "continuum" effects), the analyst will not find it easy to identify common linguistic features, and thus will have to be satisfied with showing relative similarity. But this sort of description is anathema to the NeoGrammarian view, in which languages are closed, well-structured systems. Dialectologists in the German, NeoGrammarian tradition have focus on the question of what distinguishes groups of varieties (dialect areas), and require for this a selection among linguistic features.

Heeringa and Nerbonne (2002) have also examined the issue of areas versus continua using dialectometric techniques, focusing on the issue of whether linguistic change is cumulative. The potential contribution of computational techniques to this exchange on areas and continua is to provide means of analysing large bodies of material in carefully controlled ways.

Most non-computational studies focus on a small number of features and cannot characterize AGGREGATE levels, e.g. the East Anglian dialect or the language of London teenagers, using these few characteristics. Aggregate characterizations are elusive because large data sets invariably contain counter-indicating tendencies leading to the analytical challenge of characterizing notions of aggregate levels without simply insisting on the importance of one's favorite features. Computational techniques on the one hand, and standard statistical data reduction techniques on the other, not only shed light on these classic linguistic problems, but they also suggest avenues for exploring the question at more abstract levels, and perhaps for seeking the determinants of variation. Computational and statistical analysis now makes it possible comprehensively to compare feature inventories attributively drawn from a great many locations, in order to try to solve the puzzle of linguistic systems vs. linguistic continua and address the linguistic component of our perception of dialect areas.

2. Dialectometry

The first breakthrough in techniques to characterize aggregate levels was Seguy (1971), who suggested that one simply count the number of overlapping features between any two data collection sites. This technique could be applied to the wealth of material in dialect atlas projects, which was mostly collected by questionnaires with a limited number of answers. An obvious case is lexical choice: what do you call a serving-size, unsweetened pastry? – *bun, roll, biscuit*, ... Sites that gave the same answer to a question like that are counted one point more similar than sites that give different answers. The same counting technique could be used

on pronunciation or other linguistic features once one agreed on a fixed set of categories.

Seguy effectively invented dialectometry in this step. Dialectometry is the measurement of dialect differences, i.e. linguistic differences whose distributions are determined primarily by geography. The simple step of counting differences allowed Seguy to aggregate individual differences over a large amount of material.

2.1. GOEBL

Although Seguy is rightfully credited with founding dialectometry, Chambers and Trudgill ([1]1980, 1998, p. 112 in 1st edition) could still conclude that its "utility has not been demonstrated" nine years later. By the time the second edition of their book appeared Chambers and Trudgill accept and even promote dialectometry (Chambers and Trudgill, 1998, pp. 140–148). The single person most responsible for this shift in scholarly opinion was Hans Goebl (1982, 1984), who elaborated enormously on dialectometrical ideas and demonstrated their potential much more systematically. For example, Goebl was not content with merely counting the level of overlap, but explored weightings which count overlap in infrequent words more heavily. For concept i with n responses $w_1^i, w_2^i, \ldots, w_n^i$, we let $f(w_j^i)$ be the frequency of w_j as response to query about i.

$$S(w, w') = 1 - \frac{f(w_j^i) - 1}{n \cdot w}$$

where Goebl (1984, p. 85) foresees experimentation with w.[1] In general $S(w, w')$ varies inversely with $f(w_j^i)$ so that the least frequent elements count the most in similarity. Goebl was able to obtain more satisfying analyses using this measure which counts infrequent (and therefore unlikely) matches more heavily.

These early treatments focused on categorical data, e.g. lexical variation, i.e. the question of whether the words used for a given concept varied geographically, but they also included phonological and other sorts of data treated at a categorical level.

A second major innovation of Goebl's was to investigate the degree to which a given site "fit in" with the range of other measurements, through its "relative coherence" (Goebl, 1984, p. 179ff). This has applicability to questions of how deviant a given variety is with respect to others, so that we can apply it questions of whether a given variety is a "island" or an area of "transition" between two relatively stable areas.

3. Workshop

The present special issue of *Computers and the Humanities* arose from a special session which the authors of this introduction organized at the *Methods in Dialectology XI* conference, which was organized by Prof. Markku Filppula at Joensuu,

Finland on August 5–9, 2002. The present issue would undoubtedly be better if we had been able to include more of the presentations. The following could unfortunately not be included:

Will Allen, Karen Corrigan, Hermann Moisl and Charley Rowe, Newcastle	Topographic Mapping As A Tool For Analysis and Results Visualization of Dialectal Data
Wilbert Heeringa, Groningen	The Use of Spectral Sound Distances in the Comparison and Classification of Dutch Dialects
Mika Kukkola and Päivi Nieminen, Helsinki	Electronic Morphology Archives for Finnish Dialects
Alfred Lameli, Marburg	On the Quantification of Phonetic Features in Regional Speech Forms
April McMahon, Paul Heggarty and Robert McMahon, Sheffield	Dialect Classification by Phonetic Similarity: Towards a Computational Method

On the other hand, the paper by Heeringa and Braun was presented at the main session of the conference, not the special session, and the paper by Kondrak is an outgrowth of his 2002 PhD thesis *Algorithms for Language Reconstruction.* Both are thematically so appropriate that there was no question but that including them would be beneficial.

4. Papers

In this section we place the six papers included in this special issue into the context of work in dialectometry.

4.1. HEERINGA AND BRAUN, MEASURING SEGMENT DIFFERENCES

A major limitation of existing dialectometric work was its treatment of all data as categorical. In a series of studies Nerbonne *et al.* (1996), Nerbonne *et al.* (1999), Heeringa and Nerbonne (2002) have demonstrated that appropriately modified string-distance measures made be applied to collections of phonetic transcriptions to yield numerical characterizations of pronunciation differences. These measurements are readily implemented using the LEVENSHTEIN or EDIT-DISTANCE algorithm, and they yield characterizations that are much richer that those based on categorical data, and may be analyzed in novel ways.

It is an important refinement of this line of work to show that it may be based on a phonetically defensible notion of segment distance. Heeringa and Braun's contribution applies and refines a measure of distance developed in phonetics to measure the fidelity of phonetic transcriptions – a measure that was used in the 1980's to evaluate student transcribers. It is a natural step to use FEATURES familiar from phonetics and phonology, but those features must be chosen so that feature differences contribute to segment distance. The feature [±tense], which Ladefoged (1975, p. 245) following Chomsky and Halle (1968) uses to mark the vowels most extremely front or back – in distinction to central vowels – may serve as an example of a feature that might be useful for the purpose of making phonological rule description more compact or perspicuous, but which is ill-suited as the basis for a system for determining segment similarity or dissimilarity.

Heeringa and Braun use a logarithmic correction on the sum of feature distances and test the resultant measure within a string distance framework, showing that it outperforms competitors.

4.2. KONDRAK, PHONETIC ALIGNMENT

The same algorithm used in Heeringa and Braun's work to measure string distance (given an appropriate segment distance base) is also used to ALIGN strings. Given the standard American and Bostonian pronunciations of *saw a girl* the algorithm will find the corresponding segments:

Standard American	/ s	ɔ		ə	g	I	r	l /
Bostonian	/ s	ɔ	r	ə	g	ɜ		l /

As Kondrak notes, the resulting alignments are useful in several ways. First, they provide a check on the performance of the algorithm (e.g. in its assessing of distance), and second, the alignment is a record of REGULAR CORRESPONDENCES of the sort which is the fundamental evidence linguists adduce when attributing an historical relation to two varieties – whether this be the genealogical relation, in which two varieties share an ancestor, or one of several contact relations, in which one variety is said to have borrowed from another (Thomason and Kaufmann, 1988).

But Kondrak notes a serious problem in using the edit-distance algorithm on some sorts of data: some linguistic processes radically add and delete material, e.g. entire prefixes or suffixes. Thus French *sommes* /sɔm/ is cognate with Latin *sumus* /sumus/, even though the first-person plural suffix is virtually absent (from pronunciation). Drawing inspiration from work that has been done in sequence comparison in the context of genetics, Kondrak explores LOCAL ALIGNMENT variants of the algorithm, which seek alignments which are locally optimally, sometimes ignoring very poor alignment at the beginnings and ends of strings (the more volatile parts of words). He furthermore explores the range of segment distance bases for his work, like Heeringa and Braun, and concludes that multi-valued articulatory features are the best bases from which to work.

4.3. HEERINGA AND GOOSKENS, PERCEPTUAL AND ACOUSTIC DIFFERENCES

The focus of Heeringa and Gooskens's paper is the attempt to base a measure of pronunciation difference not on phonetic transcriptions, which after all are the result of a subjective process in which a field worker interprets a respondent's utterance, but rather directly on acoustic recordings. This is very challenging for many reasons. First, the recordings must be made under very similar conditions; second, one must attempt to abstract from the personal variation which does not inform linguistic variation, e.g. the pitch with which respondents speak (and which notably differs between men and women); third, the problem of correcting for differences in the speed of speech; and fourth (and related to the third), the problem of segmenting the acoustic signal.

Recordings made by Jørn Almberg in cooperation with Kristian Skarbø and available at `http://www.ling.hf.ntnu.no.nos` appear to be of the needed quality and consistency. Using these the authors segmented the speech in a very rough fashion and likewise corrected for speed differences by differentially expanding the samples being compared. Heeringa and Gooskens conducted several experiments to determine the optimal acoustic filter needed to abstract away from individual variation, including Bark filters, formant tracks, and cochleagrams, deciding finally for formant tracks, but noting that a male/female division remained prominent in this representation.

In spite of the fact that they found no effective way to abstract away from the personal variation that a transcriber ignores automatically, however, Heeringa and Gooskens were able to modify the basic Levenshtein algorithm (again!) to obtain a reasonable measure of acoustic difference through simple curve distance, and are able to show that this correlates very significantly with psychoacoustic measures of distance which Gooskens (2003) had obtained in fieldwork. Gooskens's earlier study had simply asked subjects to judge how similar an auditorily presented variety was to their own.

4.4. SPEELMAN, GRONDELAERS AND GEERAERTS, PROFILE-BASED UNIFORMITY

Speelman, Grondelaers and Geeraerts focus on a technique to use the relative frequency of words which might be regarded alternative lexicalizations to measure the differences between varieties. English examples of such pairs might be *car* vs. *automobile*, *quiet* vs. *still* or *bike* vs. *bicycle* – assuming that one controlled for the ambiguity in the terms. A collection of frequency information about such choices is a PROFILE, and their paper aims to show the advantages of using profiles as opposed to frequencies without reference to alternatives or simply keywords.

The successful incorporation of frequency information is the realization of a long-standing wish in the measurement of linguistic distance. Goebl questioned Seguy about the need to incorporate frequency in analysis, who replied in a 1972 letter:

> Le problème des fréquences d'emploi n'a jamais cessé de me tourmenter. [. . .] Il parait certes évident qu'un lexème polyfréquent joue dans la demarcation une rôle plus puissant [. . .] Mais il est impossible de connaitre [. . .] la fréquence des lèxemes en discours [pour chaque point d'enquête . . .] Bref, j'ai adopté l'attitude de facilité: négliger les fréquences lexicales.
>
> Quoted by Goebl (1984, p. 28)

Of course frequency information remains elusive for many applications in which we should like to measure distance. But Speelman, Grondelaers and Geeraerts focus on the differences between Belgian and Netherlandic Dutch, and have been clever in collecting frequencies from shop window advertisements, newspapers (of differing stylistic levels), internet chat-rooms, and internet discussion lists. They are able to demonstrate that a reasonable choice of profiles results in a distance measure in which chat material, discussion lists and newspapers are clearly distinguished, and they show that frequency without references to alternatives (i.e. without profiles) is less successful.

4.5. NERBONNE AND KLEIWEG, LEXICAL DISTANCE

Nerbonne and Kleiweg examine the *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS), a large portion of which is available digitally at `http://hyde.park.uga.edu/lamsas/`. They apply Seguy's notion of distance in categorical data fairly directly with an eye to the question of whether Kurath was correct in postulating a "Midland" in the LAMSAS data, i.e. an area which extends from north of Philadelphia into the inland Southern states. In the course of their work they note that LAMSAS fieldworkers were inconsistent in the number of alternative lexical items they recorded, and in the number of "no response" items – perhaps suggesting an explanation for the fieldworker boundaries which earlier researchers have noted. As a result, they limit their analysis to the data collected by a single fieldworker, who, fortunately, was responsible for 71% of the records in LAMSAS.

Two refinements of basic techniques are suggested and implemented, first, a treatment of questionnaire items for which more than one response is recorded, and second, a method for dealing with related, but non-identical responses, e.g. *clears up*, *clears* and *clearing up* for which they employ a string-distance measure on spellings.

The result is an analysis which vindicates Kurath – even though the authors are careful to note that the analysis depends on clustering, an exploratory statistical technique which is potentially very sensitive to small input distinctions. And in defense of Kurath's opponents they note that the Midland area is itself divided very significantly along lines noted by Kurath and preferred by his opponents. In a response to Schneider's (1988, p. 176) criticism that dialectometric methods were unsatisfactory since they lose qualitative information about the linguistic features in the areas they characterize numerically, Nerbonne and Kleiweg establish areal

boundaries and, in a further analytical step, show which features are associated with the areas thus established.

4.6. Palander, Opas-Hännninen and Tweedie, transitional dialects

Palander, Opas-Hänninen and Tweedie are interested in what goes on in dialects at the borders between dialect areas, i.e. where some dialects do not fit neatly into a given partition of varieties, and in particular in the range of variation which these transition dialects may show. This is related to Goebl's interest in the relative coherence of a set of data collection sites (see above), but Palander, Opas-Hännninen and Tweedie focus on Finnish dialects spoken by Karelian and Savo peoples near the Russian border, and also follow an alternative analytical strategy. The authors choose ten linguistic variables as a basis for their work, and they operate not on relative frequencies (as do Speelman, Grondelaers and Geeraerts), but rather on logarithms of likelihood ratios, which they argue to be preferable mathematically. As a further methodological refinement, they normalize variables with respect not to entire distributions, but rather with respect to most frequent variants.

Palander, Opas-Hänninen and Tweedie's data consists of recordings of 198 people from nineteen parishes. The heart of the analysis is a comparison between the average feature values in parishes and the values in the speech of the individual speakers. The authors verify that the parish values cluster in ways expected on the basis of earlier work on Finnish dialects, but they then show that the variation among individual speakers is very large in the transitional areas, so large that these speakers are actually closer to other parishes in the features that were examined.

Acknowledgements

We are grateful to Prof. Filppula and other organizers of *Methods in Dialectology XI* for the opportunity to hold this one-day session in Joensuu. We particularly thank the many referees for carefully criticized submissions and led to innumerable improvements: Werner Abraham, Bridget Anderson, Harald Baayen, Walter Cichocki, David Bowie, Anders Eriksson, Ton Goeman, Charlotte Gooskens, Stephan Grondelaers, James Hammerton, Cornelius Hasselblatt, Wilbert Heeringa, Vincent van Heuven, Paul Kerswill, Greg Kondrak, Alfred Lameli, Rob Malouf, April McMahon, Hermann Moisl, Hermann Niebaum, Rogier Nieuweboer, Marjatta Palander, John Palolillo, Anneli Sarhimaa, Erik Tjong Kim Sang, David Weenink, Stephen Winters, and Menno van Zaanen.

Note

[1] Goebl refers to this weighted measure of similarity as "gewichtender Identitätswert" whenever $w = 1$.

References

Benincà P. (ed.) (1987) *Dialect Variation in the Theory of Grammar*. Foris, Dordrecht.

Bloomfield L. (1933) *Language*. Holt, Rhinehart and Winston, New York.

Chambers J., Trudgill P. ([1]1980, 1998) *Dialectology*. 2nd ed. Cambridge University Press, Cambridge.

Chomsky N. A., Halle M. (1968) *The Sound Pattern of English*. Harper and Row, New York.

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

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

Gooskens C. (2003) How Well Can Norwegians Identify Their Dialects? *Nordic Journal of Linguistics*, submitted.

Heeringa W., Nerbonne J. (2002) Dialect Areas and Dialect Continua. *Language Variation and Change*, 13, pp. 375–398.

Kretzschmar W. A. (1995) Dialectology and Sociolinguistics: Same Coin, Different Currency. *Language Sciences*, 17, pp. 271–282.

Kretzschmar W. A. (1998) Analytical Procedure and Three Technical Types of Dialect. In Montgomery M. and Nunnally T. (eds.), *From the Gulf States and Beyond: The Legacy of Lee Pederson and LAGS*. University of Alabama Press, Tuscaloosa, pp. 167–185.

Ladefoged P. (1975) *A Course in Linguistic Phonetics*. Harcourt-Brace, New York.

Nerbonne J., Heeringa W., Kleiweg P. (1999) Edit Distance and Dialect Proximity. In Sankoff D. and Kruskal J. (eds.), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed.* CSLI, Stanford, CA, pp. v–xv.

Nerbonne J., Heeringa W., van den Hout E., van der Kooi P., Otten S., van de Vis W. (1996) Phonetic Distance between Dutch Dialects. In Durieux G., Daelemans W., and Gillis S. (eds.), *CLIN VI: Proceedings from the Sixth CLIN Meeting*. Center for Dutch Language and Speech, University of Antwerpen (UIA), Antwerpen, pp. 185–202. Also available as http://www.let.rug.nl/˜nerbonne/papers/dialects.ps.

Schneider E. (1988) Qualitative vs. Quantitative Methods of Area Delimitation in Dialectology: A Comparison Based on Lexical Data from Georgia and Alabama. *Journal of English Linguistics*, 21, pp. 175–212.

Séguy J. (1971) La Relation entre la Distance Spatiale et la Distance Lexicale. *Revue de Linguistique Romane*, 35, pp. 335–357.

Tait M. (1994) North America. In Moseley C. and Asher R. (eds.), *Atlas of the World's Languages*. Routledge, London and New York, pp. 3–30.

Thomason S., Kaufmann T. (1988) *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkeley.