International Journal of Humanities and Arts Computing, 2 (1–2) 2008, Edinburgh University Press, pp.135–151

The space of Tuscan dialectal variation. A correlation study

Simonetta Montemagni
Istituto di Linguistica Computazionale – CNR
via G. Moruzzi 1
56124, Pisa – ITALY
simonetta.montemagni@ilc.cnr.it

Abstract

The paper illustrates the results of a correlation study focusing on linguistic variation in an Italian region, Tuscany. By exploiting a multi-level representation scheme of dialectal data, the study analyses attested patterns of phonetic and morpho-lexical variation with the aim of testing the degree of correlation between a) phonetic and morpho-lexical variation, and b) linguistic variation and geographic distance. The correlation analysis was performed by combining two complementary approaches proposed in dialectometric literature, namely by computing both global and place-specific correlation measures and by inspecting their spatial distribution. Achieved results demonstrate that phonetic and morpho-lexical variations in Tuscany seem to follow a different pattern than encountered in previous studies.

Keywords

Dialectometry, phonetic variation, lexical variation, language-space correlation, correlation among different linguistic levels

1. Introduction

It is a well-known fact that different types of features contribute to the linguistic distance between any two locations, which can differ for instance with respect to the word used to denote the same object or the phonetic realisation of a particular word. Yet, the correlation between different feature types in defining patterns of dialectal variation represents an area of research still unexplored. In traditional dialectology, there is no obvious way to approach this matter beyond fairly superficial and impressionistic observations. The situation changes if the same research question is addressed in the framework of dialectometric studies, where it is possible to measure dialectal distances with respect to distinct linguistic levels and to compute whether and to what extent observed distances correlate. Another related question concerns the influence of geography on linguistic variation. Answering this question can help to shed light on whether observed correlations among linguistic levels should instead be interpreted as a separate effect of the underlying geography. Over the last years, both Gröningen and Salzurg schools of dialectometry have been engaged in providing answers to these questions from different perspectives and working with different data from various languages. Concerning the former, it is worth mentioning the contributions by Nerbonne (2003), Gooskens and Heeringa (2006) and Spruit et al. (in press); the latter is represented by the “correlative dialectometry” studies of Goebl (2005, 2008). In both cases, this appears to be a promising line of research.

The main goal of this study is to gain insight into the nature of linguistic variation by investigating the degree to which a) patterns of dialectal variation computed with respect to different linguistic levels correlate in the language varieties spoken in Tuscany (a region which has a special status in the complex puzzle of linguistic variation in Italy),[1] and b) linguistic patterns of variation correlate with geographic distance. The study was performed on the corpus of dialectal data *Atlante Lessicale Toscano* (‘Lexical Atlas of Tuscany’, henceforth ALT), by combining complementary approaches proposed in the dialectometric literature: two dialectometric software packages have been used, namely RUG/L04 developed by P. Kleiweg and VDM by E. Haimerl.[2] The starting point is represented by the results of a dialectometric study focusing on phonetic and lexical variation in Tuscany (Montemagni 2007). By exploiting a multi-level representation scheme of dialectal data, the linguistic distances among the investigated locations were measured with respect to different linguistic levels. Correlational analyses were then performed on the resulting distance matrices in order to estimate the degree of association between the different levels and to evaluate the role played by geography in explaining observed correlations.

2. The data source

2.1 The *Atlante Lessicale Toscano*

ALT is a regional linguistic atlas focusing on dialectal variation throughout Tuscany, a region where both Tuscan and non-Tuscan dialects are spoken; the latter is the case of dialects in the north, namely Lunigiana and small areas of the Apennines (so-called Romagna Toscana), which rather belong to the group of Gallo-Italian dialects. ALT interviews were carried out in 224 localities of Tuscany, with 2,193 informants selected with respect to a number of parameters ranging from age and socio-economic status to education and culture. The interviews were conducted by a group of trained fieldworkers who employed a questionnaire of 745 target items, designed to elicit variation mainly in vocabulary, semantics and phonetics. A dialectal corpus with these features lends itself to investigations concerning geographic or horizontal (diatopic) variation as well as social or vertical (diastratic) variation: in this study we will focus on the diatopic dimension of linguistic variation. ALT, originally published in the year 2000 (Giacomelli *et al.* 2000) as a CD-Rom, is now available as an on-line resource, ALT-Web[3].

2.2 ALT-Web representation of dialectal data

In ALT, all dialectal items were phonetically transcribed.[4] In order to ensure a proper treatment of these data, an articulated encoding schema was devised in ALT-Web in which all dialectal items are assigned different levels of representation: a first level rendering the original phonetic transcription as recorded by fieldworkers; other levels containing representations encoded in standard Italian orthography. In this multi-level representation scheme, dialectal data are encoded in layers of progressively decreasing detail going from phonetic transcription to different levels of orthographic representations eventually abstracting away from details of the speakers' phonetic realisation.[5]

For the specific concerns of this study, we will focus on the following representation levels: phonetic transcription (henceforth, PT) and normalised representation (henceforth, NR) where the latter is the representation level meant to abstract away from within-Tuscany vital phonetic variation. At the NR level a wide range of phonetic variants is assigned the same normalised form: e.g. words such as [skja'ttʃata], [skja'ttʃaθa], [skja'ttʃada], [skja'ttʃaða], [stja'ttʃata], [stja'ttʃaθa], [stja'ttʃada], [stja'ttʃaða], [stʃa'ssɛda] etc. (denoting a traditional type of bread, flat and crispy, seasoned on top with salt and oil) are all assigned the same normalised form, SCHIACCIATA. Note that at this level neutralisation is only concerned with phonetic variants resulting from productive phonetic processes: this is the case, for instance, of variants involving spirantization or voicing of plosives like /t/, as in [skja'ttʃaθa] and [skja'ttʃada]. On the contrary, there are word forms like ['kaʎʎo] and ['gaʎʎo] (meaning 'rennet') which are assigned distinct NRs, CAGLIO and GAGLIO respectively: this follows from the fact that the [k] vs [g] alternation in word-initial context represents a no longer productive phonetic process in Tuscany. It should also be noted that the NR level does not deal with morphological variation (neither inflectional nor derivational). This entails that words such as [skjaˈttʃata] (singular) and [skjaˈttʃate] (plural) as well as [skjattʃaˈtina] (diminutive) are all assigned different NFs. Currently, NR is the most abstract representation level in ALT-Web.

3. Induction of patterns of phonetic and lexical variation

3.1 Building the experimental data sets

The representation scheme illustrated in section 2.2 proved to be particularly suitable for dialectometric analyses of dialectal data at various linguistic description levels.

First, patterns of phonetic and lexical variation could be studied with respect to different representation levels, providing orthogonal perspectives on the same set of dialectal data. In particular, the study of phonetic variation was based on PTs, whereas NRs were used as a basis for the investigation on lexical variation.

Second, the alignment of the representation levels was used to automatically extract all attested phonetic variants of the same normalised word form (henceforth, NF). In practice, the various phonetic realisations of the same lexical unit were identified by selecting all phonetically transcribed dialectal items sharing the same NF. Since the ALT-Web NR level does not abstract away from either morphological variation or no longer productive phonetic processes, we can be quite sure that phonetic distances calculated against phonetic variants of the same NF testify vital phonetic processes only, without influence from any other linguistic description level (e.g. morphology).

The experimental data used for the study of phonetic variation was thus formed by the normalised forms attested in the ALT corpus, each associated with the set of its phonetically transcribed variants; this is exemplified in the first two columns of Table 1 for the normalised form SCHIACCIATA. For the study of lexical variation we instead used ALT onomasiological questions (i.e. those looking for the attested lexicalisations of the same concept) with their associated normalised answers; this is exemplified in the last two columns of Table 1 in which the NFs collected as answers to the question n. 290 ‘schiacciata’ are reported.

NF = SCHIACCIATA	
15 Vergemoli	[scaˈttʃata]
16 Pieve Fosciana	[scaˈttʃada]
18 San Pellegrino in Alpe	[scaˈttʃata], [stʲaˈttʃata]
19 Brandeglio	[scaˈttʃaθa], [stʲaˈttʃaθa]
22 Prunetta	[stjaˈttʃaθa]
23 Orsigna	[skjaˈttʃaθa], [scaˈttʃaθa], [stjaˈttʃata]
24 Spedaletto	[stjaˈttʃaθa]
25 Castello di Sambuca	[scaˈttʃada]
28 Barberino di Mugello	[skjaˈttʃata], [stjaˈttʃata]
…	…

question = n. 290 ‘schiacciata’ (traditional type of bread, flat and crispy, seasoned on top with salt and oil)	
15 Vergemoli	FOCACCIA, FOCACCINA, **SCHIACCIATA**
16 Pieve Fosciana	FOCACCIA, **SCHIACCIATA**
17 Barga	FOCACCIA
18 San Pellegrino in Alpe	FOCACCIA, PATTONA, **SCHIACCIATA**
19 Brandeglio	FOCACCIA, SCHIACCIA, **SCHIACCIATA**
20 Rivoreta	FOCACCIA
21 Popiglio	SCHIACCIA
22 Prunetta	**SCHIACCIATA**
23 Orsigna	COFACCIA, SCHIACCIA, **SCHIACCIATA**
…	…

Table 1 – Excerpts from the experimental data sets used for the study of phonetic and lexical variation.

3.2 Measuring linguistic distances in Tuscany

3.2.1 Methodology

The linguistic distances across the locations of the ALT geographic network were calculated with the Levenshtein Distance measure (henceforth, LD), a string-distance measure originally used by Kessler (1995) as a means of calculating the distance between the phonetic realisations of corresponding words in different dialects. Kessler showed that with LD it is possible to "reliably group a language into its dialect areas, starting from nothing more than phonetic transcriptions as commonly found in linguistic surveys" (Kessler, 1995:66). The LD between two strings is given by the minimum number of operations needed to transform one string into the other; the transformation is performed through basic operations (namely the deletion or the insertion of a string character, or the substitution of one character for another), each of which is associated a cost.

With LD, comparing two dialectal varieties results in the mean distance of all performed word-pair comparisons. The use of LD in calculating the linguistic distance between language varieties was further extended and improved by Nerbonne *et al.* (1999) and Heeringa (2004) who worked on different languages and with different representation types (i.e. phone-based, feature-based and acoustic representations). In these dialectal studies based on LD, the standard measure was also refined to cope with dialectology-specific issues, dealing with: a) the normalisation of the distance measure with respect to the length of compared words (Nerbonne *et al.* 1999); b) the treatment of multiple responses (Nerbonne and Kleiweg 2003).

In the present study, we use LD to calculate linguistic distances between the ALT locations:[6] the distance between each location pair is obtained by averaging the LDs calculated for individual word pairs, be they phonetic realisations of the same NF or lexicalisations of a given concept (see section 3.1). Missing dialectal items are ignored due to their uncertain origin.[7] In what follows, we will focus on issues specific to the measure of linguistic distances with the ALT data.

3.2.2 Measuring phonetic distances

Using LD, the phonetic distance between two linguistic varieties A and B is computed by comparing the phonetic variants of NFs in A with the phonetic variants of the same NF set in B. The phonetic realisation of a given word can be represented in different ways giving rise to different approaches to the measure of phonetic distance, respectively denominated by Kessler (1995) "phone string comparison" and "feature string comparison". In the former, LD operates on sequences of phonetic symbols, whereas in the latter comparison is carried out with respect to feature-based representations. Both approaches were experimented with in the study of phonetic variation in Tuscany;[8] due to the almost equivalent results achieved in the two experiments,[9] in what follows we will focus on the distance matrix calculated on the basis of phone-based representations.

The experimental data set included only NFs having at least two phonetic variants attested in at least two locations. A collection of 9,082 NFs was thus selected, with associated 32,468 phonetic variants types: within this NF set, geographical coverage ranges between 2 and 224 and phonetic variability between 2 and 34. The resulting phonetic distance matrice was built on the basis of the 206,594 phonetic variants attested as instantiations of the selected NFs. In order to assess the reliability of the data set, we calculated the coefficient Cronbach α (Heeringa 2004:170-173) which was 0.99. This means that this data set provides a reliable basis for an analysis of phonetic differences based on LD.

The distance between the phonetic variants of the same NF in different locations was calculated on the basis of the raw LD, without any type of normalisation by the length of compared transcriptions: in this way, all sound differences add the same weight to the overall distance and are not inversely proportional to the word length as in the case of normalised distances. This choice is in line with the Heeringa *et al.* (2006) findings which notice that raw LD represents a better approximation of phonetic differences among dialects as perceived by dialect speakers than results based on normalised LD.

3.2.3 Measuring lexical distances

Whereas a study of phonetic variation based on phonetically transcribed data could only be conducted with LD, this choice is not to be taken for granted in the case of lexical distances. In fact, in the pioneering research by Seguy (1971) and Goebl (1984) the comparison between any two sites is perfomed starting from the proportion of shared answers to a given questionnaire item and of those which differ. Yet, it is often the case that answers elicited from informants are different forms of the same lexical item: typically, they are inflectional or derivational variants of the same lemma. Moreover, they can also include diacronically (e.g. etymologically) related words. By adopting a binary notion of lexical distance, related but different lexical items are treated as completely unrelated answers. To overcome this problem, in their study of lexical variation in LAMSAS Nerbonne and Kleiweg (2003) applied LD to measure also the lexical distance of the answers on the basis of the encouraging results previously obtained in the study of phonetic variation. With LD, related lexical items are no longer treated as irrelated answers and their partial similarity is taken into account.[10]

We felt that the use of LD for measuring lexical distances was also appropriate in the ALT case. This choice appears even more crucial if we consider the type of representation of dialectal data we are dealing with. Although we are using previously normalised dialectal forms, we have seen that this representation level does not abstract away from morphological variation or from no longer productive phonetic processes. To keep with the SCHIACCIATA example, the questionnaire item meant to gather lexicalisations of the concept of this traditional type of bread includes answers both in the singular and in the plural forms (e.g. *schiacciatina* vs *schiacciatine*), gender variants (e.g. *schiaccino*-masculine vs *schiaccina*-feminine), as well as derivationally related variants such as *schiaccia*, *schiaccina* and *schiacciata* or multi-word expressions like *schiacciata unta* (lit. SCHIACCIATA with oil) or *schiacciata al sale* (lit. salted SCHIACCIATA). At the NR level, all these forms still represent distinct answers. By resorting to LD, their relatedness can be accounted for in the measure of lexical distance.[11]

The present study of lexical variation in Tuscany is based on the entire set of ALT onomasiological questions (see section 3.1), namely 460 questionnaire items which gathered a total of 39,761 normalised answer types geographically distributed into 227,555 tokens. In this case, the coefficient Cronbach α was 0.97, showing that this was a sufficient basis for a reliable analysis.

Lexical distances were measured using LD operating on NFs. Given the features peculiar to the NR level, the resulting measure of lexical distance has to be seen as reflecting patterns of morphological variation as well, especially for what concerns derivation. For this reason, from now on we will refer to the distances computed against NFs as “morpho-lexical distances”. Differently from the phonetic distance computation, here it makes sense to normalise LD so that it is independent from the length of compared words (as suggested in Nerbonne *et al.* 1999). This choice follows from the fact that in the study of lexical variation words are to be considered as the linguistic units with respect to which the distance computation is performed.

4. Linguistic and geographic distances: within and between correlations

4.1 Methodology

Following Heeringa and Nerbonne (2001), the phonetic and morpho-lexical distance matrices were explored with complementary techniques, namely agglomerative hierarchical clustering and multidimensional scaling: the results of this study are reported in Montemagni (2007). Here it suffices to say that the iconic profiles of phonetic and morpho-lexical variation are visually quite different. Besides the borders identifying non-Tuscan dialects from Lunigiana and Romagna Toscana, proposed phonetic and morpho-lexical dialectal subdivisions do not overlap. This fact needs further investigation aimed at exploring the reasons underlying this state of affairs. In particular, two research questions need to be addressed:

a) whether and to what extent observed patterns of phonetic and morpho-lexical variation are associated with one another;
b) whether and to what extent phonetic and morpho-lexical distances correlate with geographic distance. In particular, if this turns out to be the case, we need to investigate whether they correlate with geography in the same way.

Following Nerbonne (2003), Goebl (2005), Gooskens (2005), Gooskens and Heeringa (2006) and Spruit et al. (in press), the correlation between the distances observed at different linguistic levels on the one hand and between linguistic and geographic distances on the other hand is calculated in terms of the Pearson's correlation coefficient. Two approaches can be recognised in the dialectometric correlation literature:

1. the correlation is measured with respect to the whole place x place matrix, thus providing a global measure of whether and to what extent the distance matrices are correlated: this is the approach followed in the Gröningen school of dialectometry;
2. the correlation is calculated separately for each of the investigated locations giving rise to place-specific measures which can then be visualised on a map highlighting the areas characterised by similar correlation patterns; this corresponds to the "correlative dialectometry" by Goebl (2005).

Interestingly enough, the two approaches complement each other nicely, providing at the same time global and place-specific correlation measures; in this study of Tuscan dialectal variation, both approaches are experimented with.

For the specific concerns of this study, we will focus on Tuscan dialects only, i.e. on the 213 out of the 224 ALT locations where Tuscan dialects are spoken.

4.2 Correlation between phonetic and morpho-lexical distances

By focussing on Tuscan dialects only, the global correlation between phonetic and morpho-lexical distances turns out to be 0.4125, with only 17% of explained variance. This situation is not reflected in the analyses of Tuscan dialects by the main scholar of Tuscan dialectology - Giannelli (2000) - whose proposed subdivision seems to result from the combination of phonetic, phonemic, morpho-syntactic and lexical features.

This global correlation value suggests that within Tuscan-speaking localities it can often be the case that two dialects differ at the level of phonetic features but still have a common vocabulary, or the other way around. In order to check whether and most importantly where this is the case, following the correlative dialectometry approach by Goebl (2005) phonetic/morpho-lexical correlation scores have been calculated separately for each of the investigated locations and then projected on a map: the result is shown in Figure 1.

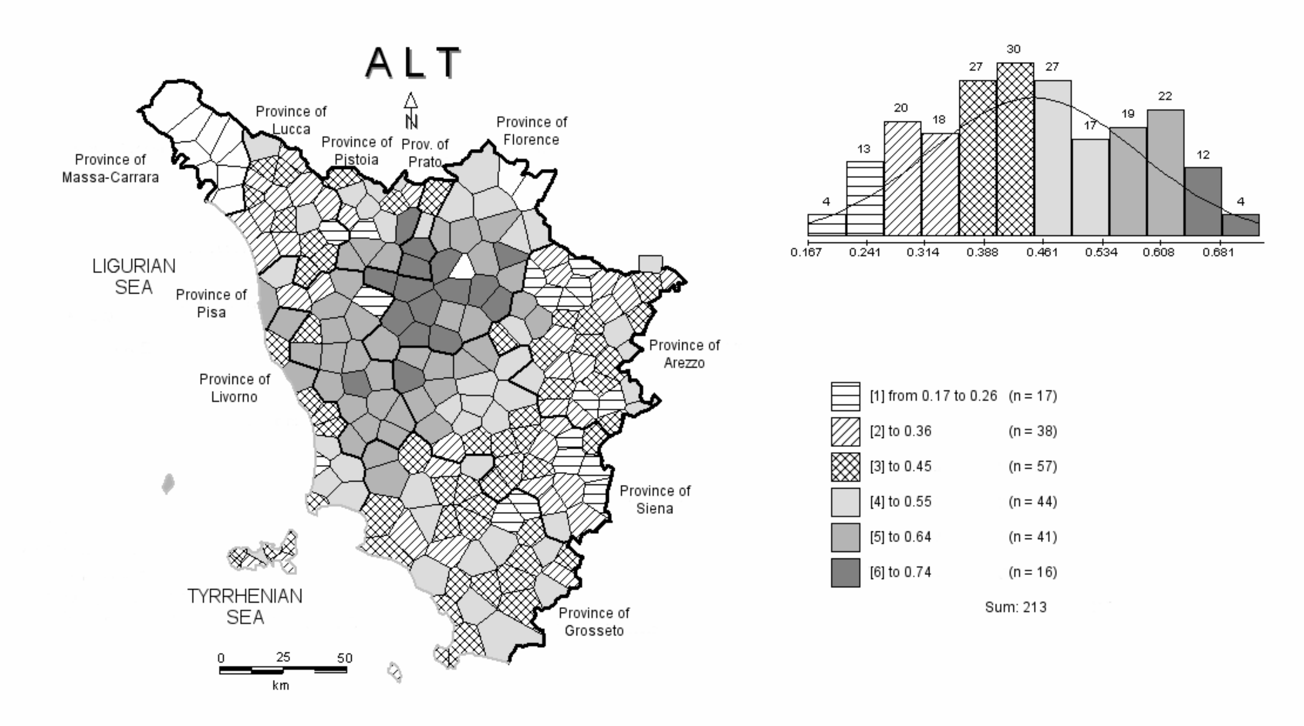

Figure 1 – Choropleth map of the correlation values between 213 phonetic proximity values and 213 morpho-lexical proximity values. Proximity = 100-distance. Algorithm of visualisation: MINMWMAX (6-tuple). Software: VDM.

Following Goebl (2005), the distance values (*dist*) obtained with LD were converted into proximity values (*prox*) with the formula *dist* + *prox* = 100. For each site, obtained phonetic and morpho-lexical proximity values were correlated showing a variability range from 0.17 to 0.74: in the map, the correlation values are organised into 6 intervals according to the MINMWMAX visualisation algorithm (Goebl 2006), where intervals 1-3 and 4-6 gather correlation values respectively below and above the arithmetic mean. The resulting spatial distribution is quite interesting: the darker zones of the map (intervals 5 and 6) indicate those areas in which phonetic variation is in lock-step with morpo-lexical variation. This happens to be the case in the area around Florence (identified in the map by the white polygon), expanding in all directions, in particular west and south. This "harmony" between phonetic and morpho-lexical variation progressively fades in the areas corresponding to intervals from 4 to 1. It is interesting to note that these results are in line with the dialectometric study of the Italian AIS atlas by Goebl (2008:55) who records relatively low phonetics-vocabulary correlation values in the peripheral areas of Tuscany.

4.3 Correlation between linguistic and geographical distances

Before drawing any conclusion, we need to take into account a third factor, geography. How much of the observed linguistic variation can be accounted for by the underlying geography? In previous dialectometric correlation studies, geography has been shown to correlate strongly with variation at different linguistic levels within the same language (Heeringa and Nerbonne 2001, Spruit *at al.* in press). This appears to hold true, with some significant differences due to the underlying geography (see below), also for other languages such as Norwegian (Gooskens 2005). Let us consider whether this is the case for Tuscany as well.

Table 2 reports observed correlations for the 213 Tuscan-speaking localities between geographical distances[12] on the one hand and phonetic and morpho-lexical distances on the other hand; note that all computed correlation coefficients are significant with p=0.0001. The results show that the differences observed at the morpho-lexical level are more strongly associated with geographic distances (r=0.6441) than variation at the phonetic level (r=0.1358). The percentages in the rightmost column indicate the amount of variation at the specified linguistic level which can be explained with geographical distance. Interestingly enough, it turned out that only 1.8% of phonetic variation can be explained with geographical distance.

	Correlation (r)	**Explained variance (r^2 * 100)**
Geography vs phonetic distances	0.1358	1.8%
Geography vs morpho-lexical distances	0.6441	41%

Table 2 – Global correlation between geographic and linguistic distances.

Tuscany presents quite a peculiar situation concerning the correlation of linguistic variation with geography, which differs in two different respects from what has been observed in the dialectometric literature so far. Consider first the association between phonetic and geographic variation: different correlations were observed in the literature, going from r=0.67 in the case of Dutch (Nerbonne *et al.* 1996) to a significantly lower value, i.e. r=0.22, in the case of Norwegian (Gooskens 2005). Gooskens (2005) explains such a different correlation as the impact of geography on dialect variation in Norway, where the central mountain range prevented direct travel until recently: she found out that in Norway travel time is correlated more strongly with linguistic distance than linear geographic distance.

Consider now the second peculiar aspect of Tuscan dialectal variation with respect to geography. Spruit *et al.* (in press) report the correlation observed for Dutch between distances at different linguistic levels and geography: such a correlation appears to be quite high and constant across all levels taken into account, namely pronunciation (r=0.68), syntax (r=0.66) and lexicon (r=0.57). The authors take these results to confirm the fundamental postulate in dialectology that language varieties are structured geographically (Nerbonne and Kleiweg 2007). This does not appear to be the case in Tuscany where significantly different correlations are recorded across distinct levels of linguistic description; because of this, the low phonetics/geography correlation observed in Tuscany cannot be explained in terms of the underlying geography.

Following the correlative dialectometry approach of Goebl, the correlation between linguistic proximity values and geographic proximity values was computed for each site. The results are summarised in Figure 2, with the left map focussing on the spatial distribution of phonetic vs geographic proximity correlation and the right one on morpho-lexical vs geographic proximity correlation. By comparing the variability range in the two maps, it can be observed that the situation differs significantly. Whereas in the case of the phonetics/geography correlation it goes from -0.35 to 0.59, in the case of morpho-lexical variation the correlation values are higher and characterised by a narrower variation span (oscillating from 0.32 to 0.83). Interestingly, the phonetics-geography variation range of intervals 5-6 (0.33-0.59) covers approximatively the variation span of intervals 1-3 (0.32-0.64) in the case of the morpho-lexical/geography correlation. The spatial distribution of phonetics/geography correlation values (Figure 2, left map) follows a pattern similar to what has been observed in Figure 1.

Again, the darker zones of the map (intervals 5 and 6), marking those areas in which phonetic and geographic proximity are "in tune", are located in the area around Florence, expanding south-west down to the coast; the surrounding areas, corresponding to intervals from 4 to 1, are characterised by progressively lower correlation scores. Again, these results are in line with Goebl (2008:54-55), whose linguistics vs geography correlation maps (namely maps 23, 24 and 25) characterise Tuscany as having low correlation scores, especially in the northern part.

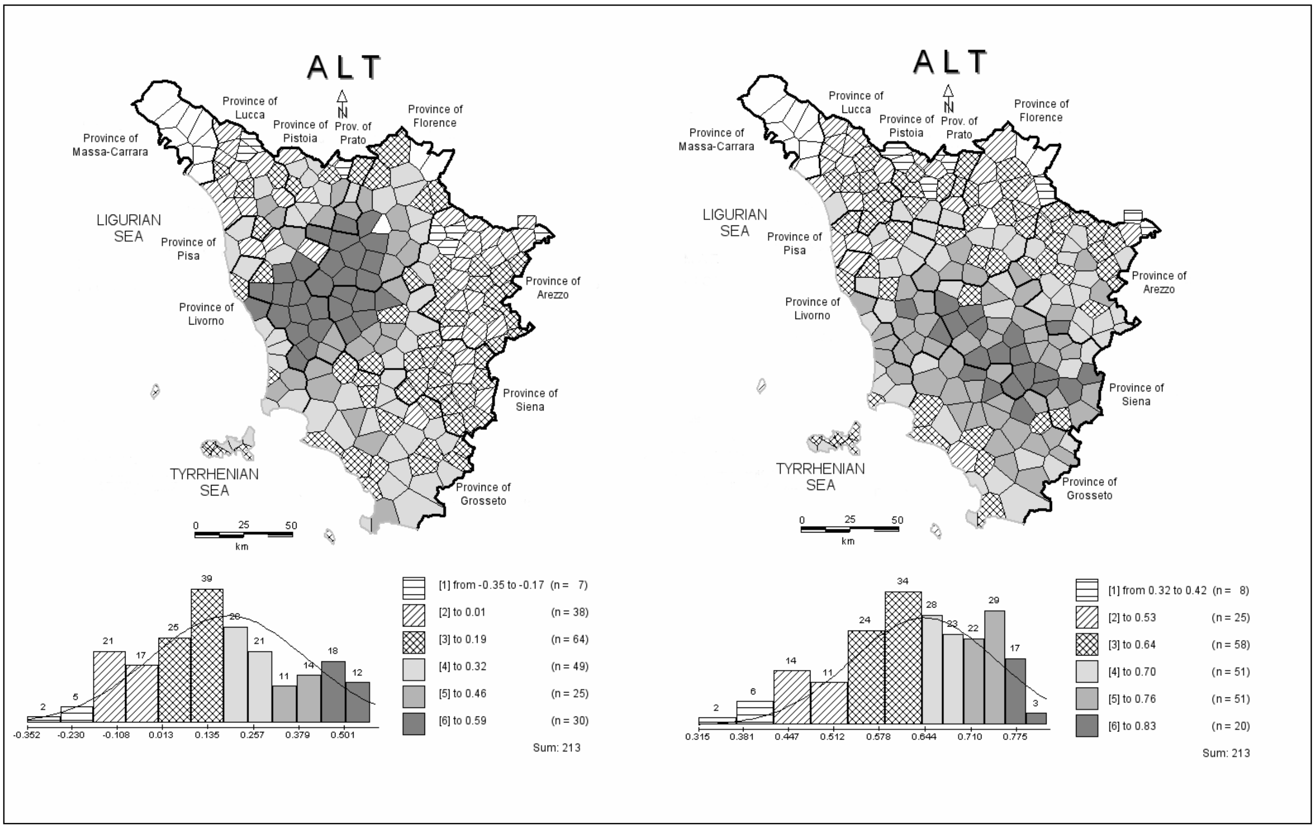

Figure 2 – Choropleth maps of the correlation values between 213 phonetic (left) and morpho-lexical (right) proximity values and 213 geographic proximity values. Algorithm of visualisation: MINMWMAX (6-tuple). Software: VDM.

4.4 Discussion

The Tuscan situation can be summarised as follows. Phonetic and morpho-lexical variation patterns do not correlate strongly (r=0.4125); the correlation between phonetic and geographic distances is much lower (r=0.1358), differing significantly from the correlation between morpho-lexical distances and geography which appears to be considerably higher (r=0.6441). Due to this combined evidence, we cannot explain the low correlation between phonetic and geographical distances in terms of the underlying geography of Tuscany as hypothesized in the case of Norway. Rather, the different correlation with respect to geography seems to suggest that phonetic and morpho-lexical variation in Tuscany is regulated by distinct patterns of linguistic diffusion.

Morpho-lexical variation in Tuscany appears to conform to the dialectological postulate that "geographically proximate varieties tend to be more similar than distant ones" (Nerbonne and Kleiweg 2007). On the contrary: Tuscan phonetic variation presents itself as an exception to the above mentioned dialectological postulate, since phonetic distances are not fully cumulative and there are geographically remote areas which appear to be linguistically similar (Montemagni 2007). Tuscan phonetic variation can thus be seen as resulting from a different pattern of linguistic diffusion: we hypothesise that it is the result of "the displacement of a formerly widespread linguistic feature by an innovation" (Chambers and Trudgill 1998:94).

In order to test this hypothesis, a closer look at phonetic variation is necessary, especially for what concerns the linguistic properties playing a major role in determining identified patterns of phonetic variation. Current research in this direction shows that among the linguistic features playing a major role in determining identified phonetic variation patterns there appears to be spirantization phenomena (so-called "Tuscan gorgia").[13] Tuscan gorgia is accepted as being a local and innovative (presumably dating back to the Middle Ages) natural phonetic phenomenon (consonantal weakening) spreading from the culturally influential center of Florence in all directions, especially southward and westward. Interestingly enough, the spatial distribution of Tuscan gorgia is very close to distribution of the darker zones in Figures 1 and 2, i.e. the areas where phonetic variation appears to correlate more strongly with morpho-lexical variation (Figure 1) and geographic proximity (Figure 2, left map). The converse is also true: surrounding areas, corresponding to the zones not affected by Tuscan gorgia, show in both cases low correlation values; this means that in these areas phonetic variation is no longer aligned with neither morpho-lexical variation nor geography.

5. Conclusions

The paper reports the results of a correlation study focusing on phonetic and morpho-lexical variation in Tuscany. The study was performed on the data extracted from the entire ALT corpus. Phonetic and morpho-lexical distances among Tuscan language varieties were calculated using LD against different representation types (PT and NR respectively). The resulting distance matrices were analysed in order to test the degree of correlation between observed patterns of phonetic and morpho-lexical variation on the one hand, and between linguistic variation and geographic distances on the other hand. The correlation analysis, restricted to the Tuscan dialects area, was performed by combining the two different but complementary approaches proposed in the dialectometric literature, namely by computing both global and place-specific correlation measures and by inspecting their spatial distribution. Differently from the results of previous correlation studies, phonetic and morpho-lexical variation in Tuscany does not appear to conform to the same pattern: whereas the latter can be taken to confirm the postulate that language varieties are structured geographically, the former rather suggests that a different pattern of linguistic variation is at work, characterised by the spread of phonetic features from a core locality to neighbouring ones and by the existence of linguistically related but geographically remote areas.

The contribution of this study is twofold. From the point of view of Tuscan dialectology, it helps gain insight into the nature of diatopic variation at different linguistic description levels, a topic which to our knowledge has never been investigated so far. From a more general dialectometric perspective, one of the innovative contributions of this study consists of identifying radically different patterns of linguistic variation for different description levels with respect to the same area. Obviously, these results need further investigation in different directions. Firstly, it would be interesting to widen the range of linguistic levels taken into account to assess whether there are levels which are more closely associated than others. First experiments in this direction suggest that morphological and lexical variation are more strongly associated than phonetic variation appears to be with them. Secondly, one could extend this correlation study by considering socio-economical factors playing a role in the linguistic variation process as well. Through this, identified variation patterns could result from the complex interaction of geographic and social factors. Note that ALT could be conveniently exploited for this purpose due to the simultaneous diatopic and diastratic characterisation of its data. Last but not least, it would be interesting to apply the adopted correlation methodology to study the relationship between patterns of linguistic variation and genetic or demographic variation, hopefully leading to a deeper understanding of the role of population movements in determining dialect diversity.

Acknowledgements

I would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper. I would like to thank John Nerbonne for his precious suggestions on previous versions of this work, and Peter Kleiweg for his support in the use of the RUG/L04 software. I would also like to thank Hans Goebl for the stimulating comments on the results of the correlational analyses illustrated in this paper, and Slavomir Sobota with Edgar Haimerl for their support with the VDM software. Finally, thanks are also due to Eva Maria Vecchi who reviewed the manuscript.

6. References

[1] J.K.Chambers, P.Trudgill (1998), *Dialectology*, 2nd Edition (Cambridge).

[2] S. Cucurullo, S. Montemagni, M. Paoli, E. Picchi, E. Sassolini (2006), ‘Dialectal resources on-line: the ALT-Web experience’, *Proceedings of LREC-2006* (May 2006, Genova, Italy).

[3] G. Giacomelli, L. Agostiniani, P. Bellucci, L. Giannelli, S. Montemagni, A. Nesi, M. Paoli, E. Picchi, T. Poggi Salani, eds. (2000), *Atlante Lessicale Toscano* (Roma).

[4] L. Giannelli (2000), *Toscana*, 2nd Edition (Pisa) (1976, 1st edition).

[5] H. Goebl (1984), *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF* (Tübingen).

[6] H. Goebl (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.

[7] H. Goebl (2006), ‘Recent Advances in Salzburg Dialectometry’, *Literary and Linguistic Computing*, 21(4), 411-435.

[8] H. Goebl (2008), ‘La dialettometrizzazione integrale dell’AIS. Presentazione di primi risultati’, *Revue de linguistique romane*, 72, 25-113.

[9] C. Gooskens (2005), ‘Traveling time as a predictor of linguistic distance’, *Dialectologia et Geolinguistica*, 13, 38-62.

[10] C. Gooskens, W. Heeringa (2006), ‘The Relative Contribution of Pronunciation, Lexical and Prosodic Differences to the Perceived Distances between Norwegian dialects’, *Literary and Linguistic Computing*, 21(4), 477-492.

[11] C. Grassi, A. Sobrero, T. Telmon (1997), *Fondamenti di Dialettologia Italiana* (Roma-Bari).

[12] B. Kessler (1995), ‘Computational Dialectology in Irish Gaelic’, *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (Dublin), 60–67.

[13] G. Kondrak (2002), *Algorithms for Language Reconstruction* (unpublished Ph.D. Thesis, University of Toronto).

[14] W. Heeringa, J. Nerbonne (2001), ‘Dialect Areas and Dialect Continua’, *Language Variation and Change*, 13, 375-400.

[15] W. Heeringa (2004), *Computational Comparison and Classification of Dialects*, Ph.D. thesis, University of Gröningen, available at http://www.let.rug.nl/~heeringa/dialectology/thesis/

[16] W. Heeringa, P. Kleiweg, C. Gooskens, J. Nerbonne (2006), ‘Evaluation of string distance algorithms for dialectology’, in J. Nerbonne and E. Hinrichs, eds., *Linguistic Distances* (Shroudsburg, PA), 51–62.

[17] S. Montemagni (2007), ’Patterns of phonetic variation in Tuscany: using dialectometric techniques on multi-level representations of dialectal data’, in P. Osenova *et al.*, eds., *Proceedings of the Workshop on Computational Phonology at RANLP-2007* (26 September 2007, Borovetz, Bulgaria), 49-60.

[18] J. Nerbonne, W. Heeringa, E. van den Hout, P. van de Kooi, S. Otten, W. van de Vis (1996), ‘Phonetic Distance between Dutch Dialects’, in G. Durieux, W. Daelemans, S. Gillis, eds., *Proceedings of the Sixth CLIN Meeting* (Antwerp, Centre for Dutch Language and Speech, UIA), 185-202.

[19] J. Nerbonne, W. Heeringa, P. Kleiweg (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), v-xv.

[20] J. Nerbonne (2003), ‘Linguistic Variation and Computation’, *Proceedings of the 10th Meeting of the European Chapter of the Association for Computational Linguistics* (15-17 April 2003, Budapest, Hungary), 3-10

[21] J. Nerbonne, P. Kleiweg (2003), ‘Lexical Distance in LAMSAS’, in J. Nerbonne, W. Kretzschmar, eds. (2003), *Computers and the Humanities (Special Issue on Computational Methods in Dialectometry)*, 37(3), 339-357.

[22] J. Nerbonne, P. Kleiweg P (2007), ‘Toward a Dialectological Yardstick’, *Journal of Quantitative Linguistics*, 14(2), 148-167.

[23] J. Prokić (2007), ‘Identifying Linguistic Structure in a Quantitative Analysis of Dialect Pronunciation’, *Proceedings of the ACL 2007 Student Research Workshop* (June 2007, Prague), 61–66.

[24] J. Séguy (1971), ‘La relation entre la distance spatiale et la distance lexicale’, *Revue de Linguistique Romane*, 35, 335–357.

[25] M.R. Spruit, W. Heeringa, J. Nerbonne (in press), ‘Associations among Linguistic Levels’, *Lingua (Special Issue on Syntactic Databases)*, available at http://marco.info/pro/pub/shn2007dh.pdf.

[1] According to the main scholar of Tuscan dialectology (Giannelli 2000), Tuscan dialects are neither northern nor southern dialects: this follows from their status as the source of Italian as well as from their representing a compromise between northern and central-southern dialects. Their linguistic characterisation is not so easy, since there appear to be very few features – if any at all – which are common to all and only Tuscan dialects. If elements of unity are hard to find, those of differentiation are present at the different levels of linguistic description.

[2] The RUG/L04 package can be downloaded from http://www.let.rug.nl/kleiweg/L04/; the Visual DialectoMetry (VDM) is a freely available software package documented at http://ald.sbg.ac.at/dm/Engl/default.htm.

[3] http://serverdbt.ilc.cnr.it/altweb/

[4] The ALT transcription system is a geographically specialised version of the *Carta dei Dialetti Italiani* (CDI) (Grassi *et al.* 1997). In what follows, for the reader's convenience phonetically transcribed data are reported in IPA notation.

[5] For more details on the representation scheme adopted in ALT-Web, see Cucurullo *et al.* 2006.

[6] Used software: RUG/L04.

[7] In principle, they could be due to the fact that interviewers did not ask the corresponding question or did not get a useful reply from informants.

[8] Two experiments were conducted, operating respectively on atomic and feature-based representations. Feature-based representations of phonetic variants were automatically generated with a software module in the RUG/L04 package on the basis of a system of 18 features, identified starting from the ALT phonetic transcription system. The adopted feature-based representation distinguishes vowel-specific features (i.e. height, advancement, length and roundedness) as well as consonantal features covering place of articulation (e.g. bilabial, dental, alveolar, velar, etc.), manner of articulation (e.g. stop, lateral, fricative, lateral, etc.) and presence/absence of voice; other features are concerned with prosodic properties such as stress and the vowel/consonant distinction. For more details, see Montemagni 2007.

[9] The distances resulting from the two experiments were compared with the Pearson's correlation coefficient which turned out to be r=0.99. This shows that when working with large data sets feature- and phone-based representations do not lead to significantly different results: if on the one hand feature-based representations do not lead to much improved analyses, on the other hand the rough measure working on phone-based representation appears to be reliable.

[10] A potential problem of this approach is to treat as lexically related accidentally close variants. However, the occurrence of cases like this one within the set of answers to the same questionnaire item is extremely rare.

[11] In principle, a viable alternative could have been resorting to lemmatisation: as Nerbonne and Kleiweg (2003) point out, the application of LD for measuring lexical distance provides "only a rough estimate of what more correctly lemmatizing ought to to". In practice, we believe that in the case of ALT data lemmatisation is not an easy solution at all, especially for what concerns derivationally related words: the question is if and when word forms such as *schiaccina* or *schiaccetta* should be lemmatised as instances of the base lemma *schiaccia* or if they represent lemmata in their own right. Lemmatisation criteria for dialectal data of this type are not easy to find and involve careful examination of the geographic distribution of words as well as of paradigmatic relations holding within the lexicon of a given locality. Therefore, recourse to LD in the ALT case should not be seen as a second best but rather as a way to overcome inherent lemmatisation problems which are not easily solvable.

[12] The geographical distances have been calculated using the "ll2dst" programme included in the RuG/L04 software package.

[13] Following Kondrak (2002) and Prokic (2007), extraction of regular sound correspondences from aligned word pairs was carried out. We focused on the aligned phonetic variants of 519 normalised forms selected on the basis of extra-linguistic criteria, namely geographical coverage and variation range. The experimental data set includes 5,218 phonetic variants types corresponding to 89,715 tokens. Attested phonetic variants were aligned using RUG/L04: alignments were induced by enforcing the syllabicity constraint. In the case of multiple alignments, only the first one was considered. From all aligned word pairs both matching and non-matching phonetic segments were extracted for a total of 25,132,756 segment pairs. A coarse-grained classification of non-matching phonetic segments (4,877,928) shows that consonants play a major role in Tuscan phonetic variation, covering the 70% of non-matching phonetic segments. A finer-grained classification of non-matching phonetic segments involving consonants demonstrates that a significant part of them (i.e. 42%) corresponds to spirantization phenomena, partitioned as follows: 35% spirantization of plosives (/k t p/ > [h ϕ θ]) and 7% weakening of palatal affricates (e.g. /ʃt/ > [ʃ]). These percentages grow further if we focus on Tuscan dialects only. We also measured the

correlation between overall phonetic distances and phonetic distances focussing on non-matching phonetic segments involving spirantization of plosives which turned out to be rather high, with r=0.61.