# Progress in Dialectometry: Toward Explanation

**John Nerbonne**
University of Groningen, 9700 AS Groningen, The Netherlands

**William Kretzschmar, Jr**
University of Georgia, Athens, Georgia 30602, USA

## Abstract

Dialectometric techniques analyze linguistic variation quantitatively, allowing one to aggregate over what are frequently rebarbative geographic patterns of individual linguistic variants, such as which word is used for a particular concept in a language area, or which sounds are used in particular words. This leads to general formulations of the relation between linguistic variation and explanatory factors. Dialectometric techniques are maturing continuously, paving the way to genuinely new opportunities for the explanation of linguistic variation. These include, most prominently, techniques for analyzing syntactic variation, techniques for comparing the relative importance of different individual linguistic variables, techniques for comparing the relative importance of linguistic levels such as pronunciation, vocabulary, and/or prosody, and many more. This article serves as an introduction to a special issue of *Literary and Linguistic Computing* devoted to presenting a new work constituting *Progress in Dialectometry: Toward Explanation*.

**Correspondence:**
John Nerbonne,
University of Groningen,
Humanities Computing,
9700 AS Groningen,
The Netherlands.
**E-mail:**
j.nerbonne@rug.nl

## 1 Introduction and Background

Linguistic variation is not merely of proverbial interest, there are even songs sung celebrating its fascination, e.g. the Gershwins *Let's call the whole thing off*: 'You say tomato [tə.ˈmeɪ.ɾoʊ], and I say tomahto [tə.ˈma.toʊ], You say potato [pə.ˈteɪ.ɾoʊ], and I say potahto [pə.ˈta.toʊ].' Variant linguistic forms constitute one of the aspects of language with the greatest popular appeal, and their systematic study has resulted in a fascinating, well-developed scholarship (Chambers and Trudgill, 1998, [1]1980] Milroy and Gordon, 2003; Niebaum and Macha, 2006). Language forms vary according to geography, social class, sex, occupation, and age, where the study of dialect geography has dominated the history of the discipline, just as it receives the lion's share of the attention in the work reported on here.

### 1.1 Motivation

Scholars and scientists generally address the subject of variation from one of three, often overlapping perspectives: social, historical, and linguistic. The social perspective addresses the synchronic function of linguistic variation to signal social identity, e.g. geographic provenance or class affiliation. The 'tomahto'-speaker signals—intentionally or unintentionally—a regional or social affiliation that interlocutors may use to classify him within the language community. The historical perspective assumes that some linguistic features are preserved differently over time, and in particular that certain features are well preserved (Labov, 1994). By studying linguistic variation, particularly patterns of unusual, shared linguistic customs, we may open a window to older forms of a language, perhaps even reconstruct prehistoric patterns of

Literary and Linguistic Computing, Vol. 21, No. 4, 2006. 
doi:10.1093/llc/fql034 Advance Access Published on 8 September 2006

shared development. Since shared linguistic developments suggest a common social history, this perspective on language variation holds the promise of insight into ancient demography. Finally, since some variation originates in linguistic processes, the study of linguistic variation may enlighten linguistic theory. Perhaps because modern linguistics has immensely broadened the range of analytical techniques we may bring to bear in variational analysis, this perspective seems to be the impetus behind a great deal of modern work on linguistic variation.

These perspectives need not agree, in the sense that the structures they rely on for explanation may not be the same. For example, for features to be effective in signaling geographic provenance, interlocutors within the language community must be sensitive to them, at least subliminally. But these need not be features that are destined to resist change and therefore provide signals of shared demographic history. So it is conceivable that the social and historical perspectives on variation could evolve into separate enterprises. And linguistic theory offers such a wealth of descriptive mechanisms that it would also be surprising if none of them were useful in describing variation. The fact that these perspectives nonetheless do often seem to agree reflects the cognitive basis of language.

For a feature to reflect linguistic history, it must be passed from one generation to the next. This transmission is not genetic but rather cultural, however, and therefore presupposes that the younger generation perceives and encodes the features being transmitted. This cognitive basis for historical transmission means that features that play a role in historical explanation must be cognitively available. And linguistic theory, even as it makes a myriad of descriptive mechanisms available, must accept the additional task of examining which descriptions most exactly fit the data.

# 2 Dialectometry

In general, individual linguistic features—words, constructions, and pronunciation variants—are associated only weakly with geography. For every promising candidate of a feature which might 'define' a dialect area, it always turns out that there are exceptional sites within and without the area which run counter to the candidate 'definition'. Dialectology was plagued with questions about the imperfect relation between form and geography until Séguy (1973) noted that, even if individual features led to imperfect characterizations, aggregates of features could reliably indicate geographic relations. This was essentially the birth of dialectometry, which especially Goebl (1982, 1984) refined and improved.

Nerbonne and Kretzschmar (2003) preceded the present volume and focused on the role of computers in dialectometry and on the refinement of dialectometric technique. In the call for articles in this volume, we therefore wished to proceed beyond the refinement of technique. We asked especially for works aiming at further explanation of language variation.

We are gratified by the response. Several studies echo the wish to involve dialectometry in deeper explanation, several extend dialectometry to new areas, and several, in particular, seek to relink the individual features which Séguy abstracted over mentioned earlier to the aggregate characterizations. We expand on these themes subsequently.

## 2.1 Organizational background

The articles in this special issue of *Literary and Linguistic Computing* arose from a special session at *Methods XII, the 12th International Conference on Methods in Dialectology* held at the *Université de Moncton* in Moncton, New Brunswick (Canada). The *Association for Literary and Linguistic Computing* (ALLC) sponsored this special session, naming Prof. Hans Goebl the ALLC invited speaker. Prof. Goebl's talk at the workshop was simultaneously a plenary address to the entire conference, and we are pleased to include it among the works here. Prof. Lisa Lena Opas-Hänninen (Oulu) also presented ALLC young researchers' awards to Dr Cynthia Clopper (Indiana University) and Mr Marco Spruit (Meertens, Amsterdam) at a well-attended reception sponsored by the ALLC.

# 3 Papers

The remainder of this introduction suggests a framework from which to understand the individual works, and it attempts to put them into a broader perspective.

## 3.1 State of the art

The first three studies may safely be said to reflect the state of the art, one (Kretzschmar's) by reflecting on that important, but often neglected question: what does the current state of the field *fail* to tell us about linguistic variation, suggesting therefore the directions the field needs to move in. Goebl's and Haimerl's studies reflect the state of the art to a great degree because Goebl's work—in which Haimerl has played an important role—has largely defined the state of the art in dialectometry. If this way of introducing the work suggests a completed, and therefore, stagnant line of research, then let us hasten to add that the research line is vibrant, and that Goebl's article in particular provides a novel insight into the role of geography in determining linguistic variation. We examine each of these in more detail.

### 3.1.1 *Kretzschmar on dialectometric reflection*

Dialectometry has involved the application of mathematical and computational techniques to the analysis of linguistic variation, and like many innovative research lines, it has been necessary first to explore many options, to refine techniques, to examine their consistency and validity, to compare different languages, different sorts of data (lexical, morphological, pronunciational, and potentially others), and different data sets. Kretzschmar sees the field progressing technically, and other articles in this volume are proofs of that progress.

But Kretzschmar, very much in the spirit of this volume, warns against focusing only on technical progress. The technically or quantitatively oriented dialectologist needs to reflect on the scientific questions being asked, and needs especially to take care that the techniques are appropriate for the questions.

Kretzschmar warns in particular of two specific dangers, first, that of imagining that the geographical reality around which linguistic variation is structured is known a priori, and second, that the linguistic reality being structured is somehow given. Dialectometrists need to be reflective both about the geography they assume and about the linguistic structure they assume. Kretzschmar admonishes against:

> [...] dialectologists leaping past interpretation of their particular data because they think they already know what it should mean [...] in order to apply it to issues of culture, language standards and even population genetics.

With regard to the tendency to be unreflective about geography, Kretzschmar notes that too many works attempts to identify dialect areas, while ideas about dialect continua also support the analysis of language variation. The issues concerning the proper organizing concepts in geography deserve further attention. Goebl (this volume, pp. 411–435) explores the link with geography in a novel way (see subsequently in this article as well).

### 3.1.2 *Goebl on current Salzburg work*

As we noted in the introduction to our last collection of articles on dialectometry (Nerbonne and Kretzschmar, 2003), Hans Goebl is the person most—indeed, almost single-handedly—responsible for the shift in scholarly opinion that has brought dialectometrical techniques to the innovative forefront of work on linguistic variation in the last decades of the twentieth century (Chambers and Trudgill, 1998, [[1]1980], pp. 140–48). Goebl (1982, 1984) elaborated extensively on basic dialectometrical ideas that Séguy had introduced, and demonstrated their scientific potential in detailed analyses of the data of the *Atlas Linguistique de France* (ALF).

It was therefore more than appropriate that Goebl was a plenary speaker at *Methods XII*, and the only invited speaker at the workshop *Progress in Dialectometry: Toward Explanation*. Goebl titles his contribution 'Recent Advances in Salzburg Dialectometry', and his article will be a most useful starting point for readers unfamiliar with the basic workings of dialectometry. Goebl reviews

the basic data normally available for analysis, its organization into a place × feature matrix in which each cell provides the realization of a linguistic feature at a particular place. He also reviews the calculation of differences between place vectors, including potential refinements. From this he introduces the notion of a similarity distribution with respect to a single place (data collection site). A site is represented by a vector of values for each of the features tested during the data collection period. We calculate the similarity of this site to every other site in the collection. Goebl then shows how the simple properties of these distributions reveal the geolinguistics of the region, focusing on the maxima of the distributions, which he interprets as 'dialect kernels', and the (right) skew, which he interprets as evidence of a transition area.

It will not be obvious to every reader that this is more than an elegant summary of a mature methodology, and the title is too modest in this respect. But Goebl continues his line of examining variation from the entire range of positions in a language area, extending the analysis now to a calculation of the correlation between linguistic distance and geography. In fact, the calculation of the correlation between linguistic and geographic distance is standard in quantitative linguistics, introduced by Séguy (1971) and Cavalli-Sforza and Wang (1986). But Goebl examines the correlation from each individual place in the distribution, demonstrating that the degree of correlation is geographically conditioned! From some perspectives the correlation is high and from others, low; moreover, these places of high and low correlations show striking geographic coherence.

### 3.1.3 *Haimerl on VDM: Visual Dialectometry*

It should be clear that Goebl's analyses—just as those in the other contributions to this volume—would not be possible without extensive computational support. The examples from Goebl's articles involve the calculation of over 600 distributions of linguistic similarity, as well as the descriptive statistics associated with them. One might indeed speculate that the rather slow reception of dialectometrical ideas in the 1970s and 1980s may have been due to the relative difficulty of performing the required analyses. Students of dialectology or language variation need no longer face such difficulties, however, thanks to the work of Edgar Haimerl.

Haimerl has developed Visual DialectoMetry (VDM), a freely available software package for the storage, management, analysis, and visualization of dialectometric data. The package is particularly aimed at supporting the exploration of dialect data, and most particularly at supporting exploration via the visualization through maps. It uses a simple database for the storage and management of the data, facilitating different views. Haimerl sketches the database design used in VDM, emphasizing its support of the needed flexibility.

In order to support efficient exploration of dialect data, it is necessary to redraw maps within a second or so, allowing the researcher to shift rapidly the perspective from one place within the language area to another. For speed in visualization, two-dimensional (place × place) similarity data is cached in packed, single-dimensional form. Support is provided for using geographic information systems such as MapInfo together with VDM.

At this moment we know only of Haimerl's VDM package and Peter Kleiweg's L04 package (www.let.rug.nl/kleiweg/L04), which are freely available for use in dialectometry. Kleiweg's Linux-based package is focused on pronunciation analysis, and Haimerl's MS Windows-based work on the analysis of categorical (nominal) data. They provide excellent opportunities for first experiments of those interested in attempting dialectometrical analyses of their own.

## 3.2 Linguistic structure

As we noted in Section 1, many linguists find dialectology fascinating for the opportunity it provides to put linguists' concepts to good use. Some linguists are therefore impatient with the tendency in dialectometry to aggregate many differences. This clarifies the relative degree of difference among the different sites and areas, but it normally assigns no pride of place to important linguistic concepts such as phonemic or structural differences, as opposed to lexical differences. But this is not inherent to dialectometry as

Shackleton (2005) and also the three contributions in the current section demonstrate.

### 3.2.1 *Clopper and Paolillo on American vowels*

Clopper and Paolillo have collected acoustic data on fourteen vowel phonemes from six regional varieties of American English as spoken by both men and women. They focus on vowel duration and on the first two formant frequencies, as is the standard. (Formants are resonant frequencies and they vary according to the shape of the mouth, which in turn varies according to the vowel being pronounced.) It is clear that there are substantial differences both among regions and between men and women, but the authors focus not on the accumulated differences (as might be expected in dialectometry) but rather on extracting the linguistic structure implicit in the data collection. They report on a preliminary study in which they verified that factor analysis would function as wished. In that study, vowels were treated as cases, and the five variables were vowel duration as well as the first two formant frequencies at two different sampling times, and factor analysis was indeed successful in reducing the five variables to two independent factors.

The second study uses the same data, but aims at detecting geographically conditioned variation, even while retaining the men and women speakers. The authors have recordings of four token sets from each speaker, where a token set included one pronunciation of each of fourteen vowels, yielding a total of 170 cases (some data were missing). They recorded values for seventy variables, viz., the frequency of the first two formants at two sample points as well as the duration. In this case, the vowel token sets, each corresponding to a single speaker's production, were analyzed as the cases.

Clopper and Paolillo were able to extract five interpretable factors, the first two of which corresponded to the first two vowel formants, confirming the importance of these vowel characteristics. These two factors together were also sufficient to distinguish men's from women's speech, and there was a little evidence of regional variation. Vowel duration was the focus of the third factor, which also reflects regional variation, Southern speakers using longer vowels. The fourth factor indicates back vowel fronting, which is a regional feature of the South and West. Finally, the fifth factor indicates the lowering of some mid and low vowels, part of a shift in the pronunciation of the northern American cities on which Labov *et al.* (2005) has written extensively. It is particularly striking to see these patterns emerge from purely acoustical data.

It is interesting to note that Clopper and Paolillo come to dialectometry from a rather different direction than most. While most researchers look to aggregation as a means of smoothing the data distributions in order to see regional tendencies (or other conditioning) more clearly, Clopper and Paolillo mention a 'structuralist' motivation—the wish to examine vowels in terms of entire vowel systems, thereby attending to the perceptual problem of how the vowels are distinguished. Studies of individual vowels, no matter how exhaustive, cannot do justice to the perceptual perspective.

### 3.2.2 *Nerbonne on Southern American vowels*

While Clopper and Paolillo work on a recently collected corpus for which acoustic data are available, the *The Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS) is a rich reservoir of data on linguistic variety in the eastern United States that were collected in the period 1933—74, and for which no recordings were made. There are transcriptions, however, and Nerbonne's analysis proceeds entirely from these.

The analysis proceeds by first characterizing vowel differences via a difference measure applied to feature descriptions of the vowels. Then for each vowel token, e.g. the first vowel in *afternoon*, Nerbonne derives a place × place matrix characterizing the dialect difference between the data collection sites with respect to that vowel token. In the spirit of dialectometry, the analysis proceeds according to linguistic distances between sites—only now, the distances represent not aggregates of hundreds of items, but rather single vowel tokens.

From such matrices, it is a direct step to calculating the vowel-token correlation matrix. This is simply the vowel-token × vowel-token matrix in which each cell represents the correlation of one vowel token with another. (This step is

implicit in Clopper and Paolillo's work, mentioned earlier.) Given this correlation matrix, Nerbonne performs factor analysis to extract those tokens which correlate highly with one another. The study thus represents an attempt to search for linguistic patterns based primarily on the function of the linguistic elements to support dialect differentiation. While Clopper and Paolillo use acoustic measurements of vowel tokens, Nerbonne effectively uses the function of the token to distinguish dialect sites, and while Shackleton (2005) uses a database of linguistic characterizations, Nerbonne's procedure works directly from the atlas transcriptions.

The results are mixed. To begin with, a great deal of the data appear to be quite unsystematic, so that one would need to examine thirteen factors in order to explain a great deal of the variance in the data (say 60%), even while their interpretation is problematic. A further indication of the complexity of the data is the fact that over sixty vowel tokens (of 200) are not associated strongly with any single factor.

But Nerbonne focuses on the three most important factors, and is able to show that they highlight a number of significant pronunciation differences in the LAMSAS. While most of those involved similar vowels in very similar environments, the factors also grouped some surprising vowels together, suggesting that this approach might support the search for more explanatory linguistic explanations.

### 3.2.3 *Gooskens and Heeringa on linguistic levels*

Anyone who works on a subject as complex as dialectology is soon impressed by the many facets (technically, DIMENSIONS) of linguistic variation. Normally, dialect speakers signal their 'linguistic provenance' in a multitude of ways, meaning that analysts have a large choice of linguistic dimensions to focus on. Dialectometric analyses often aggregate at least dozens, often hundreds of linguistic variables, so that it is never surprising to find some sort of geographic conditioning. This forces dialectometry to be reflective about the validation of its results. Which linguistic variables are most important, and which are merely correlates of the important variables?

Gooskens and Heeringa have settled on a validation step which compares dialectometric techniques to dialect speakers' judgments of dialectal affinity. They are fortunate in having at their disposal a substantial collection of data, recordings and the transcriptions of Norwegian dialects made by Jorn Almberg and Kristian Skarbo at *Norges teknisk-naturvitenskapelige universitet* (NTNU), the Norwegian University of Science and Technology, Trondheim (www.ling.hf.ntnu.no/nos/). Gooskens and Heeringa (2004) then report on lay dialect 'speakers' perceptual judgments of the linguistic proximity of those dialects to their own. Note that the brief samples played for the lay dialect speakers of necessity display differences at all linguistic levels. Gooskens and Heeringa ask how important different levels are, focusing on segmental pronunciation differences (i.e. individual sounds such as /p/ or /i/), lexical differences, and prosodic (i.e. melodic) differences.

This means that Gooskens and Heeringa begin with a place-by-place matrix reflecting how different lay speakers of place $p$ found speech samples from place $p'$. They attempt to explain this via objective measures of the three different linguistic levels. They conclude that, even though there are important correlations between subjective distance and all of the objective measures examined, only segmental pronunciation is truly explanatory since the others are too collinear (with pronunciation) to be shown to contribute independently. This is an excellent example of the sort of analysis available only to the aggregating techniques available in dialectometry.

## 3.3 New frontiers

It is most welcome to see the attention (mentioned earlier) increasingly paid to the linguistic structure that aggregate analysis implicitly builds on, but this is only one of many new developments in dialectometry. Dialectometry initially treated all linguistic levels on a par, normally focusing on lexical, morphological, and pronunciational differences. A small number of syntactic variables have been included in analyses, but Spruit's article (this volume, pp. 493–506) shows convincingly

that the basic dialectometric techniques may be fruitfully applied to purely syntactic data. Manni, Heeringa, and Nerbonne examine genetic and linguistic variation simultaneously, noting that the models are similar in showing a sublinear relation between geography and (linguistic or genetic) variation, but that there is no correlation between genetic and linguistic variations once one controls for geography. Cichocki applies correspondence analysis to data from French Canadian varieties in intense contact with English, and Gooskens and van Bezooijen apply dialectometric techniques to model the comprehensibility of Dutch and Afrikaans, demonstrating, incidentally, the asymmetry of comprehensibility.

### 3.3.1 *Spruit on syntactic variation*

Spruit is fortunate in proceeding from a large database of syntactic differences collected by the SAND project coordinated by the Meertens Institute in Amsterdam (www.meertens.nl/sand/). 'SAND' is an acronym standing for 'Syntactic Atlas of Netherlandic Dutch' (Barniers *et al.*, 2001, 2006). The data reflect the customary syntactic means of expression of 134 syntactic variables (such as the means of expressing a reflexive relation such as *he washes himself*) in 267 different data collection sites in the Netherlands. The choice of which data to collect rested with theoretical syntacticians and specialists in dialect syntax, and the collection itself was conducted by trained field workers.

Spruit is the first to apply dialectometric techniques to a large collection of syntactic data, and that makes his work especially interesting. Some of the theoreticians involved in designing the data questionnaire used to collect the data thay analyze were openly sceptical about whether geographic cohesion would turn out to be a conditioning factor at all in the case of syntactic variation, given the substantial body of work demonstrating the internal pressures on syntax to conform to typological (or universal) constraints (Comrie, 1989; Croft, 2001), and the question Spruit poses here is whether syntactic data indeed display geographic cohesion.

As the reader may see independently, Spruit's analysis leaves little doubt that there is a strong association between geography and syntax. A series of maps illustrates the tendency of the syntactic variables to be realized similarly in nearby places. This is a significant result.

But the proof of a strong association between geography and syntax is by no means last word on the conditioning of syntactic variables. In addition to the typological or universalist perspective noted earlier, there is also work proposing a historical explanation for shared syntactic similarity. Thus, Longobardi (2003) already proposes that syntactic structures might be more resistant to change than other linguistic features, and therefore provide a window to a more distant linguistic past. Dunn *et al.* (2005) borrow techniques (and even software) from phylogeny in order to reconstruct language families in New Guinea. Phylogenetics attempts to view a set of contemporary varieties as the result of divisions which are analogous to those in the spontaneous mutations found in biology, additionally allowing for some shared mutation as a result of adopting or borrowing genetic/linguistic material. Future works must certainly include comparisons between these different approaches to explaining similar linguistic structure.

### 3.3.2 *Manni, Heeringa, and Nerbonne on linguistics and genetics*

Cavalli-Sforza (1996) is famous for his speculation that the fates of populations may be traced through their genes, and in particular, how widely they are spread. It has been a part of this vision that one likewise map the spread of languages (Renfrew, 1992), but to date, most of the work has focused on shared vocabulary as evidence of linguistic relatedness, and it has focused on rather large areas.

Manni, Heeringa and Nerbonne apply dialectometric techniques for measuring the similarity of word pronunciations to obtain a characterization of the similarity of varieties in specific towns and villages, showing a novel way to characterize linguistic relatedness. They obtain a characterization of the genetic relatedness by measuring the degree to which surnames are shared. The first reaction of many to the suggestion that one use names to measure genetic relations is to think of the many ways children might not bear the names of their biological fathers, but the validity of the technique has been tested and shown to correlate highly

with independent characterizations of genetic relatedness. A second novel aspect of the study by Manni *et al.* is that they focus on the Netherlands, a relatively small area compared to those studied earlier.

It turns out that linguistic and genetic dissimilarities are highly correlated ($r = 0.4$), but it is always hasty to conclude that the influence must be direct—in this case, that might be concluding that language variation was acquired from the biological parents. In fact, once Manni *et al.* factor out geography as a common factor of influence, then the correlation disappears. So, for the Netherlands, at least, geography influences both genetic and linguistic variation massively, leading to a correlation between the two, but there appears to be no further link.

An interesting advantage of this interdisciplinary study is the applications of techniques from biology, specifically from population genetics, to linguistic problems. For example, Manni *et al.* apply the Monmonier algorithm to their linguistic data in order to identify linguistic areas. This technique tries to sketch borders wherein the linguistic distance is greater than would be expected, on the basis of a simple geographical model (on the basis of the residuals in a regression analysis). To date, most techniques for finding dialect groups suffer from statistical instability, especially clustering and self-organizing maps.

An interesting new development is the application of phylogenetic analysis as developed in biology to linguistics, in an effort to reconstruct linguistic history, but to date, most of this work has also been limited to shared vocabulary (McMahon and McMahon, 2005), in addition to a single piece of work focused on syntax (Dunn *et al.*, 2005). But the opportunities are growing for more serious interaction between linguistics and biology.

### 3.3.3 *Cichocki on Acadian French /r/*

Wladyslaw Cichocki studies the consonant /r/ in Acadian French, i.e. the French spoken in the Canadian Maritime Provinces of New Brunswick, Prince Edward Island, and Nova Scotia. He uses a corpus which was taken from fifty-four speakers in eighteen localities, and which includes over 5,000 tokens of the phoneme /r/. The data is derived from a large linguistic atlas focusing on the vocabulary derived from fishing. Pronunciations were transcribed from audio recordings under the supervision of a single phonetician. Cichocki's study is innovative in his use of correspondence analysis in order to identify the factors conditioning the various pronunciations of /r/ (so we might have discussed this study in the section, mentioned earlier, on linguistic structure), but it is also innovative in the context of dialectometry in its explicit attention to the effects of a contact language; in this case the effect of English on French.

The consonant has four variants in pronunciation, the dorsal, or uvular trill [ʀ] (pronounced at the back of the mouth), the apical trill [r] (pronounced with the tip of the tongue), the English rhotic approximant [ɹ], and the zero variant, in which the phoneme is not realized at all, and which is limited to final position. The apical [r] is the oldest Acadian pronunciation, which is being replaced by the uvular [ʀ], the standard pronunciation in France and in several other parts of Canada. The rhotic approximant is a clear interloper from neighboring English-speaking areas. But this is a very global sketch, and Cichocki demonstrates that the details are quite subtle.

Cichocki demonstrates directly that the English [ɹ] is found predominantly in English loan words—no surprise here. The same simple comparison shows that the zero variant is almost completely missing from the English loan words. The more interesting part of his study is devoted to applying correspondence analysis to the data on /r/. Correspondence analysis is applied to the vector of frequencies of fourteen different sorts of environments, namely the three variants of /r/ at the beginnings of French-word syllables, the four variants at the ends of French-word syllables, the three variants at the beginnings of English-word syllables, and finally the four variants at the ends of English-word syllables. Each vector corresponds to a place, and the differences between places can now be calculated in various ways (Cichocki uses $\chi^2$). The resulting distances are then subjected to a dimension-reducing procedure, which results in a compact characterization of the relations among the different sites. Even though Cichocki does

not speculate on applying this procedure to very large characterizations of dialect distances, it is clear that it is similar in spirit to the analyses Goebl surveys in his work, particularly if one combines these with multidimensional scaling (Kruskal *et al.*, 1971; Black, 1976), as is frequently done.

Cichocki goes on, however, to show how correspondence analysis naturally analyzes not only the sites in the site × feature matrix, but also the (linguistic) features, giving this sort of analysis a clear advantage over some of the others in its ability to recognize linguistic structure. In particular, Cichocki is able to characterize how much variance different linguistic alternations account for. Finally, Cichocki demonstrates how other explanatory hypotheses may be examined in this framework.

### 3.3.4 *Gooskens and van Bezooijen on comprehensibility*

Most dialectometrical work has focused on what one might call the SIGNAL OF LINGUISTIC PROVENANCE, i.e. the bits of variation we consciously or unconsciously use and that signal where we are from, at least in a linguistic sense. Gooskens and Heeringa (mentioned earlier) show that these signals are indeed received, and not merely available in principle. In addition, they suggest that dialectometry concentrate on the signals that lay dialect speakers are sensitive to.

A great deal of variation is mild enough that communication is unperturbed, and this allows us to concentrate on the signals that allow inferences about the speaker's background. But variation may be so extreme that it, in fact, disturbs communication, an experience many travelers have had. Everyone reading this introduction is fairly capable in English, but a trip through the English countryside, Scotland, Ireland, the American South, India, or Jamaica would normally result in several encounters in which interlocutors understand each other less than perfectly. This is the problem of COMPREHENSIBILITY, familiar in dialectology, and which Gooskens and van Bezooijen subsequently examine.

In particular, they investigate Dutch and Afrikaans, two very closely related languages (Afrikaans grew out of the Dutch spoken by the earliest settlers in South Africa, who emigrated four to five centuries ago). By focusing on the written form, the authors expect to find that speakers (readers) comprehend each others' languages fairly well. They test comprehension using newspaper articles and measure it with standard techniques from foreign language learning research. They show first that attitude plays a (weak) role in predicting comprehension, but are also able to show that several dialectometric techniques are promising, including the proportion of shared vocabulary, and the transparency with which shared vocabulary could be recognized.

It will not be apparent to nonspecialists that Gooskens and van Bezooijen's work exposes a methodological shortcoming in dialectometry, but it does, at least if one is willing to accept Dutch and Afrikaans as dialects of a single language, which is linguistically unobjectionable. Their results show that it is easier for the Dutch to understand written Afrikaans than it is for South Africans to understand written Dutch, i.e. that there is an asymmetry in comprehensibility. Dialectometrical techniques normally establish measures of varietal DISTANCE, however, which is axiomatically symmetric. The gauntlet they throw down to dialectometric theory is therefore to develop mathematical and computational models of varietal remoteness which allow for asymmetry.

## 4 Conclusions and Prospects

The articles in this volume indicate progress in dialectometry, not only in the technical means for analysis of complex variation data, but also in the application of these analyses to different kinds of problems in linguistics. We believe that such healthy beginnings, for dialectometry is still in its infancy even given such evident advances from the pioneering work of Séguy (1973), will be followed by a rich harvest in our understanding of language

in use. We look forward to continued advances in the field, both (as Kretzschmar suggests) in the techniques of its art and also as its results bear upon the science of linguistics—with far-reaching applications for geographical, historical, and social issues.

## Acknowledgements

We thank the Association for Literary and Linguistic Computing for their generous contribution toward the workshop's costs that led to this collection of works, and also for the prizes they awarded to Cynthia Clopper and Marco Spruit as 'excellent young researchers'.

We are also very grateful to our referees, Renée van Bezooijen, Isidore Dyen, Charlotte Gooskens, Wilbert Heeringa, Peter Kleiweg, Hermann Niebaum, Bob Schackleton, Marco Spruit, Nathan Vaillette, and Hans Van de Velde. This collection would not have been possible without their careful and timely work.

The Netherlands Organization for Scientific Research (NWO) subsidized the work reported here, NWO grant 360-70-120 (P.I. J. Nerbonne). Thanks, too, to Alan Wilcox for keeping the Gershwins music alive!

## References

**Barbiers, S., Cornips, L., and van de Kleij, S.** (eds), (2001). *Syntactic Microvariation*. Amsterdam: Meertens Institute. *elec. publication*.

**Barbiers, S.** *et al.* (eds) (2006). *Dynamic Syntactic Atlas of the Dutch Dialects (DynaSAND)*. Amsterdam: Meertens Institute. http://www.meertens.nl/sand/

**Black, P.** (1976). Multidimensional scaling applied to linguistic relationships. In Dyen, I. and Jucquois, G. (eds), *Lexicostatistics in Genetic Linguistics: Proceedings of the Montreal Conference*, Vol. 3/5–6 of *Cahiers de l'Institut de Linguistique de Louvain*. Leuven: Centre de Recherches Mathematiques Université de Montreal, pp. 43–92.

**Cavalli-Sforza, L. L.** (1996). *Gènes, peuples et langues*. Paris: Jacob.

**Cavalli-Sforza, L. and W. S.-Y. Wang.** (1986). 'Spatial distance and lexical replacement.' *Language,* **62**: 38–55.

**Chambers, J. and P. Trudgill.** (1998 [[1]1980]). *Dialectology*. Cambridge: Cambridge University Press.

**Comrie, B.** (1989). *Language Universals and Linguistic Typology: Syntax and Morphology*. Oxford : Basil Blackwell.

**Croft, W.** (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

**Dunn, A. M., Terrill, A., Reesink G., and Levinson, S.** (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science,* **309**(5743): 2072–75.

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

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

**Gooskens, C. and Heeringa, W.** (2004). Perceptual evaluation of levenshtein dialect distance measurements using norwegian dialect data. *Language Variation and Change,* **16**(3): 189–207.

**Kruskal, J. B., Dyen, I., and Black,** P. (1971). Some results from the vocabulary method of reconstructing languages trees. In *Lexico-Statistics in Genetic Linguistics*. New Haven: Yale University.

**Labov, W.** (1994). *Principles of linguistic change*. Vol. 1, *Internal factors*. Oxford: Blackwell.

**Labov, W., Ash, S., and Boberg, C.** (2005). *The Atlas of North American English: Phonetics, Phonology and Sound Change: A Multimedia Reference Tool*. Berlin: Mouton de Gruyter.

**Longobardi, G.** (2003). Methods in parametric linguistics and cognitive history. *Linguistic Variation Yearbook*, **3**:103–40.

**McMahon, A. and McMahon, R.** (2005). *Language Classication by the Numbers*. Oxford: Oxford University Press.

**Milroy, L. and Gordon, M.** (2003). *Sociolinguistics: Method and Interpretation*. Oxford: Blackwell.

**Nerbonne, J. and Kretzschmar, W.** (2003). Introducing computational methods in dialectometry. *Computers and the Humanities,* **37**(3):245–55. Special issue on computational methods in dialectometry, Nerbonne, J. and Kretzschmar, W.I.,Jr. (eds).

**Niebaum, H. and Macha, J.** (2006 [1999]). *Einführung in die Dialektologie des Deutschen, 2te, neubearbeitete Auflage*. Tübingen: Niemeyer.

**Renfrew, C.** (1992). Archaeology, genetics and linguistic diversity. *Man,* **27**(3): 445–78.

**Séguy, J.** (1971). La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane,* **35**(138): 335–57.

**Séguy, J.** (1973). La dialectometrie dans l'Atlas linguistique de Gascogne. *Revue de Linguistique Romane,* **37**(145): 1–24.

**Shackleton, R. G.Jr** (2005). English–American speech relationships: a quantitative approach. *Journal of English Linguistics,* **33**(2): 99–160.