Introduction to Methods section (John Nerbonne) of Charles Boberg, John Nerbonne and Dominic Watt (eds.) *Handbook of Dialectology*. (2015/2105) Wiley-Blackwell.

Methods

It is best to begin an introduction to the methods section of this handbook, which deals with gathering and analyzing data, with two near-platitudes that are not always kept in mind, first, that methods are subordinate to research questions (as well as to other factors), and second, that some methods are much more costly than others to apply.

It is easy to see that data collection depends on one's research question. To address any question in historical dialectology, we need data from different time periods – or from people of different ages. To consider a question on the diffusion of changes in dialects, we need to examine a substantial area in which diffusion patterns have a chance to emerge. These are obvious, almost commonsensical points, but others have become obvious only in the course of research. So work in social dialectology often tries to detect ongoing language changes, which (research has shown) are always accompanied by periods in which more than one linguistic form is used. To analyze such changes, frequency data is indispensable. For a further example, we note that work on diffusion has noted the importance of large population centers, meaning that such centers have to be part of the sample studied.

The basic idea that methods depend on research questions is a good principle, but not a hard and fast constraint. On the contrary, researchers have been able to exploit available material for research questions that had not even been posed when the data was collected. Nerbonne (2010) examines Trudgill's famous "gravitational view" of diffusion (Trudgill 1974) *inter alia* using data from the *Linguistic Atlas of the Middle and South Atlantic States* (Kretzschmar 1994), most of which was collected in the 1930s.[1] Eisenstein's chapter (below) analyzes the Twitter stream for evidence of geographic cohesion, and it is clear that neither the company Twitter nor the users of its services intended their communication as a contribution to dialectological data collection. Furthermore, there is also a tradition of collecting data in large cooperative efforts in dialectology in ways designed to serve many research questions – this is the tradition of the dialect atlas (see Kretzschmar's chapter, this volume). In atlas projects it is impossible to anticipate every research question, of course, and some newer questions turn out to be difficult, if not impossible to answer using atlas material. For example, the role of frequency in diffusion is theoretically interesting (Bybee 2002), but we don't have lexical frequency data for dialects. We might attempt to use frequency from standard language data (newspaper corpora), but that would entail risks.

It is also easy to understand how costs – in time and money – often play a role in data collection. For many purposes (see below), researchers would like to sample a good amount of material from many people over a large area. There are differences, of course, but in dialectology, just as in many other

[1] Nerbonne and Heeringa (2007) examine the gravity hypothesis using data from the *Reeks Nederlandse Dialektatlassen* (RND), collected 1925-1982.

modern areas of research, the best data is more data! Typical data collection efforts therefore exceed the capacity of individuals, making it necessary to seek funding. Most dialect research is financed through three- to six-year grants by research councils, foundations and funding agencies, where opportunities are scarce and competition fierce, , and where time remains limited. Scientific academies are honorable exceptions, where longer-term projects may be conducted, but even the academies' funds and time are limited. So it is small wonder that dialectologists have sought help, e.g., through crowd-sourcing. The dictionary of Flemish dialects has a team of volunteers to help in tasks such as digitizing field records (see http://www.wvd.ugent.be/). The data collection chapters also report on the use of telephones, internet and smart phones to streamline data collection (especially the chapters on written surveys and on interviews). The big promise of the new techniques is their efficiency: they allow data to be collected much more quickly and in larger amounts than old methods.

Before continuing to more specific dialectological concerns, it will be useful to recall one basic statistical concept, that of *noise*, or unsystematic variation. In variationist linguistics, encompassing dialectology and sociolinguistics, we are always interested in variation, whether it be in pronunciation, lexical choice, morphology, syntax or elsewhere. We develop and test hypotheses about the systematic variation, e.g. the variation in the pronunciation of a word such as *bike* as one travels from North to South in the US. Good hypotheses explain some of the variation we find. But in addition, we encounter variation beyond our hypotheses which we therefore cannot explain. The additional variation may depend on how carefully the speaker spoke, on when in an interview a word was elicited (in the beginning or at the end of an interview), on the linguistic context the word was used in, and on other factors that data collectors may not have even tried to control. The variation that is not part the focus of the study is noise, or apparently unsystematic variation, and we must be particularly sensitive to it given our focus on its converse, systematic variation.

Returning to data collection, and continuing beyond its costs and dependence on research questions, dialectological data collection efforts may be seen as positioned along two major dimensions, naturalness and commensurability, which of course leads to some tension. In characterizing dialects, we would like to hear how the people in an area *normally* speak, i.e. when not accommodating to visitors from outside or projecting an ideal (with respect to education, etiquette or conformance to standards) that may be quite idiosyncratic. Labov has dubbed the very presence of an interviewer "the observer's paradox" (Labov 1972:209), and noted that more formal data collection protocols were likely to influence speakers even more and noted that distortion of normal speech patterns increased as speakers' self-awareness was heightened by formal, controlled elicitation methods This brings us to the issue of ensuring that data is commensurable, i.e. that one may compare items to one other without suspecting the influence of confounding factors. Simplifying a bit, many dialect atlases aimed to provide data that one may compare with respect to geography alone – where all the other factors that influence variation (age, gender, educational level, …) are fixed. Other atlas efforts attempt to vary some of the demographic factors systematically, too. One may ensure commensurability by using very strict protocols for interviews, including lists of concepts whose lexical realization is to be noted or words whose pronunciation is to be recorded (see the Chapter below on questionnaires). This allows us to study variation along the lines used in the lists. If the items are chosen carefully, we can eliminate some sources of noise that may arise in spontaneous speech, such as the context dependence of lexical choice and pronunciation. The tension then arises as we note that strict protocols tend to dampen the spontaneity of the interview and of the speech.

Data collection

The first five chapters in this section of the book concern data collection. Data sampling concerns how to ensure even geographical coverage, the choice of respondents to interview, and how to approach and engage respondents. A major choice arises as to whether to collect data via an interview or via a written survey, which can be administered remotely (by mail or by web questionnaire), and which is therefore quicker and cheaper than interviews. Questionnaires are not restricted to use in written surveys, however; they often play a role in structuring oral interviews as well. The chapter on questionnaires focuses on the linguistic choices in the data collection effort regardless of the mode (oral or written) the collection assumes.

The field interview is a *primus inter pares* among data collection methods. In contrast to written survey techniques, the interviewer is present with the respondent, and ideally working hard to ensure that the conditions from interview to interview are commensurable. For example, an interview can note that there was an interruption in the process at some point, something written surveys have not been able to do.[2] As the chapter on field interviews documents, the interview can also steer the conversation toward topics where unselfconscious production is most likely. Phonetic researchers often prefer field interviews to other possibilities for the opportunity it provides for visual observation.

A newcomer to the team of data collection methodologies is corpus linguistics, in which data is extracted from corpora, i.e. large collections of speech and/or text. These may be corpora of dialect speech, the focus of the chapter in this section, but good work has also been done on a corpus of letters to the editor in American newspapers (Grieve 2011). In all cases one examines genuinely occurring speech or text, but many of the same methods are used regardless of the source of the corpora. While practitioners have occasionally claimed that the speech in corpora are more natural or less self-conscious than e.g. dialect atlas data, it should be clear that this varies, depending on the particular choice of corpus and atlas. Nonetheless, extracting features (e.g., lexical realizations, morphological forms) from genuinely occurring speech or text entails dealing with the great skew in word frequency distributions (Baayen 2001), and this leads to problems in identifying comparable material. Zipf (1932) noted that if one sorts words by frequency, then the frequency of the *n*-th most frequent word is roughly *1/n* times the frequency of the most frequent word. Of course, there have been subsequent attempts to reformulate and refine this (Baayen 2001), but the rough relationship suffices to make the following point. Since adults have vocabularies of tens of thousands of words, this means that most words occur rather infrequently. If we leave it to chance to elicit comparable words from speakers in different areas of a survey, then the chance of hearing any but the most common words in all of say, 20

[2] Some web-based surveys are implemented in Java programs, e.g. Charlotte Gooskens' MICReLa project on mutual comprehensibility (www.let.rug.nl/gooskens/project/), which are definitely able to keep track of time, enabling some checks on the conduct of written surveys. Still, the check is minimal when compared to the presence of an interviewer during data elicitation.

sites, is only negligibly above zero. It is no accident, therefore, that corpus-based techniques have either focused on frequent elements, or have aggregated over classes of elements, sometimes combining the two, e.g., by examining contracted vs. uncontracted forms (since contracted forms are always forms of *be, have* or one of the modals, they are fairly frequent). The chapter on social media might also be regarded as a corpus-based analysis.

Linguistic and geographical methods

Two chapters are devoted to the instrumental and computational analysis of linguistic data and one to the various sorts of mapping techniques popular in dialectology. Acoustic phonetics has long been advocated as an analytical tool in dialectology (Labov, Yaeger & Steiner 1972), especially as a way to obviate the need for phonetic transcription, which is notoriously difficult and subjective. The analysis of vowels is well established in dialectology and sociolinguistics and is now carried out automatically on large sets of vowels (Rosenfelder et al. 2011). Work on consonants is progressing and is reported on in the chapter. For text-encoded material, including phonetic transcriptions, several techniques from computational linguistics are potentially useful. Edit-distance measures are becoming standard; lemmatization may facilitate morphological studies and stimulate them further; and some simple syntactic analyses are robust enough for some statistics (Wieling & Nerbonne 2015).

We would predict that automatic analysis of the sorts presented in these two chapters will be of increasing interest to dialectologists for several reasons. One is simply that the increasing volumes of data available for analysis necessitate increasing automation in analysis. Leinonen (2010) was able to extract the formants of nearly 20,000 vowel tokens in the SweDia corpus (Eriksson 2004) but only because she applied reliable, automatic procedures for extracting formants. Second, automating (parts of) analyses improves them with respect to replicability, a requirement more difficult to fulfill in manual work, especially work requiring judgment on the part of the researcher, such as phonetic transcription or superficial syntactic analysis in terms of parts of speech. Third, the automated analyses are more and more capable of identifying the latent structure that a great deal of linguistic discourse revolves around. Identifying the parts of speech of words is an excellent example of uncovering the sort of latent structure that on the one hand may be reliably identified and on the other may suffice to support dialectological analysis (Wolk 2014).

We need not belabor the desirability of including a chapter on maps in a handbook of dialectology, except to note that automatic procedures for making good quality maps are also improving, not only in the graphic quality of the maps, but also in the ease with which the software may be used, and finally in the range of functions available for adding information to the geography – including political and physical boundaries, population sizes, densities of occurrence, optimum tiling for networks of data collection sites, and more. In short, maps are essential to dialectology.

Statistics for variationist studies

The statistical chapters may be the most challenging technically, but it is clear that a field as data-rich as dialectology could never forego statistics. Because the chapters are challenging, it will be worthwhile to remind readers of the foundation they are built on.

The chapters assume some familiarity with null hypothesis testing (NHT), a practice that has been the basis of a great deal of statistical analysis, but which has increasingly come under fire (see Field et al. 2012: Ch.2 and references there). It is still not clear what will take the place of NHT, but focus in statistical analysis is shifting toward model comparison. In NHT one contrasts the hypothesis of interest, perhaps a proposed difference in the mean recognition times for two classes of words, with a *null hypothesis*, which assumes that there is no effect, i.e. no difference. Whether such a difference is regarded as improbable will naturally depend on the (systematic and unsystematic) variation in the data, what is called the spread in the distribution of the data. For numerical data such as reaction times, the common measure of spread is the *standard deviation*. If the sample turns out to be improbable when one assumes the null hypothesis, this is regarded as evidence for the hypothesis of interest. The probability of the sample given the null hypothesis is called the *p-value*, and the lower the *p*-value, the less likely the null hypothesis, and the stronger the evidence for the alternative (the hypothesis of interest).

When the *p*-value falls below an agreed on threshold, say, 0.01, we say that the study is *statistically significant*. This use of that word must not be confused with its everyday sense, i.e. 'meaningful, having important consequences'. A p-value is a probability and is crucially influenced by how large a sample was studied. Statistically insignificant differences in small samples (say of size 10) inevitably become significant at some large sample size. This is not just a theoretical possibility, but one which is frequently seen as variationists examine ever larger data sets. But it means that the intelligent reader of statistical analyses not only examines p-values, but also tries to gauge effect size. How effect size is measured depends on the research question and the analysis technique, but for a comparison of mean values, the difference in means, expressed in terms of the number of standard deviations, is a common measure. If we notice a difference of 40 milliseconds in recognition time for two classes of words, and the standard deviation is 80 ms., then we would express the effect size as Cohen's d (= 40ms./80ms.), or about 0,5 standard deviations.

A lot of linguistic data is categorical, i.e. it occurs in various categories, where there is no intrinsic order among them. Examples are different parts of speech (Noun, Verb, etc.) or different realizations of /t/ in English ([t], [ɾ],[ʔ]). For many years, the analysis of such categorical data was limited to the χ^2 test of independence, which is ill-suited to analyzing multiple influences, which Bayley (2013) has dubbed "the principle of multiple causes." Language variationists are convinced that choices in variation may be influenced by many factors, and that combinations are possible and important. For the sake of completeness, the chapter on logistic regression also explains χ^2 analysis and its limitations.

Sociolinguists pioneered the use of logistic regression in linguistics and have continued to use it for forty years, so that it is only fitting that we include a chapter devoted primarily to the models underlying logistic regression, its most important prerequisites, its application, and the interpretation of its results.

Ignoring honorable exceptions (Woods 1979, Gregg et al. 1981), most sociolinguistic analyses focus on a small number of variables. In contrast dialectology oftenfocuses not onindividual variables (features) but on large numbers of linguistic variables simultaneously, e.g., when atlases are compiled or corpora collected. Furthermore, the individual variables are often quite noisy, and one would wish as a researcher not to hand-pick variables with all the dangers of subjectively selecting exactly those variables that support the case one is making (Nerbonne 2009). This has promoted the dialectometric perspective (see Chapter by Goebl, this volume), which has consistently emphasized the virtues of adopting an aggregate perspective on variation, and inspecting individual features only from that vantage point. The chapter on aggregate analyses also reports on first efforts at including geographical and social variables in a single statistical model, following the vision of Chambers and Trudgill (1998: Ch. 12). .

In general, dialectologists have not made use of the specialized field of spatial statistics or what is also known as geo-statistics, and this is certainly a shortcoming of the field of dialectology. Geo-statistical techniques are often used in fields that ask similar questions about diffusion and barriers to diffusion, like demography and epidemiology. One example of a point where they have gone beyond dialectometry concerns the selection of variables to analyze, where measures of spatial autocorrelation have been brought to bear. Dialectological method has much to learn from these fields.

Finally, we have included a chapter on social media in this section both because of its intrinsic interest, including the demonstration of the importance of geography even in the age of digital, world-wide communication, and because it uses usually sophisticated statistical reasoning from machine learning in order to draw conclusions from its data. We hope that its presence here will inspire more collaboration between variationist linguists and statisticians.

Future Challenges

Readers of this section should not come away with the impression that the methods in dialectology have stabilized to a point where we expect little innovation in the future. We suspect that just the opposite is the case. Data collection is likely to turn more and more to methods already in use such as web questionnaires and smart phone apps. Vaux and Golder (2003) pioneered the use of web questionnaires to collect English lexical data, and Möller & Elspass's (2008) questionnaire aims at everyday German variation not only in vocabulary, but also with respect to pronunciation and syntax. This new work will require a good deal of analysis to validate its methods and also to compare the results to older work. Work using smart phones is even newer, but Sherrer et al. (2012) and Leemann et al. (2015) (and other references they cite) describe Dialekt Äpp, an iOS application for collecting speech data that has already shown great promise.

On the analysis side, we certainly expect to see machine learning techniques make their way further into the analysis of dialect data, just as they have in other areas of statistics, especially exploratory statistics. The relation between single variable analyses and aggregate analyses also deserves further attention, as do further techniques for including geographic and social variables in single analyses (see chapter on aggregate analyses). To-date, the more encompassing analyses are regression analyses aimed at predicting the aggregate differences of a sample of varieties to a single alternative, the standard language. Techniques aimed at analyzing all pairs of varieties would improve our understanding further, as Wieling and Nerbonne (2015) also urge.

References

Baayen, R. Harald (2001) *Word frequency distributions*. Berlin: Springer (Springer Science & Business Media, Vol. 18).

Bayley, Robert (2013) The quantitative paradigm. In: J.K.Chambers & Natalie Schilling-Estes (eds.) *The handbook of language variation and change*, Boston: Wiley. 117-141.

Bybee, Joan (2002) Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language variation and change* 14(3): 261-290.

Chambers, J.K., & Peter Trudgill. *Dialectology*. Cambridge: Cambridge University Press, 1998.

Eriksson, Anders. (2004) SweDia 2000: A Swedish dialect database. In: P. J. Henrichse (ed.) *Babylonian Confusion Resolved. Proc. Nordic Symposium on the Comparison of Spoken Languages. Copenhagen Working Papers in LSP*. Copenhagen

Field, Andy, Jeremy Miles, & Zoë Field (2012) *Discovering statistics using R*. London: Sage.

Gregg, Robert J., Margaret Murdoch, Erica Hasebe-Ludt & Gaelan de Wolf (1981) An urban dialect survey of the English spoken in Vancouver. *Papers from the fourth international conference on methods in dialectology* (pp. 41-65). University of Victoria: Victoria, British Columbia.

Grieve, Jack (2011) A regional analysis of contraction rate in written Standard American English. *International Journal of Corpus Linguistics* 16(4): 514-546.

Kretzschmar, William A. (ed.) (1994) *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*. Chicago: The University of Chicago Press.

Labov, William (1972) *Sociolingustic Patterns*. Philadelphia: University of Pennsylvania.

Labov, William, Malcah Yaeger, & Richard Steiner (1972) *A quantitative study of sound change in progress*. US Regional Survey, Vol. 1.

Leemann, Adrian, Marie-José Kolly , David Britain, Ross Purves, & Elvira Glaser 2015. Documenting sound change with smartphone apps. *The Journal of the Acoustical Society of America*, 137(4): 2304-2304. http://dx.doi.org/10.1121/1.4920412

Leinonen, Therese (2010) *An acoustic analysis of vowel pronunciation in Swedish dialects*. PhD. Thesis, Groningen.

Möller, Robert & Stephan Elspaß (2008) Erhebung dialektgeographischer Daten per Internet: ein Atlasprojekt zur deutschen Alltagssprache. In: S. Elspaß & W. König (eds.) *Sprachgeographie digital. Die neue Generation der Sprachatlanten (mit 80 Karten). Hildesheim: Olms*. 115-132.

Nerbonne, John (2009) Data-Driven Dialectology. *Language and Linguistics Compass* 3(1): 175-198.

Nerbonne, John (2010) Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences*. *365*(1559): 3821-3828.

Nerbonne, John & Wilbert Heeringa (2007) Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation. In: Sam Featherston and Wolfgang Sternefeld (eds.) *Roots: Linguistics in Search of its Evidential Base* Berlin: Mouton De Gruyter, 267-297.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini and Jiahong Yuan (2011) FAVE (Forced Alignment and Vowel Extraction) Program Suite [Computer Program]. Avail. at http://fave.ling.upenn.edu/

Scherrer, Yves, Adrian Leemann, Marie-José Kolly & Iwar Werlen (2012) Dialäkt Äpp - A smartphone application for Swiss German dialects with great scientific potential, 7ème Congrès SIDG - Dialect 2.0, Vienna

Trudgill, Peter (1974) Linguistic Change and Diffusion: Description and Explanation in Sociolinguistic Dialect Geography. *Language in Society* 2:215–246.

Vaux, Bert and Scott Golder (2003) *The Harvard Dialect Survey*. Cambridge, MA: Harvard

Wieling, Martijn, and John Nerbonne (2015) Advances in dialectometry. *Annual Review of Linguistics, 1*: 243-264.

Wolk, Christoph (2014) Integrating aggregational and probabilistic approaches to language variation. PhD thesis, University of Freiburg. https://www. freidok.uni-freiburg.de/data/9656/.

Woods, Howard B. (1979) *A Socio-dialectology Survey of the English Spoken in Ottawa*. National Library of Canada.Zipf, George K. (1932) *Selected studies of the principle of relative frequency in language*. Cambridge: Harvard University Press.