View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by PUblication MAnagement

CNR Short Term Mobility Program – 2010

Simonetta Montemagni
Istituto di Linguistica Computazionale "Antonio Zampolli" – CNR – Pisa (Italy)

**Report of the joint research activity carried out
from June 13th to July 4th 2010
at the University of Groningen - Faculty of Arts – Center for Language and Cognition Groningen (CLCG) directed by Prof. John Nerbonne**

Title of Research Program
Modelli computazionali della variazione dialettale e fattori linguistici sottostanti
Computational Models of Dialectal Variation and Underlying Linguistic Features

1. Introduction

The transition from the description of the geographic distribution of individual linguistic features to a more abstract level of description intended to make generalizations on diatopic variation is now made possible by the use of dialectometric techniques that have proved particularly promising in the study of language variation in different languages and dialects, also typologically very distant: dialectometric studies have a tradition of over thirty years, since the pioneering studies of Seguy (1971) to the more recent analyses by Goebl (1984, 2005) and Nerbonne (Nerbonne et al. 1999, 2001). The greatest strength of this approach is to disregard the individual data that have contributed to the observed patterns of linguistic variation and the possibility of an "aggregate" analysis of increasingly large amounts of data, such as the entire corpus of a linguistic atlas. However, abstracting from the individual data is in danger of losing "the connection to the linguistic characterization" (Nerbonne, in press), aspect – this one - that makes dialectometric analyses not particularly interesting to the eye of the community of linguists and dialectologists. Michele Loporcaro (2009) effectively describes this view: "dialectometry measures the structural distances without passing through a rationalization of linguistic structure."

One response to such criticism has been advanced recently by Wieling and Nerbonne (2009, 2010), who used a technique of co-clustering (called "bipartite spectral graph partitioning") to identify dialects on the basis of aggregate large corpus of dialect and simultaneously identifying the underlying linguistic basis. In particular, through this technique it is possible to understand which factors underly the identified patterns of dialectal variation, the role played by each of them and the weight associated with them. In this way, the gap between models of linguistic variation based on quantitative analyses and more traditional analyses based on specific linguistic features is significantly reduced. Achieved results for Dutch dialects turned out to be particularly promising.

This report illustrates the application and specialization of the technique of "hierarchical bipartite spectral graph partitioning" (Wieling and Nerbonne, 2010) with respect to the dialectal corpus of the *Atlante Lessicale Toscano* ('Lexical Atlas of Tuscany', henceforth ALT) and discusses achieved results. The analysis focuses on the level of phonetic

variation: this is the level of analysis for which an aggregate analysis of the ALT dialectal corpus has provided divergent results compared to the analyses by Giannelli (1976, 2000) and Pellegrini (1977), as documented in Montemagni (2007, 2008). Phonetic variation in Tuscany thus provides a particularly challenging case study to test the potential of this new analysis technique to study models of linguistic variation.

2. Construction of the experimental data set

2.1. The data source

This study on Tuscan phonetic variation is based on the corpus of dialectal data of the *Atlante Lessicale Toscano* (ALT, Giacomelli et al., 2000). 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 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 is now available as an on-line resource, ALT-Web (http://serverdbt.ilc.cnr.it/altweb/; for more details see Montemagni et al. 2006).

ALT data were collected between 1974 and 1986, resulting in millions of responses (tokens) from the 2,193 speakers who were each asked 745 questions, corresponding to more than 84,000 different attested dialectal items (types). During the collection phase, the results of interviews carried out by the group of trained fielworkers were revised by the head 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 ALT, all dialectal items were phonetically transcribed. 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.

For the specific concerns of this study, we will focus on the representation levels of a) phonetic transcription (henceforth, PT), and b) 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'sseda] 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 [skja'ttʃata] (singular) and [skja'ttʃate] (plural) as well as [skjattʃa'tina] (diminutive) are all assigned different NFs. Currently, NR is the most abstract representation level in ALT-Web.

2.2. Dialectal data selection

For this study of phonetic variation, phonetic transcription was taken as the starting point. The alignment of the different representation levels was exploited 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, as exemplified in Table 1 for the normalised form SCHIACCIATA.

Location	Phonetic variant	NF
15 Vergemoli	[sca'ttʃata]	SCHIACCIATA
16 Pieve Fosciana	[sca'ttʃada]	
18 San Pellegrino in Alpe	[sca'ttʃata], [stʲa'ttʃata]	
19 Brandeglio	[sca'ttʃaθa], [stʲa'ttʃaθa]	
22 Prunetta	[stja'ttʃaθa]	
23 Orsigna	[skja'ttʃaθa], [sca'ttʃaθa], [stja'ttʃata]	
24 Spedaletto	[stja'ttʃaθa]	
25 Castello di Sambuca	[sca'ttʃada]	
28 Barberino di Mugello	[skja'ttʃata], [stja'ttʃata]	
…	…	…

Table 1 – Excerpt from the experimental data set used for this study

Since the ALT-Web normalised representation 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).

In particular, the whole set of 34,912 normalised forms attested in the ALT dialectal corpus was taken into account. For 20,671 normalised forms (59.20%) attested variation (if any) occurs within a single locality; on the other hand, there are 4,688 normalised forms

(13.42%) showing no phonetic variation at all, in spite of their being attested in different locations (with geographical coverage ranging from 2 to 206). Since both cases are of no value in assessing diatopic phonetic variation, they have been removed from the data set which served as the basis of this study. There remained 9,553 normalised forms having at least two different phonetic realisations and being attested in at least two different locations. The graph in Figure 1 shows the geographical coverage and the phonetic variability range for the selected 9,553 normalised forms.

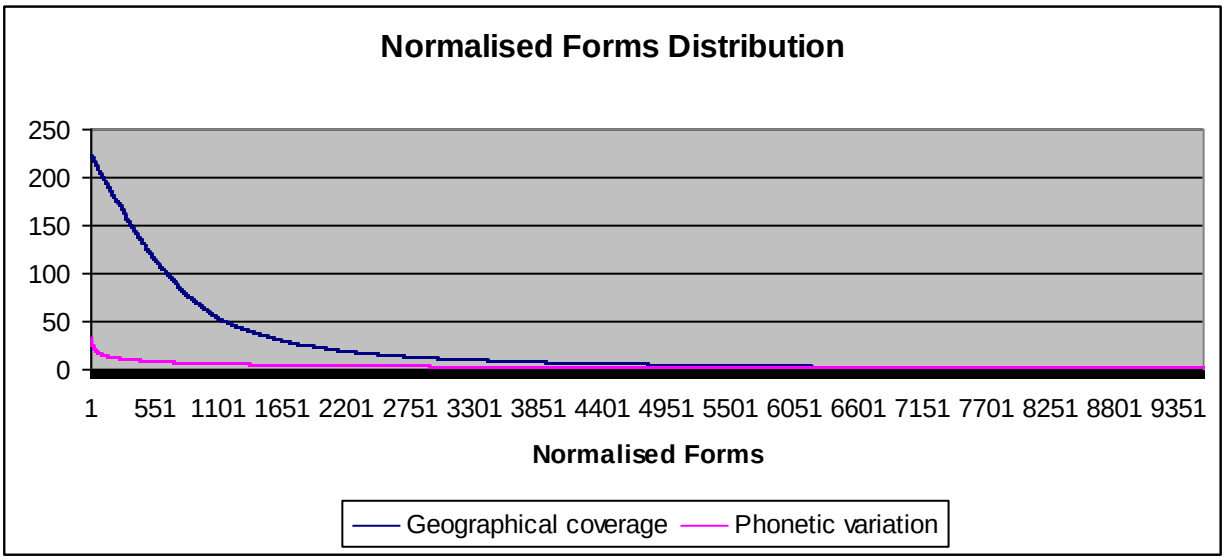

Figure 1 - Geographical coverage and phonetic variability range for the selected 9,553 normalised forms

Geographical coverage ranges between 2 and 223, and phonetic variability between 2 and 34: it should be noted, however, that within this set only 3,397 normalised forms (i.e. 35.55%) are attested in at least 10 different locations and only 1,920 show a phonetic variability range greater than 4 (corresponding to only 20.09% of NFs).

For the specific concern of this study, the following constraints have been enforced for the definition of the experimental data set: we focussed on dialectal items with a geographical coverage of at least 100 locations and showing at least 5 phonetic variants, corresponding to 523 normalised forms (5.47% of the whole sample).

The selected data set included adjectives, nouns and verbs. Due to the fact that in ALT verbal answers can be represented by different inflected forms (typically, the infinitival form, but also third person singular of the present indicative, or past participle) which are not always explicitly marked, we removed them from the experimental dataset in order to prevent potential noise deriving from verbal morphology. In this way, the set of selected normalised forms was reduced to 444 (4.64% of the whole set of diatopically varying NFs), including adjectives and nouns in the form of both single words and multi-word expressions. Note that selected multi-word expressions were represented by “frozen” word combinations, thus not showing variability due to the insertion/deletion of constituents.

In order to test the representativity of the selected sample of 444 NFs with respect to the whole set of normalised forms having at least two phonetic variants attested in at least two locations (used in Montemagni, 2008), we measured the correlation between overall phonetic distances and phonetic distances focussing on the selected sample which turned

out to be very high, with r=0.923. We can thus conclude that the selected sample can be usefully exploited to reliably study the patterns of phonetic variation in Tuscany.

2.3. Using atlas data as a corpus

From what has been said so far, it should be clear that here we are using atlas data in quite a peculiar way. Although this study is based on atlas data, it uses them as a corpus. This is to say that the dialectometric analysis of Tuscan phonetic variation here is not based on a predefined set of questionnaire items which were specifically designed to investigate the geographic distribution of phonetic features. Rather, it took the whole set of ALT attested lexical items, which were elicited from informants with quite different (mainly, lexico-semantic) purposes, and used it for studying phonetic variation.

By using atlas data as a corpus, the problem of inherently subjective feature selection is significantly reduced, thus providing a more "realistic" linguistic signal (Szmrecsanyi, to appear, p.3). On the other hand, by using atlas data as a corpus one of the main advantages usually ascribed to atlas-based studies, namely the areal coverage of dialectal items, can no longer be taken for granted. As we have seen, the areal coverage of attested NFs ranges from 1 to 223 locations: to overcome this potential problem, we enforced a minimal areal coverage threshold, corresponding to 100 locations (see section 2.2).

2.4. Dialectal data preparation

Having defined the extra-linguistic constraints which guided the definition of the experimental data set, all phonetic variants of the selected normalised forms were extracted. Extracted phonetic variants were enriched with information about the informants who attested them and were converted to IPA representation.

2.4.1. Extracting informants information

In previous studies based on ALT data (Montemagni 2007, 2008), phonetic variants of the same NF were used in a purely "categorical" way. This appears as a simplification, since the coordinates of each ALT item are not restricted to the location in which it was attested but also include the informants who testified it. This entails that for each attested phonetic variant we also know the number of informants who attested it, together with their socio-cultural profile.

For the specific concerns of this study, two different versions of the selected data set were generated, the first one containing frequency information associated with each phonetic variant token, and the second one also providing for each attested phonetic variant informants' features (age, education, profession). The format of the two versions of the data set is documented in Appendix 6.1.

2.4.2. ALT-CDI to IPA conversion

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), henceforth refererred to as CDI-ALT. This choice was in line with the Italian tradition of dialectological studies, which preferred the CDI transcription system with respect to the International

Phonetic Alphabet (IPA). Nowadays, this choice needs to be revised to make the ALT corpus usable by the wider international community of dialectologists and linguists. For this reason, the whole ALT corpus of phonetically transcribed data was converted into IPA.

Appendix 6.2 provides the correspondence table between the CDI-ALT and IPA notations. In most part of the cases, a 1:1 correspondence can be found:

- it can be the case that a CDI-ALT phonetic segment combined with one diacritic corresponds to an individual IPA phonetic segment (e.g. [ę] > [ɛ]);
- the reverse can also occur, whenever an individual CDI-ALT phonetic segment is converted into a IPA phonetic segment combined with a diacritic (e.g. [t'] > [tʲ]).

Interestingly enough, there are three different cases (highlighted in grey in the table) in which two different CDI-ALT segments are assigned the same IPA representation: this is the case of the weakened realization of palatal affricates, e.g. [ć] and [ǵ], whose representation coincides with the representation of the [š] and [ǯ] CDI-ALT segments, i.e. the voiceless and voiced postalveolar fricatives [ʃ] and [ʒ].

In ALT, phonetically transcribed data were represented through a hybrid encoding schema including both compositional and atomic representations which, depending on the task, were automatically converted into each other (Montemagni and Paoli 1989-90, pp. 36-43).

Compositional representations (see column 6 in the Table in Appendix 6.2) encode each phonetic symbol with a basic sign which may be further specified through one or more diacritics (conveying information, for instance, about stress or nasality of vowels). This representation type was particularly convenient for inputting and editing ALT data since all different phonetic symbols (about 110) are encoded by means of a restricted number of codes (36 basic signs and 9 diacritics) belonging to the first 128 ASCII codes and which can be directly accessed through the computer keyboard. To be more concrete, the compositional representation of a word like [stjaˈttʃaθa], or [stj̥aččáŧa] in terms of the CDI-ALT notation, is <sti4aCCa8t5a> where letters represent basic signs and numbers diacritics: in the case at hand, ‘5’ marks the spirantization of the voiceless dental occlusive, ‘8’ indicates the stress and ‘4’ represents a semivowel sound. This type of representation is particularly convenient for both sorting and retrieval tasks: in fact, if basic signs only are considered, it is possible to generalise over phonetic variants. Consider as an example the compositional representation of the word forms [stjaˈttʃaθa]/[stj̥aččáŧa] and [stjaˈttʃata]/[stj̥aččáta]: <sti4aCCa8t5a> and <sti4aCCa8ta> respectively. In both cases, the sequence of basic signs is the same, i.e. <stiaCCata>; this entails that a query starting from this sequence of bases will retrieve both of them.

Atomic representations (see column 2 in the Table in Appendix 6.2), on the other hand, show a 1:1 correspondence between CDI-ALT phonetic symbols and computer codes; they were typically used for on-screen display and printing. So, to keep with the <sti4aCCa8t5a> example, the combination of each base together with its diacritics is encoded through a symbol which uniquely identifies it (e.g. t5>ŧ).

The CDI-ALT to IPA conversion started from the compositional representation described above and was performed on the basis of 158 ordered conversion rules encoded as PERL regular expressions.

2.5. Reference data preparation

In the proposed analysis method, the phonetic variants recorded for each dialectal variety are compared with those attested in a reference variety (Wieling and Nerbonne, 2009), be it a standard language or a proto-language from which investigated dialects originate. The only prerequisite is that the reference language data should be available in the same transcription system as the dialectal material under study.

For the specific concerns of this study, two different reference languages have been selected, namely standard Italian and Latin, where the former is taken to originate from specific varieties of Tuscan dialects whereas the latter can be seen as a kind of proto-language from which Tuscan dialects originate. Different sets of experiments were performed by using respectively standard Italian and Latin as a reference.

For what concerns Italian, the standard Italian phonetic realization of selected normalised forms was manually encoded. Note that due to the historical relatedness between stardard Italian and Tuscan dialects all phonetic variants attested in the reference langauge are also attested as phonetic variants attested in some ALT location.

For what concerns Latin, we started from a subset of the 444 selected normalised forms (see section 2.2). In this case, the areal coverage constraint was increased to 150 locations, resulting in a 340 normalised forms from which verbs and multi-word expressions (e.g. *luna piena* ‘full moon’, *al sole* ‘in the sun’) have been removed (for a total of 40 NFs). For the remaining 300 cases, we looked for the Latin etymology, if any.

To this specific end, we used the Italian etymological dictionary by Manlio Cortelazzo and Paolo Zolli, “Dizionario Etimologico della Lingua Italiana”, 4 volumes, Bologna, Zanichelli, 1979.

All selected normalised forms were looked up in the reference dictionary and were classified as follows:

1. Latin etymology;
2. diminutive/augmentative forms of Latin words (provided that the used suffix is Latin as well);
3. complex derivative of a Latin word (e.g. *brinata* ‘hoarfrost’ from Lat. *pruinam* 'brina' Eng. ‘frost’ ; *castagnaccio* ‘cake made out of chestnut flower’ from Lat. *castăneam* ‘castagna’ Eng. ‘chestnut’);
4. complex etymology (e.g. *albicocca* ‘apricot’);
5. uncertain or unknown etymology (e.g. *afa* ‘sultriness’ or *bischero* ‘fool’);
6. non Latin etymology (e.g. *grullo* ‘silly’, *grattugia* ‘grater’);
7. onomatopoeic words.

Only cases 1 and 2 above were selected as a basis of this case study based on Latin as a proto-language.

For the IPA encoding of Latin words, the grapheme to IPA conversion rules were based on Allen (2004).

3. Method: adaptations and customizations

A general description of the bipartite spectral graph partitioning method is provided in Wieling and Nerbonne (2009, 2010). The method can be seen as articulated into the following steps:

1. obtaining, for each investigated location, the attested realization(s) of a given phonetic segment in a reference variety. The reference variety can be either a standard language or a proto-language;
2. bipartite spectral graph partitioning of the data matrix Locations x Phonetic_features resulting from step 1);
3. for each identified cluster of linguistic varieties, identify the most relevant features characterising it with respect to other clusters of varieties.

In this section, we briefly summarise the peculiarities of the ALT dialectal data set and illustrate the customizations of the method which have been performed to deal with them.

3.1. The ALT dialectal data set

The data set which has been selected for this study has been described in detail in section 2.2. Among its main features, it is worth pointing out here that the selected sample includes nouns and adjectives, both single words and multi-word expressions, testifying both productive phonetic processes and phonotactic processes. The focus is on the phonetic representation, including diacritics for a total of 109 different phonetic symbols.

3.2. Adaptations and customizations of the method

3.2.1. Enriching phonetic segment pairs with context information (Step 1)

In the adopted clustering method, each dialectal variety is described in terms of the attested phonetic realizations of a given phonetic segment with respect to a reference variety. Attested phonetic realizations are represented in terms of segment pairs where the phonetic segment attested in a given location L_x is associated with its realization in the reference variety R (either standard Italian or Latin):

[phonetic_realization_in_R]:[phonetic_realization_in_L_x].

For each selected NF (see section 2.2 above), phonetic segment pairs are obtained by aligning the phonetic realization in the reference variety R against the phonetic realizations recorded in the investigated varieties using the Levenshtein algorithm: alignments were induced by enforcing the syllabicity constraint on the basis of the PMI-based Levenshtein distance measure (Wieling et al., 2009).

Due to the fact that in the ALT dataset the same segment pair could originate from different phonetic processes, we decided to enrich the representation of segment pairs with contextual information, as exemplified below:

[L_contx|phonetic_realization_in_Ref|R_contx]:[L_contx|phonetic_realization_in_L_x|R_contx]

Consider as an example the segment pairs involving a consonant C and its geminated counterpart, i.e. [C]-[C:]. As it can be observed in the following examples, the same segment pair could result from different phonetic processes:

1. consonantal lengthening in intervocalic position
    a. [536] 225 Italiano
        [6] 101 Pieve Santo Stefano
        a _ b a ʧ ¡ o
        a _ b a ʧː ¡ o

    b. [829] 225 Italiano
        [628] 198 Piancastagnaio
        a b e t e
        a bː e t e

2. palatalization + consonantal lengthening
    a. [1013] 225 Italiano
        [847] 198 Piancastagnaio
        a l b e r o
        a j bː e r u

    b. [1141] 225 Italiano
        [872] 198 Piancastagnaio
        f a r i n a _ d o l ʧ e
        f a r i n a _ d o j ʧː e

3. phonotactic lengthening (in word initial position)
    a. [567] 225 Italiano
        [87] 107 Rosignano Marittimo
        a _ k a z o
        a _ kː a z o

    b. [536] 225 Italiano
        [24] 107 Rosignano Marittimo
        a _ b a ʧ ¡ o
        a _ bː a ʃ i o

For example, the same pair [ʧ]-[ʧː] appears both in 1.a and 2.b, as the result of different phonetic processes, namely consonantal lengthening occurring in intervocalic position and palatalization of preconsonantal /l/ followed by lengthening of the following consonant. Note that the involved phonetic phenomena show quite a different areal distribution.

For the time being, the representation of context includes:

- vowel (V);
- consonant (C);
- both (i.e. matching vowel and consonant, encoded as B);
- indel (-);
- word boundary (_);
- unknown (?).

In principle finer-grained distinctions can be resorted to in the representation of context information, with the danger of increasing the data sparseness problem.

3.2.2. Constraints on extracted segment pairs (Step 1)

In the previous section, we have seen that extracted phonetic segment pairs are formalised as follows:

[L_contx|phonetic_realization_in_Ref|R_contx]:[L_contx|phonetic_realization_in_L_x|R_contx]

Extracted segment pairs represent the basis of this study: i.e. each dialectal variety spoken in a given location is described in terms of the set of phonetic realizations of the underlying phonetic segment in the reference variety. This entails that extracted segment pairs should testify productive phonetic processes only. For this reason, from the set of extracted segment pairs we pruned out the segment pairs attested for single words only. This is the case, for instance, of the segment pair **V|n|C:V|r|C**, originating from the comparison of phonetic variants of the word *fanfarone* 'boaster' which included among its phonetic variants also *farfarone*: here, we cannot exclude that the attested variation is lexically driven, i.e. it originates from an assimilation process.

The parameters which could be used to define the set of extracted segment pairs thus include:

- the minimum number of words from which the same phonetic segment pairs could be extracted (at least 2 on the basis of what it was said above);
- the number of locations with respect to which the same phonetic segment pair has been attested.

3.2.3. Treatment of multiple responses (Step 1)

In a dialectal corpus of atlas data, in principle the same questionnaire item can be assigned multiple responses, attested either by the same informant or by different informants belonging to the same community.

In the case of ALT data, for each attested response type to the same questionnaire item within the same location we also know the number of informants who attested it. This is to say that also the frequency of occurrence of a given response in a given location is available in the ALT data set.

In previous dialectometric studies of Tuscan dialectal variation (Montemagni 2007, 2008), the treatment of multiple responses was carried out along the lines suggested by Nerbonne and Kleiweg (2003), where the distance was computed "between sets of strings where the sets represent alternative lexicalizations. The basic idea is that we average the distances between the individual strings where we consistently choose pairs in a way that minimizes the distance measure".

In this study, given the availability of token frequency information, different options have been provided to deal with multiple phonetic variants of the same NF within the same location, namely:

1. average over multiple phonetic variants tokens. In this way, the token frequency is used to determine the importance of each variant within a given location;
2. average over multiple phonetic variants types. In this way, the relative frequency of individual types is ignored and each option is weighted equally. This is the option followed in previous studies;
3. "majority vote", i.e. only the most frequent phonetic variant is considered for a given location.

The availability of these different options can help exploring the role of frequency in the study of dialectal variation, which still represents an open issue worth being investigated. As stated in Wieling and Nerbonne (2009, p.30), "while it stands to reason that more frequently encountered variation would signal dialectal affinity more strongly, it is also the case that inverse frequency weightings have occasionally been applied (Goebl, 1984), and have been shown to function well".

3.2.4. Single segment pairs (Step 1)

The analysis can be based either on all extracted segment pairs or on a subset of them. In the previous version of the method, the latter case was handles by specifying a given phonetic segment with the results that all segment pairs including it on either side (i.e. in either the target or reference location) were selected for the analysis. By doing in this way it would have been impossible to focus on specific linguistic phenomena, since the extracted data would have included pairs relating to different phonetic phenomena.

Consider, for example, the plosives in Tuscan dialectal variation: both voicing and devoicing of plosives are attested as productive processes in Tuscany, though with a different geographic distribution:

1. devoicing of plosives in intervocalic position
   V|g|V#V|k|V attested wrt the Tuscan words *aghetto* and *agaiolo*;
2. voicing of plosives in intervocalic position
   V|k|V#V|g|V: a more productive process wrt the previous one, attested wrt words such as *vicolo*, *albicocca*, *bacherozzolo*, *capocollo*, *ciuco*, *grattacacia*, *idraulico*, *oca*, *radica*, *rancico*, *ricotta*, *rustico*, *strabico*, etc..

The same applies to other phenomena such as lengthening and shortening of plosives:

3. lengthening of plosives in intervocalic position
   V|t|V#V|t：|V: attested wrt *ditale* and *sito*;

4. shortening of plosives in intervocalic position

   V|tː|V#V|t|V: a more productive process wrt the previous one, attested wrt words such as *bottiglia*, *aghetto*, *bigotto*, *bruschetta*, etc.

In order to make it possible to focus on specific phonetic phenomena, in the new version of the method when one or more phonetic segment(s) are specified, all segment pairs including them on the reference side are selected for the analysis.

3.2.5. Representativeness vs Distinctiveness (Step 3)

Wieling and Nerbonne (2009) calculate the importance of each phonetic segment pair by combining two different features, i.e. 'representativeness' and 'distinctiveness', where the former indicates the proportion of varieties in a given cluster which contain the sound correspondence and the latter indicates how prevalent a segment pair is in its own cluster as opposed to other clusters.

To be able to rank the segment pairs based on their distinctiveness and representativeness, these two values need to be combined. Different options have been experimented with by Wieling and Nerbonne (2009, 2010), namely:

a) taking the average of both values;
b) weigth distinctiveness twice wrt representativeness.

Different ways of combining the two values have been experimented with the ALT data set. Due to the strong similarity holding between the investigated dialectal varieties, it appears that option b) above leads to uninteresting results; the first option is better but still includes some noisy data. Experiments are being carried out to identify the best balance between the two scores wrt the linguistic peculiarities of ALT data.

4. Current directions of research

The report documents the activity carried out during the Short Term Mobility stay at the University of Groningen, which focussed on two main lines of research: a) preparation of the data set; b) adaptation and customization of the method with respect to the peculiar problems posed by the ALT data set. Currently, experiments are being carried out both with stardard Italian and Latin as reference varieties and results are being compared with Tuscan dialectological literature.

Preliminary results were presented at the XI Congresso della Società di Linguistica e Filologia Italiana (SILFI) which was held in Napoli on 5-7 October 2010, through a joint contribution entitled "Patterns of language variation and underlying linguistic features: a new dialectometric approach" by Simonetta Montemagni, Martijn Wieling, Bob de Jonge and John Nerbonne. The presented poster is attached at the end of the report.

Il Fruitore
Simonetta Montemagni

Il Proponente
Vito Pirrelli

5. References

Allen, W. S. (2004), *Vox Latina — a Guide to the Pronunciation of Classical Latin (2nd ed.)*. Cambridge University Press.

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

Giannelli L. (2000), *Toscana*, Pacini Editore, Pisa (1976, prima edizione).

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

Goebl H. (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, pp. 321-367.

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

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

Montemagni S., M. Paoli (1989/1990). *Dalla parola al bit (e ritorno): percorsi dall'inchiesta sul campo alla banca dati dell'ALT*, in AA.VV., "Quaderni dell'Atlante Lessicale Toscano", VI/VIII, pp. 7-52.

Montemagni S., M. Paoli, E. Picchi (2006), *ALT Web: l'Atlante Lessicale Toscano in rete*, in Bruni F., Marcato C. (a cura di), *Lessicografia dialettale: ricordando Paolo Zolli*, Atti del Convegno di Studi, Venezia, 9-11 dicembre 2004, Editrice Antenore, Roma-Padova, Tomo I, pp. 209-241.

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

Montemagni S. (2008), *The space of Tuscan dialectal variation. A correlation study*, "International Journal of Humanities and Arts Computing" (Special Issue on "Language Variation Studies and Computational Humanities"), Edinburgh University Press, Oct 2008, Vol. 2, No. 1-2, pp. 135-152.

Nerbonne J., W. Heeringa, P. Kleiweg (1999), *Edit Distance and Dialect Proximity*, in Sankoff D., Kruskal J. (a cura di), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Stanford, CSLI Press, pp.v-xv.

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

Nerbonne J., P. Kleiweg (2003), *Lexical Distance in LAMSAS*, in J. Nerbonne, W. Kretzschmar (Eds.), "Computers and the Humanities (Special Issue on Computational Methods in Dialectometry)", 37(3), pp. 339-357.

Nerbonne J. (in press), *Various Variation Aggregates in the LAMSAS South*, in stampa in C. Davis, M. Picone (eds.), *Language Variety in the South III*, Tuscaloosa, University of Alabama Press.

Pellegrini G.B. (1977), *Carta dei Dialetti d'Italia*, Pisa, Pacini Editore.

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

Szmrecsanyi B. (to appear). *Corpus-based dialectometry – a methodological sketch*, "Corpora", 6(1), Edinburgh University Press.

Wieling M., Nerbonne J. (2009), *Bipartite spectral graph partitioning to co-cluster varieties and sound correspondences in dialectology*, in M. Choudhury, S. Hassan, A. Mukherjee, S. Muresan (a cura di), Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, pp. 26–34.

Wieling M., J. Prokic, J. Nerbonne (2009), *Evaluating the pairwise alignment of pronunciations*, in L. Borin, P. Lendvai (Eds.), *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pp. 26–34.

Wieling M., Nerbonne J. (2010), *Hierarchical spectral partitioning of bipartite graphs to cluster dialects and identify distinguishing features*, in Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-5), ACL, Uppsala, Sweden, July 16, 2010.

6. Appendixes

6.1. ALT-RuG L04 data files: syntax for location and informants codes

ALT-RuG L04 data files with frequency information

: 100 Caprese Michelangelo
- aʃino
- aʃino
- aʃino
…

Location line
Es.
: 100 Caprese Michelangelo
it contains the following information types:
Numeric_location_id (100) and place_name (Caprese Michelangelo)

Phonetic variant tokens are reported one for each line starting with a dash ("-"). Frequency information can be reconstructed by counting the occurrences of the same phonetic variant token.

ALT-RuG L04 data files with informant details

: 100 Caprese Michelangelo-1-1
f;81;1898;2;9;2;1
- aʃino

Location line
Es.
: 100 Caprese Michelangelo-1-1
it contains the following information types:
Numeric_location_id (100) place_name (Caprese Michelangelo) inquiry_id (1) informant_id (1)
where

- the value of inquiry_id is numeric
- the value of informant_id is alphanumeric

Note that the numbering of informants reflects age (older informants are assigned lower alpha-numeric identifiers (i.e. informants labelled as 1 or A are older than 5 and C respectively). The ordering of identifiers follows the age ranking.

Informant line
Es.
f;81;1898;2;9;2;1
Fields separated by ";"
For each informant the following information types are provided:

- sex (f/m)
- age of the informant at the time of the interview
- year of birth

    - the two information types are useful due to the fact that ALT interviews were carried out in a time span of 20 years
- education encoded as follows:
    - 1: illiterate or semi-literate;
    - 2: primary school (not necessarily completed);
    - 3: middle school (not necessarily completed);
    - 4: so-called "Istituto Professionale" which is a type of high school providing secondary education oriented toward more practical subjects, enabling the students to start searching for a job as soon as they have completed their studies, typically after 3 years instead of 5 (not necessarily completed);
    - 5: high school (not necessarily completed);
    - 6: university degree (not necessarily completed);
- current and past profession, encoded as follows: in the first profession field it is reported the current profession, whereas the other two contain past professions, if any. The profession codes are provided below:
    - 1: farmer, farmhand, shepherd
    - 2: craftsman
    - 3: trader
    - 4: executive or auxiliary employee
    - 5: manager, concept employee, nurse
    - 6: teacher, freelance
    - 7: unskilled worker, apprentice
    - 8: skilled worker
    - 9: non-professional status (student, housewife, retired)

6.2. CDI-ALT to IPA conversion: correspondence table

	CDI-ALT	IPA-full conversion	IPA suprasegm.	IPA-simplified	CDI-ALT compositional representation	Notes
1.	ə	ə		ə	@	
2.	ə́	ə	Stress	ə	@8	
3.	a	a		a	a	
4.	ä	æ		æ	a3	
5.	a̋	æ̃		æ	a37	
6.	□	æ̃	Stress	æ	a378	
7.	a̋́	æ	Stress	æ	a38	
8.	ã	ã		a	a7	
9.	á̃	ã	Stress	a	a78	
10.	á	a	Stress	a	a8	
11.	b	b		b	b	
12.	ƀ	β		β	b5	
13.	č	ʧ		ʧ	C	
14.	ȼ	ʃ		ʃ	C5	See n.106
15.	d	d		D	d	

	CDI-ALT	IPA-full conversion	IPA suprasegm.	IPA-simplified	CDI-ALT compositional representation	Notes
16.	ď	dʲ		dʲ	D	
17.	~~đ~~	ð		ð	d5	
18.	e	e		e	e	
19.	ę	ɐ̯		e	e0	
20.	ę̃	ɐ̯̃		e	e07	
21.	ẽ́	ɐ̯̃	Stress	e	e078	
22.	ę́	ɐ̯	Stress	e	e08	
23.	ẹ	e		e	e1	
24.	ẽ	ẽ		e	e17	
25.	ẽ́	ẽ	Stress	e	e178	
26.	é	e	Stress	e	e18	
27.	ę	ɛ		ɛ	e2	
28.	ę̃	ɛ̃		ɛ	e27	
29.	ẽ́	ɛ̃	Stress	ɛ	e278	
30.	ę́	ɛ	Stress	ɛ	e28	
31.	□	ɐ̯̊		ɐ̯̊	e4	
32.	é	e	Stress	e	e8	
33.	f	f		f	f	
34.	ǧ	ʤ		ʤ	G	
35.	g	g		g	g	
36.	~~ǵ~~	ɣ		ɣ	g5	
37.	ǵ	ʒ		ʒ	G5	See n.107
38.	□	gˠ		gˠ	g6	
39.	h	h		h	h	
40.	i	ɪ		ɪ	i	
41.	į	ɪ̯		i	i2	
42.	į̃	ɪ̯̃		i	i27	
43.	ĩ́	ɪ̯̃	Stress	i	i278	
44.	į́	ɪ̯	Stress	i	i28	
45.	į̊	j		j	i4	
46.	ĩ	ɪ̃		i	i7	
47.	ĩ́	ɪ̃	Stress	i	i78	
48.	í	ɪ	Stress	i	i8	
49.	ǧ	gʲ		gʲ	J	
50.	č	kʲ		kʲ	j	
51.	k	k		k	k	

	CDI-ALT	IPA-full conversion	IPA suprasegm.	IPA-simplified	CDI-ALT compositional representation	Notes
52.	ƙ	x		x	k5	
53.	l	ɭ		ɭ	l	
54.	ľ	ʎ		ʎ	L	
55.	ɬ	ʟ		ʟ	l6	
56.	m	m		m	m	
57.	n	n		n	n	
58.	ń	ɲ		ɲ	N	
59.	ņ	ɳ		ɳ	n1	
60.	ṅ	ŋ		ŋ	n6	
61.	o	o		o	o	
62.	o̤	o̤		o	o0	
63.	õ̤	õ̤		o	o07	
64.	ő̤	õ̤	Stress	o	o078	
65.	ó̤	o̤	Stress	o	o08	
66.	ọ	o		o	o1	
67.	ọ̃	õ		o	o17	
68.	ọ̋	õ	Stress	o	o178	
69.	ọ́	o	Stress	o	o18	
70.	o̬	ɔ		ɔ	o2	
71.	õ̬	ɔ̃		ɔ	o27	
72.	ő̬	ɔ̃	Stress	ɔ	o278	
73.	ó̬	ɔ	Stress	ɔ	o28	
74.	ö	ø		ø	o3	
75.	ö̃	ø̃		ø	o37	
76.	ö̋	ø̃	Stress	ø	o378	
77.	ő	ø	Stress	ø	o38	
78.	ó	o	Stress	o	o8	
79.	o̰	ɔ̰		ɔ	o9	
80.	ó̰	ɔ̰	Stress	ɔ	o98	
81.	p	p		p	p	
82.	ꝑ	ɸ		ɸ	p5	
83.	ḍ	ɖ		ɖ	R	
84.	r	r		r	r	
85.	ṛ	ɽ		ɽ	r1	
86.	s	s		s	s	
87.	S	z		z	S	

	CDI-ALT	IPA-full conversion	IPA suprasegm.	IPA-simplified	CDI-ALT compositional representation	Notes
88.	t	t		t	t	
89.	t'	tʲ		tʲ	T	
90.	t̶	θ		θ	t5	
91.	u	u		u	u	
92.	u̯	u̯		u	u2	
93.	ü̯	ʏ̯		y	u27	
94.	ű̯	ũ̯	Stress	u	u278	
95.	ú̯	u̯	Stress	u	u28	
96.	ü	y		y	u3	
97.	ü̃	ỹ		y	u37	
98.	ű̃	ỹ	Stress	y	u378	
99.	ű	y	Stress	y	u38	
100.	û̯	w		w	u4	
101.	ũ	ũ		ũ	u7	
102.	ű̃	ũ	Stress	ũ	u78	
103.	ú	u		u	u8	
104.	v	v		v	v	
105.	ž	ʒ̶		ʒ̶	w	See n.109
106.	š	ʃ		ʃ	x	See n. 14
107.	ʃ'	ʒ		ʒ	X	See n. 37
108.	ś	ɕ		ɕ	x5	
109.	□	ʒ̶		ʒ̶	X5	See n.105
110.	ʒ	ɟ		ɟ	Y	
111.	z̨	dz		dz	Z	
112.	z	ʦ		ʦ	z	

6.3. Selected experimental data set

This Table is still under completion.

Normalised form	English translation	ALT-CDI transcription	IPA transcription	Latin etymology	IPA transcription
edera	ivy	e̯dera	edera	hĕderam	
orso	bear	o̯rso	orso		
oca	goose	o̯ka	ɔka		
occhio	eye	o̯kki̯o	ɔkːjo	ŏculum	
acino	grape	áčino	aʧino		

Normalised form	English translation	ALT-CDI transcription	IPA transcription	Latin etymology	IPA transcription
acqua cotta	Lit. ‘boiled water’, designating a traditional Tuscan soup	ákkᵤ̯a kǫ́tta	ak:wa kɔt:a		
albero	tree	álbero	albero	ărborem	
andito	passage	ándito	andito	ădĭtŭm	
ape	bee	ápe	ape	ăpem	
asino	ass	ásino	asino		
aspite	aspite	áspide	aspide		
aspito	aspito	áspido	aspido		
a bacio	to kiss	a bačío	a batʃio		
a caso	Random	a káSo	a kazo		
a cavalluccio	piggyback	a kávallúččo	a kaval:uʧ:o		
a solatio	in sunny	a solatío	a solatio		
abete	fir	abę́te	abete	abĭetem	
abeto	abeto	abę́to	abeto	abĭetem	
acquaio	sink	akkᵤ̯ái̯o	ak:wajo	ăquārĭum	
agaiolo	agaiolo	agai̯ǫ́lo	agajɔlo		
aghetto	pond	agę́tto	aget:o		
agnella	lamb	añńę́lla	aɲ:ɛl:a		
al sole	sun	al sǫ́le	al sole		
al tocco	touch	al tǫ́kko	al tok:o		
albicocca	apricot	álbikǫ́kka	albikɔk:a		
allodola	lark	allǫ́dola	al:ɔdola		
alloro	laurel	allǫ́ro	al:ɔro		
altalena	swing	általę́na	altalena		
arancio	orange	aráncho	aranʧo		
arcolaio	spinning wheel	árkolái̯o	arkolajo		
avvezzo	accustomed	avvę́zzo	av:eʦ:o	advitiātus	
becero	yahoo	bę́čero	beʧero		
borsa	bag	bǫ́rsa	borsa		
boccia	bowl	bǫ́čča	bɔʧ:a		
balzo	leap	bálzo	balʦo	bălteum	
bambola	doll	bámbola	bambola		
bazza	Bazza	báʒ̯ʒ̯a	badz:a		
baccano	din	bakkáno	bak:ano	bacchānal	
bacchettone	bigot	bakkettǫ́ne	bak:et:one		
bacherozzolo	bacherozzolo	bakerǫ́zzolo	bakerɔʦ:olo		
baciatura	baciatura	báčatúra	batʃatura		

Normalised form	English translation	ALT-CDI transcription	IPA transcription	Latin etymology	IPA transcription
balordo	stupid	balǫ́rdo	balordo		
bardella	pack-saddle	bardę́lla	bardɛlːa		
bargigli	wattles	barǧíl'l'i	bardʒiʎːi		
bastone	stick	bastǫ́ne	bastone		
bastardo	bastard	bastárdo	bastardo		
bazzone	Bazzone	baʒ́ʒ́ǫ́ne	badzːone		
bubbola	bells	búbbola	bubːola	upūpulam	
bernoccolo	bump	bernǫ́kkolo	bernɔkːolo		
beverone	mash	beverǫ́ne	beverone		
bigotto	bigot	bigótto	bigotːo		
bighellone	loafer	bigellǫ́ne	bigelːone		
birignoccolo	birignoccolo	bíriṅṅǫ́kkolo	biriɲːɔkːolo		
bindolo	waterwheel	bíndolo	bindolo		
bordello	brothel	bordę́llo	bordɛlːo		
borraccina	stonecrop	borračč́ína	borːatʃːina		
bottiglia	bottle	bottíl'l'a	botːiʎːa	butĭculam	
brace	embers	bráče	bratʃe		
braciola	chop	bračǫ́la	bratʃɔla		
braciere	brazier	bračę́re	bratʃɛre		
brinata	hoarfrost	brináta	brinata		
brindellone	Brindellone	brindellǫ́ne	brindelːone		
broccione	broccione	broč́čǫ́ne	brotʃːone		
bruscolo	mote	brúskolo	bruskolo		
bruschetta	bruschetta	bruskę́tta	brusketːa		
bischero	dawg	bískero	biskero		
biscia	snake	bišša	biʃːa	bēstiam	
butto	throw	bútto	butːo		
burrone	ravine	burrǫ́ne	burːone	burrunum	
burischio	Burisch	burískio̯	buriskjo		
buzzo	paunch	búʒ́ʒ́o	budzːo		
cenci	rags	č́ę́nči	ʧentʃi		
ceppa	log	č́ę́ppa	ʧepːa		
ceppo	log	č́ę́ppo	ʧepːo	cĭppum	
cesta	basket	č́ę́sta	ʧesta		
concio	ashlar	kǫ́nčo	kontʃo	cōmptum	
cacio	cheese	káčo	katʃo	cāseum	
caglio	rennet	kál'l'o	kaʎːo	cŏāgŭlum	
capo	head	kápo	kapo	căput	
cagnara	rumpus	kaṅṅára	kaɲːara		

Normalised form	English translation	ALT-CDI transcription	IPA transcription	Latin etymology	IPA transcription
calabrone	hornet	kalabrǫ́ne	kalabrone	crabrōnem	
caldano	brazier	kaldáno	kaldano		
calzoni	trousers	kalzǫ́ni	kaltsoni		
camomilla	chamomile	kámomílla	kamomilːa	camomĭllam	
camposanto	cemetery	kámposánto	kamposanto		
cantuccio	corner	kantúččo	kantuʧːo		
cantonata	corner	kántonáta	kantonata		
capezzolo	nipple	kapę́zzolo	kapetsːolo	capĭtiolum	
capone	capone	kapǫ́ne	kapone		
capocollo	capocollo	kápokǫ́llo	kapokɔlːo		
capomilla	capomilla	kápomílla	kapomilːa		
carbonaio	charcoal	kárbonái̯o	karbonajo	carbonārium	
casino	mess	kasíno	kasino	căsīnum	
castagneto	chestnut	kástańńę́to	kastaɲːeto		
castagnaccio	chestnut	kástańńáččo	kastaɲːaʧːo		
catasta	stack	katásta	katasta	catăstam	
cavaocchi	cavaocchi	kávaǫ́kki	kavaɔkːi		
cavalletta	grasshopper	kávallę́tta	kavalːetːa		
ciccioli	greaves	číččoli	ʧiʧːoli		
cenciaio	cenciaio	čenčái̯o	ʧenʧajo		
cenciaiolo	ragpickers	čenčái̯ǫ́lo	ʧenʧajɔlo		
cesoie	shears	čeSǫ́ie	ʧezoie		
cetriolo	cucumber	četrioǫ́lo	ʧetriɔlo	citrĭolum	
cigli	eyelashes	číl'l'i	ʧiʎːi		
ciglia	cilia	čil'l'a	ʧiʎːa		
cigna	Cigna	číńńa	ʧiɲːa		
chicco	bean	kíkko	kikːo		
chiocciola	snail	kio̯ǫ́ččola	kjɔʧːola		
chiorba	chiorba	kio̯ǫ́rba	kjɔrba		
chiasso	noise	kia̯sso	kjasːo		
chiacchierone	jay	kia̯kkia̯rǫ́ne	kjakːjerone		
chiorbone	chiorbone	kio̯rbǫ́ne	kjorbone		
ciocca	lock	čǫ́kka	ʧɔkːa		
ciabatte	slippers	čabátte	ʧabatːe		
ciucca	ciucca	čúkka	ʧukːa		
ciccione	fatty	čiččǫ́ne	ʧiʧːone		
ciucco	pacifier	čúkko	ʧukːo		
ciuco	donkey	čúko	ʧuko		
ciuffo	tuft	čúffo	ʧufːo		

Normalised form	English translation	ALT-CDI transcription	IPA transcription	Latin etymology	IPA transcription
ciliegia	cherry	čiliḛ́ğa	ʧiljɛʤa	cerĕseam	
cimitero	cemetery	čímitḛ́ro	ʧimitɛro	cimitērium	
cinquale	Cinquale	činkᵡ́ale	ʧinkwale		
cintolino	garters	číntolíno	ʧintolino		
cintura	belt	čintúra	ʧintura		
cinturino	strap	čínturíno	ʧinturino		
cipresso	cypress	čiprḛ́sso	ʧiprɛs:o	cyparĭssum	
cisoie	cisoie	čiSǫ́e	ʧizoje		
cimice	bug	čímiče	ʧimiʧe	cīmicem	
cinghia	belt	číngia	ʧingja		
cintola	waist	číntola	ʧintola	cĭnctulam	
cocomero	watermelon	kokǫ́mero	kokomero	cucŭmerem	
coccinella	ladybug	kǫ́ččinḛ́lla	koʧ:inɛl:a		
comare	Gossip	komáre	komare		
compagno	companion	kompáńńo	kompaɲ:o	companium	
compare	appear	kompáre	kompare	cŏmpatrem	
coperchio	cover	kopḛ́rkio	kopɛrkjo	copĕrculum	
covone	sheaf	kovǫ́ne	kovone		
covata	brood	kováta	kovata		
crognolo	Crognolo	krǫ́ńńolo	krɔɲ:olo		
crivello	sieve	krivḛ́llo	krivɛl:o		
crusca	bran	krúska	kruska		
cispa	blear eyes	číspa	ʧispa		
citto	citto	čítto	ʧit:o		
dolco	Pyrus	dǫ́lko	dolko		
desinare	dinner	dḛ́Sináre	dezinare		
di nascosto	secretly	di naskǫ́sto	di naskosto		
di sguincio	of sideways	di Sgᵡ́ínčo	di zgwinʧo		
di traverso	askew	di travḛ́rso	di travɛrso		
dialetto	dialect	di̯alḛ́tto	djalɛt:o	dĭălectum	
ditale	thimble	ditále	ditale	digitāle	
denti macellari	teeth butchers	dḛ́nti mačellári	dɛnti maʧel:ari		
denti occhiali	teeth glasses	dḛ́nti okki̯áli	dɛnti ok:jali		
esoso	hexose	eSǫ́So	ezɔzo	exōsum	
fosso	ditch	fǫ́sso	fɔs:o		
fango	mud	fángo	fango		
falegname	carpenter	fáleńńáme	faleɲ:ame		
fanfarone	braggart	fánfarǫ́ne	fanfarone		

Normalised form	English translation	ALT-CDI transcription	IPA transcription	Latin etymology	IPA transcription
faraona	guinea fowl	fáraǫ̣na	faraona	pharaōnem	
farina dolce	cake flour	farína dǫ́lče	farina dolʧe		
favilla	spark	favílla	favil:a		
ferraio	blacksmith	ferrái̯o	fer:ajo		
fettina	slice	fettína	fet:ina		
fiocco	bow	fi̯ǫ́kko	fjɔk:o	flŏccum	
fiammifero	match	fi̯amífero	fjamifero	flammĭferum	
fidanzato	boyfriend	fidanzáto	fidantsato		
filone	vein	filǫ́ne	filone		
filare	spin	filáre	filare		
focolare	hearth	fǫ́koláre	fokolare		
formaggio	cheese	formáǧǧo	formadʒ:o	formāticum	
formica	ant	formíka	formika	formīcam	
formicola	Tingly	formíkola	formikola	formīcolam	
fottio	fuck	fottío	fot:io		
fragola	strawberry	frágola	fragola	frāgulam	
frana	landslide	frána	frana	frăginam	
fregatura	swindle	fręgatúra	fregatura	fricatūram	
fringuello	finch	fringu̯ę́llo	fringwɛl:o		
fuliggine	soot	fulíǧǧine	fulidʒ:ine	fulīginem	
fulminante	fulminant	fulminánte	fulminante		
golpe	coup	gǫ́lpe	golpe	vŭlpem	
gota	cheek	gǫ̨ta	gɔta		
ganza	mistress	gánʐa	gandza	gangia	
ganzo	guy	gánʐo	gandzo		
gazza	magpie	gáʐʐa	gadz:a		
gazzera	Gazzera	gáʐʐera	gadz:era		
gabinetto	cabinet	gábinętto	gabinet:o		
gallinella	hen	gállinę́lla	gal:inɛl:a		
ghiaia	gravel	gi̯ái̯a	gjaja	glāream	
ghiacciaia	icebox	gi̯aččái̯a	gjaʧ:aja		
ghiandaia	jay	gi̯andái̯a	gjandaja	glandāriam	
ghiro	dormouse	gíro	giro	glīrem	
giubba	jacket	ǧúbba	dʒub:a		
ginepro	juniper	ǧinę́pro	dʒinepro	ienĭperum	
giomella	giomella	ǧomę́lla	dʒomɛl:a		
girino	tadpole	ǧiríno	dʒirino	gyrīnum	
gnocchi	gnocchi	ńńǫ́kki	ɲ:ɔk:i		
gomitolo	ball	gomítolo	gomitolo	glŏmitolum	

Normalised form	English translation	ALT-CDI transcription	IPA transcription	Latin etymology	IPA transcription
governo	government	govȩ́rno	govɛrno	gŭbernum	
greppo	chasm	grȩ́ppo	grep:o		
grappolo	cluster	gráppolo	grap:olo	clăppulum	
grasso	fat	grásso	gras:o		
grattacacia	grattacacia	grattakáča	grat:akatʃa		
grattugia	grater	grattúğa	grat:udʒa		
grembiule	apron	grembi̯úle	grembjule		
grillo	cricket	gríllo	gril:o	grīllum	
grullo	stupid	grúllo	grul:o		
gelso	mulberry	ğȩ́lso	dʒɛlso	cělsam	
gemma	gem	ğȩ́mma	dʒɛm:a		
guazza	dew	gu̯ázza	gwatsːa	acquāceam	
guercio	one-eyed	guȩ́rčo	guɛrtʃo		
idraulico	hydraulic	idráu̯liko	idrawliko	hydrāulicum	
imbroglio	cheat	imbrǫ́lʼlʼo	imbrɔʌ:o		
imbranato	clumsy	imbranáto	imbranato		
in ghingheri	dressed up	in gíngeri	in gingeri		
in proda	on shore	im prǫ́da	im prɔda		
l' anno scorso	last year 's	l ánno skǫ́rso	l an:o skorso		
locco	LOCKING	lǫ́kko	lok:o		
lodola	skylark	lǫ̨́dola	lɔdola	alāudam	
lampo	flash	lámpo	lampo		
lavatoio	wash	lavatǫ́i̯o	lavatojo	lavatōrium	
licite	Licite	líčite	litʃite		
legnaiolo	carpenter	leńńai̯ǫ́lo	leɲ:ajɔlo		
letame	manure	letáme	letame		
luna calante	waning moon	lúna kalánte	luna kalante		
luna crescente	crescent	lúna kreššȩ́nte	luna kreʃ:ɛnte		
luna piena	full moon	lúna pi̯ȩ́na	luna pjɛna		
lupo	wolf	lúpo	lupo		
lucignola	wick	lučíńńola	lutʃiɲ:ola		
lucertola	lizard	lučȩ́rtola	lutʃɛrtola		
lumaca	snail	lumáka	lumaka	limācam	
livido	livid	lívido	livido	līvidum	
mento	chin	mȩ́nto	mento	mĕntum	
moccolo	snot	mǫ́kkolo	mok:olo	mŭcculum	
mogio	dejected	mǫ́ğo	mɔdʒo		
mota	mota	mǫ̨́ta	mɔta	mältham	

Normalised form	English translation	ALT-CDI transcription	IPA transcription	Latin etymology	IPA transcription
macchia	stain	mákkia̯	mak:ja	măcŭlam	
male di capo	evil head	mál di kápo	mal di kapo		
male di testa	headache	mál di tę́sta	mal di tɛsta		
manfano	manf	mánfano	manfano		
madonnina	Madonna	madonnína	madon:ina		
maggiolino	cockchafer	maǧǧolíno	madʒ:olino		
magnano	locksmith	mańńáno	maɲ:ano		
maiale	pork	maiá̯ale	majale	maiālem	
maialino	piglet	maia̯alíno	majalino		
manciata	handful	mančáta	mantʃata		
mangiatoia	manger	manǧatǫ́ia̯	mandʒatoja		
matassa	hank	matássa	matas:a	matāxam	
materiale	material	materiá̯ale	materjale		
mattarello	mattarello	máttaré̯llo	mat:arɛl:o		
melone	melon	melǫ́ne	melone		
mestone	mestone	mestǫ́ne	mestone		
midolla	marrow	midǫ́lla	midol:a	medŭllam	
mietitura	harvest	mia̯etitúra	mjetitura		
migliaccio	Migliaccio	miľľáččo	miʎ:atʃ:o	miliācium	
mirtilli	blueberries	mirtílli	mirtil:i		
mollica	crumb	mollíka	mol:ika		
moine	moine	moíne	moine		
montone	ram	montǫ́ne	montone	multōnem	
mortadella	mortadella	mortadé̯lla	mortadɛl:a		
moscone	bluebottle	moskǫ́ne	moskone		
muschio	moss	múskia̯o	muskjo	mūsculum	
muta	pack	múta	muta	mūtam	
noccola	noccola	nǫ́kkola	nɔk:ola		
nottola	owl	nǫ̯́ttola	nɔt:ola	nŏctulam	
nottolo	owl	nǫ́ttolo	nɔt:olo		
nappone	nappone	nappǫ́ne	nap:one		
nasone	nose	nasǫ́ne	nasone		
nascondino	hide	naskondíno	naskondino		
nervoso	nervous	nervǫ́so	nervoso	nervōsum	
nevischio	sleet	nevískia̯o	neviskjo	nivisculum	
noioso	boring	noia̯ǫ́so	nojoso		
occhiali	glasses	okkiá̯ali	ok:jali		
odori	odors	odǫ́ri	odori	odōres	
orecchio	ear	orę́kkia̯o	orek:jo		

Normalised form	English translation	ALT-CDI transcription	IPA transcription	Latin etymology	IPA transcription
orzaiolo	sty	orz̨aiǫ́lo	ordzaiɔlo		
pecchia	bee	pẹ́kki̯a	pek:ja		
poggio	knoll	pǫ́ğğo	pɔdʒ:o	pŏdĭum	
palco	stage	pálko	palko		
palo	pole	pálo	palo	pālum	
pancia	belly	pánča	pantʃa	pănticem	
papera	gosling	pápera	papera		
papero	gosling	pápero	papero		
padrino	godfather	padríno	padrino	patrīnum	
pagliaia	Pagliaia	pal'l'ái̯a	paʎ:aja		
pagliuzza	mote	pal'l'úzza	paʎ:uts:a		
paletto	pole	palẹ́tto	palet:o		
pancetta	bacon	pančẹ́tta	pantʃet:a		
pancione	paunch	pančọ́ne	pantʃone		
paniere	basket	panni̯ę́re	pan:jɛre		
panzanella	panzanella	panzanę́lla	pantsanɛl:a		
papavero	poppy	papávero	papavero	papāverum	
pappavero	pappavero	pappávero	pap:avero		
parlata	speech	parláta	parlata		
pastone	mash	pastọ́ne	pastone		
pastrano	overcoat	pastráno	pastrano		
pelato	peeled	peláto	pelato	pīlātus	
pelliccia	fur	pellíčča	pel:itʃ:a		
pettata	t is	pettáta	pet:ata		
pettirosso	robin	pettirọ́sso	pet:iros:o		
pioppo	poplar	pi̯ǫ́ppo	pjɔp:o	plōppum	
piaggia	slope	pi̯áğğa	pjadʒ:a		
piattola	scraggy	pi̯áttola	pjat:ola		
pignatta	pot	pińńátta	piɲ:at:a		
pipistrello	bat	pipistrę́llo	pipistrɛl:o		
pinzo	pliers	pínzo	pintso		
pollone	sucker	pollọ́ne	pol:one		
popone	melon	popọ́ne	popone	pepōnem	
porcino	porcine	porčíno	portʃino	porcīnum	
pozzanghera	puddle	pozzángera	pots:angera	puteācula	
prese	taken	prẹ́se	preze		
proda	shore	prǫ́da	prɔda		
prezzemolo	parsley	prezzẹ́molo	prets:emolo		
primo quarto	first quarter	prímo ku̯árto	primo kwarto		

Normalised form	English translation	ALT-CDI transcription	IPA transcription	Latin etymology	IPA transcription
prete	priest	pr̨ę́te	prɛte		
pendolo	pendulum	pę́ndolo	pɛndolo		
pulcino	chick	pulčíno	pulʧino	pullicēnum	
pulenda dolce	Pulenda sweet	pulę́nda dǫ́lče	pulɛnda dolʧe		
pupilla	pupil	pupílla	pupil:a	pupĭllam	
puzzo	stink	púzzo	puʦ:o	pūtium	
puzzola	skunk	púzzola	puʦ:ola	putiolam	
rozzo	crude	rǫ́ʒ̨ʒ̨o	rodz:o		
rocchio	drum	rǫ́kkį̨o	rɔk:jo		
radica	briar root	rádika	radika		
rancico	Rancic	ráncˇiko	rantʃiko		
raspo	stalk	ráspo	raspo		
radice	root	radíče	raditʃe	radīcem	
ragazza	girl	ragázza	ragaʦ:a		
ragazzo	boy	ragázzo	ragaʦ:o		
raganella	treefrog	raganę́lla	raganɛl:a		
ramaiolo	ladle	ramaį̨ǫ́lo	ramajɔlo		
raponzoli	raponzoli	rapǫ́nzoli	raponʦoli		
ravanelli	radishes	ravanę́lli	ravanɛl:i		
riccio	hedgehog	ríčˇčo	riʧ:o	erīcium	
recinto	fence	rečínto	reʧinto		
ricotta	ricotta	rikǫ́tta	rikɔt:a		
rigatino	bacon	rigatíno	rigatino		
rimpiattino	hide and seek	rimpį̨attíno	rimpjat:ino		
ronzone	Ronzone	ronʒ̨ǫ́ne	rondzone		
rustico	rustic	rústiko	rustiko	rūsticum	
segale	rye	sę́gale	segale		
semola	semolina	sę́mola	semola	sĭmilam	
sorcio	mouse	sǫ́rčo	sorʧo	sōricem	
sodo	hard	sǫ́do	sɔdo		
soglia	threshold	sǫ́ľ'a	sɔʎ:a	sŏleam	
salcio	willow	sálčo	salʧo	sălicem	
salice	willow	sáliče	saliʧe		
sagrato	churchyard	sagráto	sagrato	sacrātum	
salamandra	salamander	salamándra	salamandra	salamăndram	
salciccia	sausage	salčíčča	salʧiʧ:a	salsīcia	
salsiccia	sausage	salsíčča	salsiʧ:a	salsīcia	
salvastrella	Burnet	salvastrę́lla	salvastrɛl:a		

Normalised form	English translation	ALT-CDI transcription	IPA transcription	Latin etymology	IPA transcription
sbornia	drunkenness	Sbǫ́rnia	zbɔrnja		
sbronza	drunk	Sbrǫ́nza	zbrondza		
scotta	sheet	skǫ́tta	skɔtːa		
scapolo	bachelor	skápolo	skapolo		
scaldaletto	warming pan	skaldalétto	skaldalɛtːo		
scaldino	Warmer	skaldíno	skaldino		
scheggia	splinter	skéǧǧa	skedʒːa	schĭdiam	
schiacciata	crushed	skiaččáta	skjatʃːata		
sciocco	silly	šš́ǫ́kko	ʃːɔkːo		
sciapo	sciapo	šápo	ʃapo		
sciamannone	sciamannone	šš́amannǫ́ne	ʃːamanːone		
sciamannato	sciamanno	šš́amannáto	ʃːamanːato		
sciapito	sciapo	šš́apíto	ʃːapito		
uscio	door	úššo	uʃːo	ōstium	
scimmia	monkey	šš́ímmia	ʃːimːja		
scoiattolo	squirrel	skoiáttolo	skojatːolo	scurĭolum	
scorciatoia	shortcut	skorčatǫ́ia	skortʃatoja		
scricciolo	wren	skríččolo	skritʃːolo		
seccatoio	squeegee	sekkatǫ́io	sekːatojo		
segatura	sawdust	segatúra	segatura		
sfoglia	puff pastry	sfǫ́ľľa	sfɔʎːa		
sugo	sauce	súgo	sugo		
somaro	ass	somáro	somaro		
soppressata	brawn	soppressáta	sopːresːata		
sottana	soutane	sottána	sotːana	subtānam	
sporta	shopping basket	spǫ́rta	spɔrta	spŏrtam	
spaccone	braggart	spakkǫ́ne	spakːone		
spazzatura	garbage	spazzatúra	spatsːatura		
spetezza	spetezza	spetézza	spetetsːa		
spigolo	corner	spígolo	spigolo	spīculum	
spranga	bar	spránga	spranga		
sputo	spit	spúto	sputo		
sedano	celery	sédano	sɛdano		
stolto	fool	stǫ́lto	stolto	stŭltum	
stoppia	stubble	stǫ́ppia	stopːja		
stalla	stable	stálla	stalːa		
stagnino	tinsmith	stańńíno	staɲːino	stāgninum	
sito	site	síto	sito	sĭtum	
strolago	loon	strǫ́lago	strɔlago		

Normalised form	English translation	ALT-CDI transcription	IPA transcription	Latin etymology	IPA transcription
strabico	squint	strábiko	strabiko		
stracco	tired	strákko	strak:o		
stradello	Stradello	stradḛ́llo	stradɛl:o		
stregone	sorcerer	stregǫ́ne	stregone		
strullo	Strullato	strúllo	strul:o		
strizza	Winking	strízza	striʦ:a		
succhiello	gimlet	sukkiḛ́llo	suk:jɛl:o		
susina	plum	suSína	suzina		
topo	mouse	tǫ́po	tɔpo		
talpa	mole	tálpa	talpa	tălpam	
tacchino	turkey	takkíno	tak:ino		
tagliere	chopping board	taľľḛ́re	taʎ:ɛre		
tarantola	tarantula	tarántola	tarantola		
tartaglione	Tartaglione	tartaľľǫ́ne	tartaʎ:one		
tartaruga	tortoise	tartarúga	tartaruga	tartarūcam	
terriccio	soil	terríččo	ter:iʧ:o		
testone	blockhead	testǫ́ne	testone		
tincone	tincone	tinkǫ́ne	tinkone		
tirato	pulled	tiráto	tirato		
topino	Mouse	topíno	topino		
trogolo	trough	trǫ́golo	trɔgolo		
trabiccolo	jalopy	trabíkkolo	trabik:olo		
tirchio	mean	tírkiǫ	tirkjo		
trucioli	chippings	trúčoli	truʧoli		
testa	head	tḛ́sta	tɛsta	tĕstam	
testo	text	tḛ́sto	tɛsto		
tuono	thunder	tuǫ́no	twɔno	tŏnum	
uggioso	dull	uğğǫ́so	uʤ:oso	odiōsum	
uncinetto	crochet	unčinḛ́tto	unʧinet:o		
unguanno	unguanno	ungu̯ánno	ungwan:o		
unguanno passo	unguanno step	ungu̯ánno pásso	ungwan:o pas:o		
volpe	fox	vǫ́lpe	volpe	vŭlpem	
vaglio	screen	váľľo	vaʎ:o	văllum	
vagabondo	vagabond	vagabǫ́ndo	vagabondo		
vicolo	alley	víkolo	vikolo	vīculum	
viottolo	path	viǫ́ttolo	vjɔt:olo		
vitalba	clematis	vitálba	vitalba		
vitellino	calf	vitellíno	vitel:ino	vitĕllinum	

Normalised form	English translation	ALT-CDI transcription	IPA transcription	Latin etymology	IPA transcription
verro	boar	vę́rro	vɛr:o		
zeppa	wedge	zzę́ppa	ʦ:ep:a		
zozzo	Zozzo	zzǫ́zzo	ʦ:oʦ:o		
zolla	clod	ʒʒǫ́lla	ʤ:ɔl:a		
zazzera	mop	zzázzera	ʦ:aʦ:era		
zittella	zittella	zzittę́lla	ʦ:it:ɛl:a		
zolfanello	match	ʒʒolfanę́llo	ʤ:olfanɛl:o		
ziro	Ziro	ʒʒíro	ʤ:iro		
zuccone	dodo	zukkǫ́ne	ʦuk:one		
zizzola	Zizzola	ʒʒízʒʒola	ʤ:iʤ:ola		
becco	beak	bę́kko	bek:o	bēccum	
caco	persimmon	kákо	kako		
ciglio	edge	číl'l'o	ʧiʎ:o	cĭlium	
grembiale	apron	grembi̯ále	grembjale		
lumacone	snail	lumakǫ́ne	lumakone		
pagliaio	haystack	pal'l'á i̯o	paʎ:ajo	paleārium	
radici	roots	radíči	raditʃi	radīces	