Linguistic Variation and Computation

John Nerbonne
Alfa-informatica, BCN, University of Groningen
9700 AS Groningen, The Netherlands
nerbonne@let.rug.nl

Abstract

Language variationists study how languages vary along geographical or social lines or along lines of age and gender. Variationist data is available and challenging, in particular for DIALECTOLOGY, the study of geographical variation, which will be the focus of this paper, although we present approaches we expect to transfer smoothly to the study of variation correlating with other extralinguistic variables. Techniques from computational linguistics on the one hand, and standard statistical data reduction techniques on the other, not only shed light on this classic linguistic problem, but they also suggest avenues for exploring the question at more abstract levels, and perhaps for seeking the determinants of variation.

1 Introduction

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 (1980, pp.111-123)). 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). The present paper sketches some ways in which techniques from computational linguistics (CL) can be put to use in the study of variation.

Language variationists study how languages vary along geographical or social lines or along lines of age and gender. Variationist data is available and challenging, in particular for DIALECTOLOGY, the study of geographical variation, which will be the focus of this paper, although we present approaches we expect to transfer smoothly to the study of variation correlating with other extralinguistic variables. Most non-computational studies focus on a small number of features and cannot characterize AGGREGATE levels, e.g., the Bavarian 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. Techniques from computational linguistics on the one hand, and standard statistical data reduction techniques on the other, not only shed light on this classic linguistic problem, but they also suggest avenues for exploring the question at more abstract levels, and

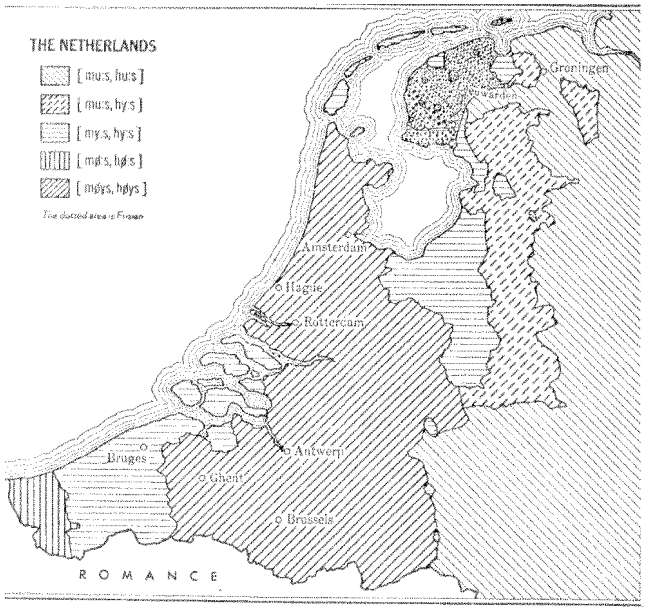

FIGURE 6. Distribution of syllabic sounds in the words *mouse* and *house* in the Netherlands. — After Klocke.

Figure 1: Bloomfield's (1933:328) 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 determine dialect areas satisfactorily.

perhaps for seeking the determinants of variation.

2 Computational Dialectology

Given a large amount of dialect data, there is a good chance that one will encounter "noise", i.e., inaccuracy, non-geographic variability, and incompatibility both in the choice of information recorded and in the level of detail at which it is recorded. There are also many linguistic features to explore, and many ways of combining them. It is in general a salutary effect of a computational approach that data is checked for conformity to specifications.

A second, much more delayed effect of a computational approach is the explicitness of data analysis techniques and the relative confidence this inspires in seeking out problems in data. Nerbonne and Kleiweg (2003) postulate that conflicting fieldworkers' methods confound studies of lexical variation based on the LAMSAS data set, available at `http://hyde.park.uga.edu/lamsas`, and they offer as confirmation the hugely varying number of responses per item which different field workers offered.

But these are general benefits of computational analysis; specific CL techniques have also been demonstrated to contribute to dialectology. We turn to these now.

2.1 Aggregate Differences

Dialectology has been studied for over a century, and several challenges are well-known. Bloomfield (1933) *inter alia* noted that ISOGLOSSES—lines dividing linguistic features on a map—often do not overlap and so do not jointly define dialect areas (see Fig. 1). Bloomfield went on to the perspicacious remark that it nonetheless seemed generally true that linguistic differences increased with geographic distance. Coseriu ([1]1956, 1975) noted that the problem of non-coinciding distributions is only aggravated as researchers examine more data in greater detail, noting a tendency toward "atomism" in the entire line of work. A major challenge therefore is to move from the level of describing the geographic distribution of individual linguistic features such as the vowel in *house*, or the word used to describe the instrument used to clear snow to a more general level, that of the

linguistic variety used in a particular area or by a particular group.

In order to rise above the atomistic level of the individual sounds or lexical items, it is beneficial to employ aggregate measures of distance, the (non-)identity of lexical items on the one hand (essentially the same measure proposed by Seguy (1971), and elaborated on by Goebl (1984)—the two major figures in DIALECTOMETRY, the exact study of dialect differences; and a string similarity measure which we apply to phonetic transcriptions on the other (Nerbonne and Heeringa, 1998; Nerbonne et al., 1999; Heeringa et al., 2002). Because the measures yield numeric characterizations of lexical/phonetic distance, it may be aggregated over many pairs of similar concepts.

But the analysis of aggregate levels of distances is problematic for non-computational work, relying on hand counts of vocabulary differences or (at best) manual alignments of phonetic segments. Computational approaches hold the promise of incorporating large amounts of data into analyses, and simultaneously remaining consistent in the application of analytical techniques.

2.2 Pronunciation Distance

Noncomputational analyses treat pronunciation data as categorical, thus blocking the way to a useful aggregation. A key to aggregating pronunciation data is to find a legitimate numeric perspective. One suitable numeric characterization of sequence difference is the standard CL algorithm for the calculation of LEVENSHTEIN distance, also known as "edit distance" or "sequence distance", which speech recognition has also used (Kruskal [1]1983, 1999). The Levenshtein algorithm assigns a measure of difference to pairs of pronunciations encoded as phonetic transcriptions. Because it is a true (numeric) measure, it is additive so that it is meaningful to examine the sum of pronunciation differences over a large collection of dialectal material. This provides a view of the *aggregate* differences we called for above. Kessler (1995) introduced the use of Levenshtein distance to Irish dialects.[1]

[1]This is the same algorithm that is also used for alignment, e.g., the alignment of bilingual texts (Gale and Church, 1993).

The Levenshtein algorithm calculates the "cost" of changing one word into another using insertions, deletions and replacements. L-distance (s_1, s_2) is the sum of the costs of the cheapest set of operations changing s_1 to s_2. The example below illustrates Levenshtein distance applied to Bostonian and standard American pronunciations of *saw a girl*.

sɔəgIrl	delete r	1
sɔəgIl	replace I/ɜ	2
sɔəgɜl	insert r	1
sərəgɜl		
	Sum distance	4

The example illustrates *one* calculation of distance. To obtain the least cost, we need to guarantee that we examine in principle all possible sequences of operations, and the Levenshtein algorithm is in fact effective for this purpose (worst-time $\mathcal{O}(mn) \approx \mathcal{O}(n^2)$ in time, where n, m are the lengths of strings to be compared. The example simplifies our procedure for clarity: refinements due to segment similarity are normally used, as Kessler (1995), Nerbonne and Heeringa (1998) and Kondrak (2000) illustrate.

Heeringa (2003) studies various ways in which sequence distances may be generated from tables of segment distances, including reference to acoustic distances (curve distance between spectrograms) and the use of different feature systems in order to induce segment distance. In all these cases replacement costs vary depending on the segments involved, and, Heeringa further investigates determining the cost of insertions and deletions via a distance between 'silence' and the segment which is inserted or deleted. Among other things, Heeringa shows that the best results are obtained using feature systems which had been developed to measure fidelity in phonetic transcription (Vieregge et al., 1984; Almeida and Braun, 1986). What distinguishes these systems from systems which are designed to facilitate the succinct statement of phonological rules (e.g., Chomsky and Halle's system in *The Sound Pattern of English*) is the following: as segments differ more perceptually, their feature descriptions tend to differ more formally.

Vieregge's (1984) system distinguishes four vowel features, advancement, height, length, and

roundedness, as well as ten consonantal features, including place, voice, nasality, height, distributiveness, and five binary manner features, including stop, glide, lateral, fricative, and flap.

In a series of experiments we have applied these techniques to data is from *Reeks Nederlands(ch)e Dialectatlassen* (Blancquaert and Pée, 1925 1982)), which contains 1,956 Netherlandic and North Belgian transcriptions of 141 sentences. We selected over 350 dialects, regularly scattered over the Dutch language area, and 150 words which appear in each dialect text, and which contain all vowels and consonants. Comparing each pair of varieties results in a sum of 150 word-pair comparisons. Because longer words tend to be separated by more distance than shorter words, the distance of each word pair is normalized by dividing it by the mean lengths of the word pair. This results in a half-matrix of distances, to which (i) clustering may be applied to CLASSIFY dialects (Aldenderfer and Blashfield, 1984); while (ii) multidimensional scaling may be applied to extract the most significant dimensions (Kruskal and Wish, 1978). A map is obtained by interpreting MDS dimensions as color intensities and mixing using inverse distance weighting (see Fig. 2). The maps that are produced distinguish Dutch "dialect areas" in a way which non-computational methods have been unable to do, giving form to the intuition of dialectologists in Dutch (and other areas) that the material is best viewed as a "continuum".

2.2.1 Results

We have confirmed the reliability of the measurements, showing that Cronbach's $\alpha > 0.95$ at 100 words, and we have validated the technique using cross-validation on unseen Dutch dialect data, and also by examining alignments, and by comparing results to expert consensus (Heeringa et al., 2002). Ongoing work applies the technique to questions of convergence/divergence of dialects using dialect data from two different periods (Heeringa et al., 2000). Finally, there have been several experiments on novel data sets, including Sardinian, Norwegian and German. See `http://www.let.rug.nl/~heeringa` for papers on these.

2.3 Lemmatizing to Ascertain Lexical Variation

A second example of the way in which CL techniques may be of service in dialectology comes from the study of lexical variation. Lexical data is obtained by asking respondents what words they use for certain concepts, e.g., by showing a picture of an object, or by describing it. For example, the fieldworkers of the *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS) asked their respondents the following:

> *"If the sun comes out after a rain, you say the weather is doing what?"*

to which question they received over 40 different answers, including:

> *clearing, clears up, clearing up, fair off, fairing, fairing off, faired off, fairs off, ...*

See `http://hyde.park.uga.edu/lamsas` for an excellent facility for browsing this dialectological data. The problem is that the data reflects not only lexical variation, but also inflection variation. Since LAMSAS consists of 1162 interviews with an average of 160 responses/interview, it is not an option to sort the response for lexical identity by hand. In this case it would have been ideal to apply the standard CL technique of lemmatizing the data. Not having a lemmatizer at hand, we applied the poor man's lemmatizer, the Porter stemmer, to extract the relevant information from the strings. It is publicly available at several places. The example below of its output shows that it filters a great deal of dialectologically interesting information, but in every case this is information about *morphological*, not *lexical* variation.

```
a hundr year    a hundred year
a hundr year    a hundred years

abat    abated
abat    abating

blew    blew
blew    blewed

ceas    cease
ceas    ceased
ceas    ceases
ceas    ceasing
```

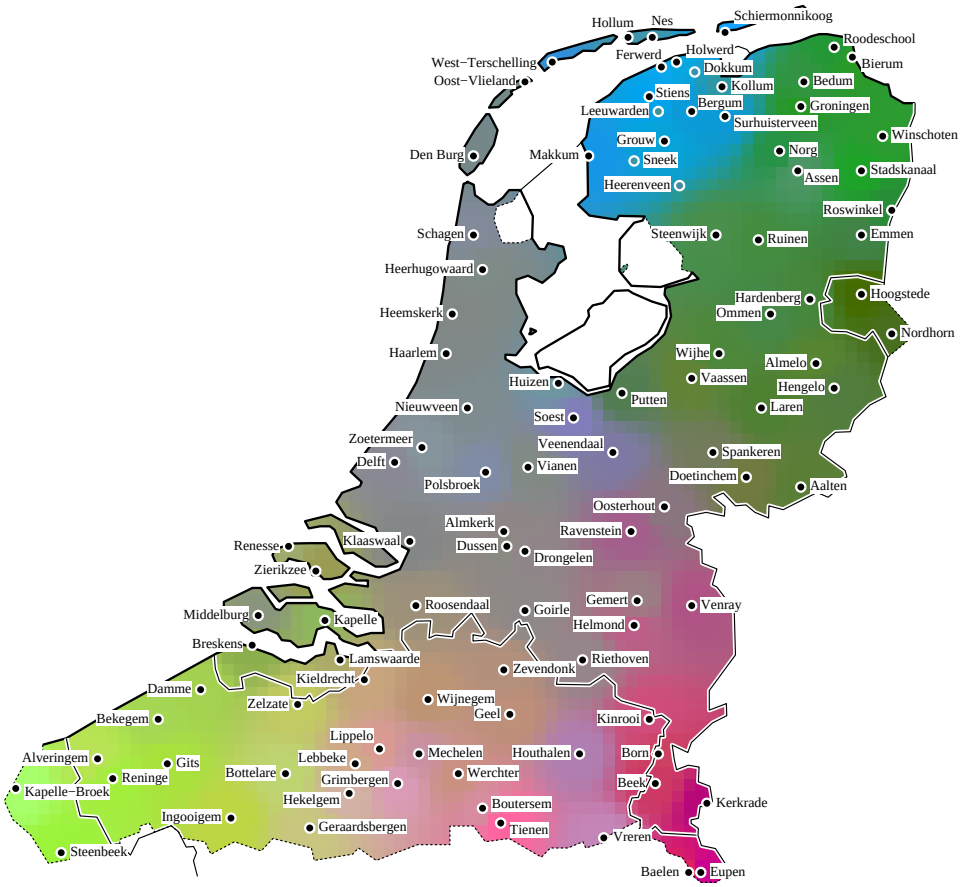

Figure 2: The most significant dimensions in average Levenshtein distance, as identified by multi-dimensional scaling, are colored red, green and blue. The map gives form to the dialectologist's intuition that dialects exist "on a continuum," within which, however significant differences emerges. The Frisian dialects (blue), Saxon (dark green), Limburg (red), and Flemish (yellow-green) are clearly distinct. (In case you have printed this on a grey-tone printer, see `http://www.let.rug.nl/~nerbonne/papers/tw-se-mm.ps`for the correct color rendition.

2.3.1 Results

It turns out that lexical variation is considerably less consistent than pronunciation variation, showing a Cronbach's $\alpha = 0.62$ at 65 words. To obtain similar levels of consistency as pronunciation, we should need an order of magnitude more data. On the one hand this reflects that fact that pronunciation contains a good deal more information that lexical identity alone, since the average word contains almost five segments. We normalize for work length, but since the normalization (an average difference per segment) stabilizes the measure, the larger number of segments still plays a role in explaining the greater consistency of pronunciation. The effect of the larger number of components is not sufficient to explain the greater consistency of pronunciation data, however. It remains the case that pronunciation is the better measure, and we suspect that this is due to the fact that lexis is simply more volatile than pronunciation. While pronunciations tend to be stable, we acquire new words easily and in great numbers.

In still unpublished work we examine the degree to which pronunciation and lexical variation correlate. Dialectologists generally claim that these two levels "coincide fairly well" (Kurath and McDavid, 1961), but when we calculate the correlation between lexical and pronunciation differ-

ences in the LAMSAS data (which Kurath and McDavid wrote about), we do not find a particularly strong correlation, viz., $r = 0.65$. If we think of a linguistic variety as a coherent collection of linguistic material subject to the same pressures to conformity—both as a sign of social belonging and more profoundly as a requisite for communication—then we might have expected the various linguistic levels such as pronunciation and lexis to correlate more closely.

3 Toward Explanation

The distance-based characterization of language variation provides a novel perspective on the characterization of the geographical distribution of linguistic variants. Although the distribution of concrete linguistic features has resisted aggregate characterization and therefore explanation, it may turn out that there are satisfying explanatory characterizations of the linguistic *distance* between variants. Fig. 3 shows the result of a regression analysis which sought to explain pronunciation distance in terms of geographical distance for a small set of Dutch dialects. The form of the question in Fig.3 was also implicit in Seguy (1971), who, however, focused on lexical variation. The sample of sites (towns) was chosen to stretch along a line from the Southwest to the Northeast of the Dutch speaking area. Pronunciation distance and the logarithm of geographic distance correlated in this sample highly ($r = 0.89$), so pure geography appears to account for 80% of the the variance in the data ($r^2 = 0.8$). Although further work has suggested that the sample of towns and villages was chosen in a way that inflated these figures (more typical levels range for the Dutch data suggest $r \approx 0.75$ as a general level, and data from other languages often yields still lower levels), still the form of the analysis is suggestive as an approach to asking for the determinants of dialectal variation.

Trudgill (1983, Chap. 3) calls for more attention to the question what determines linguistic variation—and Heeringa and Nerbonne's (2002) analysis sketched above suggests a path toward answers. We may begin to inquire about the determinants of linguistic variation from a different perspective. For a first example, if sheer geographic distance is a good predictor of linguistic (pronunciation) difference, shouldn't travel time be somewhat better, since it is likely to predict the chance of social contact more accurately? Can we show this more convincingly by examining variation in countries with varying geographies, e.g., mountain ranges? Second, older maps and discussions often partition dialects into several non-overlapping areas, suggesting that linguistic distances ought to be predicted by these. The names of areas even suggest that 'tribal' history played a role (involving, e.g., Franks, Saxons, Alemannians, and Bavarians). Which is the better predictor, geography or tribal area? Third, Trudgill's own "gravitational" model suggests that geography together with settlement size should predict best, and this is plausible, given their effect on the chance for social contact, which in turn exerts pressure to reduce variation (in order to allow communication). At least we have the methodology to address suggest questions given the background of work sketched above.

Let me emphasize that the last paragraph is intended to inspire rather than to report. We have not demonstrated what the determinants of dialect distance are, and we should not be misunderstood as claiming this. But the application of CL techniques has led to the development of a measure that can claim to reflect pronunciation and lexical differences faithfully, and this opens the way to standard quantificational analyses of these differences.

4 Conclusions

There have been several immediate benefits to approaching dialectology computationally. Several studies have involved digitizing large amounts of data and implementing software such as lexical analyzers to ensure conformity to specifications. As we have come to trust the techniques developed, we have on occasion suggested that some data is confounded in subtle ways (Nerbonne and Kleiweg, 2003).

This paper has focus on the the application of string edit distance to phonetic transcriptions on the one hand and the use of lemmatization or stemming on the other. Edit distance provides a measure of pronunciation distance which may be ag-

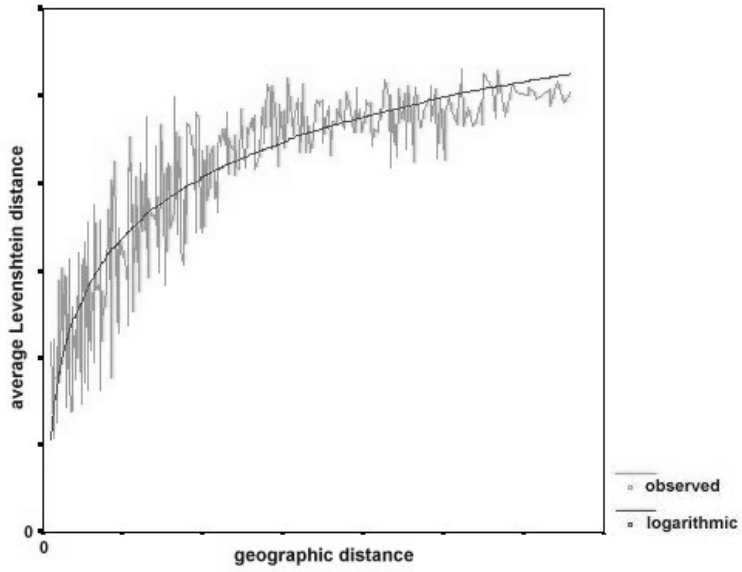

Figure 3: Average pronunciation distances as a logarithm function of geographic distances. Points are connected in order to illustrate the range of variation for average Levenshtein (pronunciation) distance. Note that the logarithmic line seems to overestimate the pronunciation differences associated with greater distances. Reproduced from Heeringa and Nerbonne (2002).

gregated over large samples of phonetic transcriptions solved a long-standing problem in the choice of features on which dialect divisions should be based and providing a firmer foundation to the frequently voiced sentiment of dialectologists that they were dealing with a "continuum" of variation. The application of stemming or lemmatization is less ambitious, but nonetheless allows a systematic view of lexical variation abstracting away from inflection that would otherwise be infeasible.

These and other computational forays enable a reformulation of key dialectological questions. We can, for example, reformulate the question of the determinants of dialectal variation in a way that focuses on distance, rather than on the concrete realization of particular linguistic variants. We illustrated this opportunity by providing the results of a regression analysis in which we predict dialectological distance as a logarithmic function of geographical distance. It is clear that similar analyses, exploring the importance of other factors, may be carried out straightforwardly.

4.1 Prospectus

Computational Linguistics is often defined as "the scientific study of language from a computational perspective" (see, e.g., the definition which the Association for Computational Linguistics offers at its web site, `http://www.aclweb.org/`), which ought to interact broadly with lots of linguistic subfields, since Linguistics is normally defined as "the study of language" (see the web site of the Linguistic Society of America at `http://www.lsadc.org/`) but in practice there's often a narrow focus on morphology and syntax, perhaps together with lexical semantics, and their processing. Worse, outsiders commonly view CL as limited to computational applications having to do with language. But there are opportunities for computational contributions in any number of subfields of Linguistics. This paper has tried to illuminate work on one such subfield, but I hope that it will encourage more.

5 Acknowledgments

Wilbert Heeringa and Peter Kleiweg have been have been intelligent co-developers of these ideas,

and they have implemented all the programs described here. Software to support dialectometry is available at `http://www.let.rug.nl/~kleiweg/levenshtein/`.

References

Mark S. Aldenderfer and Roger K. Blashfield. 1984. *Cluster Analysis*. Quantitative Applications in the Social Sciences. Sage, Beverly Hills.

Almerindo Almeida and Angelika Braun. 1986. 'Richtig' und 'Falsch' in phonetischer Transkription: Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten. *Zeitschrift für Dialektologie und Linguistik*, LIII(2):158–172.

P. Benincà, editor. 1987. *Dialect Variation in the Theory of Grammar*. Foris, Dordrecht.

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

Leonard Bloomfield. 1933. *Language*. Holt, Rinehart and Winston, New York.

Jack Chambers and Peter Trudgill. 1980. *Dialectology*. Cambridge University Press, Cambridge.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

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

Wilbert Heeringa and John Nerbonne. 2002. Dialect areas and dialect continua. *Language Variation and Change*, 13:375–398.

Wilbert Heeringa, John Nerbonne, Hermann Niebaum, and Rogier Nieuweboer. 2000. Measuring Dutch-German contact in and around Bentheim. In Dicky Gilbers, John Nerbonne, and Jos Schaeken, editors, *Languages in Contact*, pages 145–156. Rodopi, Amsterdam-Atlanta.

Wilbert Heeringa, John Nerbonne, and Peter Kleiweg. 2002. Validating dialect comparison methods. In Wolfgang Gaul and Gerd Ritter, editors, *Proceedings of the 24th Annual Meeting of the Gesellschaft für Klassifikation*, pages 445–452. Springer, Heidelberg.

Wilbert Heeringa. 2003. *Computational Comparison and Classification of Dialects*. Ph.D. thesis, University of Groningen. in preparation, 2/03.

Brett Kessler. 1995. Computational dialectology in Irish Gaelic. In *Proc. of the European ACL*, pages 60–67, Dublin.

Grzegorz Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In *Proc. 1st North American ACL*, pages 288–293, Seattle. ACL.

Joseph Kruskal and Myron Wish. 1978. *Multidimensional Scaling*. Sage, Beverly Hills.

Hans Kurath and Raven McDavid. 1961. *The Pronunciation of English in the Atlantic States : Based upon the Collections of the Linguistic Atlas of the Eastern United States*. University of Michigan Press, Ann Arbor.

John Nerbonne and Wilbert Heeringa. 1998. Computationele vergelijking en classificatie van dialecten. *Taal en Tongval*, 50(2):164–193.

John Nerbonne and Peter Kleiweg. 2003. Lexical variation in LAMSAS. *Computers and the Humanities*. Accepted to appear.

John Nerbonne, Wilbert Heeringa, and Peter Kleiweg. 1999. Edit distance and dialect proximity. In David Sankoff and Joseph Kruskal, editors, *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed.*, pages v–xv. CSLI, Stanford, CA.

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

Peter Trudgill. 1983. *On Dialect. Social and Geographical Perspectives*. Blackwell, Oxford.

Wilhelm H. Vieregge, A.C.M.Rietveld, and Carel Jansen. 1984. A distinctive feature based system for the evaluation of segmental transcription in dutch. In Marcel P.R. van den Broecke and A. Cohen, editors, *Proc. of the 10th International Congress of Phonetic Sciences*, pages 654–659, Dordrecht.