Language Variation Studies and Computational Humanities

John Nerbonne, Charlotte Gooskens, Sebastian Kürschner, and Renée van Bezooijen

The volume we are introducing here contains a selection of the papers presented at a special track on computational techniques for studying language variation held at *The Thirteenth International Conference on Methods in Dialectology* in Leeds on Aug. 4-5, 2008. We are grateful to the conference organizers, Nigel Armstrong, Joan Beal, Fiona Douglas, Barry Heselwood, Susan Lacey, Ann Thompson, and Clive Upton for their cooperation in our organization of the event. We likewise owe thanks to the referees of the present volume, who we are please to acknowledge explicitly: Agnes de Bie, Roberto Bolognesi, David Britain, Cynthia Clopper, Ken Decker, Anders Eriksson, Hans Goebl, Stefan Grondelaers, Carlos Gussenhoven, Nynke de Haas, Frans Hinskens, Angela Kluge, Gitte Kristiansen, Alexandra Lenz, Maarten Mous, Hermann Niebaum, Lisa Lena Opas-Hänninen, Petya Osenova, John Paolillo, Louis Pols, Helge Sandøy, Bob Shackleton, Felix Schaeffler, Marco Spruit, Rint Sybesma, Nathan Vaillette, Gunther de Vogelaer, and Esteve Valls

The conference track consisted of twenty-four papers and posters, including a keynote address by Vincent van Heuven on phonetic techniques for studying variation and comprehensibility. Fourteen contributions were selected for publication in this special issue of the *International Journal for Humanities and Arts Computing*, including van Heuven's. In addition the conference track featured a panel session reflecting on the introduction of computational techniques to the study of language variation and more generally, on computing and the humanities. We have prepared a report on the panel session for publication here as well.

In the remainder of this article we sketch variationist linguistics as a subfield within the discipline of linguistics and relate how we see the path that led to computational studies occupying a modest place in this branch of linguistics. Our intention is that the present introduction provide a context within which the more specialized contributions can be better appreciated.

More importantly for those especially interested in humanities computing, we sketch the contributions of this volume collectively as an example of what we might refer to as an ENGAGED HUMANITIES COMPUTING, which we intend as a contribution to the ongoing debate about how computational work can best be integrated into the humanities disciplines (Nerbonne, 2005; McCarthy, 2005). We shall elaborate on this further below, but we mean in particular that it has been the strategy of computationalists working in language variation that they primarily address existing questions from this sub-discipline, that they attempt to solve existing analytical problems, that they compare their results to those of non-computational scholars, and that they examine their methods and results critically from the perspective of the sub-discipline. The goal is to have computational techniques accepted into the toolbox that the sub-discipline normally recommends.

Variationist Linguistics

Linguistics is the scientific study of language, and VARIATIONIST LINGUISTICS studies the variation found within languages especially variation that is geographically or socially conditioned. DIALECTOLOGY is one of the oldest branches of linguistics, focusing especially on the way language varies geographically. SOCIOLINGUISTICS focuses on the social conditioning of variation. Linguistic variation was studied early on for the clues it suggested for the manner in which language changes diffuse geographically, but it is clearly fascinating to a large number of people, judging by the interest it inspires in learned and lay audiences. In the 1960s Labov demonstrated that variation not only existed along social lines (as well as along the above mentioned geographic lines), but also that these same social lines likewise demarcated the path of change for some linguistic innovations (see Chambers and Trudgill, 1998:Chap. 4). In this way the study of dialectology was expanded to include not only geographic, but also social variation. Indeed, some date the birth of sociolinguistics from this period (although there was clearly interest in issues involving language and social structure earlier, as well). Contemporary linguists who work on dialectology are normally interested in social variation as well, justifying our referring to this subfield as variationist linguistics.

Variationist linguistics sees dialects and sociolects as elements of culture, and has long been interested in the degree to which linguistic culture follows the same paths as technical, political, and aesthetic culture. Gilliéron, one of the earliest French dialectologists, linked variation in French (*langue d'oc* vs. *langue d'oil*) to differences in architecture, agricultural practice and legal institutions (Chambers and Trudgill, 1998: 95-103). This makes language variation study interesting from the point of view of more general studies in human culture, the humanities.

Data collections

A major achievement of dialectology has been the compiling of substantial data collections in structured form, normally referred to as DIALECT ATLASES. Although no one rivals Georg Wenker (1881), who collected questionnaires from over 48,500 German towns and villages on the pronunciation of words in some 40 different sentences at the end of the nineteenth century (Niebaum and Macha, 2006:58ff), still dialectologists have always been concerned to base their work on large collections of comparable data. (See the *Deutscher Sprachatlas*,1927-1956, published by Ferdinand Wrede and others, for 128 of the maps resulting from Wenker's survey, and see http://www.diwa.info/ for an internet presentation of the whole atlas.) By 'comparable' we mean, e.g., that the pronunciations of the same words are collected at a range of sites, or the words for a given concept, or the syntactic means of expressing something more abstract. Many of the papers in this volume are based on substantial samples culled from dialect atlases, and it is easy to imagine how attractive it is to apply computational techniques to such large bodies of data, and conversely, how difficult and frustrating it is to attempt to analyze such data sets without the benefit of extensive computation.

Séguy and Goebl

The early pioneers in the application of computing to problems in dialectology were Jean Séguy (1973) and Hans Goebl (1982), who knew of each other's work, but appear to have developed their ideas independently. Their emphases lay in the

measurement of dialect differences, which explains their preference for the term 'dialectometry'. Goebl also emphasized the application of novel statistical techniques, which at that time had only recently been developed for the analysis of categorical data. They pioneered not only the use of the computer in dialectology for the purpose of managing and analyzing large volumes of data, but Séguy was (and Goebl is) still very much a dialectologist himself. So computational dialectology began auspiciously, initiated by practicing dialectologists, very much familiar with problems in the field, such as the failure to develop a methodology for the selection of which linguistic material to use for samples, the lack of a theoretical foundation for commonly used notions such as 'dialect area', and the relative inattention to more abstract questions, e.g. about how geography influenced language variation and how it might be related to other issues in the diffusion of cultural innovations. All of these issues have been addressed in more detail thanks to the deployment of computational techniques in dialectology. See Nerbonne (2009) for a review of the advantages of an aggregate perspective in analyzing linguistic variation and Nerbonne and Heeringa (to appear) for a review of techniques suitable for measuring linguistic variation.

The Computational Issues

The basic computational issues facing the researcher who would like to contribute to dialectology have been addressed, but in the interest of potentially stimulating improvements, perhaps from other humanities disciplines, and in the interest of indicating where solutions in dialectology might be useful to others, we shall review the issues here briefly. To begin there are serious problems in choosing a format suitable for processing and storing research data. Modern dialectological collections should include the original recordings of fieldworkers, meaning the audio and perhaps even video recordings of linguistic data, and this needs to adhere to standards only now emerging for such multimedia data. Two projects are worth special note for their efforts in promoting standards and good practices for recording and archiving linguistic material: first, the Max Planck Institute's Dobes project (*Dokumentation Bedrohter Sprachen*, http://www.mpi.nl/DOBES/), and second, the E-MELD project, sponsored by the U.S. National Science Foundation and endorsed by the Linguist List. E-MELD stands for the "Electronic Metastructure for Endangered Languages Data" (http://emeld.org/). In addition to issues concerning multimedia data, the dialectologist often faces issues concerning multilingual data, data in various writing systems, and data in phonetic transcription. We refer to the project sites above for advice on these matters as well.

Naturally, the problems in representing and storing data are preliminaries with respect to the questions of how one should analyze dialectological data. We can usefully distinguish the basic measurement of similarity (or dissimilarity) from further steps in the analysis. Since we view similarity and dissimilarity as each other's converse, we will normally mention only one from now one, assuming we can always derive the other via the application of a suitable inverse. If we view a dialectological survey as producing an array of categorical responses **a**, then the bulk of work has been done using HAMMING DISTANCE as a measure of site distance $d(\mathbf{a},\mathbf{b}) = \sum_i \mathbf{a}_i \neq \mathbf{b}_i$, i.e. the number of survey items for which **a** and **b** differ. Goebl (1982) adds an inverse frequency weighting to this, reasoning that common infrequent responses are especially reliable indicators of dialect affinity. Naturally, this sort of measure is appropriate for categorical data of non-linguistic origin – preference for first names,

styles of clothing or architecture, religious affiliation, or any number of other cultural markers. It would therefore be an excellent candidate for measurement in other areas (see above) where one wished to gauge cultural overlap. Speelman and Geeraerts's paper in the present volume suggests that this simple Hamming measure of dialectological affinity ought to be normalized with respect to the heterogeneity one should expect on the basis of conceptual grounds.

Language is also highly structured, so more recent developments have emphasized measures of difference that are sensitive to linguistic structure. Nerbonne and Heeringa's (to appear) survey article reviews a substantial body of work applying variants of EDIT DISTANCE (also known as string distance, or Levenshtein distance) to pronunciation data. Edit distance counts the number of edit operations – insertions, deletions or substitutions – needed to transform one sequence to another. It is sensitive to linear order, and thus introduces a first level of structural sensitivity into the measurements. Several articles in the present volume use edit distance, and one (Yang and Castro) explores a refinement for detail with languages with tonal distinctions. This sort of measure might be interesting for the analysis of other material with linear structure, e.g. folk melodies or sequences of ritual actions. But linguistic structure is much deeper than sequence comparison might detect, so it is not surprising to see other attempts at structurally sensitive comparison. Resonant frequencies (FORMANTS) are well-established phonetic characterizations of vowels, which have resisted automatic extraction, but Leinonen (this volume) applies techniques to detect these with no manual intervention, paving the way to more sensitive large-scale comparisons. Maguire (this volume) proposes techniques sensitive to PHONEMIC structure (the system of sounds capable of marking differences in meaning), and Van Heuven (this volume) explores the degree to which comprehensibility can serve as a basis for studies in language variation. Gooskens, Heeringa and Beijering's and Kürschner, Gooskens and van Bezooijen's papers examine the relation between comprehensibility and other measures of linguistic distance. Montemagni's paper investigates the relation between pronunciation and vocabulary and Szmrecsanyi's uses data from MORPHOLOGY (word forms) and SYNTAX (the structure of phrases and sentences).

If the basic measurements of distance or similarity have been fairly simple throughout the history of dialectometry, the further analysis of differences has been quite varied. Clustering has been popular, as has been a range of dimension reduction techniques, including especially multidimensional scaling, principal component analysis and factor analysis. Prokić and Nerbonne (this volume) and Leino and Hyvönen (this volume) continue investigation into these areas. This line of investigation is important for several reasons: first, as means of probing how geography influences cultural variation; second, in order to explore the linguistic space with an eye toward detecting co-varying linguistic elements, perhaps with an eye toward investigating structural effects; third, for the opportunity it affords to compare the computational work with earlier, non-computational work, most of which presented its results in the form of maps showing areas of relative coherence.

Engaging the Discipline

An exciting aspect of work reported on in the current volume is the degree to which it engages other, non-computational aspects of variationist linguistics. This is interesting

in itself, offering opportunities to examine more fundamental issues in how language varieties differ, but it also illustrates one strategy for computational studies in the humanities, that of seeking close engagement with non-computational studies. Computational dialectology has regularly asked whether its techniques make sense from the perspective of the larger discipline (see McMahon and McMahon (2005) and the review by Nerbonne (2007) for an example of a dispute on whether computational techniques make sense linguistically). This is a primary point of engagement, and several papers focus on techniques that are presented and presumably evaluated on the basis of what they show about the geographic distributions of linguistic variation. Yang and Castro's paper on incorporating lexical tone differences into sequence distance measures, Prokić and Nerbonne's paper on clustering, Leino and Hyvönen's paper on statistical dimension reduction techniques, Leinonen's paper on extracting vowel system profiles from acoustic data, and Speelman and Geeraerts's paper on compensating for inherent conceptual tendencies influencing variation all fall into this category.

But naturally we are also interested in how computational results shed light not on how to detect variation and the influence geography has on it, but rather on further central questions of the sub-discipline, and this sort of question forms a second point of engagement.

One line of work which illustrates the strategy of engaged computational work is illustrated by the studies on comprehensibility, for example, Gooskens, Heeringa and Beijering's paper. The interesting added dimension in their work is the functional perspective it adopts. Linguistic variation signals social and geographic provenance, but it is overlaid on a system whose primary purpose is undoubtedly communication. If variation is too extreme, communication deteriorates. Gooskens et al. contrast computational measures of pronunciation and vocabulary difference in how well they predict difficulties in comprehension, showing that the pronunciation differences present the more serious problems (at least in this material). Maguire's paper on the lexical distribution of variation has a similar strategic aim, that of tracking which changes result in differing distributions of phonemes, i.e. sounds capable of distinguishing meaning. Maguire's work is similar to Moberg et al.'s (2007) study, but he uses a measure which allows symmetrical comparison of varieties, and applies the it to a range of English accents, while Moberg and his colleagues studied the asymmetric case of "semi-communication" among Scandinavian languages.

Montemagni's paper illustrates an avenue from variationist study into historical linguistics. The paper examines the degree to which pronunciation and vocabulary co-vary and it closes with an interesting historical conjecture. Most explanations of linguistic diffusion assume that the avenues of diffusion should be similar (Bloomfield, 1933:18.12; Trudgill, 1974), and most earlier studies have shown similar distributions of differences in the different linguistic levels (Spruit, Heeringa and Nerbonne, 2009). Montemagni shows, however, that vocabulary differences correlate better with geography in Tuscany than pronunciation differences do, and conjectures that relatively recent pronunciation changes spreading radially from Florence lie behind the current patterns. The changes have not yet reached peripheral areas which are therefore more like each other than they are like immediately adjacent areas that have already been affected by the changes.

Heeringa, Gooskens, and De Smedt's paper superficially present a methodological correction to a problem in the validation of linguistic distance measures concerning the role of preconceived opinion in the perception of linguistic distance but it also opens deeper questions concerning the relation between the perception of distance and attitudes and opinions about distance that may be the result of mere hearsay. Are our perceptions influenced by our attitudes and prejudices concerning linguistic variation?

Papers in the present volume.

We briefly discuss each of the papers in turn in this special issue. We discuss first papers which examine varietal differences via their impact on intelligibility, i.e. the point where signalling social and geographic provenance begins to impede communication. Since signalling one's provenance is surely less important than communication, we appear to be dealing with a secondary function that begins to encroach on the primary one.

In his review "Making sense of strange sounds: (mutual) intelligibility of related language varieties. A tutorial" keynote speaker Vincent van Heuven advances mutual intelligibility as a functional operationalisation of a one-dimensional notion of remoteness between closely related language varieties. The greater the intelligibility of speaker A to listener B, the smaller the distance between variety A and B. A one-dimensional yardstick for linguistic distance is invaluable if we aim to come up with computational models of linguistic distance in which the contribution of lexical, phonological and syntactic differences between language varieties are to be weighed against each other in a comparison of symbolic input (i.e. distance measures derived from comparisons of transcriptions of vocabularies, grammars and texts). Moreover, Van Heuven notes that intelligibility is asymmetrical: speaker A may be more intelligible to listener B than vice versa. Van Heuven explicitly ignores the role of experience with the non-native variety. When the aim is to exploit intelligibility as a yardstick for linguistic distance between two varieties, the listeners should have no familiarity with the non-native variety. Intelligibility should be measured before any learning effects may have taken place. Given this simplification, listening to speech in a closely related language variety is quite like listening to noisy speech. The remainder of his review deals with possibilities of measuring intelligibility and considers how low-level deviations affect the process of word recognition – which he considers to be the central skill involved in understanding input speech in a closely related language variety. The contributions of deviations at the level of vowels, consonants and of word prosody (stress, tone) are considered. Van Heuven makes a convincing claim that much insight can be gained from related fields of inquiry, specifically from the study of the effects of foreign accent on the intelligibility of second-language speakers, and from computational models predicting intelligibility from mismatches between the non-native input language system on the one hand and the sound categories and lexical structures of the target language on the other. Models of perceptual assimilation of non-native sounds to the native categories, and of word recognition are reviewed. Although it is still impossible to predict the intelligibility of deviant input speech (for instance speech in a closely related variety), the problem seems soluble in principle: the development of adequate computational models predicting the intelligibility of deviant speech – such as speech in a closely related language variety – is mainly a matter of investing time and effort.

In “Phonetic and lexical predictors of intelligibility” Charlotte Gooskens, Wilbert Heeringa and Karin Beijering proceed from the strength of dialectometric methods such as the Levenshtein algorithm, which allows the researcher to measure objective aggregate distances between language varieties. These distances can be used for dialectometric purposes and have proved to be a powerful tool for the classification of dialects. Recent research has shown that Levenshtein distance is also a good predictor of intelligibility of closely related language varieties. However, since only aggregate distance measures have been applied so far, no conclusions could be drawn about the nature of the phonetic differences that contribute most to intelligibility. In their paper, Gooskens, Heeringa and Beijering measure distances between Standard Danish and each of 17 Scandinavian language varieties at the lexical level and at various phonetic levels. In addition, they conducted a listening experiment to assess the intelligibility of the 17 varieties to young Danes from Copenhagen. In order to determine how well the linguistic levels can predict intelligibility, they correlated the intelligibility scores with the linguistic distances and carried out a number of regression analyses. The results show that for this particular set of closely related language varieties phonetic distance is a better predictor of intelligibility than lexical distance. Consonant substitutions, vowel insertions and vowel shortenings contribute significantly to the prediction of intelligibility. Knowledge about the linguistic determinants of mutual intelligibility is useful for language planning at the national and European levels. If the less frequently spoken languages are to survive in a European context, it is important to gain knowledge about the mechanisms involved in using one’s own language for communication with speakers of other, closely related European languages. Knowledge of the role of different linguistic levels for mutual intelligibility is also useful for didactic purposes. It will make it easier to give specific instructions to people trying to gain the necessary passive knowledge needed to understand a language.

A further study on intelligibility is presented by Sebastian Kürschner, Charlotte Gooskens and Renée van Bezooijen, “Linguistic Determinants of Swedish Words among Danes”. They asked how one can predict which Swedish words will and will not be understood when spoken to Danes. Many Scandinavians have the interesting custom of speaking in their own languages in conversations with those who speak other Scandinavians languages natively. So a Dane and a Swede may carry on a conversation, each speaking his own native language. It would be natural to assume that the success of such “semi-communication” (Haugen, 1966) depends on an individual’s experience with the non-native language, but probably also on the degree to which native comprehension procedures may be applied to the foreign material. In fact, the authors cite earlier work noting that experience is a relatively poor predictor of comprehensibility. Kürschner et al. focus on the comprehension of isolated spoken words in contrast to most earlier work which concentrates on textual understanding, and they examine a wide variety of potential linguistic predictors, including first and foremost how similar the Swedish pronunciation of the word is to the corresponding Danish word’s pronunciation. But Danish has a conservative orthography that reflects the historical relation to Swedish rather more faithfully than pronunciation, and this might aid comprehension in the relevant cases, so this was also examined, as was the number of syllables, the density of the neighbourhood of words competing to be understood, and the frequency with which the words are heard in everyday speech. The extensive literature on spoken word recognition had shown the last two factors to be important in the understanding of words by natives. A particularly interesting

aspect of Kürschner et al.'s work from the perspective of more general opportunities in humanities computing is their use of an internet experiment to gather data from a geographically distant group of test persons.

Two papers examine in particular national borders that may play a role in the linguistic differences. Naturally these may be relevant in the Scandinavian case noted above, but that was not the focus of the study. In the first paper, intelligibility is again the behavioural crux, while judgements of similarity are used in the second.

In "Mutual intelligibility. Standard and regional Dutch language varieties" Leen Impe, Dirk Geeraerts and Dirk Speelman investigate the mutual word intelligibility among ten Dutch language varieties, both Belgian and Netherlandic. No local dialects are included, only regional and standard varieties. Word intelligibility was tested by means of a computer-controlled lexical decision task, in which subjects had to decide as quickly as possible whether the auditorily presented word was an existing or non-existing word, and a multiple choice task, in which subjects had to select from two alternatives the one which best reflected the meaning of the stimulus word. Subjects were 641 secondary school pupils in Dutch speaking Belgium and 404 secondary school pupils in the Netherlands. At the national level, an asymmetry was found in the sense that subjects from Belgium experienced fewer intelligibility problems with words from the Netherlands, both phonetically and lexically, than the other way around. The asymmetry is explained by the (historical) language-political situation in the Dutch language area, in which the language spoken in the Netherlands figured and still figures to some extent as a model for the language spoken in Belgium. At the regional level, intelligibility differences between standard and regional varieties were found to be larger in Belgium than in the Netherlands. The experimental techniques employed in this study are very sensitive (especially the response times obtained with the lexical decision task), and they are new to the field of dialectology, where the quantification of intelligibility has been neglected. The measurement method is not only relevant to the study of the intelligibility of related languages and language varieties, but more generally also to other fields, such as speech pathology, second language acquisition and speech technology.

Of course, intelligibility is also affected by processes which impede the "density" of communication (the term is Bloomfield's). One such process is the establish of a national border. In their paper "The Dutch-German dialect border: relating linguistic, geographic and perceptual distances" Folkert de Vriendt, Charlotte Giesbers, Roeland van Hout and Louis ten Bosch examine a dialect region which straddles a national border, a particularly interesting case. Their data were collected in a sub-area of the Kleverland dialect region, extending from Duisburg in Germany to Nijmegen in the Netherlands. The dialects spoken there used to constitute a dialect continuum, marked by a direct relationship between geographic and linguistic distance. However, there is strong evidence that the Dutch-German state border established in 1830 has given rise to a linguistic gap. De Vriendt et al. try to establish the impact of the border on linguistic distance, using multidimensional scaling and correlation techniques. They compare three models to explain linguistic distance. Their first and simplest model is one where linguistic distance is a monotonic function of geographic distance. The second model is expanded by including a constant representing the state border gap. The third model is based on perceptual distances. The perceptual data were collected by means of a questionnaire in which informants were asked to indicate the linguistic

similarity between neighbouring locations and the intensity of contacts with friends, relatives, and shops. The study reveals that a model based on perceptual distance, including both perceptual linguistic distance and perceptual socio-geographic contact distances, explains linguistic distance much better than a continuum or gap model based on geographic distance. The method applied and the results obtained are relevant not only for the study of linguistic variation in other dialect areas intersected by state borders but more generally for the investigation of the role of state borders for the variation and transmission of cultural products in a rapidly globalizing world.

We now turn to a wider range of topics, including a historical conjecture about an unusual divergence in geographical diffusion between two levels of linguistics, several examinations of dimension-reducing techniques for their utility in the study of variation, and a proposal for incorporating the tone in languages such as Chinese into existing measures of pronunciation difference.

It is clear that de Vriendt et al. examine a well-documented historical situation with respect to its linguistic consequences. Simonetta Montemagni examines a curiously divergent pattern in Tuscany and suggest that there may be an historical explanation. She examines the relation between two different levels of linguistic variation in "The space of Tuscan dialectal variation. A correlation study". Using L04, software developed in Groningen, as well as VDM, developed in Salzburg, she measures on the one hand the degree to which pronunciation differs from town to town in Tuscany, and on the other the degree to which vocabulary (the lexicon) varies, i.e. the degree to which different words are used for the same concepts. By measuring phonetic and lexical differences at the same set of sites, she is able to calculate the degree of correlation between the two. Further, she measures the degree to which geographic distance can predict phonetic and lexical distances. The linguistic assumptions behind these questions are definitely of interest to students of cultural transmission in general. We assume in linguistics that variation in both pronunciation and vocabulary is diffused through social contact. Linguists have also observed that the lexicon is quite easily changed: words come and go in languages in large numbers. If we regard proper names as words as well, which is justified, then it is clear that individuals adopt new words quite easily as well. Pronunciations are less volatile, perhaps for cognitive reasons having to do with the way first languages are imprinted in the mind, or perhaps for social reasons having to do with guaranteeing enough comprehensibility for communication. Whatever the reason, we expect a less than perfect correlation between pronunciation and the lexicon, and we expect lexical variation to be the more variable of the two, *inter alia* as it is the more sensitive indicator of social contact. It is simply easier to change one's vocabulary than one's pronunciation. If we operationalise the chance of social contact via simple geographic distance, then we expect the greater variability of lexical differences to translate to a correlation between geography and lexical differences that is weaker than the correlation between geography and pronunciation differences, and this expectation has been borne out in the past (see references in Montemagni's paper). Montemagni notes that correlations between linguistic variation and geography may be complicated by Tuscany's hilly terrain, which is difficult for travel, but this does not explain the difference between pronunciation and vocabulary. She closes with a speculation that ongoing pronunciation changes spreading radially from Florence, which moreover have not reached the various borders, may have left a ring of

linguistic similarity. This might result in a large number of similar sites which are quite distant from one another.

In “Recognizing Groups among Dialects” Jelena Prokić and John Nerbonne review clustering as means of recognizing groups in data, in particular groups of language varieties that might be said to constitute a common dialect. As the authors note, finding groups among data is a problem linguistics shares with other humanities disciplines, at least with history, musicology and archaeology. Focusing on pronunciation data from Bulgarian, and noting the embarrassment of riches in the form of different clustering algorithms, and the problem of clustering’s instability with respect to small differences in input data, they review various means of comparing clustering results. These results may be visualized as DENDROGRAMS, trees which group more similar varieties more closely, and which display reconstructed (clustered) distance in the branch length between varieties, i.e. COPHENETIC DISTANCE. They compare the results of clustering to those of the more stable multi-dimensional scaling, and they examine COPHENETIC CORRELATION, i.e. the degree to which input distances and the distances in dendrograms coincide, a comparison, and the information-theoretic measures of purity and entropy as “internal comparisons” that give a sense of the quality of clustering solutions independent of external information. They note, however, that there is a substantial literature on the dialect landscapes of many modern European (and Asian) languages, and there are formal measures with allow one to gauge how well one classification coincides with another. Prokić and Nerbonne settle on the MODIFIED RAND INDEX, which assigns a score between zero and one indicating how well one partition (due to clustering) overlaps with another (due to independent scholarship). As the humanities disciplines focus even more on the history of scholarship than other sciences, this is a useful measure for other purposes, as well. Methodologically, the paper concludes that the determination of groups is easily confounded by unclear borders, even when these are only sparsely instantiated, which means that some judgment is always sensible. The study also draws surprising conclusions about Bulgarian dialects, seeing only two very distinct and internally coherent groups and a third, Rupskian group which is extremely diverse. Earlier scholarship had distinguished six groups where Prokić and Nerbonne see two.

Antti Leino and Saara Hyvönen’s paper “Comparison of Component Models in Analyzing Dialect Features” may be understood as an exploration into alternatives to the clustering techniques explored in Prokić and Nerbonne’s contribution. Instead of searching for groups in the sample sites, Leino and Hyvönen examine dialect variation more generally as a range of distributions of linguistic material not necessarily organized neatly into groups or areas. It is a technical paper comparing various statistical techniques with respect to their success in uncovering linguistic structure in large atlases (or other collections of linguistically varying material). The statistical techniques all attempt to identify the linguistic “components” or factors in the atlas material, and they are all applied to two Finnish databases, one phonetic and one lexical. They authors conclude that factor analysis is the most robust of the techniques, that non-negative matrix factorization and aspect Bernoulli are most sensitive to possible flaws in the data, that independent component analysis is most likely to provide new insights, and, finally that principal component analysis most capable of providing a “layered” view of dialect structure. The focus on the paper is

the methodological comparison of the various statistical techniques, but the argument is carried out using the dialect atlas material. Given the abstract level of the presentation, it is to be expected that the discussion could be of interest to other areas of scholarship, and, in fact the authors are interested not only in dialects but also in "similar cultural aggregates".

The "Factor Analysis of Vowel Pronunciation in Swedish Dialects" by Therese Leinonen attempts to secure insight into linguistic details even while retaining the advantages of aggregate analyses. Dialectometric research avoids the problem of having to choose which linguistic factors to use as the basis for dialect areas by aggregating over large data sets. These methods have mostly been based on pair-wise distances between dialects, obtained by aggregating over large amounts of linguistic material, but the aggregation step has made it difficult to trace the underlying linguistic features determining the dialect division. This paper expands the horizons of dialectometry by acoustic measurements of vowel pronunciations in Swedish dialects. Vowel spectra were analysed by means of principal component analysis. Two principal components were extracted explaining more than three quarters of the total variance in the vowel spectra. Plotting vowels in the PC1-PC2 plane showed a solution with strong resemblance to vowels in a formant plane. For the analysis of dialectal variation factor analysis was used. Nine factors were extracted showing different geographical distribution patterns in the data, without, however, suggesting distinct dialect areas. Results are visualised using coloured maps showing the Swedish dialect area as a continuum. The study is based on acoustic measurements of the vowels of more than 1,014 speakers: 91 sites spread over 6 traditional Swedish dialect regions, 12 speakers per site (3 elderly and 3 younger women, 3 elderly and 3 younger men). Per speaker 18 words repeated 3-5 times were recorded. The approach taken is valuable for other high-dimensional data involving computers and the humanities. It may be important not only for the analysis of language, but also in the larger context of cultural history and the analysis of other human activities as investigated by disciplines like archaeology, ethnology and musicology.

Cathryn Yang and Andy Castro examine pronunciation differences in so-called TONE LANGUAGES in "Representing tone in Levenshtein distance". Linguists distinguish SEGMENTAL components of pronunciation, i.e. what is represented in a good alphabet, from so-called SUPRASEGMENTAL components, which include in particular tone (or pitch) and duration. All languages, including English, use pitch at a phrasal or sentential level to distinguish different structures, but many other languages, including Chinese, but also Bai and Zhuang, which Yang and Castro study, also use tone extensively as a means of distinguishing words. LEVENSHTEIN DISTANCE has come to be accepted as a computational means of measuring pronunciation distance between words, but it is limited to processing sequences of symbols, i.e. segmental information. Yang and Castro focus on Bai and Zhuang, in which different atomic tonal elements are always associated with single syllables. Their proposal is essentially to measure tonal differences separately from segmental differences. For both levels they measure differences in sequences, so they in addition compare several different tonal representations including e.g. sequences of pitches, or alternatively initial pitches plus contours (melodies). They report on a series of measurements of pronunciation differences which they validate via a comparison to intelligibility measured behaviourally. They are able to demonstrate in two in-depth studies first, that tone is in one case just as important and in a second case even *more* important

than segmental aspects of pronunciation in determining intelligibility, a definitely surprising result, and second, that the representation of tone using initial tone level and contour is most informative. A natural next step in this research line would be to verify that the segmental and suprasegmental levels complement each other in assaying linguistic distance (perhaps using the same intelligibility tests), and then examine means of combining the segmental and suprasegmental levels, e.g. by simply adding the two.

A final group of four papers examines alternatives and corrections to existing methodologies, including normalising for the effect of inherent variability in some concepts, the potential effect of second-hand opinions on the perception of linguistic differentiation, a proposal on how to analyse differing sound systems (rather than simply different sounds), and a proposal for obtaining and analysing data that reflects language use more directly than questionnaires.

Dirk Speelman and Dirk Geeraerts's paper, "Concept Characteristics and Lexical Heterogeneity in Dialects" proceeds from the suspicion that genuine geographic influence may be signalled better by some concepts than by others. The idea behind their essay is that some concepts lend themselves to heterogeneous expression and that this second, conceptual source of variation may confound the measures of geographic variation currently used in dialectology. If there were attempts to measure cultural affinity on non-linguistic material, then one would expect similar sorts of issues to arise. Just as concepts such as 'money' are lexically variable in many languages and dialects, and thus seems to be inherently heterogeneous in expression, others, e.g. 'eyebrow', are not (usually). But if we were to examine the physical realisations of the same concept ('money'), any attempt to gauge the affinity of one settlement with another based on the concept's physical realization would likewise run the risk of being confounded by the variability of the physical realizations. It is clear that linguistic concepts need to be a bit heterogeneous for their variability to signal geographic provenance, but extremely heterogeneous concepts also do not function well if the subdialectal variation is likewise large. Speelman and Geeraerts examine several factors which contribute to conceptual heterogeneity, including the unfamiliarity of concepts, their tendency not to be verbalized (and therefore to be missing from surveys at many locations), their tendency to multi-word expression, their sheer variety of lexicalizations, and whether the concept is negatively "charged." They examine a substantial sample of words with respect to these properties asking whether they contribute to the concept's heterogeneity as reflected in the diversity of words used to lexicalize the concept, how limited the use of the words are geographically and finally, how compact the region is in which a given word is used. Speelman and Geeraerts test their ideas on dialect atlas material from Limburg.

In "What role does dialect knowledge play in the perception of linguistic distances?" Wilbert Heeringa, Charlotte Gooskens, and Koenraad De Smedt raise the question whether naive listeners base their judgments of linguistic distances between their own dialect and other dialects of the same language solely on what they hear during an experiment or whether they also generalise from the knowledge that they have from previous experience with the dialects. In order to answer this question the authors first performed an experiment among Norwegian dialect listeners in order to measure perceptual linguistic distances between 15 Norwegian dialects. These perceptual distances were correlated with objective phonetic distances measured on the basis of

the transcriptions of the recordings used in the perception experiment. In addition, they correlated the perceptual distances with objective distances between the same dialects but based on other datasets. On the basis of the correlation results and multiple regression analyses they conclude that the listeners did *not* base their judgments solely on information that they heard during the experiments but also on their general knowledge of the dialects. This conclusion is confirmed by the fact that the effect is stronger for the group of listeners who recognized the dialects on tape than for listeners who did not. This dialectometric study is of interest to (computational) dialectologists, sociolinguists and psycholinguists. The results are important to scholars who want to understand how dialect speakers perceive dialect pronunciation differences and may give more insight in the mechanisms behind the way in which linguistic variation is experienced.

"Quantifying dialect similarity by comparison of the lexical distribution of phonemes" by Warren Maguire introduces a method for quantifying the similarity of the lexical distribution of phonemes in different language varieties exemplified by standard and dialectal varieties of English. He takes dialectometry beyond lexicostatistic comparison and a purely phonetic comparison of words by considering lexical distributions (i.e. which cognate pairs have the same phone in a given position and which have a different phone). He integrates the lexical distribution of phonemes in the varieties in the measurements by examining the vowel correspondences that result when one aligns the 1,000 most frequent English words phonemically. Similarity is calculated by taking the most frequent correspondence per phoneme into account, and dividing the frequency by the number of comparisons made (i.e. including all correspondences with less frequent corresponding items). The method assesses to what degree dialects have differentiated from each other in the course of history and provides a means of examining historical connections between language varieties which are not obvious from surface inspection. Since this method is aimed at uncovering historical connections between varieties, it has implications for the interaction between dialects and historical factors of all kinds, including migration, standardisation, language death/replacement, historical boundaries, etc. Methods of this sort should add to our understanding of history. In addition, it allows for correlations between geography and linguistic variation, and may well allow us to look into recent changes in languages which are the result of wider societal pressures.

Benedikt Szmrecsanyi's contribution to this volume is noteworthy in several ways. First, while the great majority of studies in variation both in this volume and in more general scholarship focuses on pronunciation and vocabulary, Szmrecsanyi's piece "Corpus-based dialectometry: aggregate morphosyntactic variability in British English dialects" concentrates rather on morphology and syntax. Second, while most variation studies draw their data from the responses to the carefully designed questionnaires, Szmrecsanyi looks rather to naturally recorded collections of material, so-called CORPORA. It is certain that naturally occurring data reflect genuine speech habits more faithfully than do the responses to questionnaires, so the attempt to use this sort of data is worthwhile, but it is also more difficult to control the sorts of material one collects. Third, Szmrecsanyi also compares his findings to those in earlier literature which relied on selected features rather than larger aggregates. A first question is whether this naturalistic data will submit profitably to the usual sorts of analyses, and this question is answered in the positive (but see below). Even though there have been earlier aggregate studies of syntactic variation (see Szmrecsanyi's paper for references), there have not been many, making the fundamental question of

geographical structuring of the data a very interesting second question. In particular typologists and syntactic theorists have been sceptical about the role of geography as an influence on linguistic variation, conjecturing that structural constraints would prove to be the more influential elements in explaining syntactic variation. And indeed, Szmrecsanyi's analysis displays rather less geographic conditioning than we are accustomed to in the study of linguistic variation, in particular in a rather weak correlation between geographic and linguistic (syntactic) distance ($r=0.22$). It remains therefore to be seen whether this reflects a difficulty with the use of the naturalistic data or whether it reflects the typologist's conjecture that syntax offers less affordance for the expression of affinity.

Conspectus and prospectus.

What does the present collection contain in the way of indications for humanities computing more generally? We select two prominent elements, namely the need for statistics on the one hand, and the virtues of deeper engagement with noncomputational work on the other.

The present volume presents a range of papers applying computational, and almost always statistical techniques to topics in the study of language variation. We suspect that the extensive and serious use of statistics is a benefit not only to our computational humanities discipline, but to others as well. Since one of the greatest advantages of the computer is its ability to process very large amounts of material, and since that material is usually empirical for most humanities disciplines, excepting perhaps philosophy, it is also variable. We simply do not see clean categorical generalizations in the data. Statistics allow us to wean lawful regularities from the complexity of noise, error and counterindicating factors. We suspect that the computational turn in the humanities will be accompanied by a statistical turn as well.

In addition to this technical reflection we return to the theme of "engaged computational humanities" we mentioned in the introduction. Our goal is to have computational techniques accepted beside the range of techniques already available to the researcher in language variation, and the papers in this volume illustrate some of the consequences of this choice. As we noted at the beginning of this introduction, this entails our addressing questions from this sub-discipline, attempting to solve existing analytical problems, and comparing our results to those of non-computational scholars, and examining our methods and results critically from a variationist perspective. The papers in this volume illustrate these consequences abundantly, e.g. in the large number of papers, including the paper arising from Van Heuven's keynote, exploring how to validate computational techniques using alternatives known from non-computational studies, including in particular behavioural studies. Leinonen's paper, developing a computational technique for automatically comparing large numbers of vowel pronunciations, is also careful to show how the innovative technique relates to well-established techniques which require manual intervention.

One way to summarize the strategy of engaged computational humanities is to note that computational humanities scholarship remains humanities scholarship, meaning in particular that it must be compared to non-computational scholarship and that apparently conflicting results always stimulate reflection.

References

L. Bloomfield (1933), *Language* (New York).

J. K. Chambers and P. Trudgill (1998), *Dialectology,* 2nd Ed (Cambridge).

Deutscher Sprachatlas (DSA) (1927–1956), Auf Grund des Sprachatlas des deutschen Reichs von Georg Wenker, begonnen von Ferdinand Wrede, fortgesetzt von Walther Mitzka und Bernhard Martin (Marburg).

H. Goebl (1982), *Dialektometrie: Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie.* Vol. 157 of *Philosophisch-Historische Klasse Denkschriften* (Vienna).

E. Haugen (1966), 'Semi-communication: The language gap in Scandinavia', *Sociological Inquiry*, 36 (2), 280-297.

W. McCarty (2005), *Humanities Computing* (New York).

A. McMahon and R. McMahon (2005), *Language Classification by Numbers* (Oxford).

J. Moberg, C. Gooskens, J. Nerbonne and N. Vaillette (2007), 'Conditional entropy measures intelligibility among related languages', in P. Dirix, I. Schuurman, V. Vandeghinste and F. Van Eynde, eds, *Computational Linguistics in the Netherlands 2006* (Utrecht), 51-66

J. Nerbonne (2005), 'Computational contributions to humanities', *Linguistic and Literary Computing* 20(1), 25-40.

J. Nerbonne (2007), 'Review of McMahon and McMahon (2005)', *Linguistic Typology* 11, 425-436.

J. Nerbonne (2009), 'Data-driven dialectology', *Language and Linguistics Compass* 3(1), 175-198. doi: 10.1111/j.1749-818x2008.00114.x

J. Nerbonne and W. Heeringa (to appear), 'Measuring dialect differences', in J. E. Schmidt and P. Auer, eds, *Theories and Methods* Vol. in series *Language and Space* (Berlin).

H. Niebaum and J. Macha (2006), *Einführung in die Dialektologie des Deutschen* (Tübingen).

J. Séguy (1973), 'La dialectométrie dans l'Atlas linguistique de la Gascogne', *Revue de linguistique romane*, 37, 1-24.

M. Spruit, W. Heeringa and J. Nerbonne (2009), 'Associations among linguistic levels', *Lingua*. To appear. doi:10.1016/j.lingua.2009.02.001

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

G. Wenker (1881), *Sprachatlas von Nord- und Mitteldeutschland. Auf Grund von systematisch mit Hülfe der Volksschullehrer gesammelten Material aus circa 30000 Orten. Abtheilung I Lieferung 1 (6 Karten und Textheft)* (Strasbourg/London).