Further contributions to Romance dialectometry

Oxford Research Encyclopedia of Linguistics

John Nerbonne, Groningen, Freiburg & Tübingen

Table of Contents

Summary 1
Keywords 2
1. Further contributions to Romance dialectometry 2
2. Methods 3
2.1 Data collection 3
2.3 Clustering 3
2.3 Edit distance 3
2.4 Multidimensional Scaling 4
2.5 Mixed-Effects Modeling 5
2.6 Geographic Influence and Generalized Additive Modeling 6
3. Novel Research Questions 7
4. Social Media 8
5. Languages 8
6. Prospects 9
7. Scholarship and Further Reading 9
8. Links to Digital Material 9
References 10

Summary

Jean Séguy and Hans Goebl were the founders both of Romance dialectometry and of dialectometry in general, which focused largely on Romance languages in its early years. While other attention to dialects had appealed to scholarly intuition to adduce the principles behind the geographic distribution of linguistic variation, dialectometry insisted on employing exact methods and on basing analyses on entire large samples of material. The samples may often be found in dialect atlases, which pre-dialectometry scholars had assiduously compiled.

Dialectometry thus continues the work of dialectology but always proceeding from entire large data collections and using more exact methods. It would nonetheless be a mistake to regard dialectometry purely as a methodological contribution, for dialectometry enables new research questions, as well as sharper versions of traditional ones, which it has also pursued.

Most centrally, dialectometry asks how geography influences linguistic variation, contrasting, for example, the perspective of discrete dialect areas with that of continua, or by examining the influence of distance or dialect areas on differences among varieties. It further seeks to characterize the sorts of variation involved, i.e., which sounds, words, or grammatical elements are involved, and aspires to characterize linguistic variation in ways that facilitate comparison to variation in other cultural dimensions such as, e.g., religion, ethnicity, or mobility.

Further work on Romance dialectometry has built on Seguy's and Goebl's innovative foundations (see ORE article on the Salzburg school) and has expanded the empirical scope of the research line to include non-geographic influences as well. Further contributions to this area of study have conducted analyses on different Romance language areas and have incorporated novel data collection protocols, new measures of pronunciation differences as encoded in phonetic transcription (edit distance), and novel statistical analyses, notably the application of multidimensional scaling, mixed-effects regression techniques and generalized additive models.

Emerging questions in dialectometry include attention to linguistic levels beyond phonetics and lexicology, the stricter validation of its techniques, more effective means of identifying the most important linguistic bases exploited in dialectal differentiation and naturally, the continued research into the enormous range of linguistic variation and its geographic distribution.

Keywords
Romance dialects, French, Italian, Catalan, Romanian, Portuguese, Sardinian, edit distance, multi-dimensional scaling, mixed-effects regression, generalized additive models

1. Further contributions to Romance dialectometry

The lion's share of attention will be devoted to methods, which have been emphasized in dialectometry. The further methodological contributions to Romance dialectometry build on Séguy's approach in proceeding from the measurement of differences in comparable material (e.g., words for the same concepts) between all the pairs of sites a substantial selection of comparable material, e.g., a list of many dozens of words in at least twenty data collection sites. The differences between each pair of comparable items at each site is determined, and the sum of these differences is taken to characterize the difference between sites, and a table of site X site differences is then analyzed to reveal the geographical (and perhaps other) influences on the variation. See the "Salzburg school" article in this encyclopedia for more detail on the procedures commonly used. We note here that the focus on aggregate differences was criticized by Woolhiser (2005), but also by Loporcaro (2009), who both found fault with its missing

attention to the linguistic bases of the differences. We return to this below in the sections on "Edit Distance" and "Mixed-Effects Regression".

2. Methods

2.1 Data collection

Bolognesi & Heeringa (2002, 2005) report on a project which collected data from Sardinia and analyzed it dialectometrically. The project was methodologically innovative in randomly selecting words for which pronunciations were elicited, unusually following statistical advice. Because a corpus was first collected, the investigators were also able to apply a weighting to the selection scheme so that more frequently used words were more likely to be selected. They applied edit-distance using phonological features to weight the segment operations (see below, section on "Edit Distance").

2.3 Clustering

Goebl (1982) had introduced clustering to dialectometry as a means of detecting groups which correspond to the dialect areas. But clustering without resampling is unstable, meaning that small differences in input can lead to very different results, which led Mucha & Haimerl (2005) to suggest so-called bootstrap methods, further developed by Nerbonne et al. (2008). Wieling & Nerbonne (2011) introduced a further refinement, namely bipartite spectral graph partitioning, where varieties together with their features are partitioned into groups. These can usually be mapped directly to dialect areas.

2.3 Edit distance

A very significant addition to the dialectometric toolbox has been the development of edit distance algorithms to measure the difference between the pronunciation of comparable words as rendered in phonetic transcription, where Kessler (1995) was the first to apply the technique to modern Irish dialects. While a full explanation of the algorithm would go beyond this article, it will be valuable to explain some aspects. The algorithm maps one pronunciation to another, using a limited set of operations, but always including at least insertion, deletion and substitution, and always seeking the least costly set of operations. A by-product of this procedure is an alignment of the phonetic transcriptions, which, when given the pronunciations of *albicocca* 'apricot' in a Tuscan dialect and in standard Italian, take the form shown in Table 1. Note the second segment, [r] or [l], and the fifth segment in Italian:

Putignano	a	r	b	i		ɔ	k:	a:
Standard Italian	a	l	b	i	k	ɔ	k:	a:

Table 1. The alignment produced by applying the edit distance algorithm to the transcription of the word for 'apricot' in the Tuscan dialect spoken in Putignano and standard Italian. Note that the words are the same except where standard Italian includes a [k] not used in Putignano and and where standard Italian [l] has been substituted for Putignano's [r].

All versions of edit distance assign a zero cost to the operation of substituting a segment for itself, the identity substitution. The Levenshtein-Damerau algorithm allows transposition of adjacent segments as an additional operation, which would correspond to linguistic metathesis. The basic algorithm assigns a cost of '1' to insertions, deletions and non-identify substitutions (resulting in a distance of two between the pronunciations in Table 1), but various other versions of the algorithm differ a good deal in the costs assigned to other operations. Heeringa (2004) showed how to ban consonant-vowel correspondences in alignments (with obvious exceptions for glides and the like) and experimented with substitution costs based on the segments' phonological features (Chap.3), and alternatively, with costs based on spectrogram distances (Chap.4), but found it difficult to improve much on the simple version when comparing it to the aggregate distances between pairs of sites in Norway. A data-driven alternative to determining costs was suggested in Wieling et al (2009), where an entire data set was first aligned using a simple version of edit distance, after which all the correspondences are collected and counted. Operations weights were determined by the relative frequency of the correspondence using a measure from information theory, pointwise mutual information (PMI). The result, PMI-based edit distance, could be shown to be a bit superior to simpler versions of edit distance, which we mention to note that even the simple versions of edit distance are not terribly inferior.

The additional value of having edit distance in the dialectometric arsenal lies first in the capability it confers to analyze phonetic transcriptions. Naturally, this material had not been ignored, and both Séguy and Goebl include individual pronunciation differences in their analyses. But this required manual intervention to isolate the sounds of interest, a step Goebl refers to as taxation. A second advantage of applying edit distance is thus to obviate the need for manual steps, streamlining analyses and allowing the expansion to entire phonetic transcriptions rather than select examples. Third, and finally, it adds computational rigor to the phonetic aspects of the research.

It is therefore not surprising that applications of edit distance have been popular in Romance dialectometry. Montemagni et al. (2012) used the PMI algorithm to align the material in the *Atlante Lessicale Toscano* (see links to digital materials), and used bipartite spectral graph partitioning (see Section above on "Clustering") to cluster Tuscan varieties together with their most characteristic elements. These turned out to be correspondences due to *gorgia Toscana* 'Tuscan throat', i.e., the spirantization and lenition of stops for which Toscana is linguistically known. This demonstrates that contemporary dialectometry is no longer subject to the criticism leveled by Woolhiser (2005) and Loporcarno (2009). The relation between aggregate analyses and their linguistic foundations can be adduced and is supported by one popular web application for dialectometry, Gabmap (Nerbonne et al. 2011).

2.4 Multidimensional Scaling

A further important development has been the use and further development of multidimensional scaling (MDS), first introduced to dialectometry by Embleton (1993). Given a set of data

collection sites and the distance between all the pairs of sites, MDS provides as good a representation of the differences among them as possible in a small number of dimensions. In the experience of dialectometrists using MDS, it is normally possible to represent large sets of sites faithfully (i.e., representing 80% and more of the variance) in only three dimensions, leading to insightful three-color maps (introduced in Nerbonne et al. 1999). MDS is often misunderstood to be simply an alternative to clustering and the dialect area maps which clustering produces. But the application of MDS allows probes into the degree to which the geographic distribution of variation is continuous, as opposed to categorical, as suggested by the partitioning of collection sites induced by clustering. Some modern dialectometric tools allow researchers to compare the results of clustering with those of MDS, e.g., Gabmap (Nerbonne et al. 2011). As noted above, this is one of dialectometry's central research questions.

2.5 Mixed-Effects Modeling

Johnson (2009) had introduced mixed-effects modeling to the other variationist linguistics subfield, sociolinguistics, demonstrating several advantages over the logistic regression model commonly applied in sociolinguistics. A mixed-effects model can use all the independent variables ("fixed effects") used in a standard regression model, to which so-called random effects are then added. Fixed effects distinguish only a few classes, such as gender or educational level, while random effects are used to model individual elements, for example speakers, data collection sites or individual linguistic items such as words.

Wieling et al. (2014) analyzed the lexicalizations of 170 concept in the Atlante Lessicale Toscano, with respect to standard Italian, including individual lexical items as random effects. The authors were able, for example, to display the effect of educational level on the lexical choices of their respondents (i.e., a by-concept random slope for community size). Highly educated speakers were more likely to use the standard Italian words for the bird species *upupa* 'Eurasian hoopoe' (bird species), *abete* 'fir tree', *allodola* 'sky lark' or *orzaiolo* 'stye' (eyelid infection), while less educated ones preferred the dialect words for *verro* 'male swine, boar', *cocca* 'end', braciola 'cutlet', *cascino* 'mold'. See Fig. 1 for a graphic representation. Mixed-effects modeling thus offers a further answer to Woolhiser's (2005) criticism that dialectometry focuses too much on aggregate relations while ignoring their linguistic bases.

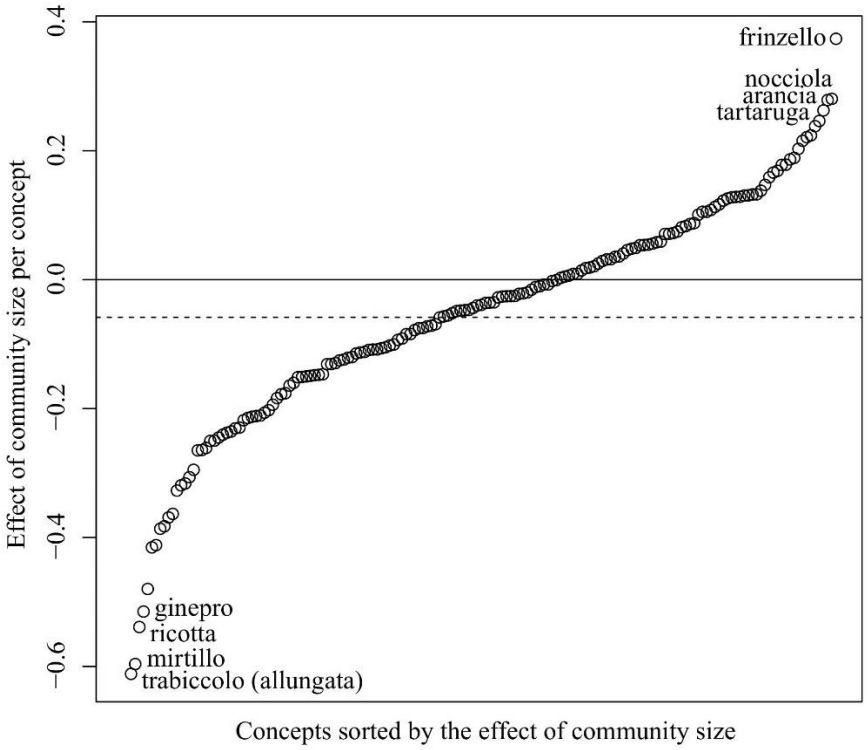

Figure 1. The effect of education on the use of standard Italian (lower on the y-axis) vs. Tuscan dialect (higher). The y coordinate represents the coefficient assigned to the variable for community size, and the x-axis represents educational level, sorted from higher to lower. The dotted line represents what the coefficient for community size would have been if it had not been treated as a random effect. See text for further comments. Figure taken from Wieling et al. (2014).

2.6 Geographic Influence and Generalized Additive Modeling

As noted above, dialectometry has made the fundamental dialectological insight – the dependence of variation on geography – more exact by analyzing the aggregate sums of distances (between all pairs of sites), and their dependence on geographic distance, e.g., using linear regression (or a transformation) (Séguy 1973). Naturally, no one imagines there to be a direct effect of space on language, but distance is a good proxy for the likelihood of contact. Thus, it was not surprising that Gooskens (2005) could show that travel time was a better predictor than simple distance. Nerbonne (2010) examined six language areas, showing that aggregate linguistic differences were predictable based on the logarithm of geographic distance $0.16 \leq r^2 \leq 0.37$, and Embleton et al. (2018) examine as-the-crow-flies geographic distance, travel distance and travel time as predictors of Romanian dialect differences, concluding that travel distance and travel time were generally, but not always the better predictors.

Distance is a simplified, one-dimensional reduction of geography, so it is interesting to examine a more sophisticated view introduced in Wieling et al. (2014), who apply generalized additive modeling (GAM) to Tuscan lexical differences. GAMs allow one to examine explanatory variables in potentially non-linear combinations. Technically, functions representing the interaction of the individual variables are added and optimized, in our case modeling the interaction between longitude and latitude. As one can readily see, the simple distance from

Florence, the center from which many a Tuscan innovation proceeds, is anything but constant in the degree of difference with respect to the standard. Thus, Sienna dialect (in beige) is among the most different, but to the east and west of Sienna, dialects more similar to Florence may be seen (in yellow).

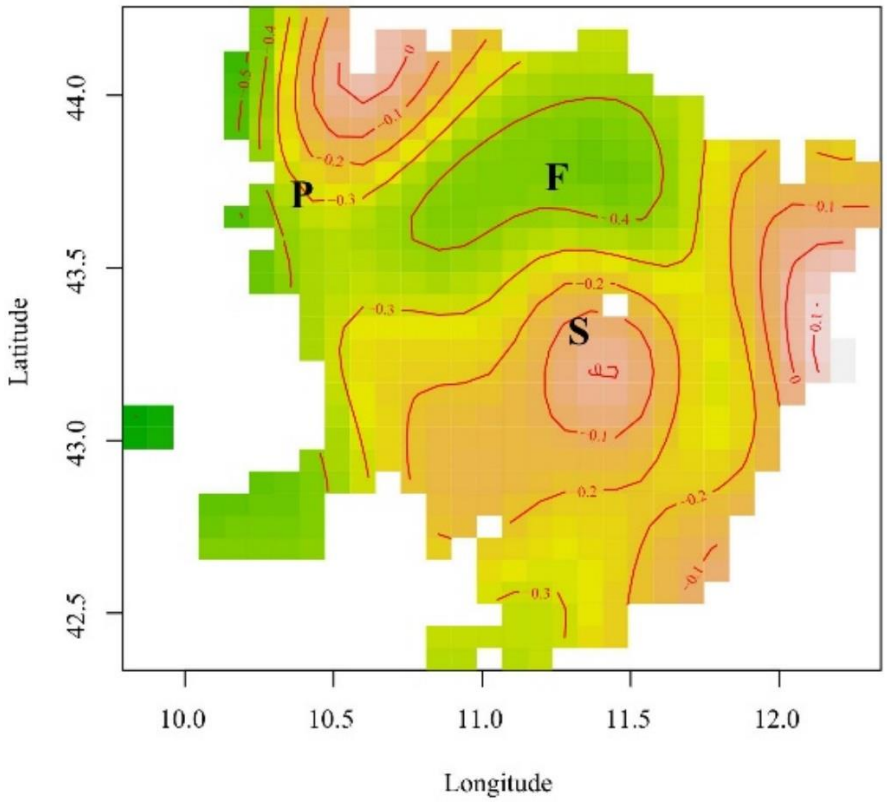

Figure 2 Contour plot for the combination of longitude and latitude predicting differences between standard Italian and local dialects. In the green areas standard forms dominate, in yellow less, and in beige the least (so that, conversely, the beige areas are most likely to use dialectal forms). 'P', 'F' and 'S' mark Pisa, Florence and Sienna, respectively. Data was missing from the white squares. The red "isolines" represent coefficients in the regression analysis marking the likelihood of using standard Italian rather than a dialectal form. Figure taken from Wieling et al. (2014).

3. Novel Research Questions

Dialectologists are typically interested in a range of related subjects and seek ways of linking these interests to dialectology. Dialectometrists are no exception. Dialectometric techniques have been applied to questions of linguistic genealogy and contact influence. Jäger (2015) adapted the PMI-based edit distance to detect historical relations among languages, where it is still being used. Bolognesi & Heeringa (2005) compared 59 Sardinian dialects to Italian, Catalan, Spanish and classical Latin to see which modern Romance language might have influenced Sardinian the most. Using an edit-distance measure combined with feature-based operation costs, they obtained results showing that Latin was closest to Italian, after which Spanish, Sardinian and Catalan followed. All the modern Sardinian varieties differed more from Latin than Italian did, which indeed turned out to be closer to Spanish than to Sardinian. Tamburelli & Brasca

(2018) examine the Italian varieties spoken between the Alps and the La Spezia-Rimini line known generally as “Gallo-Italic” (other than Veneto), but which are classified by some as part of Italo-Romance and by others as Western Romance, which also includes Rate-Romance.

A common fusion of interests involves socially motivated variation, sometimes referred to as “diastratic”. Indeed, Chambers and Trudgill suggest that practitioners in the two fields pursue “variationist” linguistics, a fusion of dialectology and sociolinguistics, and some researchers have applied dialectometric techniques to sociolinguistics questions. Given the sociolinguistic wish to detect (individual) changes in progress, it might seem surprising to look to dialectometry, with its focus on aggregate analyses. But Leinonen (2010) could show that age was an important in predicting leveling processes in Swedish; and Nerbonne et al. (2013) investigated the relation between regiolects and base dialects. Wieling et al. (2018) investigated the introduction of a standard for Catalan in the Catalonian schools and were able to show earlier studies on single variables had confounded effects which their mixed-effects regression could uncover. All these efforts examined effects that one might expect to be quite general, i.e., to affect many aspects of the language being studied, explaining the added value of the aggregate perspective.

A perennial question regarding dialectometry is its relation to linguistic theory, where dialectometrists have often seen their limited theoretical commitments as an advantage, i.e., namely in making the analyses less sensitive to theoretical differences, while many linguists would like to see dialectology illuminate issues in linguistic theory. But nothing stands in the way of, e.g., counting the number of differences in underlying representations for a corresponding samples of dialect material, which is exactly what Valls et al. (2012) do for Catalan, contrasting a phonologically inspired analysis with a simple edit-distance analysis. The phonologically well-informed results correlated highly with the results using the simple edit-distance measure ($r = 0.88$).

4. Social Media

Eisenstein et al. (2014) is a computationally sophisticated study of twitter that examines lexical diffusion in Twitter, and it seems to have spurred a good deal of work among computational linguists, most of which focuses on sociolinguistic rather than dialectological topics. Nguyen et al. (2020) surveys this work.

5. Languages

Several Romance dialect areas have been added to the already long list examined in Seguy and Goebl’s work. Embleton et al. (2007) report on work done beginning in 2003 compiling an online version of an existing Romanian dialect atlas, which one of the authors had been involved in (the late Dorin Uritescu). In addition to the language areas already noted, Ashby et al. (2012) focus on Portuguese in three different countries, and Galician has been studied a good deal (Dubert & Sousa 2016 and references there). French was studied intensively by Goebl (see his

article in the ORE), but it continues to engage new dialectometry (Chagnaud et al. 2021; Brun-Trigaud et al. 2017), including Québecois French (Cichocki & Perreault, to appear).

6. Prospects

Dialectometry is well established with respect to its methods both within Romance scholarship and elsewhere, but many language areas have yet to be studied, and syntax, morphology have garnered to-date too little dialectometric attention, even with the honorable exception of Dunn's (2017) *cri de coeur* for syntactic dialectometry.

Validation likewise remains underdeveloped. Reflecting the view that variation serves as an indicator of provenance, Gooskens & Heeringa (2004) introduced a perception-based validation of edit distance, effectively seeking correlations of edit distances with lay judgments of Norwegian dialect similarity. Wieling et al. (2014) used a similar scheme to demonstrate the superiority of PMI-based edit distance (see above in Section "Edit Distance"). But so far, validation has been applied only to the aggregate levels of human judgments and edit-distance measures. Data sets based on single-word comparisons would enable more sensitive comparison, and the validation of lexical and syntactic measures has yet to begin!

While most dialect atlases contain phonetic transcriptions which can be used as input for the edit distance approach, recent work (Bartelds et al., 2020) has successfully sought to use automatic machine-learning-based techniques to automate determining pronunciation differences of the basis of acoustic recordings. This beneficially allows for dialectometric analysis when acoustic data has been collected but not transcribed yet.

Dialectologists are typically interested in a range of related subjects and seek ways of linking these interests to dialectology. Dialectometrists are no exception, as the section on "Novel Research Questions" expounds on with respect to historical linguistics and sociolinguistics. Nerbonne (2021) speculates on further application of methods that have been useful in dialectometry.

7. Scholarship and Further Reading

Wieling & Nerbonne (2015) is a good overview of developments in dialectometry until 2014.

8. Links to Digital Material

Atlante Lessicale Toscano: ALT-Web, http://serverdbt.ilc.cnr.it/ALTWEB/.

Gabmap, a web application for conducting dialectometrical analyses: Gabmap

References

Ashby, S., Viaro, M. E., Barbosa, S. & Campaniço, N. (2012) Modeling phonetic variation in pluricentric languages: an integrative approach. *Dialectologia: revista electrònica*, 1-26.

Bartelds, M., Richter, C., Liberman, M., & Wieling, M. (2020) A new acoustic-based pronunciation distance measure. *Frontiers in Artificial Intelligence*, *3*, 39.

Bolognesi, R. & Heeringa, W. (2002) De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten. *Gramma/TTT*, *1*.

Bolognesi, R., & Heeringa, W. (2005) *Sardegna fra tante lingue. Il contatto linguistico in Sardegna dal Medioevo a oggi*. Cagliari: Condaghes.

Brun-Trigaud, G., Malfatto, A. & Sauzet, M. (2017) Essaie des aires lexicales occitanes: regards dialectometriques. In: J.-F. Courouau (ed.) *Actes du XIIe congrès de l'Association Internationale d'Études Occitanes*, 169-179.

Chagnaud, C., Brun-Trigaud, G., & Garat, P. (2021). Identification of clusters of lexical areas using geographical factors. In: Van der Velde, H. et al. *Language Variation–European Perspectives VIII*, 209-226.

Cichocki, W. & Y. Perreault (2020) Vers une analyse dialectométique du français parlé au Nouveau-Brunswick: l'apport de la variation phonétique. In D.Bigot, D.Liakin, R.Papen, A.Jebali, & M.Tremblay (eds) *« Les français d'ici en perspective »*. Québec: Presses de l'Université Laval, 7-34.

Dubert, F. & Sousa, X. (2016) On quantitative geolinguistics: An illustration from Galician dialectology. *Dialectologia: revista electronica*, 191-221.

Dunn, J. (2019) Global syntactic variation in seven languages: Toward a computational dialectology. *Frontiers in artificial intelligence*, *2*, 15.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PloS one*, *9*(11), e113114.

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

Gooskens, C. (2005) Travel time as a predictor of linguistic distance. *Dialectologia et Geolinguistica*, 13: 38-68.

Gooskens, C., & Heeringa, W. (2004) Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language variation and change*, *16*(3), 189-207.

Heeringa, W. J. (2004) *Measuring dialect pronunciation differences using Levenshtein distance* (Doctoral dissertation, University of Groningen. Avail. http://www.wjheeringa.nl/thesis/

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

Embleton, S. (1993) Multidimensional scaling as a dialectometrical technique: Outline of a research project. In: Köhler, R. & Rieger, B.B. (eds.) *Contributions to quantitative linguistics*. 267-276. Springer, Dordrecht. https://doi.org/10.1007/978-94-011-1769-2_19

Embleton, S., Uritescu, D., & Wheeler, E. (2007) Data capture and presentation in the Romanian Online Dialect Atlas. *Linguistica Atlantica*, 37-39.

Embleton, S., Uritescu, D., & Wheeler, E. S. (2018) An Expanded Quantitative Study of Linguistic vs. Geographic Distance Using Romanian Dialect Data. In: Wang, L. et al. *Structure, Function and Process in Texts*.

Johnson, D. E. (2009) Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and linguistics compass 3*(1): 359-383.

Loporcaro, M. (2009) *Profilo linguistico dei dialetti italiani.* Roma/Bari: Laterza.

Montemagni, S., Wieling, M., De Jonge, B., & Nerbonne, J. (2012) *Patterns of language variation and underlying linguistic features: A new dialectometric approach*. In: P.Bianchi et al. (eds.) *La variazione nell'italiano e nella sua storia. Varietà e varianti linguistiche e testuali*. Proc. *XI Congresso SILFI*, Naples, 2010. 879-889. Florence: Franco Cesati Editore

Mucha, H. J., & Haimerl, E. (2005) Automatic validation of hierarchical cluster analysis with application in dialectometry. In *Classification—the Ubiquitous Challenge* (pp. 513-520). Springer, Berlin, Heidelberg.

Nerbonne, J. (2021) Studying language differences - an intellectual ecology In: A. Arejita (ed.) *Variability: Words and voices of linguistic varieties. 2nd Int. Conf. Dialectology, Royal Academy of the Basque Language* Euskaltzainda: Bilbao. 243-261.

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., & Leinonen, T. (2011) Gabmap-a web application for dialectology. *Dialectologia: revista electrònica*, 65-89.

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*. Stanford: CSLI Press. v-xv.

Nerbonne, J., Kleiweg, P., Heeringa, W., & Manni, F. (2008) Projecting dialect distances to geography: Bootstrap clustering vs. noisy clustering. In *Data analysis, machine learning and applications* (pp. 647-654). Springer, Berlin, Heidelberg.

Nerbonne, J., van Ommen, S., Gooskens, C., & Wieling, M. (2013) Measuring socially motivated pronunciation differences. Lars Borin and Anju Saxena (eds.) *Approaches to Measuring Linguistic Differences*. Mouton De Gruyter: Boston & Berlin. 107-140.

Nguyen, D., Liakata, M., DeDeo, S., Eisenstein, J., Mimno, D., Tromble, R., & Winters, J. (2020) How we do things with words: Analyzing text as social and cultural data. *Frontiers in Artificial Intelligence*, *3*, 62.

Séguy, Jean (1971) La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35, 335-357.

Tamburelli, M., & Brasca, L. (2018) "Revisiting the classification of Gallo-Italic: a dialectometric approach." *Digital Scholarship in the Humanities* 33.2: 442-455.

Valls, E., Nerbonne, J., Prokic, J., Wieling, M., Clue, E., & Lloret, M. R. (2012) Applying the Levenshtein Distance to Catalan dialects: A brief comparison of two dialectometric approaches. *Verba: Anuario galego de filoloxía*, *39*.

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., & Nerbonne, J. (2015) Advances in dialectometry. *Ann. Review Linguistics*, 1(1): 243-264.

Wieling, M. & Nerbonne, J. (2011) Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech & Language*, *25*(3), 700-715.

Wieling, M., Nerbonne, J., Montemagni, S., & Baayen, R. H. (2014) Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language*, 669-692.

Wieling, M., Prokić, J., & Nerbonne, J. (2009) Evaluating the pairwise string alignment of pronunciations. *Proc. EACL Workshop (LaTeCH–SHELT&R 2009)* Athens: ACL. 26-34. https://aclanthology.org/W09-0304

Wieling, M., Valls, E., Baayen, R. H., & Nerbonne, J. (2018) Border effects among Catalan dialects. In *Mixed-effects regression models in linguistics*. 71-97. Springer, Cham.

Woolhiser, C. (2005) Political borders and dialect divergence/convergence in Europe. In Auer, P. et al. (eds.) *Dialect Change. Convergence and divergence in European languages*, CUP: New York, 236-262.