Respecting local variation[1]

John Nerbonne
University of Groningen and University of Freiburg
J.Nerbonne@RuG.NL

0. Abstract. Data collection in dialectology is focused on the geographic distribution of variation at a fairly large scale, e.g. across entire language areas or across a large portion of one. The normal procedure is to choose a set of data collection sites fairly evenly distributed over (inhabited) areas and to analyze the variation among the sites, even while ignoring variation within a given site. 'Local variation' refers to the variation within the site, i.e., below the level of the data collection site, and is also referred to as 'multiple responses'. We recall how common variation at levels is, including variation at sites, justifying the development of measures that compare sets of responses and even multisets. The focus is on categorical data, but I will pay attention to the situation in which string data (pronunciation transcriptions) or numeric data (e.g., vowel formants) is to be analyzed.

Keywords: dialectology, dialectometry, multiple responses, Jaccard distance, Canberra distance

1. Background and an appreciation

Gotzon Aurrekoetxea has long championed 'Diatech', which he and colleagues at the University of the Basque Country, Vitoria-Gasteiz, have developed (Aurrekoetxea et al. 2013, 2016). Diatech has the potential to enable dialectologists to conduct technically difficult analyses without requiring them to acquire computational skills. Diatech supports not just analyses of Basque, but potentially of any language. Naturally users must be knowledgeable about data analysis to interpret difficult analyses intelligently, but they are spared the burden of implementation. Computationally less skilled dialectologists benefit the most, but everyone benefits from publicly available tools based on "common scientific assumptions" since these are well understood, and since analyses using common tools are more easily compared.

Hans Goebl began this effort with Visual Dialectometry (VDM) in collaboration with Edgar Haimerl (Haimerl 2006), and there has been an effort in Groningen, as well, with an emphasis on providing a web-application, and on analyzing phonetic transcriptions using edit distance (Nerbonne et al. 2011; Leinonen et al. 2016). It's a sign of the health of dialectometry that three different packages – all sharing the dialectometric emphasis on aggregate analyses – but each with its own specific focus. If nothing else, the existence of the

[1] I am indebted to Prof. Hans Goebl, Salzburg, for exchanges on the topic of this paper, which, however, justifies no inference about the degree of agreement between us.

three packages should make it more likely that some software will continue to be available in spite of all the difficulties of maintaining software long after projects have finished.

2. The reality of local variation

Prof. Aurrekoetxea and I not only share a strong interest in dialectology, a methodological commitment to dialectometry, and a programmatic view on the modern importance of providing software to fellow researchers, we have also both been concerned with the analysis of data containing local variation, or multiple responses. Examples of this local, even individual variation occur at all linguistic levels. Lexically, informants may not be able to choose between synonyms when shown pictures of objects ('big' vs. 'large') or scenes ("Is the weather 'clearing up' or 'clearing off'?). In syntax there are book-length studies on variation in which elements appear in fronted position (Bouma 2008), and on variation in the order of elements in verb phrases – 'send Mary a note' vs. 'send a note to Mary' (Van der Beek 2008). In phonology informants may for example reduce or not reduce consonant clusters /lɛts.go/ vs. /lɛs.go/ 'let's go' and optional rules will always lead to potential variation, while at a finer level it would be surprising to encounter any fine acoustic data that was identical to the single Herz level when repeated, even for a single speaker. In summary, there are any number of data collections where variation extends well below the level of the collection site.

2.1 Goebl's scruples

As Aurrekoetxea et al. (2013: 24-25) note, however, Hans Goebl has been critical of applying dialectometric techniques to data collections that include multiple responses:

> With regard to the taxometric examination of the multiple responses, I should give a word of warning to all dialectologists. In linguistic geography the phenomenon of multiple responses is primarily considered as a qualitative problem. Dialectometry, however, necessitates converting this qualitative problem into quantitative relations, […] the 'beautiful' qualities of multiple responses are 'dissolved into the melting pot' of quantification. As it can be assumed that multiple responses contain sociolinguistic rather than purely linguistic information, the taxometric inclusion of multiple responses must be seen as a blending of sociolinguistic and geolinguistic information. (Goebl 1997: 28)[2]

This is essentially the common criticism of aggregating techniques, namely that the elements within the aggregation may be confounded by variables that the analysis ignores,

[2] Incorrect internal quotation marks crept into the quotation as Aurrekoetxea et al. (2013: 24-25) produce it. They have been removed here.

and like the common criticism, it is an admonition that one should be cautious about. But for several reasons the subject cannot simply be closed due to this danger.

First, the reproach that aggravating over large amounts of data accrues to virtually all dialectometric work, in which location is normally the only independent variable which is analyzed even though there are well-known effects of sex/gender, age, educational level, history of residence and occupation. Compilers of atlases normally try to hold such influences constant, e.g. by focusing on non-mobile, old, rural males (NORMs, see Chambers & Trudgill 1998: 29ff), but they are not always successful, and even among this group potential confounds remain – e.g., those involve occupation, degree of mobility, and upbringing. For example, 60 of Edmont's 700 informants (for the *Atlas Linguistique de France*) were women, and 200 of 700 were educated (Chambers & Trudgill 1998:29) It is often of interest to ignore these potential confounds and focus on the effect of geography.

Second, modern variationist linguistics insists that dialectology and sociolinguistics should be regarded as a single discipline (Chambers & Trudgill 1998, [1]1980), both concerned with language variation, its limits and distribution, its methods, and its role in reflecting the social identity of language users.

Third, some "sociolinguistic" variables are emphatically of dialectological interest. If dialect change is a dialectological topic, then it is interesting to include age among one's independent variables. This follows directly from Labov's frequently quoted view that language change is always preceded by a period of variation.

Fourth, the "sociolinguistic" variables need not remain confounds. On the contrary, they may play an explicit and enlightening role in analyses. SweDia is a database of 1,200 speakers – 600 nearly 65 years old (on average) and 600 of 27 years of age – from about 100 sites throughout Sweden and Swedish-speaking Finland (Eriksson 2004). Speakers' ages are explicitly recorded in the database, and each speaker pronounced five tokens of each of the 19 vowels in Swedish. This enabled Therese Leinonen (2010) to sketch the "flattening" of the Swedish dialect landscape in an insightful way. She extracted the first two formants of all more than 100,000 vowel tokens automatically and ran the entire aggregate set in a single multidimensional scaling (MDS) analysis, assigning colors to the dimensions in the way that has become popular in dialectology (Nerbonne 2010). She then separated the older from the younger speakers for the purpose of visualizing the distinction. The result may be seen in Figure 1. The left-hand side of the figure shows a great deal of variation among the older speakers that is simply missing from the younger ones, shown on the right. Finally, note that

to obtain these maps it was essential that the data collection contained geolinguistic information (location) as well as sociolinguistic information (age).

2.2 Multiple responses

In summary, there is good reason to include multiple responses in survey material. The multiple responses may reflect genuine variation among or within individual speakers, making it entirely appropriate to include it, and there are genuine research questions for which we should expect to encounter a variety of responses. These may involve contrasts with respect to sociolinguistic variables such as age or sex/gender, but it is also legitimate to describe overall variation, abstracting away from various influences other than location. Once the reality of local variation is acknowledged, it is then clear that we should prefer samples of local speech to single data points, obtaining thereby a more secure view of the variation.

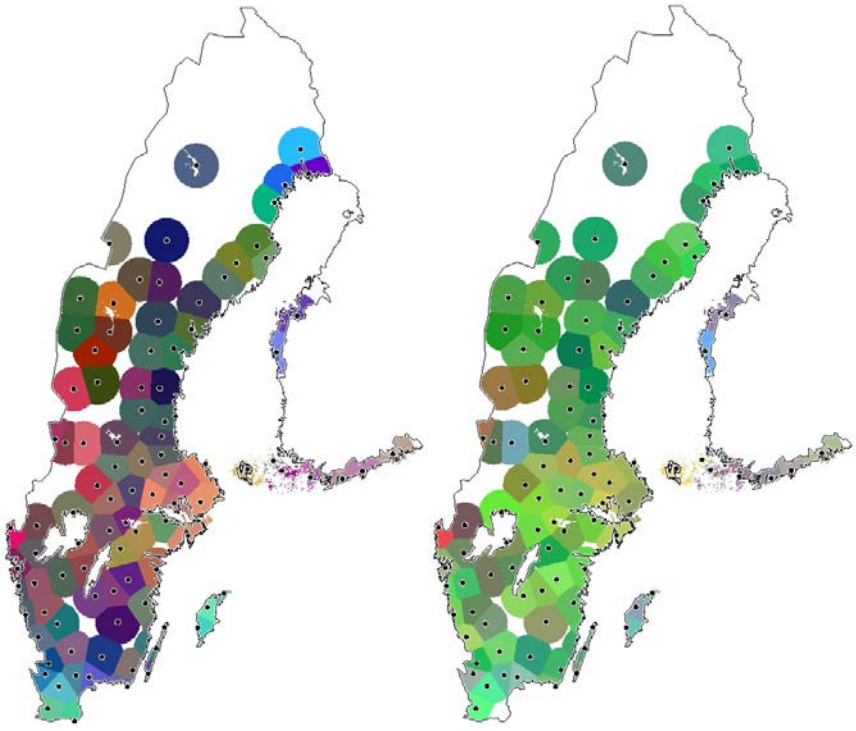

Figure 1. The result of a analyzing a database of mixed geolinguistic and sociolinguistic data. Leinonen (2010) analyzed the entire SweDia database of over 100,000 vowel tokens using MDS and assigned colors to the dimensions. She then separated the older speakers (left) from the younger speakers (right) demonstrating graphically the degree to which the Swedish dialects have been "leveled". The map on the left shows much more variation than does the one on the right.

3. Measuring the (dis-)similarity of features with multiple values

The question arises then of how best to analyze data in which local variation (multiple responses) occurs. We proceed, with most practitioners of dialectometry, from the case where

single values are counted as either the same or different, even while we note that various weighting schemes may be worthwhile. We consider therefore first the case of sets of categorical data and consider generalizations (to multi-sets or alternative comparisons) only later.

I have occasionally heard colleagues say that there's no real analytical problem here, since it is trivial to simply calculate the mean of all pairs of comparisons in case there are multiple responses. They must be thinking of simple cases such as comparing {*a*} at one site to {*b,c*} at another, in which case their procedure would yield *½ * (d(a,b)+d(a,c))*. In the absence of a difference metric other than (non-)identity, this would be just zero, since we assume that neither *b* nor *c* is identical to *a*. This procedure seems sound in this case and in the case of comparing {*a*} to {*a,b*}, where it would yield *0.5*, but it fails badly if applied to sites containing {a,b} on the one hand and {b,a} on the other. Since these two sets are identical, their difference is zero, but the mean difference of the pairs would be *0.33*, since one of the three pairs of individual items is <*a,b*>, which is non-zero.

3.1 Dice similarity coefficient and Jaccard measures

Aurrekoetxea et al. (2016:8) suggest therefore that an inverse of the 'dice' similarity coefficient might do service here. Dice is defined as follows to gauge similarity between pairs of sets (Manning & Schütze 1999:299):

$$\text{Dice-sim}(A,B) = 2*|A \cap B|/(|A|+|B|)$$

$$\text{Dice-diff}(A,B) = 1 - 2*|A \cap B|/(|A|+|B|)$$

The reader may quickly verify that *Dice-diff({a},{a,b}) = ⅓*, and *Dice-diff ({a,b},{b,a}) = 0,* suggesting that this is a serious contender.

There are two reasons indicating that we might wish to look further, however. First, the Dice measure is based purely on set cardinality so that there is no chance of basing feature differences on more sensitive pairwise measures, such as Goebl's inverse frequency weighting (*GIW*, see Goebl 1984) or, e.g., a measure that gauged pronunciations as more or less similar – as opposed to simply identical or not – such as the edit distance measure applied to phonetic transcriptions (Heeringa et al. 2002). Second, the difference coefficient is not a proper 'distance' measure, since it does not satisfy the triangle inequality. To see this, consider the three sets {*a*}, {*b*} and {*a,b*}. To satisfy the triangle inequality, the sum of the differences between any two of these sets must be greater than the difference between the

remaining pair of sets. But Dice-diff($\{a\},\{a,b\}$) = ⅓ = Dice-diff($\{b\},\{a,b\}$ (as noted above), while the sum (⅔) is less than Dice-diff($\{a\},\{b\}$) = 1. Since some procedures in dialectometry assume that distances are under analysis (multi-dimensional scaling), this counts against dice as a measure of difference between set-valued data cells. I hasten to add that MDS is often applied to difference measures that would not qualify as genuine distances, so that this point is not crippling, but it still counts against Dice.

As we look further for promising possibilities, it will be useful to abstract two elements from this discussion. We should prefer candidate measures not to be purely cardinality based, but rather to be based on the comparison of set members, allowing us to continue to use Goebl's inverse frequency weighting or the edit-distance measure of pronunciation similarity. And we prefer genuine distance measures, if possible.

Jaccard distances are genuine distances but since they are also based purely on cardinality, just as the Dice measure is, they also do not offer the flexibility we are looking for (Manning & Schütze 1999:299).

$$\text{Jacc-sim}(A,B) = |A \cap B|/(|A \cup B|)$$

$$\text{Jacc-diff}(A,B) = 1 - |A \cap B|/(|A \cup B|)$$

3.2 Manhattan measures

'Manhattan' measures derive their name from their applicability to deciding how far two points are in a plane where discrete steps are taken (Manning & Schütze 1999:304). This is like figuring out how far a taxi would need to travel in place like Manhattan. The taxi might drive five blocks horizontally, then two vertically, another one horizontally and finally another four vertically (see the yellow line in Fig. 2), but any path that always brings the taxi closer will finally travel (5+1) blocks horizontally and six blocks vertically, or twelve in total. It is defined as follows:

$$\text{Manh-Dist}(A,B) = \sum_{i=1}^{n} |A_i - B_i|$$

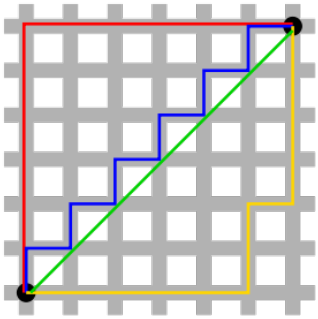

Figure 2 Any non-backtracking path from the origin (lower left) to the goal (upper right) will travel the same number of blocks, six horizontal and six vertical. 'Manhattan distance' is the sum of these, so-called for its similarity to city travel. Graphic from Wikipedia 'Taxicab geometry' (en.wikipedia.org/wiki/Taxicab_geometry, downloaded 27 Sept. 2016).

This sort of measure might not seem immediately useful, but, for example, if we are to analyze data accompanied by frequency information, we might accept the total difference in frequency to reflect the difference in the sets. If A={a,b,c} with relative frequencies <50%, 35%, 15%> and B={a,b,d} with frequencies <40%, 55%, 5%>, then the sum of the differences (= |50-40| + |35-55| + |15-0| + |0-5| = 50%) reflects the contribution of this feature to overall differences. One can thus compare histograms in this very simple fashion, assuming that enough values have been collected to provide a reliable picture.

But if it seems reasonable to compare the distribution of values in this way, then a rougher version of the same would be to simply count any non-zero frequency as 1 and zero frequencies as zero, and then to use the same sum of absolute differences. In the case under discussion this would yield a calculation of (= |1-1| + |1-1| + |1-0| + |0-1| = 2). If we want to normalize the measure to yield values between zero and one, we might then divide by the total number of different variants. We note that this yields the same value as the Jaccard difference measure, 1 - |A∩B|/(|A∪B|), so that the Manhattan measure used this way gives us no purchase on the problem of founding our measure of set differences on the differences between the pairs.

Manhattan distance is defined as Manh-Dist(A,B) = $\sum_{i=1}^{n}|A_i - B_i|$, i.e., for each dimension i, we sum the differences between A and B. For the sake of completeness, we note that 'Canberra distance' relativizes this to the magnitudes at each dimension, making it a kind of weighted Manhattan distance:

$$\text{Canb-Dist}(A,B) = \sum_{i=1}^{n}\frac{|A_i - B_i|}{|A_i| + |B_i|}$$

We shall not pursue this sort of weighting here, however.

3.3 A measure based on 'cover sets'

In comparing sets of linguistic responses, it is often difficult to identify dimensions for comparison ahead of time even though this is crucial for applying a Manhattan distance measure. If we compare two pronunciations for the English word 'thought' at one site, {[θɑt^{h}], [θɑt^{-}]} (where the final [t] may be released and aspirated or, alternatively, unreleased), with two at another {[θɑt^{h}], [θɔt^{h}]}, the first identical to one element in the first pair, and the second involving an 'open o' ([ɔ]) in place of the [ɑ], then we should wish to include both the [ɔ]/[ɑ] difference and the [t^{h}]/[t^{-}] difference (note that the second site has only released and aspirated t's), but in a third site we might find {[t̪^{h}ɔt^{h}], [θɔ^{u}t^{h}]} or {[t̪^{w}ɔt^{h}], [θɑ^{u}t^{h}]}, which seem to call for attention to different dimensions.

3.3.1 Covering

Proceeding with this line of thought, first developed in Nerbonne and Kleiweg (2003:§3.2)we do not attempt to define the dimensions of comparison ahead of time, but only insist that the set comparison is based on pairwise comparisons of all the elements encountered in both of the set. Given the pair of sets *A* and *B*, we will examine a subset C of ordered pairs where the first element is from A and the second from B. That is, we examine $C \subseteq A \times B$. Informally, we want to require that every element in A and every element in B play a role in the measure of the set difference. Formally, we can do this if we are able to refer to *C*'s first and second projections $C^1 = \{a_i | \langle a_i, b \rangle \in C\}$, $C^2 = \{b_i | \langle a, b_i \rangle \in C\}$, since we can then state the requirement simply that we want to base our calculations on a *C* that 'covers' *A* and *B*, i.e. where $C^1 = A$ and $C^2 = B$. We then take the mean of the distances of the minimal set of ordered pairs to be the distance between the two sets:

$$\text{CoverDist}(A, B) = \frac{1}{|C|} Min \sum_{i=1}^{|C|} dist(a_i, b_i), \text{ where } C \text{ covers } A \text{ and } B.$$

The intuition behind this proposal was not spelled out in Nerbonne & Kleiweg (2003), but the definition ensures that the distance between the two sets is strictly based on the distances between the individual elements and not merely on the presence or absence of elements and also that every element in the two sets must be involved.

3.3.2 Examples

To develop this intuition, we examine some examples. If A={a}, and B={a,b}, and distances are zero for identical elements and 1 for differing elements, the minimal cover is {<a,a>,<a,b>} (where the first pair is needed to cover B), and the cover distance is 0.5. If

instead A={a,b}, and B={a,b}, then the minimal cover is {<a,a>,<b,b>} and the cover distance is zero. Finally, let's consider an example where a finer distance measure between elements is available. Let $A = \{a_1, a_2, b\}$ and $B = \{a_3, c\}$, and let the element distances be given by

$$dist(x,y) = \begin{cases} 0 & \text{if } x = y \\ 0.1 & \text{if } x, y \in a_{i\in\{1,2,3\}} \\ 1 & \text{otherwise} \end{cases}$$

We note that this sort of range of distance values occurs frequently using Goebl's inverse frequency weighting (see above) or using an edit distance measure between pronunciation transcriptions. In this case the covering set is $C = \{< a_1, a_3 >, < a_2, a_3 >, < b, c >\}$ and the cover distance is 1.2/3 = 0.4. The dice distance would be one since the intersection of A and B is null, and in general, the cover distance discriminates more finely than dice does.

3.3.3 Aurrekoetxea et al.'s objection

Aurrekoetxea et al. (2013:26) ask us to consider the case where $A = \{a, b\}$ and $B = \{a, c\}$ $d(b,c) < d(a,c) < d(a,b)$. In this case, they correctly conclude that the minimal covering set would be $C = \{< a, a >, < b, c >\}$, and the cover distance would be $d(b,c)/2$. But, they go on to object, if a slightly different *c* were present, so that $d(a,c) < d(b,c) < d(a,b)$, then the covering distance would be, they maintain, $(d(a,c) + d(b,c))/3$. This conclusion is clearly wrong, however. The covering set would still be $C = \{< a, a >, < b, c >\}$, and the covering distance would still be $d(b,c)/2$. The first element in C still "covers" a, and since $d(a,a) = 0$, there can be no need to include $d(a,c)$ in the covering set solution.[3]

3.3.4 Efficiency and metatheoretic properties

The covering set solution seems well motivated, but the search for the minimal-distance covering set can involve examining all the subsets of $A \times B$, or at least all the subsets that cover A and B, and there might be $2^{|A\times B|}$ of these, which is computationally daunting. For large A and B, the computation would not be feasible. Adding a single element to A, would increase the time required by a factor of $2^{|B|}$. A greedy algorithm that attempts to cover the one set in a minimal fashion, then the next, has been serviceable, but would clearly not be suited for large sets. A stochastic approximation might then be most sensible. We noted above that we should prefer out measures to be distances in the mathematical sense, i.e. the value

[3] They also object that they'd expect distances between $A = \{a, b\}$ and $B = \{a, c\}$ to be always ≤ 0.5 (Auurekoetxea et al. 2013: 26), assuming apparently that $d(b,c) \leq 1$, and this indeed will also hold.

should be zero if, and only if we are dealing with identical sets, the measure should be symmetrical, and the triangle inequality should hold: $\text{CoverDist}(A, C) \leq \text{CoverDist}(A, B) + \text{CoverDist}(B, C)$, for all A, B and C. It is clear that $\text{CoverDist}(A, B) = 0$ if $A = B$, since the cover set will contain pairs of identical elements in this case, all contributing zero to the set distance. And conversely if $\text{CoverDist}(A, B) = 0$, then all the pairs in the covering set must also be of zero distance, which is only possible if they are all identical pairs, implying that $A = B$. It is also clear that $\text{CoverDist}(A, B) = \text{CoverDist}(B, A)$, since they are based on the same symmetrical distance function for elements. One might also be optimistic about the triangle inequality, since it would only seem possible to find $\text{CoverDist}(A, B) + \text{CoverDist}(B, C) < \text{CoverDist}(A, C)$ in case there were elements of these sets that failed to satisfy the triangle inequality, which should not happen as long as we are building on a genuine distance function. But optimism is not a rigorous proof, and I am at a loss how to provide one.

4. Further work

Neither inverse frequency weightings nor edit-distance measures on pronunciation transcriptions constitute genuine extensions to this measure of distance between sets, since we can incorporate their effects by using them in combination with the distance function on elements. Challenges for the future include the metatheoretic question of whether the cover difference measure is mathematically a distance, as well as its relation to other, better known measures (Manning & Schütze 1999:304).

In view of the great interest in the geolinguistics of social media (Nguyen et al. 2016), a pressing problem is the development of measures suitable for detecting affinities in very large sets of words – often derived from millions of messages.

5. Conclusions and prospects

The present paper confirms Prof. Aurrekoetxea's concern for developing measures of response similarity (or difference) for sets of responses rather than single responses, suggesting first how natural it is to find multiple responses in some circumstances and noting the scientific benefit that accrues to working with samples larger than one response per informant. This not only makes analyses more reliable, but also enables us to address a wide range of questions.

Bibliography

Aurrekoetxea, Gotzon, Fernandez-Aguirre, Karmele, Rubio, Jesús, Ruiz, Borja, & Sánchez, Jon, 2013, "'DiaTech': A new tool for dialectology". *LLC: Journal of Digital Scholarship in the Humanities 28*(1), 23-30.

Aurrekoetxea, Gotzon, Santander, Gotzon, Usobiaga, Iker, & Iglesias, Aitor, 2016 "Diatech: tool for making dialectometry easier". *Dialectologia: revista electronica* 17, 1-22.

Van der Beek, Leonoor, 2005, *Topics in corpus-based Dutch syntax*. Diss., University of Groningen.

Bouma, Gerlof, 2008, *Starting a sentence in Dutch: A corpus study of subject-and object-fronting*. Diss., University of Groningen.

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

Eriksson, Anders, 2004, "SweDia 2000: A Swedish dialect database." In: P. J. Henrichsen (ed.), *Babylonian Confusion Resolved. Proc. Nordic Symposium on the Comparison of Spoken Languages, Copenhagen Working Papers in LSP*. 33-48.

Goebl, Hans, 1997, "Some dendrographic classifications of the data of CLAE 1 and CLAE 2". In: W. Viereck & H. Ramisch (eds.) *The computer developed linguistic atlas of England*, App. 9, Tübingen: Max Niemeyer, 23-32.

Haimerl, Edgar, 2006, "Database design and technical solutions for the management, calculation, and visualization of dialect mass data". *Literary and Linguistic Computing* 21(4), 437-444.

Heeringa, Wilbert, Nerbonne, John, & Kleiweg, Peter, 2002, "Validating dialect comparison methods". In: W.Gaul & G.Ritter (eds.) *Classification, automation, and new media. Proc. 24th Ann. Conf. of the Gesellschaft für Klassifikation* Berlin: Springer. 445-452.

Leinonen, Therese, 2010, *An acoustic analysis of vowel pronunciation in Swedish dialects*. Ph.D. Diss., University of Groningen.

Leinonen, Therese, Çöltekin, Çağrı & Nerbonne, John, 2016, "Using Gabmap". *Lingua* 178, 71-83.

Manning, Christopher, & Schütze, Hinrich, 1999, *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Nerbonne, John, 2010, "Mapping aggregate variation". In A.Lameli, R.Kehrein & S.Rabanus (eds.) *Language Mapping. An International Handbook of Linguistic Variation*, Berlin: Mouton de Gruyter, Vol. 2, 476-495.

Nerbonne, John, Colen, Rinke, Gooskens, Charlotte, Kleiweg, Peter, & Leinonen, Therese, 2011, "Gabmap – a web application for dialectology". *Dialectologia: revista electronica*, Special Issue 2, 65-89.

Nerbonne, John, & Kleiweg, Peter, 2003, “Lexical distance in LAMSAS”. *Computers and the Humanities* 37(3), 339-357. doi:10.1023/A:1025042402655

Nguyen, Dong, Doğruöz, A. Seza, Rosé, Carolyn, de Jong, Franziska, 2016, “Computational sociolinguistics: A survey”. *Computational Linguistics* 42(3), 537-593.