Chapter 25

Exploring the role of extra-linguistic factors in defining dialectal variation patterns through cluster comparison

Simonetta Montemagni
Institute for Computational Linguistics "Antonio Zampolli"

Martijn Wieling
University of Groningen

This paper contributes to two open issues in the dialectometric literature, i.e. i) whether and how patterns of linguistic variation are influenced by extra-linguistic features such as the geomorphology of the area, or cultural, administrative and political boundaries, and ii) whether and how the influence of extra-linguistic factors remains stable across linguistically-grounded partitions of data. To investigate these issues, a case study focusing on lexical variation has been carried out on a regional lexical atlas of Tuscan dialects. A variety of extra-inguistic features was taken into account, whose impact and role has been evaluated with respect to both the whole dialectal dataset and across different semantic fields.

1 Introduction

In the variationist literature, it is a widely acknowledged fact that (bundles of) isoglosses correlate fairly closely with non-linguistic boundaries: in other words, language is reported to correlate with other aspects of culture (Chambers & Trudgill 1998). What underlies observed correlations remains however an important issue, which is still worth being explored. According to the notion by Bloomfield (1933) of "density of communication", linguistic variation depends primarily on the frequency of communication within the network of speakers of a given language, which is in turn influenced by a variety of extra-linguistic factors, ranging from physical features and population density of the investigated area, to cultural and demographic patterns as well as administrative and/or political boundaries. Due to their role in

limiting or promoting communication across space, extra-linguistic factors such as these can be seen as influencing linguistic variation diatopically and diachronically.

In traditional dialectology, there is no obvious way to explore this relationship beyond superficial and impressionistic observations. In dialectometric studies, the issue of the influence of extra-linguistic features on linguistic variation has been tackled for what concerns geography and population genetics.

Whether and to what extent geographic distance influences aggregate linguistic differences among language varieties has been investigated since early dialectometry (Heeringa & Nerbonne 2001; Gooskens 2005). In these studies, geographical distance, including derivatives of the notion, such as travel time, is regarded as an operationalization of the chance of social contact. Geography has been shown to correlate strongly with linguistic variation. This type of analysis, carried out for typologically distant languages (Bantu, Bulgarian, Dutch, English, German, Norwegian) with respect to phonetic variation, showed that geography accounts for 16% to about 37% of the linguistic variation (Nerbonne 2013).

On the population genetics front, Manni et al. (2008) conclude that the social contacts reflected by dialect varieties do not seem to be related to the demographic history of the populations speaking those dialects. Linguistic and genetic patterns of variation (the latter reconstructed from people's surnames) turned out to be different, even if both are strongly conditioned by geography.

Despite the contrasting conclusions, these studies share the methodology of analysis, i.e. the comparison is carried out with respect to computed distances, be they linguistic, genetic or geographic. The variety of extra-linguistic factors potentially influencing linguistic variation, however, cannot be always modeled in terms of distances. Consider, for instance, the case of physical features (e.g., mountain ranges or river basins) of the investigated area, or its administrative or political organization (e.g., borders of the state or smaller administrative subdivisions). In these cases, the comparison cannot be carried out with respect to distances, but it is rather concerned with clusterings of the investigated locations, based e.g., on their belonging to a given state or other administrative unit, or to their being located in the valley of a given river. To our knowledge, dialectometric studies so far did not tackle this kind of analysis.

In this paper, we will investigate the relationship between linguistic and non-linguistic boundaries through a clustering comparison method, with the final aim of trying to reconstruct role and impact of extra-linguistic features in shaping patterns of linguistic variation. To this specific end, we will use an information theoretic criterion for comparing partitions of the same data set, proposed by Meilă (2003). The criterion is called VARIATION OF INFORMATION (VI), and quantifies the "distance" between the two clusters. In contrast to the modified Rand Index (Hubert & Arabie 1985) used by Prokić, Wieling & Nerbonne (2009) to compare partitions, the VI measure is a true metric (i.e. satisfying the triangle inequality).

Let us focus now on linguistic variation patterns observed with respect to a given area. The question which naturally arises at this point is whether they remain stable between and within levels of linguistic description. In dialectometry, correlation

studies focusing on dialectal variation recorded with respect to distinct linguistic description levels (e.g., phonetic, morphological, lexical) have been carried out with different methodologies and with respect to different dialects. See, for example, Goebl (2005), Spruit, Heeringa & Nerbonne (2009) and Montemagni (2008). Among them, Montemagni (2008) reports that phonetic and morpho-lexical variation in Tuscany does not appear to conform to the same pattern. This result, which is in contrast with the outcome of the other studies, is complemented by a significantly different degree of correlation between geography on the one hand and observed patterns of phonetic vs. morpho-lexical variation on the other hand (r = 0.1358 vs. r = 0.6441).

But what happens if within the same level of description different linguistically-grounded partitions of the data are considered? For instance, if the focus is on lexical variation, will the resulting partitions be the same across different semantic fields? Within a dialectometric study of Tuscan dialectal variation, Montemagni (2010) reports that patterns of variation identified with respect to different semantic domains (e.g., agriculture, weather, house, stockbreeding) differ significantly from the overall picture, suggesting the influence of extra-linguistic factors other than geography (e.g., climate, geomorphology, history). However, this is an impressionistic analysis which needs to be investigated further. Stronger evidence in this direction emerges from Franco, Geeraerts & Speelman (2015) who, building on previous studies (Speelman & Geeraerts 2008; Geeraerts & Speelman 2010), demonstrated that semantic field influences so-called 'onomasiological heterogeneity', i.e. the fact of being significantly more prone to variation in concept naming.

Within the wider context of this study, in this paper we also intend to contribute to the issue of whether and how the influence of extra-linguistic factors varies across linguistically-grounded partitions of data. In particular, whether there are extra-linguistic factors playing a stronger role in the definition of specific clusters. The aim of this paper can thus be summarized by the following two research questions:

- are patterns of linguistic variation influenced by extra-linguistic features such as geomorphology of the area, or cultural, administrative or political boundaries?
- To what extent does the role and impact of these features remain stable across linguistically-grounded partitions of data?

To answer these questions a case study focusing on lexical variation has been carried out on a regional lexical atlas of Tuscan dialects. A variety of extra-linguistic features was taken into account, whose impact and role has been evaluated with respect to both the whole dialectal dataset and across semantic fields.

2 Data

2.1 Dialectal data

The dialectal corpus of the *Atlante Lessicale Toscano* ('Lexical Atlas of Tuscany', ALT; Giacomelli et al. 2000) was used.[1] ALT is an Italian regional lexical atlas focusing on dialectal variation throughout Tuscany, where both Tuscan and non-Tuscan dialects are spoken. In this paper, we focused on Tuscan dialects only, recorded in 213 localities by a total of 2060 informants, socio-demographically selected with respect to parameters such as age, education and gender.

ALT interviews were carried out on the basis of a questionnaire including onomasiological questions, i.e. looking for concept lexicalizations, and organized into semantic domains (e.g., agriculture, food, wild animals, weather, house etc.). Out of the 460 onomasiological questions, we selected only those focusing on nominal concepts and characterized by lower 'onomasiological heterogeneity' (in the case at hand, showing 50 or fewer answer types). The resulting subset consists of 170 questionnaire items for which a total of 5,174 normalized answers types were given, corresponding to 61,496 geo-localized responses and 384,454 individual ones. Based on the results by Wieling & Montemagni (2016) who demonstrated that cluster quality improves when the analysis is based on all data, we used unfiltered data, with no pruning of infrequent variants.

To abstract away from productive phonetic variation, we used the normalized representation of ALT dialectal items (Cucurullo et al. 2006). The representativeness of the selected sample with respect to the whole set of ALT onomasiological questions was assayed using the correlation between overall lexical distances and lexical distances obtained from the selected sample, which turned out to be high ($r = 0.94$). Note that the same set of questions was used in different studies, by Montemagni & Wieling (2016) on Tuscan lexical variation, and by Wieling et al. (2014) on the relationship between Tuscan dialects and standard Italian.

2.2 Extra-linguistic data

For this study, we focused on the following typology of extra-linguistic features:

- geomorphology of the area, described in terms of hydrographic basins;
- religious subdivisions, corresponding to dioceses in turn aggregated into archdioceses: these represent territorial units of administration of the Catholic Church whose origin dates back to centuries ago, when a formal church hierarchy was set up, parallel to the civil administration (whose areas of responsibility often coincided);
- political and administrative subdivisions, such as state or province.

[1] ALT is available as an online resource at the following address: `http://serverdbt.ilc.cnr.it/ALTWEB`.

Figure 1 shows the clusterings of Tuscan ALT locations according to the selected extra-linguistic features. Some of these were obtained from the online version of the *Geographical, physical and historical dictionary of Tuscany* by Emanuele Repetti (1833-1843), which is an encyclopedic collection of information about Tuscany published in the 19th century concerning notable places, from large towns to small villages, providing historical, archaeological and artistic information as well as physical land attributes (e.g., mountains, rivers, lakes, etc.).[2] Information about state, archdiocese and basin authority (the administrative counterpart of hydrographic basin) was extracted from Repetti's Dictionary. State refers to the political organization of Tuscany in 1833; archdiocese and basin authority refer to current geographical–administrative subdivisions, which were reconstructed from the diocese and valley information respectively, reported in the dictionary. The province subdivision refers to the current administrative organization of Tuscany.

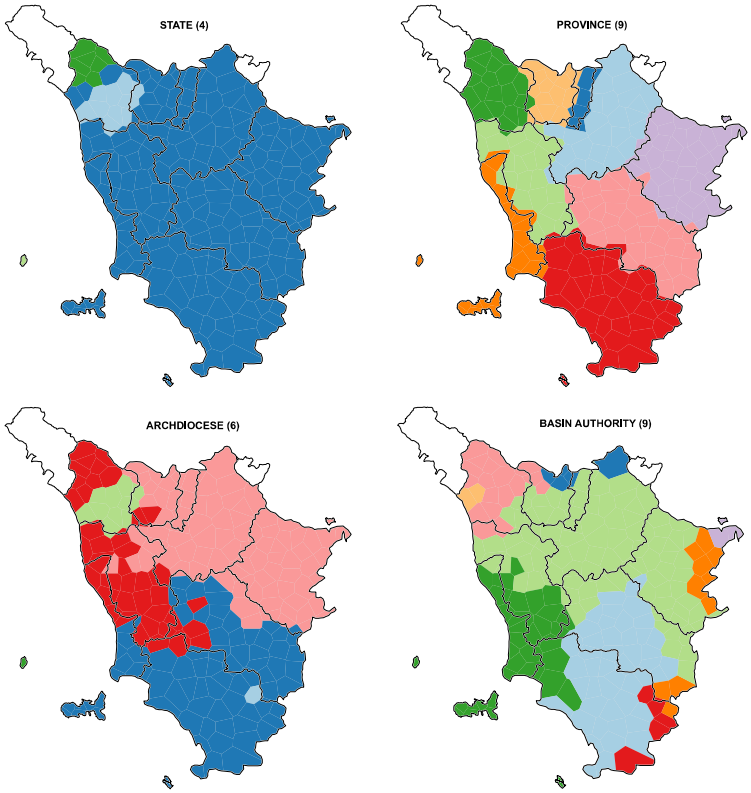

Figure 1: Geographical clustering of Tuscan ALT locations on the basis of extra-linguistic criteria (state, province, archdiocese, basin authority).

[2] The dictionary, published online in 2005, can be accessed at the following address: `http://www.archeogr.unisi.it/repetti`.

3 Methods

3.1 Clustering of dialectal data

For the clustering of the dialect data, we use bipartite spectral graph partitioning (Dhillon 2001). This algorithm has been used for the clustering of Tuscan dialect data before (Montemagni & Wieling 2016) and was introduced to dialectometry by Wieling & Nerbonne (2011). Bipartite spectral graph partitioning simultaneously clusters geographic locations together with their associated (characteristic) linguistic features. In short, the method functions by computing the singular value decomposition of the input matrix (of geographical locations and linguistic features), and subsequently uses k-means to (recursively) obtain a partitioning in two groups. Consequently, the clustering consists of locations together with their associated linguistic features. While Montemagni & Wieling (2016) focused on investigating and discussing the underlying features of the different clusters, here we only use the resulting geographical clustering. As we use non-linguistic data organized into either 4, 6 or 9 groups, we use linguistic clusterings having a similar number of clusters.[3] Besides including a clustering in 6 and 8 groups on the basis of all data, we include clusterings generated on the basis of including separate semantic fields. These consist of agriculture (6 and 9 clusters), animals (3 and 4 clusters), food (6 and 8 clusters), and house (6 and 8 clusters).

3.2 Clustering comparison

As indicated above, we use the VARIATION OF INFORMATION criterion created by Meilă (2003). This measure is an information theoretic criterion related to mutual information to compare two clusterings of the same data set. The advantage of the approach is that it "makes no assumptions about how the clusterings were generated" (Meilă 2003: 173) and that the measure is a "true metric on the space of clusterings" (Meilă 2003: 173). It is defined as follows:

$$V(X;Y) = -\Sigma_{i,j} r_{ij}(\log(r_{ij}/p_i) + \log(r_{ij}/q_j))$$

with $p_i = |X_i|/n$ and $q_i = |Y_i|/n$ (n equals the complete number of data points in the clustering). In the following, we will use this measure as implemented in the R package `mcclust` (Fraley & Raftery 2002).

Since there are several clusterings we need to compare, we will summarize the results visually by using multidimensional scaling (MDS) in two dimensions. In this way we are able to visualize the 10 linguistic clusters (2 on the basis of all data, and 8 on the basis of separate semantic fields) together with 4 extra-linguistic clusterings.

[3] Importantly, the VI measure we use to obtain a quantification of the difference of two clusters, does not require the clusterings to have the same number of clusters. This is fortunate as the bipartite spectral graph partitioning method does not result in a pre-specified number of clusters, as not always a division in two groups may be possible.

4 Results

Figure 2 visualizes the linguistic clustering for the complete data set, as well as separated by semantic domain. By visually comparing the clusterings, it is clear there are similarities. For example, the south area corresponding to the Grosseto and Siena provinces is always clustered together, but there are also clear differences at the level of the north area, going from Lucca to Arezzo (the names of the aforementioned provinces are marked in the top-left graph in Figure 2). The question which is being investigated is whether and how these clusters correlate with those reported in Figure 1, based on extra-linguistic criteria.

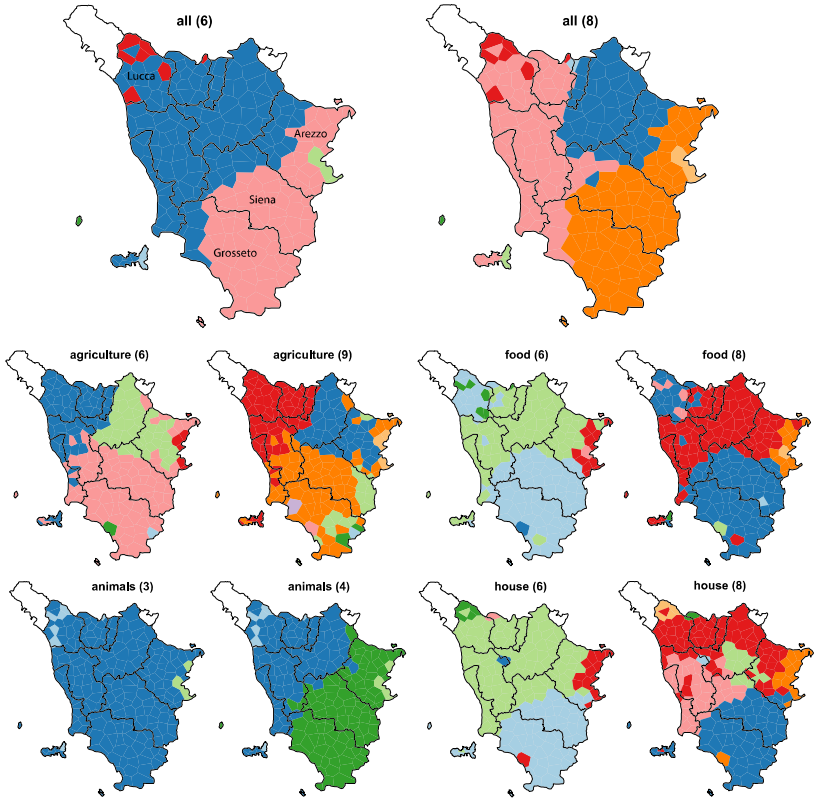

Figure 2: Geographical clustering of linguistic data.

Figure 3 represents the MDS visualization of the similarity of linguistic and extra-linguistic clusters on the basis of the VARIATION OF INFORMATION criterion: note that capitalized cluster names refer to extra-linguistic partitions of locations and that the suffixed number indicates the number of clusters. The visualization in two dimensions was adequate, as stress was only 0.16.

Consider first the relationship extra-linguistic vs. linguistic clusterings. With re-

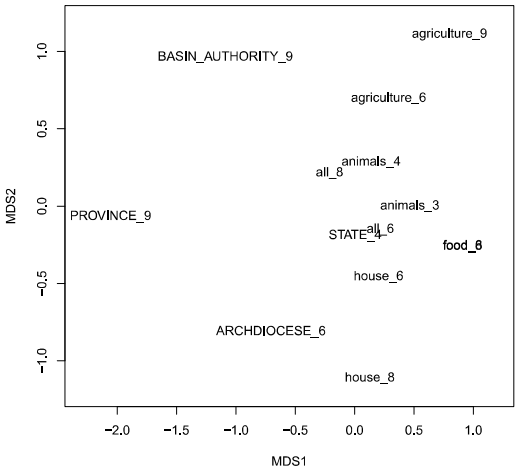

Figure 3: MDS visualization of similarity of clusterings.

spect to political and administrative subdivisions, it can be noticed that the STATE clustering seems to be most similar to the dialect results, in particular to those obtained with respect to the *animals*, *house* and *food* domains as well as to the complete dataset (named *all*). Interestingly, the current administrative subdivision, i.e. PROVINCE, seems to be far away from all linguistic subdivisions. For what concerns ARCHDIOCESE, the stronger similarity concerns the *house* and *food* domains. Last but not least, the clustering based on geomorphology, i.e. BASIN_AUTHORITY, shows a stronger similarity with respect to the *agriculture* and *animals* domains. Among all, the closest distances are observed with respect to STATE, followed by ARCHDIOCESE, BASIN_AUTHORITY and lastly by PROVINCE (average VI scores with respect to each of them are 1.76, 2.42, 2.59 and 2.85, respectively). For what concerns the linguistically-based partitions, it is interesting to note the stronger similarity between the clusterings in the *agriculture* and *animals* domains on the one hand, and between *food* and *house* on the other hand, with *food* functioning as a transition between nature-based vs. culture-based partitions.

5 Conclusion

Going back to the research questions we started with, the results of this study clearly show that the patterns of lexical variation identified through bipartite spectral graph partitioning are influenced by external factors limiting the communication across the region. These factors range from political (states and provinces) and cultural (archdioceses) subdivisions to physical ones (river basins). This provides strong support

to Bloomfield's theory of linguistic variation which is seen to be driven by so-called "density of communication", predictable from a variety of external factors.

We have seen that STATE and ARCHDIOCESE, reflecting pre-unitarian[4] subdivisions dating back to centuries ago, appear to play a central role in defining patterns of dialectal variation. In this respect, it is interesting to go back to Bloomfield (1933: 343), who claims that the primary correlation is with political subdivisions: "The important lines of dialectal division run close to political lines'.' Similarly, Kurath (1954) reports the coalescence of Middle English dialect boundaries with county lines. Interestingly, more recent administrative subdivisions (provinces) turned out to play a very small role in defining the Tuscan dialectal landscape. Again, these results are in line with the claim by Bloomfield (1933: 343) that "isoglosses along a political boundary of long standing [...] would persist, with little shifting, for some two-hundred years after the boundary had been abolished".

BASIN_AUTHORITY, in spite of representing a sort of artificial subdivision collapsing adjacent river basins for administrative reasons, seems to play an important role in shaping patterns of lexical variation, especially for what concerns the semantic fields of *agriculture* and *animals*. This leads to the second research question, investigating whether and to what extent the impact of external features remains stable across linguistically-grounded partitions of data, semantic fields in the case at hand. On the basis of these results, the answer is positive. The influence of the different extra-linguistic factors taken into account in this study turned out to differ across semantic fields. Physical subdivisions are more relevant for what concerns the *agriculture* and *animals* domains, while political and cultural borders appear to play a stronger role with more culturally-oriented semantic domains (i.e. *house* and *food*).

References

Bloomfield, Leonard. 1933. *Language*. New York: Holt.

Chambers, J. K. & Peter Trudgill. 1998. *Dialectology*. 2nd edn. Cambridge, New York: Cambridge University Press.

Cucurullo, Nella, Simonetta Montemagni, Matilde Paoli, Eugenio Picchi & Eva Sassolini. 2006. Dialectal resources on-line: the ALT-Web experience. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, 1846–1851. European Language Resources Association (ELRA).

Dhillon, Inderjit. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 269–274. ACM New York, NY, USA.

Fraley, Chris & Adrian E. Raftery. 2002. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association* 97. 611–631.

[4] Italian political unification was in 1861.

Franco, Karlien, Dirk Geeraerts & Dirk Speelman. 2015. Why dialects differ: the influence of concept features on lexical geographical variation. In *Proceedings of the International Conference on Language Variation in Europe (ICLaVE 8), Universität Leipzig, Leipzig, Germany, May 27-29, 2015.*

Geeraerts, Dirk & Dirk Speelman. 2010. Heterodox concept features and onomasiological heterogeneity in dialects. In D. Geeraerts, G. Kristiansen & Y. Peirsman (eds.), *Advances in Cognitive Sociolinguistics*, 23–40. Berlin, New York: De Gruyter Mouton.

Giacomelli, Gabriella, Luciano Agostiniani, Patrizia Bellucci, Luciano Giannelli, Simonetta Montemagni, Annalisa Nesi, Matilde Paoli, Eugenio Picchi & Teresa Poggi Salani. 2000. *Atlante Lessicale Toscano.* Roma: Lexis Progetti Editoriali.

Goebl, Hans. 2005. La dialectométrie corrélative: un nouvel outil pour l'étude de l'aménagement dialectal de l'espace par l'homme. *Revue de Linguistique Romane* 69. 321–367.

Gooskens, Charlotte. 2005. Traveling time as a predictor of linguistic distance. *Dialectologia et Geolinguistica* 13. 38–62.

Heeringa, Wilbert & John Nerbonne. 2001. Dialect areas and dialect continua. *Language Variation and Change* 13(3). 375–400.

Hubert, Lawrence & Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2(1). 193–218.

Kurath, Hans. 1954. *Middle English Dictionary, Plan and Bibliography.* Ann Arbor: University of Michigan Press.

Manni, Franz, Wilbert Heeringa, Bruno Toupance & John Nerbonne. 2008. Do surname differences mirror dialect variation? *Human Biology* 80(1). 41–64.

Meilă, Marina. 2003. Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines*, 173–187. Springer.

Montemagni, Simonetta. 2008. The space of tuscan dialectal variation: a correlation study. *International Journal of Humanities and Arts Computing* 2(1–2). 135–152.

Montemagni, Simonetta. 2010. Esplorazioni computazionali nello spazio della variazione lessicale in Toscana. In *Atti del convegno 'parole. Il lessico come strumento per organizzare e trasmettere gli etnosaperi', 2–4 luglio 2009, Rende*, 619–644. Centro Editoriale e Librario dell'Università della Calabria.

Montemagni, Simonetta & Martijn Wieling. 2016. Tracking linguistic features underlying lexical variation patterns: a case study on tuscan dialects. In *The Future of Dialects: Selected Papers from Methods in Dialectology XV*, 117–134. Language Science Press.

Nerbonne, John. 2013. How much does geography influence language variation? In Peter Auer, Martin Hilpert, Anja Stukenbrock & Benedikt Szmrecsanyi (eds.), *Space in language and linguistics: Geographical, Interactional, and Cognitive Perspectives*, 220–236. Berlin, New York: Walter de Gruyter.

Prokić, Jelena, Martijn Wieling & John Nerbonne. 2009. Multiple sequence alignments in linguistics. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, 18–25. Association for Computational Linguistics.

Speelman, Dirk & Dirk Geeraerts. 2008. The role of concept characteristics in lexical dialectometry. *International Journal of Humanities and Arts Computing* 2(1–2). 221–242.

Spruit, Marco René, Wilbert Heeringa & John Nerbonne. 2009. Associations among Linguistic Levels. *Lingua* 119(11). 1624–1642.

Wieling, Martijn & Simonetta Montemagni. 2016. Infrequent forms: noise or not? In *The Future of Dialects: Selected Papers from Methods in Dialectology XV*, 215–224. Language Science Press.

Wieling, Martijn, Simonetta Montemagni, John Nerbonne & R. Harald Baayen. 2014. Lexical differences between Tuscan dialects and standard Italian: accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language* 90(3). 669–692.

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