Studying language differences - an intellectual ecology

John Nerbonne[1]
Universities of Groningen and Freiburg

Introduction

My goals in this paper are first to suggest that our focus on linguistic diversity, i.e. differences among languages and dialects, is more broadly relevant than would be apparent at most conferences on dialectology. I'll do this by examining application of dialectological methods in related sub-disciplines of linguistics. From that basis I'll also secondly suggest that the current self-identification of dialectologists might profitably be revised. Right now, most dialectologists see their work as most closely related to sociolinguistics. Since dialectology studies variation with respect to geography, and sociolinguistics studies variation with respect to social factors such as class or occupation, the close relation seems sensible. But the current research foci of the two sub-disciplines belie the apparent similarity in the definitions. To lend some authority to what I say, but also to make my computational and mathematical perspective clear, I'll first review my own contributions to dialectology. These lead me to see similar efforts in other fields.

We turn now to that background, i.e. work in dialectology.[2]

Background

I'll review some of the work that I see as relevant for linguistics outside of dialectology. Because the related work I'll discuss is that closest to my own, this background section may also seem skewed. It emphasizes my contributions more than their prominence in the field as a whole could justify.

As many know, dialectology has a "dusty", "stuffy" image in many places, no more so than in the Netherlands, where I spent most of my career, and where the former director of the foremost research institute for dialectology, Voskuil, published a 7-volume *roman à clef* about his institute, its eccentric staff, their arcane pursuits, and especially about their never-

[1] I wish to offer my thanks to the Royal Academy of the Basque Language for the opportunity to speak in Bilbao and to hear about current work on Basque dialectology. I am further indebted to the DFG Science Center "Words, Bones, Genes and Tools" at the University of Tübingen for a stay as a research fellow, which gave me a chance to write this article.

[2] In the section on background, I borrow directly from my valedictory lecture in Groningen (Nerbonne, 2017b).

ending intrigues. I hope that an overview of some contemporary work can dispel that dusty image a bit.

Let's nonetheless emphasize that dialectology *is* the oldest study of language variation, i.e. how it varies and where. We also have an advantage over other sub-disciplines of linguistics, as there is an enormous lay interest in our work, guaranteeing public interest in what we do. Of course, we do more than just answer the layperson's questions, since, as scientists we are dissatisfied with long lists of varying words, pronunciations and constructions; instead, and we seek some generality in our linguistic characterizations. The hope is that this can lead to an explanation of why some things vary in certain places while others, elsewhere do not. Similarly, we seek explanations regarding the geography of the distribution of linguistic variation.

My own work in this sub-field of linguistics began when I suggested to seminar students in 1996 that we try to replicate some work that Brett Kessler had presented at the 1995 meeting of the European Association for Computational Linguistics (Kessler 1995). The students were eager and bright, so that the seminar work led to a publication, and, as is so often the case, increased curiosity.

When we started the work (Nerbonne et al. 1996), the dominant approach in dialectology involved a lot of manual labor, sensitive analysis and enormous data collection efforts. It was non-computational and inexact.

A shining exception to the dominant paradigms was the work of Jean Séguy in Gascogne, who was frustrated by the noisiness of his data – where exceptions seem to appear at all levels. He suggested that one might begin to see the forest and not just the individual trees if one shifted focus to the aggregate differences. So he added up the number of differences between each pair of sites in his data and published a curve showing a sub-linear increase in differences with respect to geography (Seguy 1971). The Austrian Hans Goebl was in close contact with Séguy before his untimely death (around 1974) and followed his ideas, but adding clustering – to detect groups of sites, i.e. dialect areas – as well as an inverse frequency weighting, which has also survived replication. Goebl continued Séguy's perspective of viewing the data as categorical – i.e. as either the same or different (Goebl 1982).

We in Groningen[3] continued this tradition in dialectometry, adding techniques to provide stability in clustering and reacting critically to Goebl's position of regarding all multiple responses as confusion (Nerbonne & Kleiweg 2003). We inevitably found multiple responses in our data, and we enjoyed exchanges on the correct mathematics of the necessary treatment with Gotzon Aurrekoetxea, the leading Basque dialectologist of his generation (Aurrekoetxea et al. 2013). It pleases me to return to the Basque country, where this paper was first read. to continue this discussion.

We just noted that Séguy published a curve showing a sub-linear increase in differences with respect to geography. At my suggestion, this curve, which has been replicated dozens of times, is now known as SÉGUY'S CURVE. Gabon Bantu, Bulgarian, German, American English, Dutch and Norwegian all appear to show a sublinear growth in linguistic differences with respect to geographic distance (Nerbonne 2010).

Measuring pronunciation difference

But we also saw opportunities to analyze the data in a more discriminating way using a technique developed by the Russian information scientist, Valdimir Levenshtein in 1966 to improve self-correcting codes. Levenshtein developed an algorithm which seeks the optimal alignment between two strings of discrete elements. He did this by assigning costs to the editing operations of substitution, insertion and deletion. As noted above, Brett Kessler first applied this to the phonetic transcriptions in a dialect atlas in 1995, and my students and I did a follow-up in 1996. Here's an example to illustrate the ideas more concretely. We see in a single example each of the three operations, and the distance between the strings is the sum of the edit costs, here three.

(Grouw) [mɔlkə]	m	ɔ	l		k	ə
Harlem) [mɛlək]	m	ɛ	l	ə	k	
∑ = 3		1		1		1

Table 1 The alignment resulting from the application of the Levenshtein algorithm to the phonetic transcriptions of the word for 'milk' in the Dutch towns of Grouw (in Friesland) and Harlem (in North Holland). Note that the alignment involves the insertion, deletion and the substitution of phonetic segments.

[3] The 'we' is emphatically editorial, not a *pluralis majetatis*. See below for a long list of collaborators without whom this research line could never have taken off.

Introducing edit distance as a means of comparing pronunciations provides a lever to solve two specific problems in dialectology.

First, where earlier work had catalogued differences at a categorical level (same vs. different), as noted above, edit distance assigns a numerical value to the difference, one that can safely be summed and analyzed in the aggregate. The need for this may be appreciated by taking even a cursory look at data in a dialect atlas. Nerbonne (2009) shows an excerpt from a dialect atlas consisting of data from fewer than 200 sites that recorded 87 different pronunciations of the two-sound word *ich*, with standard pronunciation [iç], meaning 'I' in German. Table 2 provides a small selection of the different transcriptions.

iç	ɐɪç	ʕɪk	əɪʃ	əɪg	ç	ɛɪʃ	ɛçk	ɪʔ	…

Table 2 A small selection of the 87 different phonetic transcriptions in the Phonetic Atlas of Germany, illustrating the daunting variety -- and detail -- of the phonetic transcriptions found in dialect atlases. From Nerbonne (2009).

Table 2 is illustrative not only of the variety of the data catalogued in atlases, but also of the fine discrimination that field workers bring to the task of collecting and recording data. This led our group in Groningen to experiment with a myriad of variations on Levenshtein's algorithm with the goal of treating individual segments in transcriptions such as [i,ɪ,e,ɛ] or /æ/nor merely as the same or different, but rather as similar to a certain degree. We wished to provide more finely differentiated costs in the algorithm, and Heeringa's (2004) dissertation alone examines several hundred. Wherever we evaluated 100 words as pronounced and recorded at dozens (or, better, even hundreds) of different data collection sites, then the result was reliable, even with only rough measures. For all those purposes where we wish to characterize the variety and not the individual word, a simple version is quite satisfactory. But we wished to examine individuals word pronunciations as well.

Detecting geographical distributions.

We need some introduction to present the second problem and its solution. Traditional dialectology typically analyzed the variation it found in areas, less frequently in continua. One might aim to detect areas by focusing on the borders between them, where one needed to find instances where borders between individual variants ("isoglosses") roughly coincided. The problem with this clear idea is that the borders of the distributions of variants often did not coincide (or otherwise roughly align), which led researchers to distinguish dialect areas (or continua) by appealing to a selection of features. For

illustration, we turn to a *locus classicus* from Bloomfield's *Language* (1933:328) concerning Dutch dialects. The map shows that the vowel in the Dutch words for *house* and *mouse* are realized in five different combinations of sounds, only two of which coincide. In three cases the vowels, which were the same in Proto-Germanic, are realized differently.

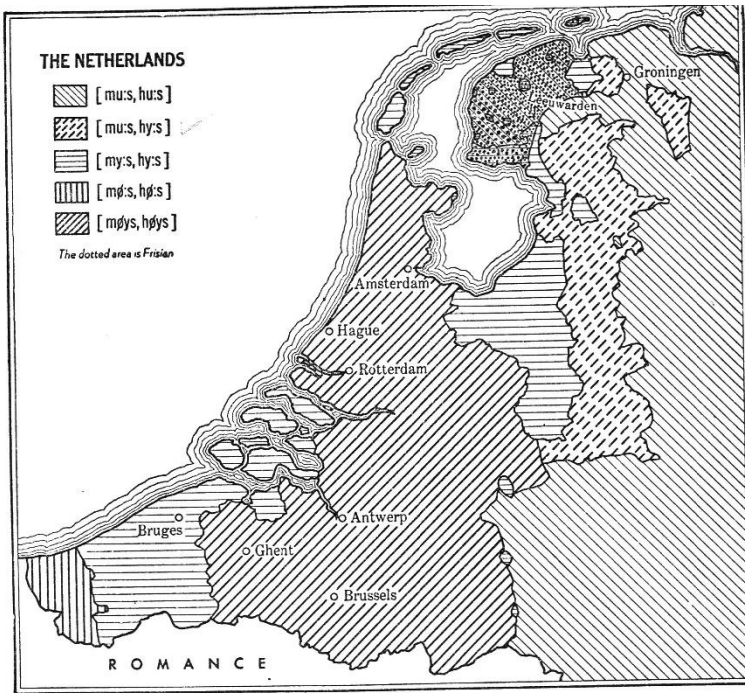

Figure 1 Bloomfield's (1933: 328) snapshot of the difficulty of relying on overlapping borders between different forms to identify dialect areas. Although the vowels in the two words arose from the same historical source, the borders do not overlap well. It is the experience of virtually all dialectologists that, although one can find overlapping borders in a large selection of material, they are atypical.

Another related, frequently noted problem is that earlier maps tended to show the boundaries as sharp even though researchers were quick to note that they were normally gradual.

Detecting gradual distributions of language variation

By measuring differences and not merely cataloguing them, we were able to apply a powerful statistical technique, multi-dimensional scaling (MDS),[4] in a further analysis step (Nerbonne et al 1999). Nerbonne (2009) compares the individual distributions of nine well-studied features in German dialectology to an aggregate analysis of nearly two hundred words, focusing a three-dimensional MDS analysis, which, incidentally accounted for over 85% of the variance in the data. In the interest of space, we will not reproduce the maps

[4] The late Joe Kruskal suggested this in an email around 1998. Only later did we discover that Embleton (1993) had made a earlier plea for the use of MDS in dialectometry. Incidentally, Wilbert Heeringa and Peter Kleiweg each credit the other for mapping the MDS dimensions to colors.

here. Wilbert Heeringa suggested a way to detect typical dialectal phenomena, namely by testing which phenomena correlated with the individual MDS dimensions.

Elaborations and excursions[5]

Wilbert Heeringa (2004) conducted the first extended study using the computational measure, and he and Charlotte Gooskens inspected the reliability and validity of the measure, introducing a more reflective element into general dialectology (Gooskens and Heeringa 2004). Heeringa and Nerbonne (2007) and Nerbonne (2010) exploited the numerical nature of the measure to re-examine theories of diffusion, showing that differences accumulated in a sub-linear fashion that contradicted earlier "gravity models", concluding that these had overestimated the influence of geographic distance.[6] Marco Spruit (2008) examined syntax and also examined the degree to which various syntactic features intercorrelate – a topic worth returning to. Bob Shackleton (2010) probed American and British dialect atlases trying to determine the English sources of American dialects and also showed that both geographic distance and traditional dialect areas predict linguistic differences. The combined model is better than either a purely areal model or one based solely on distance (Nerbonne 2013).

Jelena Prokić (2010) worked on Bulgarian, focusing on the relation between MDS (above) and clustering, which had been introduced to dialectometry earlier, and examining phylogenetic inference algorithms in addition, i.e. algorithms that attempt historical inference. Peter Nabende (2011) examined a practical application of measuring pronunciation distance, namely recognizing names from other writing systems (e.g., Urdu or Russian). These are always transliterated in a way that suggests the pronunciation. Martijn Wieling (2012) showed how to make the edit distance measure more sensitive in a data-driven way, and by using (mixed-effects) non-linear regression was able to to gauge both the influence of geography and that of social factors, deriving a single statistical model for both. Sandra Hansen (2016) examined the relation between elicited and spontaneous data in Baden, and Franz Manni (2017) researched the relation between linguistic and genetic diversity.

[5] This section rephrases the valedictory lecture (Nerbonne 2017) reviewing this research line.
[6] And not quadratic as the r^2 in the gravity equation would suggest.

Therese Leinonen (2010) combined a measure of vowel differences in formant space with dialectometric analysis, introducing a correlative perspective. Leinonen measured the speech of young adults (27 yr. old) and the speech of those nearly 40 years older, and extracted the most important dimensions. She then projected the results onto two maps, one for the older group and one for the younger one (Leinonen 2010:14). The result provides a striking picture of dialect leveling in modern Sweden, where dialect differences are rapidly being lost, as in the rest of Europe, too. Note that Leinonen's work simultaneously explores geographical (diatopic) as well as the influence of age.[7]

I won't attempt to sketch open problems in this research program at this point because I want to turn to remarks on related work in neighboring disciplines. I hope that the summary above of my work will be helpful as one assesses my further remarks. It should be clear that my perspective has focused on methods, and in particular on applying mathematical and computational tools to the study of language variation. The survey of related work will reflect this perspective.

Similar work in related (sub)disciplines

Let's therefore return to the plan for the paper, and in particular to the second part, dialectology's intellectual neighbors.

We can begin by noting that some of the continuations pursued in Groningen, which we just discussed, already forayed into adjacent fields. This work has already crossed borders, venturing afield from dialectology proper. In all cases we shall discuss, there is a need to understand, theorize about, or operationalize the notion LINGUISTICALLY SIMILAR. We note from the outset therefore that we survey these areas well aware that similarity is not a relation between two things, but rather among three. Two things are always similar to each other with respect to some property, which, importantly, doesn't imply similarity with respect to

[7] I've concentrated above mostly on the further achievements in dissertation-length projects, but colleagues in Groningen and elsewhere collaborated a great deal in other ways, too. First and foremost among these was Charlotte Gooskens who applied dialectometric methods to the question of the intelligibility of (closely related) languages.She in turn collaborated with Vincent van Heuven, Renée van Bezooijen, Femke Swarte and Jelena Golubovic. We discuss this work in more detail below. In Groningen some of the other colleagues who got involved in the collaboration were Leonie Bosveld, Çağrı Çöltekin, Bob de Jonge, Peter Houtzagers, Remco Knooihuizen, Sebastian Kürschner, Hermann Niebaum, and Ernst Wit. The colleagues elsewhere included Harald Baayen, Erhard Hinrichs, Bill Kretzschmar, Timo Lauttamus, Philippe Mennecier, Simonetta Montemagni, Lisa Lena Opas-Hänninen, Petya Osenova, Esteve Valls, and Vladimir Zhobov. The work benefited greatly from all this collaboration!

other dimensions. Put in another way, there are many notions of similarity, and we aim to be clear about what sorts we discuss.

We'll turn now to a discussion of eight (or perhaps nine) areas where the linguistic similarity we investigate in dialectology also plays a role, namely detecting and referencing transliterations of foreign names; avoiding confusable names of medicines; (spoken) word recognition; the intelligibility of closely related languages; sociolinguistics, including the study of regiolects; language contact; language education and language therapy; and diachronic (genealogical) linguistics.

Practical applications

Let's begin with two practical applications. As modern scientists, we are proud not only of our calculations and explanations but also of our contributions to practical products and services.

As noted above, Peter Nabende wrote a PhD thesis in information retrieval in 2011 focusing on how to search for names in western (Latin-based) texts even when those names come from languages with non-Latin writing systems, such as Urdu, Hindi, Russian or Chinese. He began by investigating the usefulness of the edit distance measure used in dialectology (and discussed above) in recognizing what specialists call transliterations, but turned quickly to dynamic Bayesian nets for his solution. The point here is that it can be very useful to be able to judge how similar two pronunciations are. The Urdu name پرویز مشرّف is generally transliterated as 'Musharraf' in English, but Nabende could point to several other variants. He set out to identify these by checking whether their pronunciation resembled the Urdu.

Greg Kondrak and Bonnie Dorr conducted a study in 2006 on the usefulness of edit-distance in detecting potential drug names that might be confusing if admitted by the US Food and Drug Administration (Kondrak & Dorr 2006). The confusion could be life threatening, for example, when a patient needed an injection of *Narcanbut* instead got the drug *Norcuron* and went into cardiac arrest. Using variants of the same edit distance measure discussed above the authors were able to detect names that might be confusing either in writing or in pronunciation. Their best-performing system used a combination of various string comparison algorithms.

It is noteworthy that both applications involving determining the degree to which words are similar with respect to pronunciation and/or orthography. In fact, both applications investigated both pronunciation and orthography, reasoning that they might both be relevant to the task at hand.

Word recognition

Psycholinguists study word recognition using the notion of a NEIGHBORHOOD. Landauer & Streeter (1973), working on written word recognition, are often credited with this idea. In their work every word has a neighborhood which consists of all the words in the lexicon which differ from that word by exactly one letter, meaning that one could obtain a neighboring word by replacing a single letter in the word in focus. They showed that misperceptions of written words tended to identify words in the stimulus's (graphemic) neighborhood. Luce & Pisoni (1998) shifted the focus to spoken word recognition and applied the edit distance algorithm, thus generalizing the one-letter substitution to a one-phoneme difference, where the difference might arise not only via substitution, but also through insertion and deletion (see Table 1 above). This was at roughly the same time that edit distance was introduced into dialectology (Kessler 1995; Nerbonne et al. 1996).

Of course it only makes sense that misperceptions would tend to target phonologically similar words. It would therefore be most interesting to test the more sensitive versions of edit distance that have been developed in dialectology (Heeringa 2004; Wieling et al. 2012) on the word recognition problem.

Again, the studies wished to determine the degree to which words are similar with respect to pronunciation and/or orthography.

Intelligibility

Charlotte Gooskens has applied dialectology-inspired language similarity measures, including edit distance for pronunciation to attempt to predict when closely related languages should be mutually intelligible. She and her students worked on German, Romance and Slavic languages, always taking a subset of five related languages for research. Swarte (2016), working on Germanic, found unsurprisingly that experience in a foreign language was the best prediction of intelligibility, followed by lexical overlap, and then pronunciation similarity as measured by edit distance. Gooskens and her group have also

considered the potential consequences of their work for language education, exploring the idea that focusing on language comprehension (as opposed to production) might effectively enable higher levels of proficiency. So this neighboring field has practical applications as well – even if its primary motivation is theoretical, not practical.

These studies of intelligibility obviously generalize the notion of similarity – from similarity with respect to pronunciation and/or orthography to similarity with respect to other levels of linguistic structure, especially the lexicon but also syntax.

It is worth noting that mutual intelligibility has also been discussed in dialectology, especially in conjunction with the idea that one might identify the dialects of a language as the mutually intelligible varieties. Given this idea, varieties that are not mutually intelligible would be regarded as (belonging to) different languages.

But the idea doesn't have many adherents, first because the notion "intelligible" itself is not sharp. One has to specify the degree to which a foreign variety might be intelligible, which suggests that varieties are not sharply divided into separate groups of mutually intelligible sub-varieties. The second problem with the idea is that, by itself, it doesn't say how one should construe asymmetric intelligibility, cases such as the intelligibility of Spanish for Portuguese speakers of that of Swedish for Danish speakers.

Sociolinguistics

As noted in the introduction, sociolinguistics shares dialectology's focus on language variation, replacing the geographic perspective with a social one, so applications of dialectological ideas in sociolinguistics would seem congenial. Social perspectives often involve class, race, age, occupation, sex/gender, the networks of friends and acquaintances that people interact with, and the groups with which people identify. It's a lively neighboring field!

One application of dialectological techniques was noted earlier, Leinonen's (2010) demonstration of how Swedish dialect diversity is declining. We remark here that her study goes beyond dialectology proper by including an investigation of the effects of age, but we emphasize here that she applied the same techniques for measuring pronunciation differences developed in dialectology.

Another sociolinguistic study using these techniques investigated whether the degree to which the formation of so-called regiolects is proceeding as Auer & Hinskens (1996) have claimed. The claim is illustrated below (Figure 2), namely that European regiolects are found in an intermediate position between the base dialects of the village in the area and the standard language of the country. This makes perfect sense. Speaking a variety closer to the standard should improve the chances of intelligibility, and speaking a variety with more local flavor than the standard signals an identification with the region. The cone provides a picture of this.

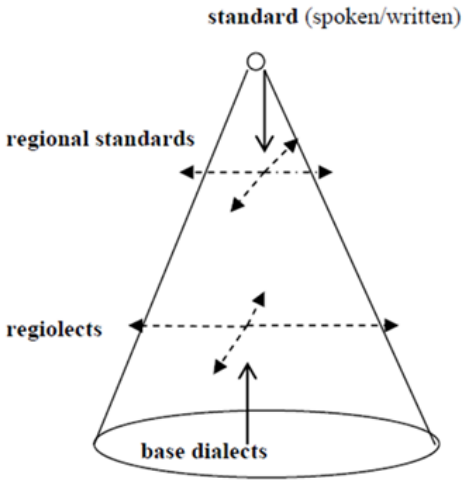

Figure 2 Hinkens and Auer's cone of regiolect formation. Base dialects (at the base of the cone) are atrophying due to mobility, compulsory education in the standard language, and (perhaps) mass media. A standard language is normally available, and people are expected to function in it. Various processes of accommodation lead to intermediate forms which may become conventional, i.e. "standard' in some (non-explicit) sense.

When this idea was tested on Dutch regional radio broadcasters in the Netherlands and Belgium, the results were surprising (Nerbonne et al. 2013). While the West Flemish broadcaster spoke in conformance with the picture, two broadcasters used speech that was extremely different from the standard, more than most of the base dialects were! These might be pictured "below" the base of the cone. In the other cases that were examined the regional speech was not further from the standard than the base dialects, but they were also not closer to the base dialects than the standard itself. They might be pictured outside the cone but clearly above the base dialects.

The conclusion that the speech of the eight radio broadcasters studied did not conform to the Auer-Hinskens conical model is solid, but since the broadcasters cannot be taken as representative, no conclusion is warranted about the general model.

The study of an area closer to here focused on the effects of a linguistic reform in Catalonia, where the reform was also used in education (Wieling et al. 2018). The authors could indeed show that language patterns had changed more in Catalonia than in neighboring Aragon – where the reform had not been effected. The analysis was illustrated by a gradient from Aragon, where no reform had been instituted, to Catalonia in the east, where more standard forms have become more popular.

In this case we were also interested in whether the aggregate, dialectometric analysis yielded the same picture of standardization as more standard sociolinguistic views that had focused on single features. As expected, the aggregate view is normally quite similar, but one of the single-feature studies we examined, Recasens' (1991) study of how the Catalan palatal lateral [ʎ] is giving way to the approximant [j], does not reflect the general picture. The choice of feature is important.

Let's note that all these studies have focus on similarity with respect to pronunciation, just as (most of) the dialectology studies noted in the introductory sections.

Let's also make time for a brief diatribe, noting that all three sociolinguistic studies have concerned, not the prototypical focus on a single sound change in progress (which has justifiably earned a stable position in current methods for studying language change) but rather large-scale changes.

- All the changes influenced by dialect leveling in Swedish
- All the changes influenced by regiolect formation in Dutch
- All the changes influenced by the introduction of the standard in Catalonian schools

Since dialectometry is uniquely equipped for studying large aggregates of data, the list above suggests a role for dialectometric-like work in sociolinguistics that might provide a complementary perspective to that of most current theorizing. Single-feature analyses – whether in dialectology or sociolinguistics – run the risk of hastily generalizing from that single feature. The situation undoubtedly improves in work based on a set of ten or twenty features, but it is difficult to argue for representativeness for any particular choice of features.

Contact languages

Zhang et al. (2021) have shown in unpublished work that loan words may be detected fairly well using data from field work in Tajikistan, Kyrghyzstan and Uzbekhistan based on the Swadesh list. Words from different language families (Turkic vs. Indo-Iranian) that were responses to the same question (given in Russian) were regarded as meaning (roughly) the same. When the pronunciations were shown to be more similar than expected, they were hypothesized to be loans, and this was right about 75-85% of the time.

In this case a special variant of comparison, List's (2012) SOUND-CLASS ALGORITHM, which ignores small differences, proved to be superior to the simpler (and more sensitive) edit-distance measures. The authors speculate that the superiority is due to sounds needing to conform to the phonology of the borrowing language. These sorts of differences should not be regarded as evidence against loan-word status.

The dialectological focus on phonological similarity in the realization of comparable concepts is obviously very close to the focus needed to detect loan words in the study of language contact. Zhang et al. (2020) appears to be the first paper to try to exploit this shared focus, but other opportunities abound.

Speech disorders and foreign accents

In the study of speech disorders and in the study of foreign accent, it is often useful to be able to measure how normal or how native-like pronunciations are. Sanders and Chin (2009) showed that the deviations in the speech of cochlear implant bearers correlate well with the judgments of experts on language disorders. The deviations in the speech of cochlear implant bearers were measured using a variant of the edit distance algorithm, and the authors cite dialectological work as their inspiration for this idea. Since expert judgments need not be applied in every case, the algorithm will not be of practical value to cochlear implant bearers, but the demonstration suggests that a broader range of speech disorders might be detected automatically using the algorithm made popular in dialectology.

Wieling et al (2014) showed that an edit-distance measure of foreign accent strength correlates nearly as well with informed judgment (*r=0.81*) as the judgments correlated with one another (*r=0.84*) in the speech of foreign learners of English. Although the focus of the paper was on the validation of the algorithm as a measure of pronunciation difference, the application area, foreign accents, suggests opportunities for further deployment in second-

language acquisition, which is in turn closely related to those branches of applied linguistics concerned with learning second languages. Bloem et al. (2016) develop this idea, showing that it is not only possible to measure how strongly foreign speech is accented, but that it also possible to identify the elements (phonemes) responsible for the accent.

In both of these cases, dissimilarity with respect to pronunciation was the key concept. In Sanders & Chin's work this functioned as a sign of a hearing disorder, and in Wieling et al.'s work it signaled non-native speech.

Historical Linguistics

For more than fifteen years, researchers have been applying phylogenetic algorithms developed to model evolution in biology linguistic material, but using data at a categorical level – lists of cognates which are or are not shared, or other features manually prepared (Atkinson & Gray 2005). Judging whether or not two words are cognate cannot be relegated to untrained staff, and even expert historical linguists need to study the languages being analyzed before they can decide on whether words are cognate.

Jäger (2015) began applying an edit-distance algorithm to word lists in historical linguistics, eliminating the need for expert cognate judgment and providing a more sensitive measure of word-relatedness. Interestingly, he independently developed a very similar segment-weighting scheme developed earlier in dialectometry, and he employed a measure applied earlier in dialectology to assay how reliable estimates of relatedness are.

Of course, the path between dialectology and historical linguistics is well trodden; Schirmunski (1962:79ff) relates a nineteenth century dispute between Bremer, a Neogrammarian (historical linguist) and Wenker, one of the first researchers to tackle dialectology systematically. It is clear that the relation of similarity between cognates must be modulated by the regular sound correspondences among the languages, but it is a further application of dialectological ideas in a different area.

Let's note that a fairly specialized sort of phonological similarity is crucial in historical work, namely phonological similarity with respect to regular sound changes. As Hock & Joseph (2009) and many other textbooks on historical linguistics take pains to note, chance similarity doesn't signal historical relatedness at all; instead, it adds sand to the gears.

Dialectology and Sociolinguistics

I promised to reflect on the sociolinguistics-dialectological fusion initiated by Chambers and Trudgill, and it's come time to do that.

As we've noted, **dialectology** – at least as practiced in dialectometry – focuses macrolinguistically on aggregate differences, while Labovian sociolinguistics has emphasized the importance of capturing sound changes in progress together with their social motivation; the latter has come to be referred to as the social meaning attached to the sound change. The Labovian program has been very successful, not only in sociolinguistics proper, but in understanding the influence of social factors on language change (Labov 2001). But the tradition has been to focus on one sound change at a time, for example, pronouncing syllable-final /r/ in New York English in the 1960's (Labov 1986).

Dialect differences are seldom, if ever, restricted to single features. Instead dialects normally differ with respect to a large number of features. Moreover, dialect speakers born into a community may attach no particular social meaning to the differences of their speech (with respect to others), nor must they perceive any. Dialects are accumulations of many differences, which might all have arisen accompanied by social identifications, but which new generations accept without appreciating their origins.

I argued in the section above on "Sociolinguistics" that several topics involving the social distribution of linguistic variation benefit from an approach that seeks to understand the large range of elements that may vary, even at the cost of ignoring some of the individual variation. These included the effects of dialect leveling in Swedish, the changes influenced by regiolect formation in Dutch, and the changes brought about by the introduction of the standard language in Catalonian schools. These studies – and similar topics – suggest that research techniques aimed at analyzing the aggregate differences among varieties serve a useful purpose. These sorts of techniques have been developed over the past several decades in dialectology.

If these studies and topics indeed belong to sociolinguistics, then the difference with respect to dialectology is perhaps primarily a difference in the central position of the Labovian emphasis on the importance of studying individual changes as they occur.

Conclusion

This paper has reviewed some of the "excursions" of dialectological ideas and techniques into neighboring linguistic sub-disciplines and into the applied sciences. The paper is not all comprehensive with respect to applications: there are, for example, applications of dialectology in forensic linguistics and in speech training for actors (Watt 2018)

The existence of a range of disciplines that involve linguistic similarity suggests an alternative "intellectual ecology" for dialectology, one that emphasizes its relations to other disciplines studying linguistic similarity. Chambers and Trudgill suggested that dialectology and sociolinguistics be viewed as sister disciplines within the common field of VARIATIONIST LINGUISTICS (Chambers & Trudgill 1980). This perspective has served well, but it should not distract us from their differences (discussed in the last section) nor from the substantial interests that dialectology shares with other fields, which we have surveyed here.

It is clear that the notion LINGUISTICALLY SIMILAR has been central in our review of dialectology-related work, and it should play a role in how we reflect on the relations among the various linguistic sub-fields and areas of application. Szmrescanyi & Walchi (2014) plead for a similar perspective in which dialectology, typology and register analysis are the sub-fields in question. Some fields outside linguistics likewise focus on linguistic similarity. Non-traditional authorship attribution views similarity in the relative frequency of the most frequent words (function words) as the most important evidence of authorship (Nerbonne 2007). The point of noting the common questions that the various subfields in are posing and the common methods they are applying, is, of course, to benefit from one another's work more regularly and more systematically. One might speculate about the need to view all the work studying similarity as justifying a disciplinary umbrella conceived as "comparative linguistics", but this term is usually take to be synonymous with comparative grammar,[8] the name the 19th century used to refer to historical linguistics. Still, if the term weren't already in use for another purpose, the focus on a comparative approach, which is needed to determine similarity, would fit quite well.

References

[8] See "Comparative linguistics". *Encyclopaedia Britannica*. Britannica.com. 2011. Retrieved 29 June 2019.https://www.britannica.com/science/comparative-linguistics

Atkinson, Q. D., & Gray, R. D. 2005. “Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics” *Systematic Biology*, *54*(4), 513-526.

Aurrekoetxea, G., Fernandez-Aguirre, K., Rubio, J., Ruiz, B., & Sánchez, J. 2013. “’DiaTech’: A new tool for dialectology” *Literary and Linguistic Computing*, *28*(1), 23-30.

Bloem, J., Wieling, M., & Nerbonne, J. 2016. “Automatically identifying characteristic features of non-native English accents” In: Côté, M. H., Knooihuizen, R., & Nerbonne, J. (eds.) *The future of dialects: Selected papers from Methods in Dialectology XV*. (Language Variation 1), Berlin: Language Science Press. 153-170.

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

Chambers, J. K., & Trudgill, P. [1]1980, 1998. *Dialectology*. Cambridge: Cambridge University Press.

Embleton, S. 1993, “Multidimensional scaling as a dialectometrical technique: Outline of a research project” In: Köhler, R. and Rieger, B. (eds.) *Contributions to quantitative linguistics*. 267-276. Dordrecht: Kluwer.

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

Gooskens, Ch. & Heeringa, W. 2004. “Perceptual evaluation of Levenshtein dialect distance Measurements using Norwegian dialect data.” *Language Variation and Change* 16(3), 189–207.

Hansen-Morath, S. 2016. *Regionale und soziolinguistische Variation im alemannischen Dreiländereck. Quantitative Studien zum Dialektwandel*. PhD Diss. Albert-Ludwigs Universität, Freiburg.

Heeringa, W. 2004. *Measuring dialect pronunciation differences using Levenshtein distance.* PhD Diss. University of Groningen.

Jäger, G. 2015. “Support for linguistic macrofamilies from weighted sequence alignment” *Proceedings of the National Academy of Sciences*, *112*(41), 12752-12757.

Kessler, B. 1995 "Computational dialectology in Irish Gaelic" *Proc. 7th conf. European Chap. Association for Computational Linguistics*. Burlington, Mass.: Morgan Kaufmann Publishers.

Kondrak, G., & Dorr, B. 2006. Automatic identification of confusable drug names. *Artificial Intelligence in Medicine*, *36*(1), 29-42.

Labov, W. 1986. “The social stratification of (r) in New York City department stores”. In Allen, H. B., & Linn, M. D. (eds.) *Dialect and language variation*. New York: Academic Press. 304-329.

Labov, W. 2001. *Principles of linguistic change Vol. 2: Social factors*. (Language in Society *29)*. Oxford: Oxford University Press.

Landauer, T. K. & Streeter, L. A. 1973. ”Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition” *Journal of Verbal Learning and Verbal Behavior*, *12*(2), 119-131.

Luce, P. A., & Pisoni, D. B. (1998). “Recognizing spoken words: The neighborhood activation model” *Ear and hearing*, *19*(1), 1-36.

Leinonen, Th. 2010. *An acoustic analysis of vowel pronunciation in Swedish dialects*. PhD Diss., University of Groningen. hdl.handle.net/11370/51096135- 2965-4d55-882e-4f078cd52057

List, J.-M. 2012. “SCA: phonetic alignment based on sound classes” In: Lassiter, D. & Slavkovik, M. (Eds.).*New directions in logic, language and computation*. *ESSLLI 2010 and ESSLLI 2011 student sessions, selected papers*. Vol. 7415. Springer, Berlin, Heidelberg. 32-51.

Manni, F. 2017. *Linguistic probes into human history*. PhD Diss. University Groningen.

Nabende, P. 2011 *Applying dyanamic Bayesian networks in transliteration and detection*. PhD Diss. University Groningen. hdl.handle.net/11370/2eb9dbf9-0e9c4b8c-a3e2-410b30215b16

Nerbonne, J. 2007. “The exact analysis of text” Foreword to the 3rd edition of Frederick Mosteller and David Wallace’s *Inference and disputed authorship: The federalist papers.* Stanford: CSLI.

Nerbonne, J. 2009. “Data-driven dialectology” *Language and Linguistics Compass*, *3*(1), 175-198.

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

Nerbonne, J. 2013. “How much does geography influence language variation?”. In: P. Auer, M. Hilpert, A. Stukenbrock, & B. Szmrecsanyi (eds.) *Space in language and linguistics: Geographical, interactional, and cognitive perspectives*, *220*-236.

Nerbonne, J. 2017. “Respecting local variation” In: A. Iglesias & A. Ensunza (eds.) *Gotzon Aurrekoetxea lagunarterik hara* (*Festschrift for Gotzon Aurrekoetxea*). Victoria Gasteiz:UPV/EHU. 13-24.

Nerbonne, J. 2017b. *Humanities, exactly!/Letteren, exact!.* Text of valedictory lecture. Available at http://www.let.rug.nl/nerbonne/papers/HumanitiesExactly-2017-May.pdf

Nerbonne, J., Heeringa, W. & Kleiweg, P. 1999. Edit distance and dialect proximity. In D.Sankoff & J. Kruskal (eds.) *Time warps, string edits and macromolecules: The theory and practice of sequence comparison*, 2nd ed., Stanford, CA: CSLI. pp. v–xv.

Nerbonne, J. & Heeringa, W. 2007. "Geographic distributions of linguistic variation reflect dynamics of differentiation." In: S. Featherston and W. Sternefeld (eds.) *Roots: Linguistics in search of its evidential base*. 267-297.

Nerbonne, J., Heeringa, W., Van den Hout, E., Van der Kooi, P., Otten, S., & Van de Vis, W. (1996). "Phonetic distance between Dutch dialects". In Durieux, G., Daelemans, W.& Gillis, S. (eds.) *CLIN VI: proceedings of the sixth CLIN meeting*. 185-202.

Nerbonne, J., & Kleiweg, P. 2003. "Lexical distance in LAMSAS" *Computers and the Humanities*, *37*(3), 339-357.

Nerbonne, J., van Ommen, S., Gooskens, C., & Wieling, M. 2013. Measuring socially motivated pronunciation differences. In L. Borin & A. Saxena (eds.), *Approaches to measuring linguistic differences*. Boston & Berlin: Mouton De Gruyter. 107-140.

Prokić, J. 2010. *Families and resemblances*. PhD Diss. University Groningen. hdl.handle.net/11370/be920d0e-2f88-417a-89d2-7704f962b9e4

Recasens, D. 1991. *Fonètica descriptiva del català: (assaig de caracterització de la pronúncia del vocalisme i consonantisme del català al segle XX*) (Vol. 21). Institut d'estudis catalans.

Sanders, N. C., & Chin, S. B. 2009. "Phonological distance measures" *Journal of Quantitative Linguistics*, *16*(1), 96-114.

Shackleton Jr, R. G. 2010. *Quantitative assessment of English-American speech relationships*. PhD Diss. University Groningen. hdl.handle.net/11370/b8a69e64-7f7f-4643-98b4-aa0097c5cf20

Schirmunski, V. M. 1962. *Deutsche Mundartkunde. Vergleichende Laut-und Formenlehre der deutschen Mundarten.* Berlin: Akademieverlag.

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

Szmrecsanyi, B., & Wälchli, B. (eds.). 2014. *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*. Berlin: Walter de Gruyter.

Spruit, M. R.2008. *Quantitative perspectives on syntactic variation in Dutch dialects*. PhD Diss. University Amsterdam.

Swarte, F. 2016. *Predicting the mutual intelligibility of Germanic languages from linguistic and extra-linguistic factors*. PhD Diss. University Groningen. http://hdl.handle.net/11370/2b6e7325-761b-4056-b883-48ed53808715

Voskuil, J.J. 1996-2000. *Het bureau*. Vol.1-7. Amsterdam: Van Oorschot.

Watt, D. 2018. "Applied dialectology: dialect coaching, dialect reduction, and forensic phonetics." In: Boberg, C., Nerbonne, J. & Watt, D. (Eds.) *The handbook of dialectology*, Boston: Wiley-Blackwell, 219-232.

Wieling, Martijn B. (2012) *A quantitative approach to social and geographical dialect variation*. PhD Diss. University Groningen. http://hdl.handle.net/11370/cd637817-572f-4826-98c1-08272775fb64

Wieling, M., Margaretha, E., & Nerbonne, J. 2012. Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, *40*(2), 307-314.

Wieling, M., Bloem, J., Mignella, K., Timmermeister, M., & Nerbonne, J. 2014. "Measuring foreign accent strength in English: Validating Levenshtein distance as a measure" *Language Dynamics and Change*, *4*(2), 253-269.

Wieling, M., Valls, E., Baayen, R. H., & Nerbonne, J. 2018. "Border effects among Catalan dialects". In Speelman, D., Heylen, K., & Geeraerts, D. (eds.) *Mixed-Effects regression models in linguistics*. Springer: Berlin. 71-97.

Zhang, L., Manni, F., Fabri, R., & Nerbonne, J. 2021. "Detecting loan words computationally" Accepted to appear in Aboh, E., DeGraff, M., & Vigouroux, C. et al. (eds.) *Ecology roles the dice: Festschrift for NN.*