In *Proceedings of the International Workshop on Computational Phonology* held in conjunction of RANLP 2007, 26 September 2007, Borovets, Bulgaria, pp. 49-60

Patterns of phonetic variation in Tuscany: using dialectometric techniques on multi-level representations of dialectal data

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 first experiments carried out on the corpus of dialectal data of an online dialectal resource documenting the language varieties spoken in an Italian region, Tuscany, with the RUG/L04 software for Dialectometrics and Cartography. By exploiting a multi-level representation model of dialectal data, the study focuses on patterns of phonetic variation attested in Tuscany and tries to shed light on the determinants of pronunciation variation and on the degree to which pronunciation and lexical variation correlate in Tuscan dialects.

Keywords

Dialectometry, phonetic variation, lexical variation, representation of dialectal data

1. Introduction

Although the birth of dialectometry dates back to Séguy (1973) and Goebl (1984), efforts in the last twelve years have led to significant progress in the application of mathematical and computational techniques to the analysis of linguistic variation of different languages, with respect to different sorts of data (going from pronunciational to morphological, lexical and – most recently – also syntactical), different data sets and languages. The state of the art is well documented in the Special issue of *Computers and the Humanities* on Computational Methods in Dialectometry (Nerbonne and Kretzschmar 2003), and – for what concerns more recent developments – in the Special Issue on Progress in Dialectometry of *Literary and Linguistic Computing* (Nerbonne and Kretzschmar 2006). This paper intends to contribute to this line of research by presenting the results of first dialectometric experiments carried out on a corpus of Italian dialectal data for what concerns phonetic variation. For these experiments, we used the corpus of dialectal data of ALT-Web, an online resource giving access to the *Atlante Lessicale Toscano* 'Lexical Atlas of Tuscany', and the RUG/L04 software for Dialectometrics and Cartography developed by P. Kleiweg.

Studies on linguistic variation carried out with dialectometric techniques require three basic ingredients: i) a computable representation of dialectal data; ii) the definition of a suitable distance function measuring how close any two such representations are; iii) a way to turn distance relations into similarity-based partitions and/or relations. In the article, these ingredients are illustrated in detail. After providing a brief overview of the dialects spoken in Tuscany (section2), we describe the background dialectal resource which has been used for this study, with particular emphasis on the representation model adopted for dialectal data (section 3). In section 4, we describe how we measured the pronunciation distances between the language varieties attested in Tuscany. The resulting distance measures were analysed by means of explorative statistical techniques looking for patterns of phonetic variation in Tuscany: achieved results are illustrated in section 5, where more linguistic explanations are also looked for, in particular for what concerns the determinants of pronunciation variation and the interplay between pronunciation and lexical variation.

2. The dialects of Tuscany

Tuscany is a region which has a special status in the complex puzzle of Italian dialects. According to the main scholars of Tuscan dialectology (Giacomelli 1975, 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. However, 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. Giannelli (2000) proposes the following subdivision of Tuscan dialects (represented in Figure 1), based on phonetic, phonemic and morpho-syntactic features:

1. dialects from the Florentine area (Fiorentino);
2. dialects from Siena area (Senese);
3. dialects from Pisa and Livorno areas (pisano-livornese);
4. dialects from Lucca area (Lucchese);
5. dialects from Elba island (Elbano);
6. dialects from Arezzo area (Aretino);
7. dialects from Mount Amiata (Amiatino);
8. dialects from Garfagnana and Versilia (subdvided into Upper-Garfagnino and Low-Garfagnino/Upper-Versiliese);
9. dialects from Massa (Massese).

Tuscany is a region where also non-Tuscan dialects are spoken: this is the case of dialects spoken in Lunigiana and in small areas of the Apennines (Romagna Toscana), which are strongly influenced by northern dialects. Besides dialects, Giannelli (2000) also identifies different transition zones, wich are marked in grey in Figure 1.

Figure 1. The dialects spoken in Tuscany according to Giannelli (2000)

3. The background resource: ALT-Web

3.1 The Atlante Lessicale Toscano

The *Atlante Lessicale Toscano* (henceforth ALT) is a specially designed linguistic atlas in which dialectal data have both a diatopic and diastratic characterization. The adjectives qualifying this linguistic atlas in its name are "lexical" and "Tuscan". ALT is lexical in the sense that its main focus is on lexical variation but this does not exclude that it contains valuable information for what concerns e.g. phonetic or morphological variation. ALT is Tuscan in the sense that it is a regional atlas focusing on dialectal variation within Tuscany, a region which we have seen to have a peculiar status among the 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, socio-economic status to education and culture by a group of trained fiedworkers who employed a questionnaire of 745 target items, designed to elicit variation mainly in vocabulary, semantics and pronunciation. In particular, informants were asked two basic types of questions:

- "onomasiological questions" starting from concepts and asking for the lexical items designating them (e.g. "What terms do you use to name 'bread crumbs'?");
- "semasiological questions" starting from terms and asking what they mean or what concepts they refer to (e.g. "What is the meaning of the term *ceppo*?" whose possible answers in Tuscany include 'tree stump', 'log', as well as 'Christmas present'). Note that in ALT the outcome of semasiological questions also includes the recording of the pronunciation of the target word.

ALT data were collected between 1974 and 1986, resulting in millions of responses from the 2,193 speakers who were each asked 745 questions, corresponding to more than 84,000 different attested dialectal items. During the collection phase, the results of interviews carried out by the group of trained fielworkers were revised by the director of the project, Gabriella Giacomelli, in order to guarantee comparability of collected data and reduce as much as possible potentially misleading effects deriving from fieldworker's collection techniques or transcription peculiarities.

In 1985, the digitization of the huge corpus of dialectal data collected through fieldwork started. The entire ALT corpus was compacted into about 380,000 entries partitioned in about 350,000 entries containing canonical responses to the questionnaire items attested in different locations (including typical contexts of use and informants' comments), and about 30,000 entries recording dialectal items collected as additional material which emerged in the course of interviews.

ALT was published in year 2000 (Giacomelli *et al.* 2000) as a CD-Rom where dialectal data can be retrieved through complex queries taking into account a wide range of parameters interactively defined by the user. With the advent of Internet, the CD-Rom version of the Lexical Atlas of Tuscany was replaced by ALT-Web (Cucurullo *et al.* 2006), an on-line dialectal resource which gives access to the entire corpus of linguistic data gathered for the *Atlante Lessicale Toscano* to a widened target audience ranging from professionals to citizens interested by Tuscan dialectology related topics.[1]

3.2 ALT-Web data representation model

In ALT all dialectal responses, be they individual lexical items or short ethnotexts, were phonetically transcribed. The phonetic alphabet used in the ALT project was a geographically specialized version of the *Carta dei Dialetti Italiani* (CDI) transcription system (Grassi *et al*. 1997). In

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

what follows, for the reader's convenience phonetically transcribed dialectal items are reported in IPA notation.

The encoding of phonetically transcribed data is one of the major problems that has to be faced in the construction of computational dialectal resources based on oral interviews. Solutions may differ, depending on the types of analyses phonetically transcribed data should be subjected to. On the one hand, there is the need to ensure a proper treatment of phonetically transcribed data during different automatic analysis stages including editing, sorting, retrieval, on-screen display and printing. On the other hand, there are the specific problems of retrieving phonetically transcribed data: in spite of the fact that, in principle, computers facilitate access to data, narrowness of phonetic transcription may constitute a major difficulty for what concerns their recovery. To fulfill the specific requirements of the different processing tasks, a complex and articulated encoding schema was designed in ALT-Web.

Table 1. The multi-level representation model of ALT-Web data: an example

Phonetic representation	Orthographically transcribed form	
	Basic orthographic transcription	Normalized representation
[skja'ttʃɛda]	schiaccéda	schiacciata
[skja'ttʃɛta]	schiaccèta	
[skja'ttʃaða]	schiacciàda	
[skja'ttʃada]	schiacciàda	
[skja'ttʃaha]	schiacciàha	
[skja'ttʃaθa]	schiacciàta	
[skja'ttʃata]	schiacciàta	
[skja'tʃata]	schiaciàta	
[stʃa'sseda]	s-ciasséda	
[stja'ttʃada]	stiacciàda	
[stja'ttʃaha]	stiacciàha	
[stja'ttʃaθa]	stiacciàta	
[stja'ttʃata]	stiacciàta	
[ʃkja'ttʃata]	schiacciàta	
[ʃtja'ttʃata]	stiacciàta	
[ʃtja'ttʃaθa]	stiacciàta	
…	…	

In the ALT-Web data bank, all dialectal responses are assigned different levels of representation: a first level rendering the original phonetic transcription; other levels containing normalized representations of the original form encoded in standard Italian orthography. In this multi-level representation model, dialectal data are encoded in layers of progressively decreasing detail going from phonetic transcription to different levels of normalized representations abstracting away from details of speakers' pronunciation. This is exemplified in Table 1 where for each phonetically transcribed form the corresponding orthographic representations are reported: this excerpt is constituted by different pronunciation variants of the same word *schiacciata* denoting a traditional type of bread, flat and crispy, seasoned on top with salt and oil.

At the first level (column 1), there is the phonetic representation of dialectal items as transcribed by fieldworkers. In ALT-Web phonetically transcribed data are represented through an ad hoc hybrid encoding schema including both compositional and atomic representations which, depending on the task, are automatically converted into each other (for more details see Cucurullo *et al.* 2006, § 3.2). For the specific concerns of this study, we used atomic representations showing a 1:1 correspondence between ALT phonetic symbols and computer codes.

Each phonetically transcribed dialectal item is then assigned two different types of orthographically transcribed forms, respectively referred to as "basic orthographic transcription" and "normalized representation" (reported in the second and third columns of Table 1). At the first orthographic representation level, attested dialectal data are encoded according to standard Italian orthography: this level of representation is designed to help the non-expert user to understand the phonetically transcribed form. From this it follows that this level of representation seeks to account for the variety of phonetic realizations attested by informants. Yet, Italian orthographical conventions imposed some unavoidable neutralizations, due to the unavailability of the corresponding graphemes. For instance, in Table 1 it can be noticed that [skja'ttʃada] and [skja'ttʃaða] are both assigned the same word *schiacciada* as the corresponding orthographically transcribed form.[2]

Normalized representation of dialectal items represents the first representation level abstracting away from pronunciation details: in particular, it is meant to abstract away from within-Tuscany vital phonetic variation. This entails that at this level a wider range of variants, if compared with the previous level, is assigned the same normalised form: this is clearly represented in Table 1 where it can be noticed that all different phonetically transcribed dialectal items are 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 voicing or

[2] To check how close basic orthographic transcription was with respect to the originally attested forms, a normalization factor was calculated as the ratio between the number of different phonetically transcribed forms and the number of different orthographically transcribed forms: the result is 1.13, showing that neutralized representations are resorted to in a quite reduced number of cases.

spirantization of plosives like /t/, e.g. [skja'ttʃaθa] and [skja'ttʃada]. On the contrary, there are word forms like ['kaʎʎo] and ['gaʎʎo] (meaning 'rennet') which are assigned two different normalised representations, *caglio* and *gaglio* respectively: the reason for this lies in the fact that this alternation represents a no longer productive phonetic process in Tuscany. It should also be noted that this representation level does not deal with morphological variation (neither inflectional nor derivational): this entails that words such as [skja'ttʃata] (singular) and [skja'ttʃate] (plural) as well as [skjattʃa'tina] (diminutive) are all assigned different normalized forms. Currently, this represents the most abstract representation level neutralizing productive phonetic variation phenomena; more abstract normalization levels (e.g. lemmatization) are envisaged for future developments.

Even if this articulated representation model was originally devised with a view to editing, sorting and especially retrieval problems, it seems to us to be particularly suitable and flexible also for dialectometric analyses of dialectal data at different linguistic description levels. For the specific concerns of this study on linguistic variation in Tuscany, we focussed on the levels of the phonetic transcription and normalised representation. In particular, this articulated representation scheme was exploited in different ways. First, the alignment of the representation levels was used to automatically extract all phonetic realizations attested in Tuscany for the same abstract normalized word form. Second, patterns of phonetic and lexical variation could be studied with respect to different representation levels of the same dialectal data.

4. Measuring linguistic distance

Starting from the data of a linguistic atlas there is a number of different ways for measuring the linguistic distance between any two locations belonging to the atlas geographic network. Following Seguy (1971), who is recognised to be the founder of dialectometry, one could count the overlapping features (typically, but not limited to, lexical items provided as answers to a questionnaire) between the data collected in any two sites. Individual differences between two locations can be aggregated over a large amount of material thus resulting in a global and reliable measure of linguistic distance. Goebl (1984) further developed and improved these ideas (to which he arrived independently of Seguy) demonstrating their potential to the wider community of dialectologists. In both cases the measure of linguistic distance between any two locations was based on categorical distinctions. By comparing two different data sets, only two different distance types are recognised: the distance is set to 0 when compared items coincide, and to 1 when the they do not.

The categorical treatment of dialectal data advocated in these dialectometric approaches to the study of linguistic variation was soon felt as a major limitation, especially for what concerns pronunciation. In fact, the proposed similarity measure is not sensitive to partial overlap between dialectal data. An interesting solution to this problem was proposed by Kessler (1995) who resorted to the use of a string-distance measure, the Levenshtein distance (henceforth referred to as LD), as a means of calculating the distance between the pronunciations of corresponding words in different dialects (his study was based on Irish Gaelic dialects). The basic idea underlying LD is to imagine that one is rewriting one string into another. The rewriting is carried out through basic operations: the deletion of a string character; the insertion of a string character; the substitution of one character for another. To each of these operations is associated a cost. The LD between two strings is the least costly sum of costs needed to trasform one string into another (for more details on the LD algorithm for dialectological studies see Nerbonne *et al*. 1999a).

With LD, comparing two dialectal varieties results in a sum 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*. (1999a) and Heeringa (2004) who worked on different languages and with different representation types. 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*. 1999a); b) the treatment of multiple responses provided as dialectally appropriate either by the same informant or by different informants belonging to the same community (Nerbonne and Kleiweg 2003).

In what follows, we will focus on issues specific to the measure of phonetic distances with the ALT data.

4.1 Measuring phonetic distances

Using LD, the distance between two linguistic varieties A and B is computed by comparing the pronunciation of words in A with the pronunciation of the corresponding words in B. The pronunciation 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 first one, LD operates on sequences of phonetic symbols, whereas in the second one comparison is carried out with respect to feature-based representations. The main drawback of so-called phone string comparison is that the underlying notion of phonetic distance is binary: non-identical phones contribute to phonetic distance, identical ones do not. According to this approach minor phonetic differences, such as that holding between /t/ and its fricative realization /θ/, count the same as major

differences, such as the one between a vowel and a consonant. Higher accuracy in the measure of phonetic distances can in principle be achieved by applying LD to representations in which each phonetic symbol is described in terms of a bundle of features: in this way major phonetic distinctions can be assigned greater distance than minor ones, i.e. /t/ and /θ/ are closer than /a/ and /t/. Obviously, in this case the selection of features to be used for the representation of basic sounds is a crucial issue.

Both approaches were experimented with for this study of phonetic variation in Tuscany. The data set used in our experiments was built as follows: the different phonetic realisations of the same lexical unit were identified by selecting all phonetically transcribed dialectal items associated with the same normalised form. Since the ALT-Web normalised representation level does not abstract away from morphological variation nor from no longer productive phonetic processes, we can be quite sure that phonetic distances calculated on these data testify vital phonetic processes only, without interference from any other linguistic description level (e.g. morphology). In fact, we have seen that different inflectional variants of the same lemma give rise to different normalised forms: *schiacciata* (singular) and *schiacciate* (plural) are different normalised forms whose pronunciation variants are considered separately; the same holds for derivationally related words such as *schiacciatina* or *schiacciatello* each of which is considered separately from its base form *schiacciata*.

As a basis for this study, the whole set of 33,094 normalised forms attested as answers to either onomasiological or semasiological questions was taken into account. Of the whole set of attested normalised forms, 23,637 show no phonetic variation at all, and for another 618 attested variation occurs within a single locality. Since both cases are of no value in assessing phonetic variation across Tuscany, they have been eliminated from the data set which served as the basis of these dialectometric analyses. There remained 8,839 normalised forms having at least two different phonetic realisations and being attested in at least two different locations. The graph in Figure 2 shows the geographical coverage and the phonetic variability range for the selected 8,839 normalised forms. Geographical coverage ranges between 2 (with normalised forms being attested in only one location being excluded) and 224: it should be noted, however, that only 980 normalised forms (i.e. 11%) are attested in at least 10 different locations. Phonetic variability ranges between 2 and 34: the number of normalised forms for which the range of phonetic variability is greater than 5 is however quite low, corresponding to 13.78% of the cases (namely, 1218).

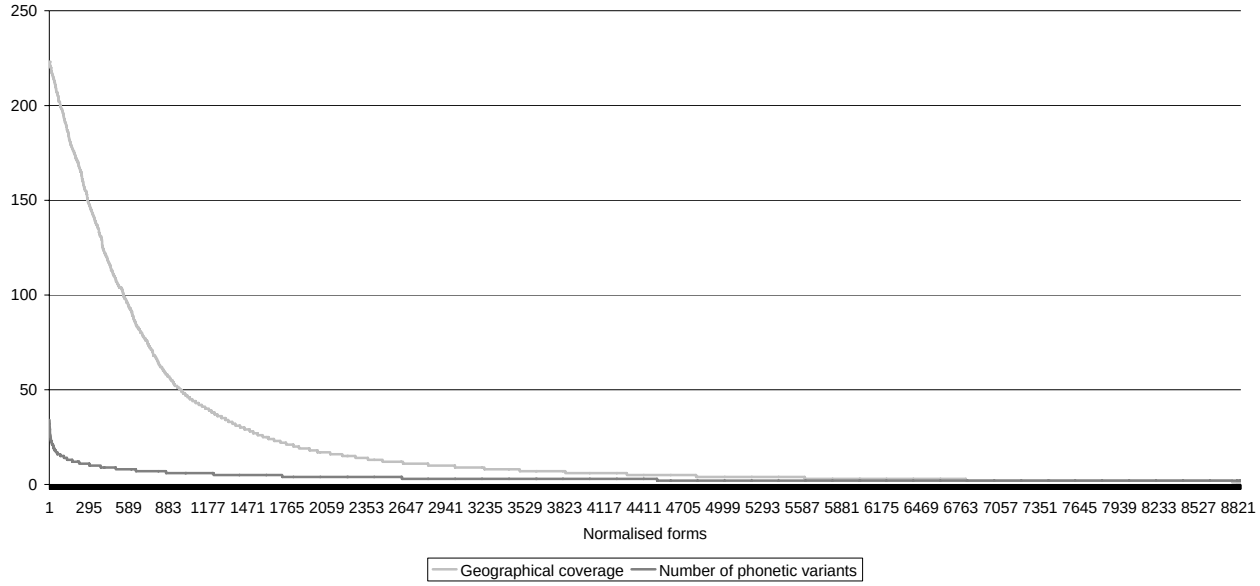

Figure 2. ALT normalised forms: geographical coverage and phonetic variability range

Two different experiments were carried out on the selected data set, operating respectively on atomic and feature-based representations of phonetically trascribed data. Feature-based representation of phonetic variants were automatically generated with a software module included in the RUG/L04 package on the basis of a system of 18 features, identified starting from the phonetic transcription system adopted by ALT. 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. Such a feature system was used to encode 86 different phonetic tokens which were attested in the whole selected data set.

The resulting phonetic distance matrices were built on the basis of 195,124 different phonetic variants attested in Tuscany for the selected normalised dialectal items. In order to assess the reliability of the data set, we calculated the coefficient Cronbach α (for details see Heeringa 2004, pp. 170-173) which was 0.99 in both experiments. This means that this data set provides a reliable basis for an analysis of pronunciation differences based on LD. We also compared the distances resulting from the two experiments with a correlation coefficient: for this specific aim, we used the Pearson's correlation coefficient which turned out to be r=0.99. On the basis of this, the distances identified on the basis of phone-based and feature-based representations appear to be very close.

5. Turning distance relations into similarity-based partitions and relations

Using the distance matrices generated during the previous step, we can now try to characterise the relation among the attested language varieties from the pronunciation point of view. Following Heeringa and Nerbonne (2001), the distance matrices were explored with two different but complementary techniques: hierarchical clustering, aimed at classifying dialects into relatively close groups, and multidimensional scaling (MDS), to go beyond identified dialect areas and to explore the transition modality from one dialectal variety to another.

5.1 Patterns of phonetic variation

The results of clustering applied to the distance matrices obtained in the phone-based and feature-based experiments (see § 4.1) coincide. In what follows we report the results obtained with phone-based representations.

The seven most significant groups which emerged from clustering are reported in Figure 3, where it can be noticed that identified phonetic areas are arranged in an onion-like shape built around a central area covering the province of Florence and propagating in different directions, towards south (in the province of Siena) and west (covering the provinces of Pistoia, Lucca up to some areas of Pisa and Livorno). Around this central area, there is a transition layer separating the core from an external layer within which non-Tuscan dialects (Lunigiana and Romagna Toscana) as well as the east side of the province of Arezzo (Chiana Valley and Upper Tiber Valley) can be clearly detected.

In order to test the salience of identified borders, a "cluster composite map" (Kleiweg *et al.* 2004) was generated (Figure 4), providing a more articulated picture than simple clustering divisions. In this type of map, the salience degree of borders is represented in terms of increasing darkness, i.e. the most salient borders are represented as dark lines, whereas the reverse holds for less sharp dialectal distinctions. It came out that the most significant border is the one which separates the central area (also including the transition area) from the outer most layer; there follows the borders, ordered by relevance, identifying respectively the non-Tuscan dialects, the east-side of the province of Arezzo and the border separating the core from the transition area.

To go beyond identified linguistic borders, we explored the phonetic distance matrix through MDS techniques. We generated a map[3] which was obtained by "translating" the MDS coordinates identifying each dialectal variety into different mixtures of colors (Nerbonne *et al.* 1999b, Heeringa and Nerbonne 2001) reflecting the gradual changes from one dialect to another. In this type of map, the extent to which colors contrast reflects the extent to which dialects differ: the map highlights three areas only corresponding respectively to linguistic varieties spoken in Lunigiana and in Romagna Toscana and to a wider area covering all Tuscan dialects. If on the one hand non-Tuscan dialects strongly constrast with respect to other areas, on the other hand Tuscan dialects appear to present themselves as a rather uniform area. It is interesting to point out that, according to Giannelli (2000), whereas the areas characterised by strong contrasts of colors (i.e. Lunigiana and Romagna Toscana) differ at the level of their phonemic inventories, underlying the rather uniform area of Tuscan dialects there is the same phonemic inventory and identified variation patterns appear to originate from phonetic differences only.

5.2 Behind patterns of phonetic variation

In the previous section we illustrated the first results on phonetic variation within Tuscany which were achieved on the basis of an aggregate analysis carried out on a consistent set of available data. The used data set resulted from a selection process which was guided from extra-liguistic criteria (minimal geographic coverage and minimal range of variability equal to 2). This is to say that the selection of data was not guided by any predefined linguistic assumption, i.e. they were not selected as representative of arbitrarily selected linguistic features which we knew in advance to play a role in the definition of patterns of phonetic variation in Tuscany. This is in line

[3] The MDS map can be found at the following address http://webilc.ilc.cnr.it/~montemagni/mds_fon_all.pdf

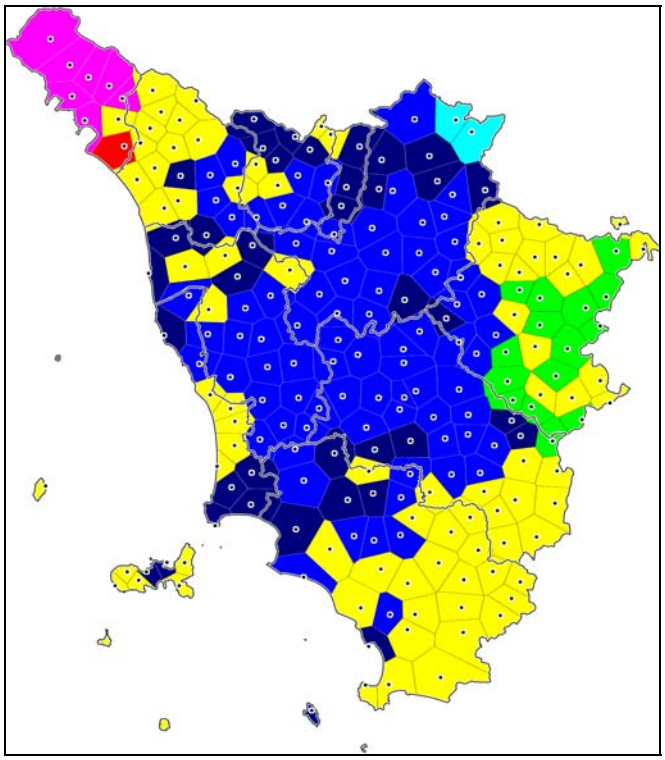

Figure 3. The seven most significant phonetic areas identified through clustering. Around a central area covering the province of Florence and propagating towards south, west and east there is a transition layer separating the core from an external layer within which non-Tuscan dialects (Lunigiana and Romagna Toscana) as well as the east side of the province of Arezzo are identified

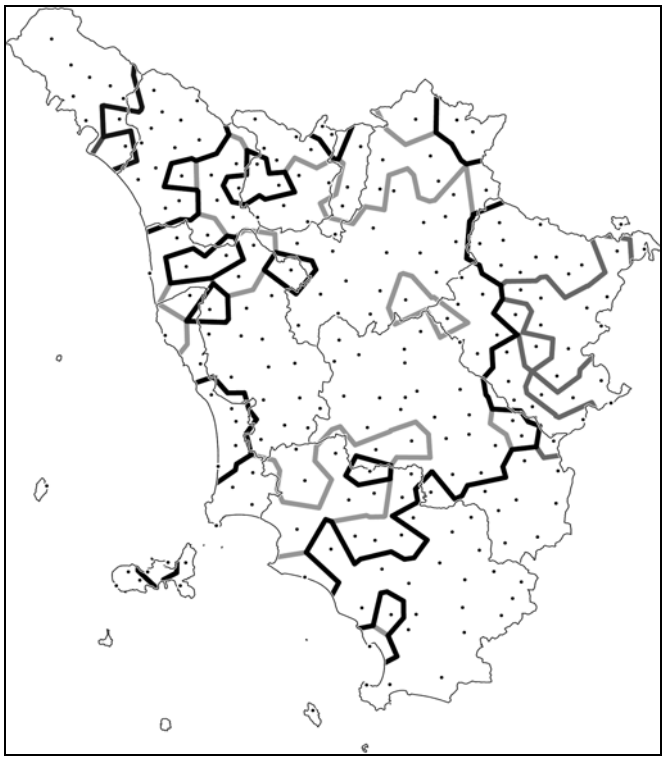

Figure 4. This map shows the salience degree of identified borders in terms of increasing darkness of lines. The most significant border appears to be the one separating the central area (also including the transition area) from the outer most layer. There follow the borders, ordered by relevance, identifying non-Tuscan dialects, the east-side of the province of Arezzo and the border separating the core from the transition area.

with the general approach of dialectometric studies, where the aggregation of linguistic differences in a given dialectal variety is taken to provide the most reliable basis for characterising its relations to the other varieties. As pointed out by Nerbonne (2005, p. 4), "in measuring differences, the dialectometrist deliberately abstracts away from the details of what has contributed the difference, in an abstraction step that is inherent to the strength of the approach, but which at the same time loses the connection to the linguistic characterization". As a matter of facts, identified dialect areas as well as dialect continua provide general characterisations of linguistic variation without accounting for the linguistic features which contributed to it. In this section we report the results of first experiments carried out along the lines suggested in Nerbonne (2005, manuscript) to relate general dialectometric characterisations with partial but linguistically-oriented ones. Our purpose here is to attempt to identify some of the linguistic features contributing to the overall chacterisation of the Tuscan dialectal landscape depicted in § 5.1. To this end, we focussed on patterns of vocalic and consonantal variation to see whether and to what extent they contributed to the overall picture.

In order to establish the role of vowels and consonants in determining dialectal differences in Tuscany, patterns of vocalic and consonantal variation were first identified and then compared with the results of the analysis of the entire set of phonetically transcribed data. The patterns of vocalic and consonantal variation were obtained by carrying out the types of dialectometric analyses described in the previous sections on restricted data sets, consisting respectively of vowels and consonants extracted from the original data set. In this case we used a measure of phonetic distance based on phone-based representations.

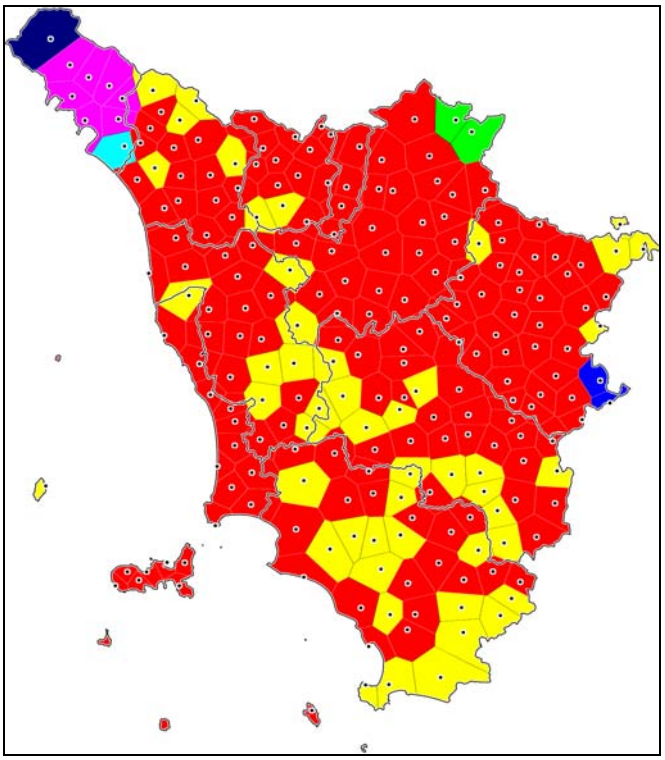

Figure 5. The seven most significant phonetic areas identified through clustering based on vocalic variation patterns. Non-Tuscan dialects are partitioned in four different areas, whereas Tuscan dialects are partitioned into three clusters which do not appear to form geograpically well-defined dialectal areas

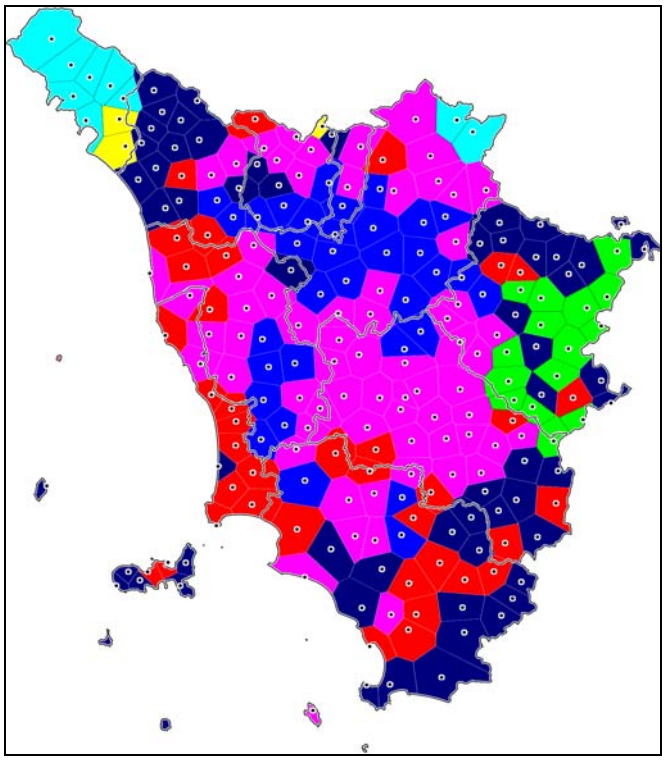

Figure 6. The seven most significant phonetic areas identified through clustering based on consonantal variation patterns. Areas are arranged in an onion-like shape built around a central area (province of Florence), with a transition layer expanding mainly towards south and west and separating the core from an external layer including non-Tuscan dialects and part of the province of Arezzo

Let us first consider the distance matrices obtained through these linguistically-oriented analyses. It is interesting to note that the distances between dialectal varieties obtained using vowels only correlate closely with the distances assigned through LD using the original corpus of phonetic transcriptions (r=0.9483 with p=0.0001). The same comparison carried out between the distances obtained using consonants only and the entire corpus shows a slightly lower but still significant correlation, with r=0.8837 (and p=0.0001). From this we can conclude that both vowel and consonant pronunciation play a major role in accounting for phonetic variation within Tuscany: the former account for 89.92% of the variance in pronunciation, whereas the latter for 78.09%. The correlation does not appear so strong when we compare distance measures based on vowels and consonants respectively: in this case, the correlation value is much lower, with r=0.6912 (p=0.0001).

Distance matrices were first explored through clustering. The results are reported in Figure 5 where the map on the left shows dialectal areas identified on the basis of vowels, and the right one those identified on the basis of consonants. The two maps show major differences. For instance, in the vowel-based map the non-Tuscan dialects spoken in Lunigiana and Romagna Toscana represent different dialectal areas, whereas this is not the case with consonantal variation: in the right map, Lunigiana and Romagna Toscana are clustered together. As to the rest of the region, vocalic patterns of variation do not appear to form geograpically well-defined dialectal areas; consider for instance the cluster gathering locations most part of which are scattered around the west side of the region.

Concerning consonantal variation, besides the non-Tuscan cluster of dialects there are other dialectal areas which can be clearly identified (see Figure 6). This is the case of the east-side of the province of Arezzo, an area which was already present more or less with the same extension in the general map (Figure 3). Interestingly enough, the shape of dialectal areas identified here is quite similar to that observed in the general map, where phonetic areas are arranged in an onion-like shape built around a central area, which here corresponds to a more restricted

area mainly concentrated in the province of Florence. Around this central area, there is a transition layer expanding mainly towards south and west and separating the core from an external layer including non-Tuscan dialects and the east side of the province of Arezzo.

With these experiments, we tried to identify the determinants of phonetic variation in Tuscany: if on the one hand the two areas of non-Tuscan dialects (Lunigiana and Romagna Toscana) appear to originate from patterns of vocalic variation, on the other hand it seems that consonantal variation underlies the general onion-like shape of identified phonetic areas as well as the identification of the area corresponding to the east-side of the province of Arezzo. It would be interesting to proceed in these types of analyses by progressively restricting the number of parameters taken into account up to specific features of Tuscan dialects, e.g. the spirantization of plosives (so-called "gorgia toscana") for what concerns consonantal variation.

5.3 Comparing patterns of phonetic and lexical variation

In the previous section we made first attempts to discover the linguistic properties playing a major role in determining identified patterns of phonetic variation. It would also be interesting to go beyond identified patterns and check whether and to what extent observed pronunciation variation correlates with linguistic variation observed with respect to other levels of linguistic description. In this section, we will focus on the correlation between pronunciation and lexical variation.

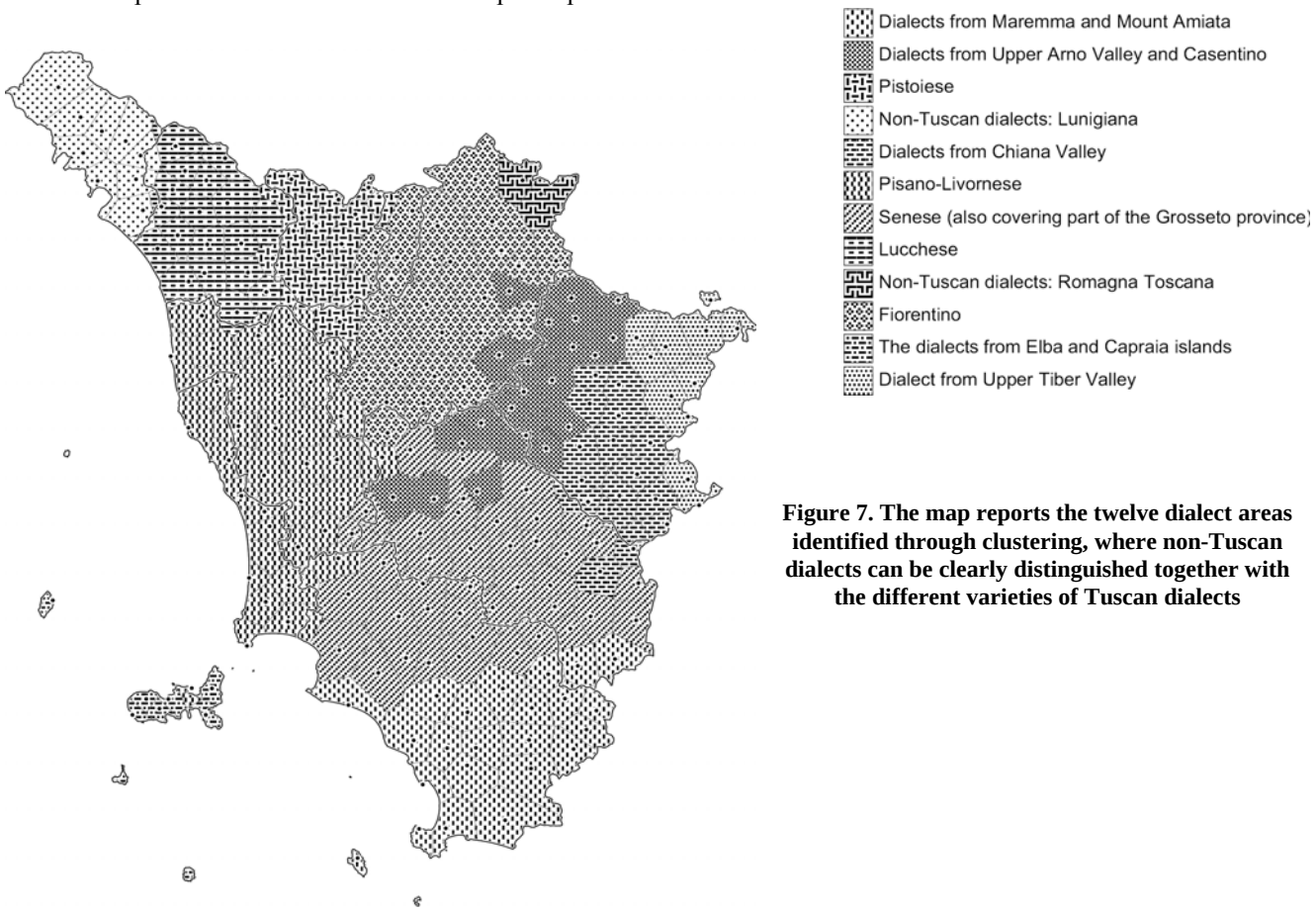

Figure 7. The map reports the twelve dialect areas identified through clustering, where non-Tuscan dialects can be clearly distinguished together with the different varieties of Tuscan dialects

5.3.1 Lexical variation in Tuscany

Whereas a study of phonetic variation based on phonetically transcribed data could only be carried out with LD, this choice is not to be taken for granted in the case of lexical distances. In fact, the pioneering research by Seguy and Goebl mainly focussed on lexical variation, i.e. on whether and to what extent words denoting the same concept vary geographically. Basically, in these studies the comparison between any two sites is carried out 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 different inflectional or derivational variants of the same

lemma. Moreover, they can also include diacronically (e.g. ethimologically) related words. By adopting a binary notion of lexical distance, related lexical items are treated as independent and totally unrelated answers.To overcome this problem, Nerbonne and Kleiweg (2003) in their study of lexical variation in LAMSAS applied LD to measure also the lexical distance of the answers on the basis of the encouraging results previously obtained in the study of dialectal pronunciation. With LD, related lexical items are no longer treated as different and irrelated answers and their partial similarity is accounted for. 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, and this is even more unlikely to occur in linguistically close dialectal varieties such as the Tuscan dialects.

We felt that the use of LD for measuring lexical distances was appropriate also 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 layer does not abstract away from morphological variation nor from no longer productive phonetic processes. To keep with the *schiacciata* example (§ 3.2), the questionnaire item meant to gather all attested lexicalizations of the concept of 'traditional type of bread, flat and crispy, seasoned on top with salt and oil' 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*, *schiaccetta* e *schiacciata* or multi-word expressions like *schiacciata unta* or *schiacciata al sale*. At the normalised representation level, all these forms still represent different answers to the same questionnaire item. By resorting to LD, their partial overlap can be accounted for in the measure of lexical distance.

.

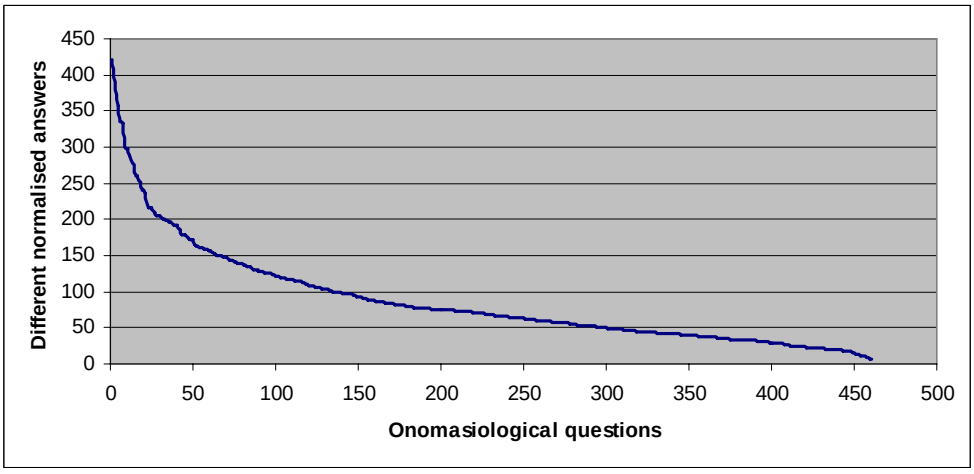

Figure 8. Number of different normalised answers per onomasiological question in ALT data

In principle, a viable alternative could have been resorting to lemmatization: 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 lemmatization 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 lemmatized as instances of the base lemma *schiaccia* or if they represent lemmata in their own right. Lemmatization 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 lemmatization problems which are not easily solvable.

For the study of lexical variation in Tuscany we used the whole set of normalised answers to a subset of onomasiological questions (i.e. those looking for the attested lexicalizations of the same concept). This choice follows from the fact that the number of different normalised answers per question in ALT shows a quite wide range of variation (see Figure 8), going from a minimum of 6 different normalised answers to a maximum of 421. At closer inspection, however, it appears that highly productive questionnaire items include many hapaxes which do not appear to be lexicalised answers. For

instance, the questionnaire item looking for denominations for 'stupid' gathered 372 different normalised answers, 122 of which are hapaxes representing productive figurative usages (e.g. metaphors) like *cetriolo* 'cucumber' and *carciofo* 'artichoke' or originating from productive derivational processes (this is the case of answers like *scemaccio*, *scemalone*, *scemarotto*, *scemarano*, *scemarlotto*, etc.) or multi-word expressions like *mezzo scemo* 'half stupid', *mezzo spostato* 'half maladjusted', *puro locco* 'pure stupid' and the like. In order to to prevent noisy effects deriving from these particularly productive questionnaire items, the data set for this experiment was built by selecting onomasiological questions showing a "middle" range of variation, i.e. only questions whose range of variability was between 6 and 50 were selected as a basis for this study.

The selected subset turned out to include 165 different questionnaire items, for a total of 61,714 different normalised answers, with a rather high Cronbach α (0.94) showing that this was a sufficient basis for a consistent and reliable Levenshtein analysis. It should be noted that, given the peculiar features of the normalised representation level (see above), the resulting measure of lexical distance has to be seen as also reflecting patterns of morphological variation, especially for what concerns derivation.

The obtained lexical distance matrix was explored with the same types of analyses were carried out for the study of phonetic variation, i.e. clustering and MDS. Through clustering we obtained an excellent view into the nature of morpho-lexical variation in Tuscany. Figure 7 reports the twelve identified dialect areas, where non-Tuscan dialects can be clearly distinguished (Lunigiana and Romagna Toscana) together with the different varieties of Tuscan dialects articulated into: the Fiorentino, the Pistoiese, the Lucchese, the Pisano-Livornese, the dialect from Elba island, the Aretino (with its subdivisions), the Senese (also covering part of the Grosseto province) and the dialect from Maremma and Mount Amiata. 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 Giannelli (2000). Identified linguistic varieties include both dialects in their own right as well as transitional varieties like the Pistoiese, the dialects from Chiana Valley or the dialects from Maremma and Mount Amiata.

The cluster composite map built on top of the lexical distance matrix confirms a widely acknowledged fact in Tuscan dialectology, i.e. that the main subdivision is between Northern Tuscan dialects and Southern ones (the latter corresponding to the Senese, Maremmano-Amiatino and Arezzo's dialects). But salience of borders is not the only issue worth being explored; using MDS and by projecting its results on a map in terms of mixtures of colors,[4] it can be noticed that the transition from one area to another is gradual. As in the previous case, non-Tuscan dialects (also including the east part of the Arezzo province) emerge clearly, being characterised by strong constrasts of colors. For what concerns Tuscan dialects, the transition is gradual reflecting a dialect continuum, which however does not appear so uniform as observed in the case of phonetic variation.

5.3.2 Phonetic vs morpho-lexical variation

By comparing the identified phonetic and morpho-lexical patterns of variation there appears to be a discrepancy which is worth being explored: in fact, identified dialectal areas differ significantly in the phonetic and lexical case, and also the underlying "continuum" map appears to show different degrees of contrast between language varieties. The correlation between the phonetic and lexical levels thus represents an interesting topic to be explored to contribute to the study of both Tuscan dialects and – more generally - of the interplay between patterns of language variation at different levels of linguistic description. In this specific case, we did not find a particularly strong correlation between the phonetic and lexical distance matrices, with r=0.7039 (p=0.0001). If on the one hand this is in line with Chambers and Trudgill (1998, p.97) assumption that lexical differences do not necessarily coincide with pronunciation differences "because the former are more subject to self-conscious control or change by speakers than the latter", on the other hand it seems that this situation is not reflected in the analyses of Tuscan dialects by Giacomelli and Giannelli. We believe that dialectometry can help to understand better the interplay between patterns of phonetic and morpho-lexical variation by untangling what in today's studies appears as a complex puzzle.

6. Conclusions

This paper reports the first results of a dialectometric study focussing on pronunciation variation in Tuscany: it is the first time that the whole corpus of ALT data is explored by means of computational techniques. Pronunciation distances among attested language varieties were calculated through LD against phone-based and feature-based representations and the resulting distance matrices were explored by means of statistical tecniques (clustering and MDS) to identify underlying dialectal areas and continua. We also tried to go behind and beyond identified patterns of pronunciation variation. First, we made preliminary attempts to discover the linguistic properties playing a major role in determining identified variation patterns: in particular, dialectometric analyses were carried out against restricted data sets, consisting only of vowels and

[4] The MDS map can be found at the following address http://webilc.ilc.cnr.it/~montemagni/mdslex.pdf

consonants, and achieved results were compared with the general pronunciations patterns emerged from the analysis of the entire set of selected data with promising results: these experiments are worth being pursued by progressively restricting the number of phonetic features taken into account. Another promising line of research is concerned with the correlation between variation patterns emerged with respect to different levels of linguistic description; from the results of our study it appears that pronunciation and morpho-lexical variation do not correlate perfectly. This is an issue which is worth being further explored both from the Tuscan dialectology point of view and from a more general methodological perspective. Last but not least, it should be pointed out that the corpus of ALT dialectal data can be used to study not only patterns of diatopic variation, but also patterns of diastratic variation: in fact, interviews were carried out with more than 2,000 informants selected with respect to different parameters such as age, socio-economic status, education and culture. In this study, we considered all answers provided by all informants, abstracting away from their socio-cultural status. However, such a data set should also be exploited to study patterns of subdialectal variation and their interaction – if any - with diatopic variation.

7. References

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

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

[3] Giacomelli G., 1975, *Aree lessicali toscane*. "La ricerca dialettale", I, pp. 115-152.

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

[5] Giannelli L., 2000, *Toscana*. Pacini Editore, Pisa.

[6] Goebl H., 1984, *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Max Niemeyer, Tübingen.

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

[8] Kessler B., 1995, *Computational Dialectology in Irish Gaelic*. In: Proc. of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Dublin, pp. 60–67.

[9] Kleiweg P., Nerbonne J., Bosveld L., 2004, *Geographic Projection of Cluster Composites*. In: Blackwell A. *et al.* (eds.), Diagrammatic Representation and Inference. Third International Conference, Diagrams 2004. Cambridge, UK, March 2004 Lecture Notes in Artificial Intelligence 2980. Springer, Berlin. pp. 392-394.

[10] Heeringa W., Nerbonne J., 2001, *Dialect Areas and Dialect Continua*. In: "Language Variation and Change", 13, 2001, pp.375-400.

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

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

[13] Nerbonne J., Heeringa W., Kleiweg P., 1999a, *Edit Distance and Dialect Proximity*. In: Sankoff D., Kruskal J. (eds.), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Stanford, CSLI Press.

[14] Nerbonne J., Heeringa W., Kleiweg P., 1999b, *Comparison and classification of dialects*. In: Proceedings of the 9th Meeting of the European Chapter of the Association for Computational Linguistics, pp. 281-282, available at citeseer.ist.psu.edu/512989.html.

[15] Nerbonne J., Heeringa W., 2001, *Computational comparison and classification of dialects*. "Dialectologia et Geolinguistica. Journal of the International Society for Dialectology and Geolinguistics", 9, pp. 69–83.

[16] Nerbonne J., Kleiweg P., 2003, *Lexical Distance in LAMSAS*. In: Nerbonne J., Kretzschmar W. (2003), pp. 339-357.

[17] Nerbonne J., Kretzschmar W. (eds.), 2003, *Computational Methods in Dialectometry.* Special issue of *Computers and the Humanities*, 37(3).

[18] Nerbonne J., Kretzschmar W. (eds.), 2006, *Progress in Dialectometry: Toward Explanation. Literary and Linguistic Computing* 21(4).

[19] Nerbonne J., 2005, *Various Variation Aggregates in the LAMSAS South*, in C. Davis and M. Picone (eds.), *Language Variety in the South III*, Tuscaloosa, Univ. of Alabama Press.

[20] Nerbonne J., manuscript, *Variation in the Aggregate: An Alternative Perspective for Variationist Linguistics*, available at http://www.let.rug.nl/~nerbonne/papers/Nerbonne-Aggregating-2007.pdf

[21] Séguy J., 1971, *La relation entre la distance spatiale et la distance lexicale*. In: "Revue de Linguistique Romane", 35, pp. 335–357.