Chapter 7

Tracking linguistic features underlying lexical variation patterns: A case study on Tuscan dialects

Simonetta Montemagni
Istituto di Linguistica Computazionale "Antonio Zampolli", ILC-CNR

Martijn Wieling
University of Groningen, CLCG

In this paper, we illustrate the application of hierarchical spectral partitioning of bipartite graphs in the study of lexical variation in Tuscany based on the data from a regional linguistic atlas. This method makes it possible not only to identify existing patterns of lexical variation in Tuscany, but also to uncover the underlying lexical features in terms of the most characteristic concept-lexicalization pairs. The results are promising, demonstrating the potential of the method for tracking the linguistic features underlying identified patterns of lexical variation and change across generations.

1 Introduction

In dialectometry (Séguy 1971) the focus lies on the aggregate analysis of dialect variation. In contrast to "cherry-picking" a few linguistic items confirming the analysis one wishes to settle on (Nerbonne 2009), the advantage of the aggregate approach is that it offers a more objective view of dialect variation. Unfortunately, many studies focusing on the aggregate pattern of dialect variation have disregarded the underlying linguistic basis. As a consequence, linguists have remained critical of the dialectometric approach (Schneider 1988; Woolhiser 2005; Loporcaro 2009).

To counter this criticism, various new dialectometric methods have been developed aimed at identifying the linguistic basis of dialectal variation (as reviewed in Wieling & Nerbonne 2015). For example, Nerbonne (2006) and Pröll, Pickl

Simonetta Montemagni & Martijn Wieling. 2016. Tracking linguistic features underlying lexical variation patterns: A case study on Tuscan dialects. In Marie-Hélène Côté, Remco Knooihuizen & John Nerbonne (eds.), *The future of dialects*, 117–135. Berlin: Language Science Press. DOI:10.17169/langsci.b81.146

& Spettl (in press) use an approach based on factor analysis, whereas Shackleton (2005) uses principal component analysis. Grieve, Speelman & Geeraerts (2011) follow the workflow of traditional dialectology (i.e. identifying isoglosses, bundling isoglosses and cluster analysis) by using multivariate spatial analysis.

The method we will apply here, Hierarchical Bipartite Spectral Graph Partitioning (HBSGP), has been developed by Wieling & Nerbonne (2009; 2010; 2011), who adopted it from information retrieval (Dhillon 2001) and applied it to dialectology. HBSGP results in a clustering of geographical varieties while *simultaneously* providing a linguistic basis for each of the identified clusters. The approach of Wieling & Nerbonne (2011) has been successfully applied to study phonetic variation in Dutch dialects (Wieling & Nerbonne 2011), English dialects (Wieling, Shackleton & Nerbonne 2013) and Tuscan dialects (Montemagni et al. 2012; 2013). More recently, the method has also been applied to investigate lexical variation in contemporary English dialects on the basis of the BBC *Voices* data (Wieling et al. 2014).

In this study, we focus on lexical variation. Our dataset, a regional lexical atlas of Tuscan dialects whose data have a diatopic and diachronic characterization, allows us to explore the potential of the HBSGP method in the study of lexical variation. In particular, it enables us to identify lexical features and their relationships on the one hand and to reconstruct the dynamics of lexical change across generations on the other hand. Technically, a new measure is proposed for determining the most important lexical features associated with the identified dialectal areas.

2 Data

We investigate Tuscan lexical variation on the basis of a linguistic atlas of Tuscany, the *Atlante Lessicale Toscano* (ALT, Giacomelli et al. 2000), now available as an online resource (http://serverdbt.ilc.cnr.it/ALTWEB). ALT is a regional Italian lexical atlas focusing on dialectal variation throughout Tuscany, where both Tuscan and non-Tuscan dialects are spoken. In this paper we focus on Tuscan dialects only, recorded in 213 localities by a total of 2060 informants who were selected with respect to various socio-demographic parameters (such as age, education and gender).

ALT interviews were carried out on the basis of a questionnaire of 745 target items, designed to elicit mainly lexical, but also semantic and phonetic variation. This study is based on the results of onomasiological questions, i.e. starting from concepts and looking for their lexicalizations. A typical onomasiological ques-

tion asks how a given concept is designated or named, e.g. "what is the name for flat and crispy bread, seasoned with salt and oil?". To avoid interference with non-lexicalized answers, we excluded questions prompting 50 or more distinct lexical items. Furthermore, we only considered nouns (the large majority of items of ALT questionnaire) in this study. The resulting subset consists of 170 questionnaire items for which a total of 5,174 distinct normalized answers were given (on average 30 lexical variants per concept) distributed into 61,496 geo-referenced responses (i.e. associated with locations). The total number of speaker-responses was 384,454.

To abstract away from phonetic variation, we used the most abstract representation level present in ALT (Cucurullo et al. 2006). This normalized representation was meant to abstract from phonetic variation (caused by productive phonetic processes), but did not remove morphological variation or variation caused by unproductive phonetic processes. In this study we used the normalized lexical answers to the selected subset of 170 onomasiological questions. The same set of questions has also been used by Wieling, Upton & Thompson (2014) in a study of lexical differences between Tuscan dialects and standard Italian.

The representativeness of the selected sample with respect to the whole set of ALT onomasiological questions (i.e. a total of 460 questionnaire items) was assayed using the correlation between overall lexical distances and lexical distances obtained from the selected sample (Wieling, Upton & Thompson 2014). The Pearson's correlation coefficient was $r = 0.94$, showing the representativeness of the selected sample with respect to the whole set of onomasiological questions.

3 Methods

In this study, we use hierarchical bipartite spectral graph partitioning as our method of choice (Wieling & Nerbonne (2011)). As mentioned before, this approach simultaneously clusters the geographic locations together with the linguistic features characterizing them. In this case, a cluster of locations is characterized by a linguistic basis expressed in terms of the most salient lexical features. These lexical features can be seen as a proxy of the traditional notion of lexical isoglosses, establishing the boundaries of dialectal areas.

Every variety attested in a given location is described in terms of Concept-Lexicalization (CL) pairs linking each of the 170 selected concepts with its lexicalization(s) (reported in the normalized form) in the specific location. CL frequencies are normalized by dividing the number of recorded answers by the number of informants in a given location, with their value ranging between 0 and 1. Since

there was a socio-demographically differentiated group of informants potentially giving rise to multiple responses to denote the same concept for each location, the sum of normalized frequencies of lexical variants associated with the same concept in a certain location can be greater than 1.

The input for the HBSGP method is a bipartite graph which contains two sets of vertices, locations and CL pairs, connected by lines. There exists a line between a location and a CL pair whenever at least one of the speakers in the location uses the lexical variant. The lines are weighted between 0 and 1. A value of 0 indicates that no speakers in the location use the lexical variant (and thus equals the absence of a line), whereas a value of 1 indicates that all speakers in the location use the lexical variant to denote the concept being investigated. Table 1 gives an example of (a tabular representation of) the bipartite graph, with the rows corresponding to the locations and the columns to the CL pairs. About 80% of the speakers in Caprese Michelangelo use the form *aràncio* to denote an ORANGE (henceforth, concept denominations are represented by small caps). A similar number of speakers also uses *melàngola* to denote the same (speakers frequently provided multiple lexicalizations to denote a certain concept).

The input matrix is then subjected to Singular Value Decomposition (SVD), and the *k*-means clustering algorithm (with *k* equals 2) is applied to the results of the SVD resulting in a two-way clustering. The *k*-means clustering was repeated 1000 times for robustness. As the output of the SVD combines the locations with the CL pairs, the clustering likewise groups locations and CL pairs. Consequently, lexical variants grouped with locations can be seen as characteristic elements of those locations. For more mathematical details, we refer the interested reader to Wieling & Nerbonne (2011).

In order to identify the most characteristic linguistic features for a group of locations, Wieling & Nerbonne (2011) combined two different criteria which were implemented in two different and complementary measures: representativeness and distinctiveness. Representativeness measures the relative frequency of the lexicalization of a given concept in the locations in the cluster. For example, if the cluster contains ten locations and all speakers in seven locations use the lexical variant, the representativeness is 0.7. Distinctiveness measures how frequently the lexical variant occurs within as opposed to outside of the cluster (corrected for the relative size of the cluster, which is calculated by dividing the number of locations in the cluster by the total number of locations in the dataset). A distinctiveness of 1 indicates that the lexical variant is only used inside the cluster. The distinctiveness equals 0 when the relative frequency of the lexical variant in the cluster is equal to the relative size of the cluster (i.e. it is not distinctive). Inter-

Table 1: Tabular representation of a bipartite graph. The numbers represent the normalized frequency (obtained by dividing by the number of speakers) of the lexical variant associated with a given concept in the different locations which ranges between 0 and 1. As the speakers may use multiple variants to denote a concept, the normalized frequencies associated with a concept in a certain location do not have to sum to 1.

Location	ORANGE-*arància*	ORANGE-*aràncio*	ORANGE-*melàngola*
Caprese Michelangelo	0.1379	0.7931	0.7931
Pieve Santo Stefano	0.4000	0.7333	0.2000
Anghiari	0.0000	0.7059	1.0000
Sansepolcro	0.0000	1.0000	1.0000

estingly, the measures of representativeness and distinctiveness are reminiscent of the "consistency" and "homogeneity" measures introduced by Labov and colleagues for the construction of isoglosses in the *Atlas of North American English* (Labov, Ash & Boberg 2006). Homogeneity measures how much variation exists within the region defined by the isogloss (i.e. corresponding to a non-chance corrected variant of distinctiveness) and consistency (i.e. corresponding to representativeness) measures how strongly the variable is concentrated within a given region.

The two measures capture two different equally important desiderata of isoglosses: to put it in the words of Labov, Ash & Boberg (2006), "First, we want the area defined to be as uniform as possible [...]. Second, we want as high a proportion of hits as possible to be located within the isogloss". For this reason they need to be combined. Wieling & Nerbonne (2011) combined representativeness and distinctiveness measures by averaging them, yielding the importance score. Here, we propose that to determine the relevance of CL pairs in the characterization of identified lexical areas it is better to multiply the two values. The advantage of this approach is that it is not possible to assign high importance values to lexical variants which score high on a single measure only. For example, lexical variants occurring in all locations are highly representative, but not distinctive. Similarly, a lexical variant only occurring in a single location is highly distinctive, but not representative (unless the cluster contains a single location). Note that constraints on isogloss construction were also foreseen by Labov, Ash & Boberg (2006) by enforcing frequency thresholds. However, the advantage of the approach proposed by Wieling & Nerbonne (2011) and its evolution presented

here consists in the fact that no *a priori* constraints on the values of individual measures are defined.

4 Results

In this section, we report the results of applying the HBSGP method to the selected ALT dataset. The results obtained are based on 5,174 CL pairs and 213 locations, which correspond to all lexical data gathered through fieldwork (as opposed to a dataset in which infrequent lexical variants are filtered out) for the 170 selected concepts. See Wieling & Montemagni (2015) for a discussion of the advantages connected with this dataset.

The map in Figure 1 shows the geographic visualization of the clustering of Tuscan varieties into seven groups designated as follows: the Florence area (A), the western Tuscan area (C) and the dialects from Arezzo, Siena, Grosseto and

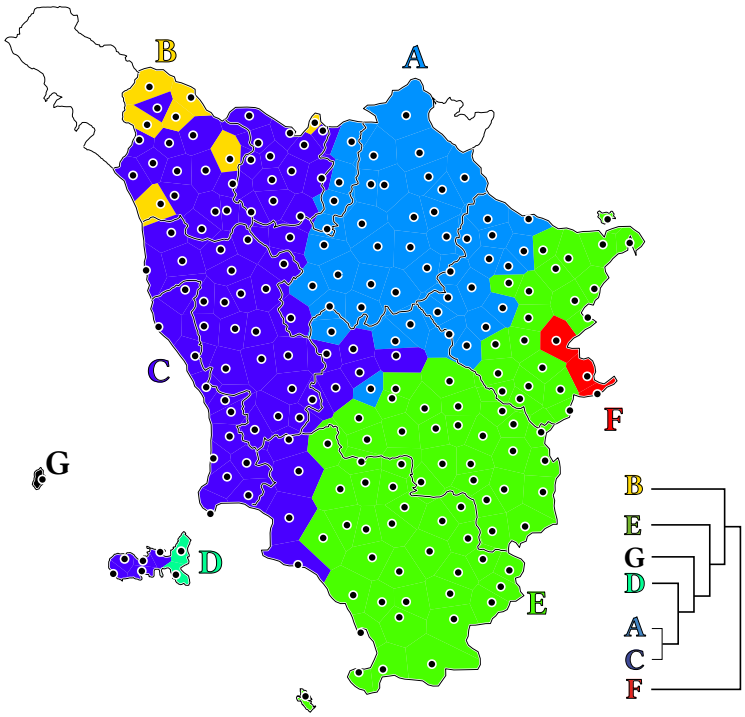

Figure 1: Geographic visualization of the clustering of Tuscan varieties into seven groups.

Mount Amiata (E) which represent the three main groupings, together with the dialects from Elba island (D), Chiana Valley (F), Capraia Island (G) and Apuan Alps (B) which are minor but clearly distinct dialectal areas.

It is interesting to note that this result is in line with the classifications of Tuscan dialects proposed by Giacomelli (1975) for what concerns the lexicon, and by Giannelli (1976 [2010]) which is based instead on phonetic, phonemic, morpho-syntactic and lexical features. It is also in line with the subdivision of Tuscan dialects by Pellegrini (1977), in spite of it being mainly based on the distribution of phonetic phenomena.

4.1 Linguistic features underlying identified lexical areas

For what concerns the underlying lexical features, we first focus on the three main dialectal clusters (A, C and E). Table 2 reports for each cluster the five most important CL pairs with associated values of representativeness, distinctiveness and importance.

The relevance of the lexical features with respect to the dialectal subdivision emerges clearly from the value maps in Figure 2, which show the geographic distribution of the first and second topmost lexical features of each of the three

Table 2: The five topmost lexical variants for the three main clusters of Tuscan dialects.

Cluster	Concept-Lexicalization pair	Representativeness	Distinctiveness	Importance
E	TURKEY-*billo*	0.863	0.700	0.604
	CORNER OF TISSUE-*pìnzo*	0.724	0.795	0.576
	EYE GUM-*cipìcchia*	0.624	0.920	0.574
	OIL JAR-*zìro*	0.879	0.609	0.535
	VAT-*bigónzo*	0.649	0.821	0.533
A	ORANGE-*arància*	0.779	0.675	0.526
	LADLE-*romaiòlo*	0.788	0.536	0.423
	OIL JAR-*órcio*	0.671	0.590	0.396
	TURKEY-*tàcco*	0.390	1.000	0.390
	BRAWN-*capofréddo*	0.432	0.900	0.389
C	OIL JAR-*cóppo*	0.749	0.696	0.522
	EYE GUM-*cìspia*	0.702	0.676	0.474
	BREAST-*pùppa*	0.649	0.717	0.466
	FLEA-*pùce*	0.602	0.686	0.413
	CLUSTER OF GRAPES-*pìgna*	0.570	0.701	0.400

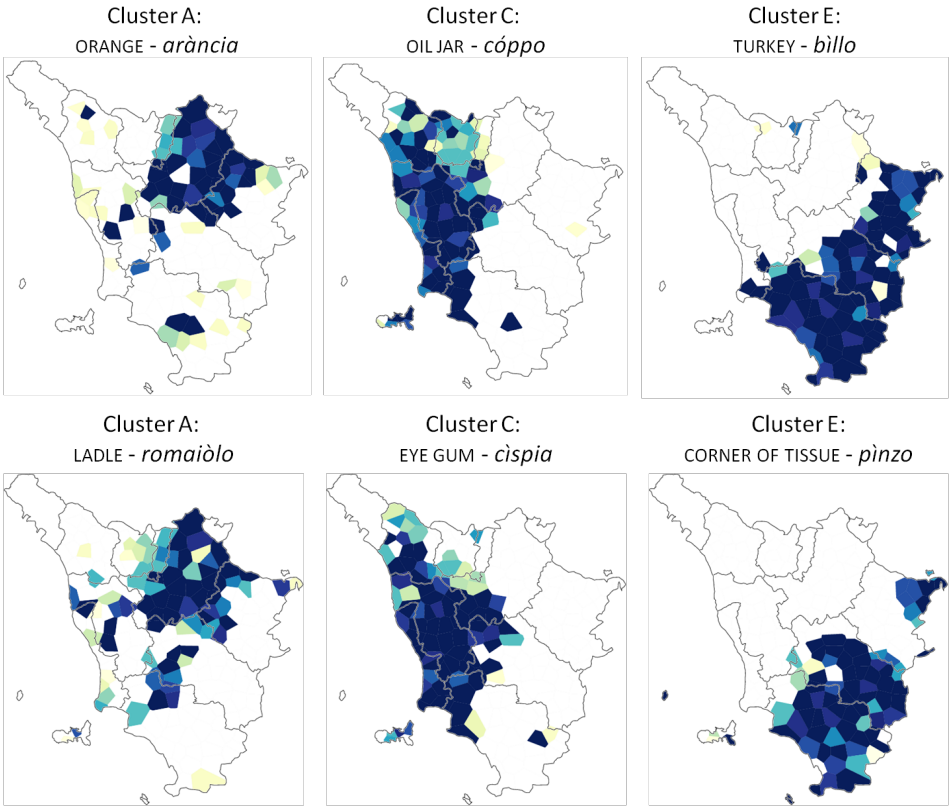

Figure 2: Value maps of the first (row 1) and second (row 2) topmost CL pairs for the A, C and E dialectal clusters. Areas with darker (blue) color denote a greater frequency of occurrence of the selected lexical variant; lighter colors denote a lower frequency, while no coloring (white) denotes the absence of the variant.

main identified clusters (A, C and E). The topmost lexical features associated with each identified cluster can be assimilated with the traditional notion of bundle of isoglosses, which have long been considered a major criterion for the definition of dialect areas: as Chambers & Trudgill (1998) put it, "the significance of a dialect area increases as more and more isoglosses are found which separate it from adjoining areas".

By comparing the maps of Figure 2, we can observe that the geographic distribution of the topmost CL pairs of the E, A and C clusters does not cover all and

only the locations in the cluster. Each of them can be seen as a quantitative visualization of individual isoglosses, where darkness of color denotes the frequency of occurrence of the represented lexical variant (dark colors denote a greater frequency, lighter colors lower frequency, and no coloring indicates the absence of the variant). As can be observed, lexical variants shown in Table 2 may occur beyond the border of the cluster area, thus lowering the distinctiveness value of the CL pair, or they may not occur in the whole cluster area resulting in a lower representativeness. For instance, in cluster A comparable representativeness values are observed for the two topmost CL pairs (0.77-0.78), whereas the CL ranked in second place, i.e. LADLE-*romaiòlo*, has a lower distinctiveness value (0.53) than the topmost CL (i.e. whose distinctiveness value is 0.67). Different patterns can be observed in clusters E and C, with decreasing representativeness and increasing distinctiveness in the former case, and with both of them decreasing in the latter case. Despite these slight differences, in all cases representativeness and distinctiveness show relatively high values which never reach the value of 1 (with the only exception of the CL pair TURKEY-*tàcco* in cluster A whose distinctiveness is equal to 1). The average values of the five topmost lexical features for representativeness and distinctiveness range between 0.61 and 0.74, and 0.69 and 0.77 respectively, demonstrating that the corresponding dialect areas are not marked by very clear and strong dialect borders.

Different distinctiveness-representativeness patterns are observed in the case of the smaller peripheral areas B, D, F and G (see Table 3). Here, the most salient CL pairs are highly distinctive (their average values ranges from 0.84 to 1), with the average representativeness ranging from 0.49 to 1. Thus smaller dialect areas are characterized by much more distinctive features than the larger areas.

Besides the strength of dialectal borders, granularity of the identified dialectal areas is another open issue in the study of dialectal variation. Consider, for instance, the traditional dialectal subdivision of Tuscan dialects by Pellegrini (1977) and Giannelli (1976 [2010]). In his *Carta dei Dialetti d'Italia,* Pellegrini (1977) identifies a western variety of Tuscan which is further subdivided into Pisano-Livornese-Elbano, and Pistoiese and Lucchese. On the other hand, Giannelli (1976 [2010]) identifies Pisano-Livornese, Lucchese, Elbano and Pistoiese as independent dialectal varieties in his seminal work *Toscana.* The two subdivisions are compatible with each other but adopt different levels of granularity, i.e. they are seen through lenses differing in their magnifying power. Depending on the specific goals of a study, different levels of granularity of the dialectal landscape may be appropriate. By exploiting the hierarchical clustering results, the HBSGP method can also be used to identify increasingly smaller dialectal areas associ-

Table 3: The five topmost lexical variants for the smaller peripheral areas F, B, D and G.

Cluster	Concept-Lexicalization pair	Representativeness	Distinctiveness	Importance
F	FINCH-*frenguéllo*	1.000	1.000	1.000
	CUCUMBER-*citróne*	1.000	0.973	0.973
	HAIL-*granìschia*	0.667	1.000	0.667
	GOOSE-*ciucióne*	0.667	1.000	0.667
	LIZARD-*racanàccio*	0.667	1.000	0.667
B	SNOW-*gnéva*	0.429	1.000	0.429
	ROLLING PIN-*canèlla*	0.429	1.000	0.429
	STYE-*orzaiolo*	0.653	0.633	0.414
	GARBAGE-*rùsco*	0.531	0.734	0.389
	LIZARD-*ciortellóne*	0.430	0.853	0.367
D	HORNET-*buffóne*	0.950	1.000	0.950
	KHAKIS-*cicàchi*	0.500	1.000	0.500
	KHAKIS-*cicàco*	0.500	1.000	0.500
	PINE CONE-*pignòcca*	0.500	1.000	0.500
	TROUGH-*tròlego*	0.500	1.000	0.500
G	WATERMELON-*patècca*	1.000	1.000	1.000
	MELON-*melòne*	1.000	1.000	1.000
	CLUSTER-*raspòllo*	1.000	1.000	1.000
	SQUIRREL-*miseràngolo*	1.000	1.000	1.000
	LIZARD-*bìscia*	1.000	1.000	1.000

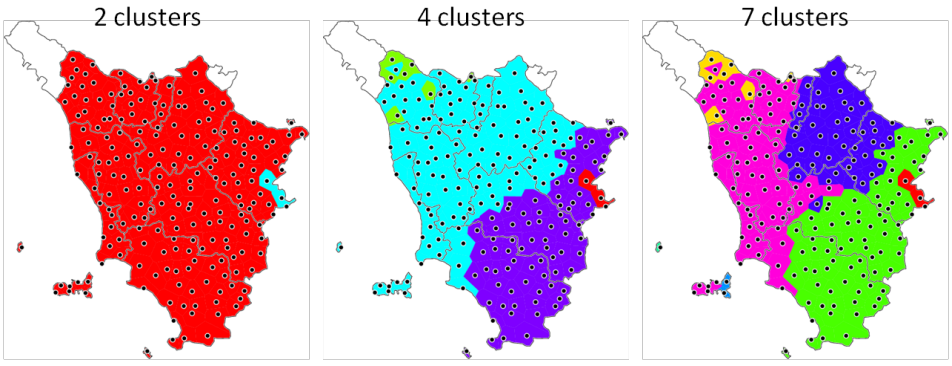

Figure 3: Geographic visualization of the clustering of Tuscan varieties into two, four and seven groups.

Table 4: The five topmost lexical variants of the red, cyan and pink areas in the two, four and seven-cluster maps of Tuscan dialects.

Cluster	Concept-Lexicalization pair	Representativeness	Distinctiveness	Importance
Two-cluster map: Red	SINK-*acquàio*	0.909	1.000	0.909
	CELERY-*sèdano*	0.853	1.000	0.853
	MELON-*popóne*	0.844	1.000	0.844
	LAUREL-*allòro*	0.801	1.000	0.801
	WATERMELON-*cocómero*	0.794	1.000	0.794
Four-cluster map: Cyan	THIMBLE-*anèllo*	0.495	0.857	0.424
	OIL JAR-*cóppo*	0.525	0.808	0.424
	CATERPILLAR-*brùcio*	0.448	0.928	0.416
	EYE GUM-*cìspia*	0.498	0.798	0.397
	TURKEY-*lùcio*	0.445	0.872	0.388
Seven-cluster map: Pink	OIL JAR-*cóppo*	0.749	0.696	0.522
	EYE GUM-*cìspia*	0.702	0.676	0.474
	BREAST-*pùppa*	0.649	0.717	0.466
	FLEA-*pùce*	0.602	0.686	0.413
	CLUSTER OF GRAPES-*pìgna*	0.570	0.701	0.400

ated with progressively more specific lexical features. These nested dialect areas are characterized by nested isoglosses (i.e. the spatial distribution of one feature is entirely contained within that of another). To assess these nested isoglosses, we compare the geographical and linguistic results obtained by clustering the selected dataset into two, four and seven groups (with the latter representing the clustering discussed so far).

Figure 3 reports the geographic visualization of clustering the Tuscan varieties into two, four and seven groups. In the map with two clusters (Figure 3, left), the large red cluster corresponds to the composite set of Tuscan dialects, excluding only the Chiana Valley dialects (cyan cluster). The map with four clusters (Figure 3, middle) shows the main subdivision of Tuscan dialects between Northern dialects (cyan and green clusters), covering (from east to west) Fiorentino, Pistoiese, Lucchese and Pisano-Livornese, and Southern dialects (violet and red clusters), i.e. (from east to west) the dialect from Arezzo, Siena and Grosseto (violet cluster) and from the Chiana valley (red cluster). The map containing seven clusters (Figure 3, right) has already been discussed above.

Table 4 shows the lexical features characterizing the red, cyan and pink clusters in the first, second and third map, respectively. These clusters cover a progres-

sively restricted area. Table 4 reports, for each of these clusters, the five topmost lexical variants with their associated scores. The most salient CL pairs characterizing the red cluster of the two-clusters map coincide with pan-Tuscan words well known from the literature (Giacomelli & Poggi Salani 1984): they show a distinctiveness value equal to 1 and very high representativeness values (≥ 0.79). Similar observations hold for the cluster corresponding to the set of Northern Tuscan dialects (the cyan cluster in Figure 3, middle) with one main difference: all values are considerably lower, with a general reduction observed at the level of representativeness. This illustrates that the cyan cluster is a heterogeneous area. However, by comparing the CL pairs underlying the cyan cluster in the second map and the pink cluster in the third map, we can also see there are two shared lexical variants, namely OIL JAR-*cóppo* and EYE GUM-*cìspia*, which appear among the topmost features whose importance values in the smaller pink cluster are higher (determining a higher ranking), despite their unavoidably lower distinctiveness. In this case, these CL pairs are more characteristic of the smaller cluster, whereas a word such as THIMBLE-*anèllo* is more characteristic of the larger cluster (in the pink cluster it appears in a lower position with much lower values). This suggests that whenever the same features appear to qualify nested clusters, they should be taken as relevant features for the cluster in which they play a more prominent role (i.e. having a higher importance value). Consequently, OIL JAR-*cóppo* and EYE GUM-*cìspia* should be removed from the most salient features of the cyan cluster due to the lower importance (0.424 against 0.522 for the former, and 0.397 against 0.474 for the latter) with respect to the nested pink cluster.

In sum, these results show that hierarchical spectral partitioning can be usefully exploited to identify dialectal areas at different levels of granularity with their associated lexical features. In particular, the method may help in the selection of the most appropriate isoglosses for each dialectal area and in the reconstruction of nested isoglosses.

4.2 Reconstructing the dynamics of lexical change

The hierarchical spectral partitioning method can also be used for studying the dynamics of lexical change across generations. For this purpose, ALT speakers were grouped in an old age group (born in 1930 or earlier – 1930 was the median year of birth) and a young age group (born after 1930). To guarantee comparability of results, we focused on two maps each having four clusters. As Figure 4 shows, the analysis of the two datasets results in slightly different, partially overlapping lexical areas, with the area corresponding to the southeastern (cyan) cluster being more restricted for the older speakers. Major differences, however, are

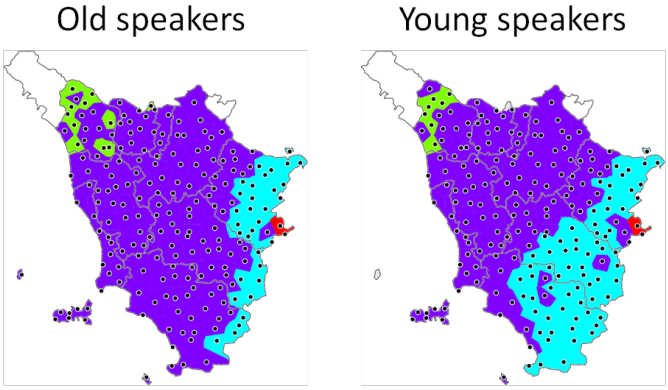

Figure 4: Geographic visualization of a four-way clustering of Tuscan varieties on the basis of data from young vs. old speakers.

explicitly clear at the level of the underlying lexical features. In particular, the central blue area is more restricted (and also linked with fewer CL pairs: 881 vs. 1193) in the map built on the basis of the answers by the young speakers.

Besides the different size of the set of associated linguistic features (i.e. more reduced in the case of young speakers), it is interesting to note that 424 salient lexical features underlying the old speakers map do not appear among the features underlying the young speakers map. These CL pairs emerging from old speakers correspond typically to old-fashioned and traditional notions as well as less common plants and animals. Examples include STRUCTURE FOR BED WARMER-*prète*, POPPY-*ròsolo*, MUTTON-*bìrro*, SET OF POPLARS-*alborellàia*. These CL pairs can be seen as lexical variants which are no longer being used by younger speakers, and these are likely to disappear altogether.

The number of CL pairs restricted to young speakers is much lower (112) than the number of CL pairs restricted to the old speakers. In this case, the CL pairs correspond to standard Italian words (e.g., CLOSET-*ripostìglio*, WEEPING WILLOW-*sàlice piangènte*, HARVEST-*mietitùra*), generic terms (e.g., AFTERNOON-*dópo mangiàto*, SLUG-*lumàca ignùda*) or "distorted" (i.e. deviant with respect to traditional pronunciation) variants of dialectal terms (e.g., TUSCAN COLD CUT FROM PORK SHOULDER-*capricòllo*). The typology of these lexical variants shows the dynamics of lexical change ongoing in younger Tuscan generations, characterized by the loss of local features in favor of generic or standard terms, and by the creative distortion of dialectal words.

In both cases, however, these CL pairs are not highly ranked (i.e. not the most

important) for the associated old and young clusters. Instead, the CL pairs underlying both maps (a total of 769) show clear differences with respect to their ranking. For example, the 1st, 10th, 20th and 50th lexical variants in the ranked list of CL pairs underlying the old speakers map correspond to the 60th, 809th, 59th and 818th position in the young CL pairs list, respectively. Similarly, the 1st, 10th, 20th and 50th ranked lexical variants of the young speakers are ranked (respectively) in the 100th, 13th, 17th and 69th position in the old speakers list. The asymmetry between the old-young vs. young-old correspondences can be seen as the result of a dialect leveling process, causing the lower importance of old-fashioned lexical variants for the young speakers (which are top-ranked for the old speaker). Seen from the perspective of young speakers, the disalignment of the ranking is more reduced, reflecting an additional shared set of dialectal lexical items.

Table 5 reports the five topmost CL pairs underlying the blue cluster in the two maps. Clearly, the importance values associated with the blue cluster of the old speakers are higher than those associated with the blue cluster of the young speakers. This pattern is confirmed by comparing the average importance scores of the top-10 and top-100 CL pairs in the two lists, which are much higher for the old speakers (0.42 vs. 0.34 for the top-10 and 0.26 vs. 0.17 for the top-100). This may also be seen as evidence in support of dialect leveling: lexical areas inferred from young speakers data are characterized by less distinctive and/or representative features.

Table 5: The five topmost lexical variants of the blue cluster in the young vs. old speakers maps of Tuscan dialects.

Cluster	Concept-Lexicalization pair	Representativeness	Distinctiveness	Importance
Old speakers: Blue cluster	GRAPE-*chìcco*	0.721	0.828	0.597
	CHESTNUT HUSK-*rìccio*	0.706	0.661	0.467
	EMBERS-*bràce*	0.673	0.632	0.425
	BRAZIER-*bracière*	0.596	0.680	0.405
	HAZELNUT-*nocciòla*	0.794	0.507	0.403
Young speakers: Blue cluster	BAT-*pipistrèllo*	0.736	0.538	0.396
	BREAST-*pùppa*	0.428	0.900	0.385
	THIMBLE-*anèllo*	0.394	0.893	0.352
	OIL JAR-*cóppo*	0.437	0.772	0.337
	EYE GUM-*cìspia*	0.431	0.779	0.335

5 Conclusion

In this paper, we illustrated the application of hierarchical spectral partitioning of bipartite graphs in the study of lexical variation in Tuscany based on the dialectal corpus of the *Atlante Lessicale Toscano*. Our results demonstrate the potential of the method in bridging the gap between models of linguistic variation based on aggregate analyses and more traditional analyses based on individual linguistic features.

By using the HBSGP method, we not only identified existing patterns of lexical variation in Tuscany on the basis of the whole dialectal corpus, but also uncovered the underlying lexical features in terms of the characterizing concept-lexicalization pairs. The most relevant CL pairs represent the features used to classify and define each identified lexical area. To put it in more traditional terms, they can be seen as a proxy of lexical isoglosses marking both the qualitative and quantitative distribution of the lexical variants identified as discriminating features of a given lexical dialect area. This entails that the set of the topmost CL pairs associated with each identified lexical dialect area acts as a proxy of bundles of isoglosses, where the grading of individual isoglosses within the bundle is determined on the basis of the combination of representativeness and distinctiveness. If the representativeness score associated with identified isoglosses (CL pairs) can help to shed light on how much variation exists within the area defined by a given isogloss, the distinctiveness score reflects how strongly the lexical variant is concentrated within that area. By comparing the results obtained for different dialect areas, we have seen that different stages of the process of dialect differentiation can be inferred from the different values of these two measures: dialectal subdivisions range from clearly defined areas to areas characterized by fuzzy borders.

We also investigated whether and to what extent patterns of lexical variation and their associated features varied with respect to the granularity of the identified dialectal areas and with the age of informants, revealing interesting results. The possibility of exploring linguistic variation at different levels of granularity makes it possible to customize the analysis with respect to the user's needs. The linguistic features associated with increasingly smaller areas can be seen as nested isoglosses, occurring when the spatial distribution of one feature is contained entirely within that of another and establishing an implicational relationship between the two.

The analysis and comparison of lexical variation patterns and associated features across generations showed that the method can also be usefully exploited

to track the change in the typology of features in young vs. old informants and to monitor the vitality of a dialect in a given area. In particular, the HBSGP method turned out to effectively capture the dynamics of lexical change in Tuscany, by highlighting the emergence of lexical innovations and the obsolescence of old-fashioned traditional dialectal words.

Current directions of research include testing the robustness of these results by noisy clustering and the analysis of lexical variation patterns across semantic domains.

Acknowledgements

The research reported in this article was carried out in the framework of a Short Term Mobility program of international exchanges funded by the National Council of Research (CNR, Italy). The authors thank the anonymous reviewers for their comments, which have helped to improve this article.

References

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. Genova.

Dhillon, Inderjit S. 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* (KDD '01), 269–274. New York, NY, USA: ACM. DOI:10.1145/502512.502550

Giacomelli, Gabriella. 1975. Aree lessicali toscane. *La ricerca dialettale* 1. 115–152.

Giacomelli, Gabriella & Teresa Poggi Salani. 1984. Parole toscane. *Quaderni dell'Atlante Lessicale Toscano* 2(3). 123–229.

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.

Giannelli, Luciano. 1976 [2010]. *Toscana*. Pisa: Pacini Editore.

Grieve, Jack, Dirk Speelman & Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23(2). 193–221.

Labov, William, Sharon Ash & Charles Boberg. 2006. *The atlas of North American English: Phonetics, phonology and sound change*. Berlin, New York: Mouton de Gruyter.

Loporcaro, Michele. 2009. *Profilo linguistico dei dialetti italiani*. Roma, Bari: Laterza.

Montemagni, Simonetta, Martijn Wieling, Bob De Jonge & John Nerbonne. 2012. Patterns of language variation and underlying linguistic features: A new dialectometric approach. In Patricia Bianchi, Nicola De Blasi, Chiara De Caprio & Francesco Montuori (eds.), *La variazione nell'italiano e nella sua storia. Varietà e varianti linguistiche e testuali. Atti dell'XI congresso SILFI (Società Internazionale di Linguistica e Filologia Italiana)*, 879–889. Firenze: Franco Cesati.

Montemagni, Simonetta, Martijn Wieling, Bob De Jonge & John Nerbonne. 2013. Synchronic patterns of Tuscan phonetic variation and diachronic change: Evidence from a dialectometric study. *Literary and Linguistic Computing* 28(1). 157–172.

Nerbonne, John. 2006. Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing* 21(4). 463–475.

Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3(1). 175–198.

Pellegrini, Giovanni Battista. 1977. *Carta dei dialetti d'Italia*. Pisa: Pacini.

Pröll, Simon, Simon Pickl & Aaron Spettl. in press. Latente Strukturen in geolinguistischen Korpora. In Michael Elmentaler, Markus Hundt & Jürgen Erich Schmidt (eds.), *Deutsche Dialekte - Konzepte, Probleme, Handlungsfelder*. Stuttgart: Steiner.

Schneider, Edgar W. 1988. Qualitative vs. quantitative methods of area delimitation in dialectology: A comparison based on lexical data from Georgia and Alabama. *Journal of English Linguistics* 21. 175–212.

Shackleton, Robert. 2005. English-American speech relationships. *Journal of English Linguistics* 33(2). 99–160.

Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35(138). 335–357.

Wieling, Martijn & Simonetta Montemagni. 2015. Infrequent forms: Noise or not? In Marie-Hélène Côte, Remco Knooihuizen & John Nerbonne (eds.), *The Future of Dialects*. Berlin: Language Science Press.

Wieling, Martijn & John Nerbonne. 2009. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. In *Proceedings of the 2009 Workshop on Graph-Based Methods for Natural Language Processes*, 14–22. Stroudsburg, PA: ACL.

Wieling, Martijn & John Nerbonne. 2010. Hierarchical spectral partitioning of bipartite graphs to cluster dialects and identify distinguishing features. In *Proceedings of the 2010 workshop on graph-based methods for natural language processing*, 33–41. Stroudsburg, PA: ACL.

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

Wieling, Martijn & John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics* 1. 243–264.

Wieling, Martijn, Robert Shackleton & John Nerbonne. 2013. Analyzing phonetic variation in the traditional English dialects: Simultaneously clustering dialects and phonetic features. *Literary and Linguistic Computing* 28(1). 31–41.

Wieling, Martijn, Clive Upton & Ann Thompson. 2014. Analyzing the BBC Voices data: Contemporary English dialect areas and their characteristic lexical variants. *Literary and Linguistic Computing* 29(1). 107–117.

Wieling, Martijn, Simonetta Montemagni, John Nerbonne & R. Harald Baayen. 2014. Lexical differences between Tuscan dialects and Standard Italian: A sociolinguistic analysis using generalized additive mixed modeling. *Language* 90(3). 669–692.

Woolhiser, Curt. 2005. Political borders and dialect divergence/convergence in europe. In Peter Auer, Frans Hinskens & Paul Kerswill (eds.), *A handbook of varieties of English*, 236–262. New York: Cambridge University Press.