Synchronic patterns of Tuscan phonetic variation and diachronic change: Evidence from a dialectometric study

SIMONETTA MONTEMAGNI* / MARTIJN WIELING[+] / BOB DE JONGE[+] / JOHN NERBONNE[+]

*Istituto di Linguistica Computazionale "Antonio Zampolli"– CNR (Italy)

[+]Center for Language and Cognition Groningen, University of Groningen (The Netherlands)

1. INTRODUCTION

A careful investigation of synchronic patterns of linguistic variation with underlying linguistic features can lead to important insights into the comprehension of diachronic processes. In this paper, starting from the analysis of synchronic patterns of phonetic variation in Tuscany we would like to tackle one of the main and most debated features of Tuscan dialects, the phenomenon of spirantization with a specific view to the so-called "Gorgia Toscana" whose earliest reference dates back to the beginning of the 16th century.

The study is based on the dialectal data of the *Atlante Lessicale Toscano* ('Lexical Atlas of Tuscany', ALT, Giacomelli *et al.*, 2000; http://serverdbt.ilc.cnr.it/ALTWEB). Even though the published atlas documents lexical variation, the atlas material contains phonetic transcriptions on which we based the present study. Initial dialectometric investigations of this dialectal corpus with respect to phonetic variation (Montemagni, 2007, 2008) provided divergent results compared to the analyses by the main scholars of Tuscan (Giannelli, 2000) and Italian (Pellegrini, 1977) dialectology. Montemagni (2008) conjectured that pronunciation changes (corresponding to spirantization phenomena) spreading radially from Florence underlie the observed variation patterns.

This initial hypothesis, however, needed further investigation, which is possible thanks to a newly proposed dialectometric technique of co-clustering (called "bipartite spectral graph partitioning”) advanced by Wieling and Nerbonne (2010, 2011). This technique identifies groups of dialects on the basis of the aggregate analysis of a large corpus of dialectal data and simultaneously reconstructs the underlying linguistic basis. Through this technique it is possible to understand which factors underlie the identified patterns of variation, the role played by each of them and their interaction. Montemagni *et al.* (forthcoming) apply the hierarchical spectral partitioning (of bipartite graphs) to the ALT dialectal corpus. Their results helped gain insight into the nature of phonetic variation in Tuscany, by simultaneously providing a classification of dialectal varieties and their underlying linguistic basis. In particular, they demonstrate that patterns of phonetic variation in Tuscan dialects appear to be mainly due to spirantization phenomena, which originally arose in Florence and spread rapidly geographically and generalized phonologically.

The present study is aimed at further investigating spirantization phenomena across Tuscany with the final goal of exploring whether and how diatopic linguistic variation, which is detected in a synchronic dialectometric analysis, can be used to shed light on diachronic phonetic processes. On the methodological side, this study gave us the opportunity to assess the impact of a contextualised representation of sound correspondences in tracking down the evolution and diffusion of phonetic phenomena (in this case, spirantization).

2. The *gorgia toscana*

The phenomenon commonly known as *Gorgia Toscana* (literally,’Tuscan throat’, henceforth referred to as Tuscan *gorgia*) is a phonetic process belonging to the general class of lenitions, i.e., consonantal weakening phenomena. In particular, it refers to the intervocalic weakening of the voiceless stop consonants /p/, /t/ and /k/ taking the form of spirantization to [ɸ], [θ], and [x/h] respectively, none of which occur in the consonant inventory of Italian. The spirantization of /k/ and /t/ can even extend as far as deletion. Tuscan *gorgia* typically occurs intervocalically, both within and across words in continuous speech. Typical *gorgia* examples include [ˈbaxo], [ˈbaho] and [ˈbao] for /ˈbako/ ‘worm’ or [paˈθaθa] for /paˈtata/ ’potato’ in word-internal position, as well as [la ˈxaza], [la ˈhaza] and [la ˈaza] for /la ˈkaza/ ‘the house’ in sandhi (across the word boundary).

Extensive variation in the frequency and extent of spirantization of /p/, /t/ and /k/ is reported in the literature since the beginning of the 20th century. According to Rohlfs (1930) and Hall (1949), the geolinguistic extension of spirantization varies across the involved consonants: /k/ is affected in a wider area than /t/, which is in turn subject to spirantization over a larger area than /p/. Giannelli and Savoia (1978) report that Florentine speakers experience decreasing difficulty in pronouncing /k/, /t/ and /p/ respectively, thus reflecting decreasing levels of spirantization for/k/, /t/ and /p/, respectively. Bafile (1997) reports that the occurrence of less spirantized (or non-spirantized) forms becomes more frequent, passing from the velar to the dental, and then to the labial. Sorianello (2001) finds that /k/ is the primary target of the *gorgia*, followed in frequency by /t/ and /p/. The observed asymmetry in presence and extent of synchronic spirantization is also reported to hold diachronically. Izzo (1972) provides evidence that velars spirantized at least several generations before non-velars did.

Original Tuscan *gorgia* was accompanied by a rapid spread of spirantization through the Tuscan consonants, i.e. spirantization of /p/, /t/ and /k/ was extended to non-intervocalic contexts as well as to voiced stops /b/, /d/, /g/. Giannelli and Savoia (1978), Marotta (2001) and Sorianello (2001) all note that the voiced stops /b/, /d/ and /g/ are also involved in the process of weakening, surfacing as [β], [ð], and [ɣ/ɦ]. See, for instance, [ˈdaðo] for /ˈdado/ ‘dice’ or [la ˈɣamba] for /la ˈgamba/ ‘the leg’. Note that in this case, the resulting surface realizations do not occur in the Italian consonant inventory. Note further, however, that the presence and extent of spirantization of /b d g/ differ significantly when compared to their voiceless counterparts. Diachronically, spirantization of /b d g/ appeared later than voiceless stop spirantization; diatopically, it shows a much more restricted diffusion throughout the region.

Tuscan *gorgia* is far from being an obligatory rule. Acoustic studies performed by Marotta (2001) and Sorianello (2001) show that stops do, in fact, surface among the allophonic variants. This fact combined with the asymmetric distribution of spirantization phenomena across Tuscan consonants shows that Tuscan *gorgia* presents itself as a gradient process exhibiting rich variation, both diachronically and synchronically; in the synchronic domain further variation is found, diatopically and also socially (Giannelli and Savoia, 1978, 1980).

Figure 1 – Geographic diffusion of Tuscan *gorgia*

With respect to the origin of Tuscan *gorgia*, different attempts have been put forward in the literature to justify and explain it, both substratist or anti-substratist in character. In more recent studies, the substratist hypothesis has been rejected (Hajek, 1996). Instead, Tuscany is placed squarely within the framework of ordinary consonantal weakening in Romània (the Romance languages). Within this context, Tuscan *gorgia* is increasingly accepted as a local and innovative natural phenomenon spreading from the influential center of Florence, traditionally viewed as the epicenter, in all directions (as depicted in Figure 1). From Florence, the *gorgia* spreads along the entire Arno valley, losing strength closer to the coast. It is also present to some extent in the northwest and the northeast, with the Apennines representing the northern border of the phenomenon, as well as in Siena and further south (but not in far southern Tuscany).

Last but not least, a terminological remark is in order here. A careful distinction must be made between the term *gorgia*, traditionally referring only to the voiceless spirantization of intervocalic /k t p/, and the more generic term "spirantization" which can affect any consonantal phoneme in any type of context.

3. THE DATA SOURCE

3.1. The *Atlante Lessicale Toscano*

ALT is an especially 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 lexico-semantic variation, but this did not preclude its containing valuable information with respect to phonetic variation. ALT is Tuscan in the sense that it is a regional atlas focusing on dialectal variation throughout Tuscany, a region where both Tuscan and non-Tuscan[1] dialects are spoken. ALT data were collected between 1974 and 1986 from 2,193 speakers who were first selected with respect to a number of parameters including age, socio-economic status, education and culture and who were then each asked 745 questions.

In ALT, each dialectal item is assigned different levels of representation organized in layers of progressively decreasing detail going from phonetic transcription to different levels of orthographic representations (Cucurullo *et al.* 2006). For this study, we focused on phonetic transcription[2] and the normalized (orthographic) representation levels, where the latter was useful

---

[1] This is the case for dialects in the north, namely Lunigiana and small areas of the Apennines (so-called Romagna Toscana), which belong to the group of Gallo-Italian dialects.

[2] The phonetic alphabet used in the ALT project was a geographically specialized version of the "Carta dei Dialetti

for its elimination of the effects of productive phonetic processes even while representing distinct morphological (both inflectional and derivational) variants. To illustrate this more concretely, phonetic variants originating from spirantization or voicing phenomena such as [skjaˈtʃːa**θ**a] or [skjaˈtʃːa**d**a] are both assigned the same normalised form, i.e. *schiacciata*, whereas words such as [skjaˈtʃːat**a**] (singular) and [skjaˈtʃːat**e**] (plural) as well as [skjatʃːaˈt**ina**] (diminutive) are all assigned distinct normalized forms.

The alignment of the different representation levels was exploited to automatically extract all phonetic variants (henceforth, PV) of the same normalized form (henceforth, NF). Due to the features of the ALT normalized representation, a study based on the analysis of PVs of the same NF should document only genuine phonetic processes, without interference from any other linguistic description level (e.g. morphology).

3.2. The experimental dataset

For the specific concerns of this study, we used ALT dialectal data in a somewhat peculiar way: namely, we started from the attested phonetic variants which were elicited from speakers for lexico-semantic purposes (see above) without any a priori phonetically-driven selection. Szmrecsanyi (submitted) argues that compared to linguistic atlas material dialectal corpora yield a more realistic linguistic signal, with two main advantages. First, they provide graded frequency information better matching the perceptual reality of linguistic input than discrete atlas classifications. Second, while the atlas signal is non-naturalistic, metalinguistic, and competence-based in nature, text corpora provide more direct, performance-based access to language. In order to reduce some of the inherent problems of atlas data we used lexical answers to questionnaire items organised by type (corresponding to the normalised form) to infer the production frequency of specific phonetic features as a proxy for their salience in determining the observed patterns of phonetic variation.

Italiani” (CDI) transcription system. For this study, the whole ALT corpus of phonetically transcribed data was converted to the International Phonetic Alphabet (IPA).

However, in this way one of the main advantages of atlas-based studies, i.e. the areal coverage of collected dialectal evidence, can no longer be taken for granted. This potential problem was overcome by enforcing a minimal areal coverage threshold for what concerns the selection of NFs (see below).

In particular, data selection from the ALT dialectal corpus was carried out by combining linguistic and geographical criteria. With respect to the former, only nouns and adjectives were selected,[3] attested either as single words or as multi-word expressions.[4] Phonetic variability represented the other linguistic criterion on the basis of which we performed data selection. The number of phonetic variants associated with the same NF was enforced to range between a minimum of 5 and 34 (in ALT, this is the maximum number of PVs associated with the same NF). Geographical criteria were concerned with a) the network of the locations investigated which was restricted to the 213 (out of the 224) locations where Tuscan dialects are spoken and b) the areal coverage of selected NFs, which was required to be higher or equal to 100 (out of 213) locations. The resulting experimental dataset was composed of 444 NFs (covering the 4.64% of the whole set of diatopically varying NFs) including a total of 502,799 phonetic variant tokens.

Since we did not know in advance whether the selected sample of 444 NFs with associated phonetic variants was representative of the whole set of NFs having at least two PVs attested in at least two locations (used by Montemagni, 2008), we measured the correlation between phonetic distances in the overall dataset and in the sample selected, which turned out to be very high (r = 0.994). It follows that the sample selected can be reliably exploited to study phonetic variation across Tuscany.

---

[3] As in ALT verbal answers represented by different inflected forms are not always explicitly marked, verbs were excluded from the experimental dataset to prevent potential noise deriving from verbal morphology.

[4] Note that multi-word expressions selected were represented by “frozen” word combinations, showing no variability due to the insertion/deletion of constituents.

Since the analysis method proposed (see Section 4.1) is based on the comparison between PVs collected through fieldwork and their phonetic realization in a reference variety, the experimental dataset underlying this study also includes the phonetic transcription of the selected NFs according to standard Italian pronunciation.

4. METHODS

4.1. Obtaining sound correspondences

In order to obtain sound correspondences (henceforth, SCs) linking the dialectal allophone with its realization in our reference variety (standard Italian), we aligned the phonetic variants with their phonetically transcribed standard counterpart. We obtained the alignments using an adapted Levenshtein algorithm (Levenshtein 1965). The standard Levenshtein algorithm obtains the alignment between two strings by minimizing the number of deletions, insertions and substitutions needed to transform one string into the other. For example, the Levenshtein distance between the standard and a dialectal realization of *albicocca* 'apricot' is 2, since we need two operations (one substitution and one deletion) as shown below.

Standard Italian	a	l	b	i	k	ɔ	kː	a
Putignano	a	r	b	i		ɔ	kː	a
		1			1			

Instead of the standard Levenshtein algorithm which assigns the same cost to substitutions involving similar sounds (such as [u] and [o) as opposed to different sounds (such as [a] and [i]), we used a version which employs automatically determined segment distances based on Pointwise Mutual Information (PMI; Church and Hanks, 1990). This method was introduced by Wieling *et al.* (2009) and found to yield superior alignments as well as acoustically sensible sound correspondences (Wieling *et al.*, to appear).[5] As multiple speakers were interviewed in every

[5] For the technical details of the procedure, we refer to Wieling *et al.* (to appear).

location, we used the most frequent phonetic variant as representative of all attested PVs for every normalized form.

As we are interested in spirantization in this study, we focus on phonetic correspondences involving both identical and non-identical segments, also including insertions and deletions, with a stop on the reference (standard) side and with either an occlusive or a spirantized (including absent) realization on the allophonic (dialectal) side.

As in the ALT dataset the same sound correspondence could originate from different phonetic processes, Montemagni *et al.* (forthcoming) in their initial investigation of the ALT dataset enriched the representation of the sound correspondences with contextual information. In this study, we follow their approach and for each sound correspondence we identify the left and right (single segment) context. We use a rough context, as we only distinguish consonants, vowels, semi-vowels, gaps (i.e. for an insertion or deletion) and the word boundary (i.e. for the first and last segment of a word). For example, the SC /k/:[-] (i.e. the [-] indicates a gap) in the alignment above is recorded as V/k/V: V[-]V. This indicates that there are vowels to the left and right of both /k/ and [-]. This enriched representation of sound correspondences is crucial in the framework of the present study focusing on the propagation of a specific phonetic process (spirantization) across different types of context in Tuscan dialects.

After obtaining the sound correspondences, we count the frequency of each of these. We normalize these values by dividing by the number of words whose alignments include the specific sound correspondence, as not all words are attested in every variety. The normalized frequencies are stored in a matrix (exemplified in Table 1), where rows represent locations and the columns distinct (contextualized) sound correspondences.

Table 1 – Excerpt of the matrix with locations x sound correspondences

	V/k/V:V[x]V	V/k/V:V[k]V	V/k/V:V[-]V	V/k/V:V[h]V
Pieve Santo Stefano	0.0000	0.0593	0.0000	0.0000

	V/k/V:V[x]V	V/k/V:V[k]V	V/k/V:V[-]V	V/k/V:V[h]V
Anghiari	0.0000	0.0683	0.0000	0.0000
Antignano	0.0000	0.0132	0.0219	0.0219
Rosignano Marittimo'	0.0181	0.0090	0.0226	0.0136
Cecina'	0.0000	0.0000	0.0000	0.0528

This matrix is then used as input for the approach (illustrated below) we employ to cluster locations together with their characteristic sound correspondences.

4.2. Clustering sound correspondences and locations simultaneously

We use hierarchical spectral partitioning of bipartite graphs (Wieling and Nerbonne, 2010) to *simultaneously* identify the geographical clusters as well as their characteristic phonetic features in the ALT dataset. A bipartite graph has two sets of vertices (representing locations and sound correspondences) and a set of edges connecting vertices from one set to the other (i.e. an edge represents the occurrence of a sound correspondence in a location). By calculating the singular value decomposition (SVD) of the input matrix and repeatedly applying the *k*-means (with *k* equals 2) algorithm to these results, a hierarchical clustering is obtained in which sound correspondences are clustered together with locations. The mathematical details of the procedure are discussed in detail by Wieling and Nerbonne (2010).

Following Montemagni *et al.* (forthcoming), we scaled all columns of the input matrix (see above) between zero and one, to ensure that every sound correspondence was equally important. After applying the hierarchical spectral partitioning method to this scaled input matrix, we obtain a hierarchical clustering where locations are clustered together with the sound correspondences.

4.3. Determining the most important sound correspondences for every cluster

The hierarchical spectral partitioning method yields many sound correspondences in every geographical cluster. To identify the most important ones, we use the method introduced by Wieling and Nerbonne (2011) which is based on calculating the DISTINCTIVENESS and REPRESENTATIVENESS of each SC.

REPRESENTATIVENESS (of a sound correspondence) simply measures how frequently the sound correspondence occurs in the cluster. For example, if there are ten locations in the cluster and the sum of the normalized frequencies equals five, the representativeness is 0.5 (5 divided by 10).

In contrast, the DISTINCTIVENESS of a sound correspondence measures how frequently the sound correspondence occurs within as opposed to outside of the cluster. It also takes the relative size of the cluster into account. For example, if the SC does not occur outside of the cluster, the distinctiveness is maximal (i.e. 1; the sound correspondence perfectly distinguishes the cluster from the others), irrespective of the size of the cluster. Alternatively, if a cluster contains half of the total number of locations and half (or less) of the total sum of the normalized frequencies, the distinctiveness is minimal (i.e. 0; the sound correspondence does not distinguish the cluster at all).

The values of both distinctiveness and representativeness range between zero and one (in the cases we are interested in). In line with Wieling and Nerbonne (2010) we average the representativeness and distinctiveness to obtain the IMPORTANCE value for every individual sound correspondence.

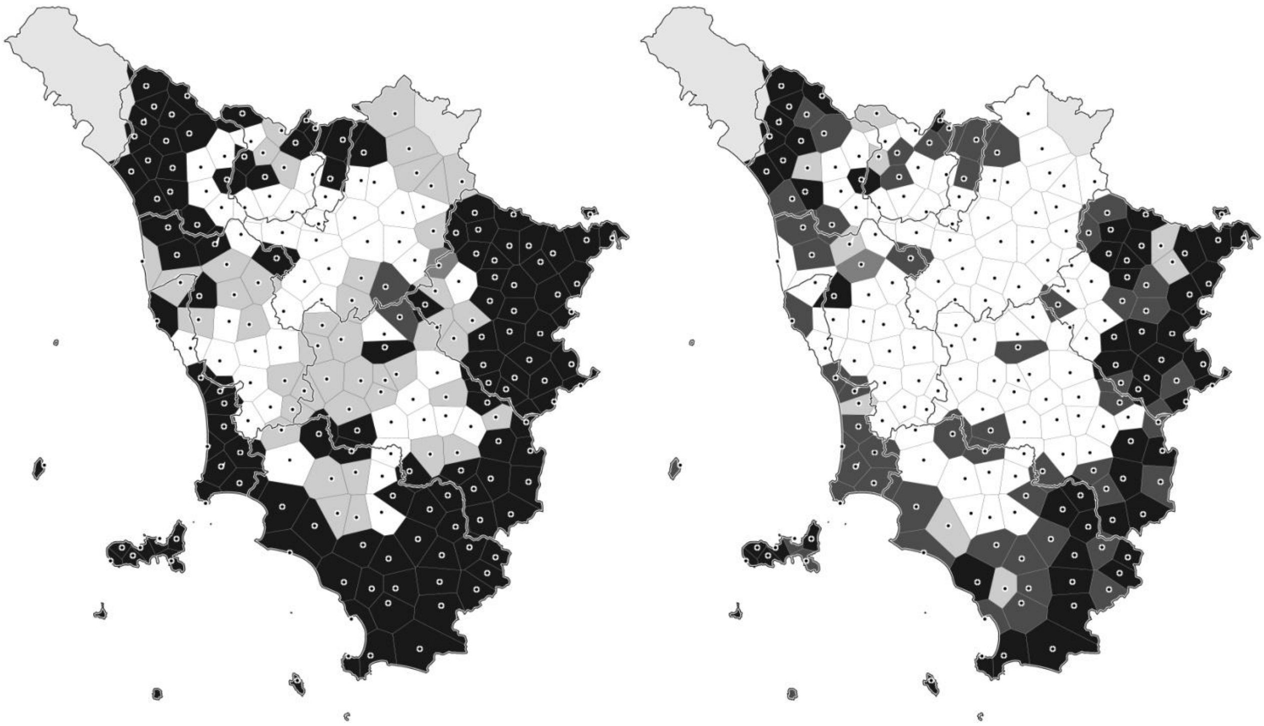
5. RESULTS

In this section, we report the results of applying the hierarchical spectral partitioning method to the selected dataset. In order to assess the contribution of contextual information in reconstructing the spreading of phonetic phenomena, two different experiments have been carried out, respectively based on 16 context-free and on 84 contextualized sound correspondences extracted from the alignments involving the 444 selected NFs. The comparison of these results is expected to shed light on the impact and role, if any, of contextual information in reconstructing the diffusion of spirantization across Tuscany.

5.1. Geographical results

In Figure 2, the maps show the geographic visualization of the clustering of Tuscan varieties into five groups obtained by using contextualized SCs (left map) and context-free SCs (right map).

Figure 2 - Geographic visualization of the clusters obtained with contextualized SCs (left map) and context-free SCs (right map). Different shades of darkness indicate different clusters, which also correspond to different steps in the generalization of Tuscan *gorgia*. See text for further remarks.



It is clear that the general pattern is similar in the two maps: the phonetic areas identified are arranged in an onion-like shape built around a big central area covering the province of Florence and propagating in different directions, towards the south (in the province of Siena), east (in the province of Arezzo) and west (covering the provinces of Prato, Pistoia, Lucca up to most part of Pisa and Livorno). Around this central area, there is an external layer.

However, major differences can also be observed between the two maps. Let us first focus on the central area. In the left map, based on contextualized SCs, this area is articulated into different clusters (white, light grey, and two smaller darker grey regions). In the right map, based on context-free SCs, this corresponds to the white cluster with the light grey one (sparse locations around the white core) acting as a kind of transition zone. The reverse appears to hold at the level of the outer layer which has approximately the same geographical coverage in both maps with the following differences: while the left map presents a unique cluster (black), the right map presents a further fragmentation into two clusters (black and dark grey).

5.2. Linguistic results

We turn now to a discussion of the linguistic features underlying the clusters just discussed. In both maps the central area, with Florence as epicenter, is characterized by spirantization phenomena, whereas the features underlying the outer layer correspond to retention of (some) occlusives. Besides this general common trend, the most salient features underlying the clusters identified in the two maps differ significantly.

Let us first focus on the left, contextualized map. Table 2 reports the ranked SCs with associated importance scores for the three main clusters identified. In the first column it can be seen that underlying the white core cluster there are SCs corresponding to spirantization phenomena involving voiced stops /b d g/ as well as /k/ in word initial position. It is also interesting to note the ranking of features here inversely follows the spirantization hierarchy (i.e. with velars being associated with greater importance with respect to dentals and bilabials, and with voiced stops being more important than voiceless ones). The second layer cluster (light grey), in contrast, is characterized by SCs corresponding to prototypical Tuscan *gorgia* (involving voiceless stops in intervocalic context), with extensions to other contexts, e.g. word initial position (second column of Table 2). Interestingly, SCs with /t p/ on the reference side (i.e. standard Italian) are assigned a greater importance than those involving /k/. With respect to the outer layer (third column), corresponding to the white cluster, we observe mainly SCs with occlusive realization: the feature list includes only two spirantized SCs involving the voiceless stop /k/ with a less spirant outcome [x] which are however assigned quite a low rank in the list of ordered sound correspondences (i.e. they appear in the 25th and 27th position respectively).

Table 2 - Ranked contextualized spirantization-related SCs with associated importance scores (between parentheses) for the three main clusters identified.

Core cluster (white)	Second layer cluster (light grey)	Marginal cluster (black)
1. V/g/V:V[ɤ]V (0.319) 2. V/d/V:V[ð]V (0.281) 3. _/k/C:_[h]C (0.210) 4. _/k/V:_[h]V (0.126) 5. V/b/C:V[β]C (0.112)	1. V/t/V:V[θ]V (0.192) 2. _/p/V:_[ɸ]V (0.164) 3. V/p/V:V[ɸ]V (0.152) 4. V/p/C:V[ɸ]C (0.144) 5. _/t/C:_[θ]C (0.130) 6. V/t/C:V[θ]C (0.130) 7. _/p/B:_[ɸ]B (0.128) 8. V/k/V:V[h]V (0.112)	… 25. _/k/V:_[x]V (0.134) 26. ... 27. V/k/V:V[x]V (0.117)

Let us now turn to the clusters of the map on the right, based on context-free SCs. The features underlying the white central cluster correspond to spirantization phenomena involving both voiceless and voiced stops. They are reported in Table 3 (first column) with the associated importance scores:

Table 3 - Ranked context-free spirantization-related SCs with associated importance scores (between parentheses) for the three main clusters identified.

Core cluster (white)	Second layer cluster (light grey)	Marginal cluster (dark grey)	Marginal cluster (black)
1. /t/:[h] (0.500) 2. /d/:[ð] (0.484)	1. /k/:[x] (0.197)	1. /t/:[t] (0.196) 2. /p/:[p] (0.148)	1. /k/:[k] (0.178)

Core cluster (white)	Second layer cluster (light grey)	Marginal cluster (dark grey)	Marginal cluster (black)
3. /t/:[θ] (0.449) 4. /p/:[ɸ] (0.421) 5. /b/:[β] (0.421) 6. /g/:[ɤ] (0.405) 7. /k/:[h] (0.259) 8. /t/:[ø] (0.178)		3. /b/:[b] (0.096) 4. /d/:[d] (0.094) 5. /g/:[g] (0.091)	

The light grey cluster (covering sparse locations around the white core) is characterized by just one SC involving /k/ with a less spirant outcome (see the second column). Proceeding towards the external area, the features underlying the black and dark grey external clusters correspond to retention of stop consonants (occlusives) (see columns 4 and 3 respectively). In particular, whereas the most external cluster is characterized by a single SC only, which is /k/:[k], corresponding to the first stop originally affected by Tuscan *gorgia*, the ranked SCs of the dark grey cluster involve the remaining voiced and voiceless stops. Contrary to the results obtained starting from contextualized SCs, in Table 3 it can be noticed that the ordering of features within each cluster does not appear to reflect the spirantization hierarchy.

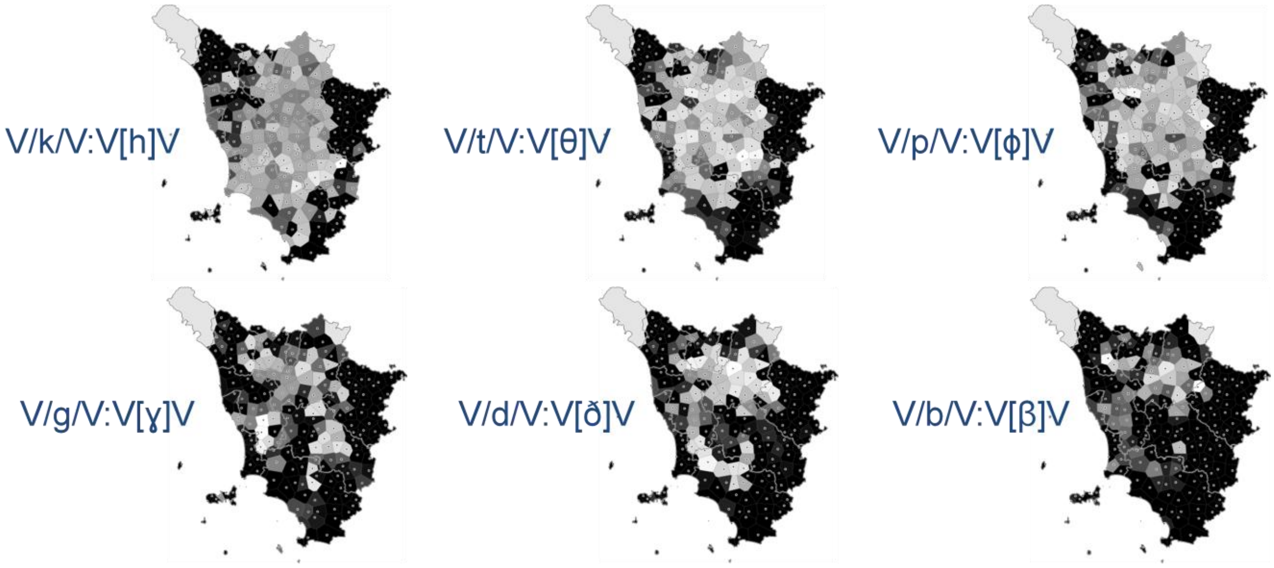
By comparing the linguistic results obtained with and without contextual information it is now possible to better assess its role in the reconstruction of the spreading of spirantization phenomena across Tuscany. With contextualized SCs (left map), the spirantization area (with Florence as epicenter) is articulated into different clusters, inversely reflecting the evolution of the phenomenon across the Tuscan consonantal phonology. The most important features distinguishing the core cluster (white) correspond to the most recent spirantization phenomena involving voiced stops and voiceless ones (only /k/) in non-canonical contexts, whereas the cluster immediately surrounding the core cluster is characterized by spirantization of voiceless stops (intervocalically and elsewhere). Using context information, the marginal area presents itself as a unique cluster, mainly characterized by retention of occlusive outcomes.

The result obtained without context information combines a compact spirantization area (with Florence as epicenter), involving both voiceless and voiced stops, with a marginal area characterized by retention of occlusive outcomes articulated into two clusters (the most internal cluster is characterized by the retention of occlusive outcome for all stops except /k/, which is associated with the most external cluster).

The results sketched above show that context information plays a central role: sound changes are recognized to be conditioned by phonetic context, as we saw in the case of Tuscan *gorgia*. Methodologically, we note that the approach used in this study has successfully detected the influence of context automatically. Contextualized SCs enable the detection of a linguistically well-founded and articulated picture, both at the level of regional coherence and the underlying linguistic features. In particular, using contextualized SCs we were able to "reconstruct" the spreading of spirantization phenomena through Tuscan consonantal phonology, i) by originally involving the velar stop /k/, then /p/ and /t/ up to the voiced stops /b/, /d/, and /g/, and ii) through different types of contexts, i.e. intervocally in medial word position but also as realization of a sandhi effect. In fact, going from the core to the outer layer, the important features associated with each cluster inversely follow the spirantization hierarchy. While the core area is characterized by the most advanced and recent spirantization phenomena, in the other layers spirantization is progressively restricted to voiceless plosives, with only /k/ being involved with quite a low salience in the external layer. By gradually moving away from the core, we first observe clusters characterised by SCs involving /p/, /t/, then by /k/ with progressively fewer spirant outcomes. Without contextual information, a more static picture emerges, with a single spirantization cluster.

In order to assess the reliability of the features identified, consider now the geographic distribution of SCs instantiating spirantization phenomena involving voiceless and voiced stops and their spirantized counterpart intervocalically. In the maps in Figure 3, the frequency of occurrence of each SC class is represented in terms of increasing darkness: areas characterized by a higher frequency are colored with light greys, whereas the reverse holds for less frequent features. The maps in the two rows report the distribution of voiceless and voiced spirantization, respectively, occurring in intervocalic context. One can see that the core area around Florence is characterized by the spirantization of both voiceless and voiced stops, whereas the area surrounding it is characterized by voiceless spirantization only.

Figure 3 - Geographic distribution of SCs involving voiceless and voiced stops and their spirantized counterpart intervocalically

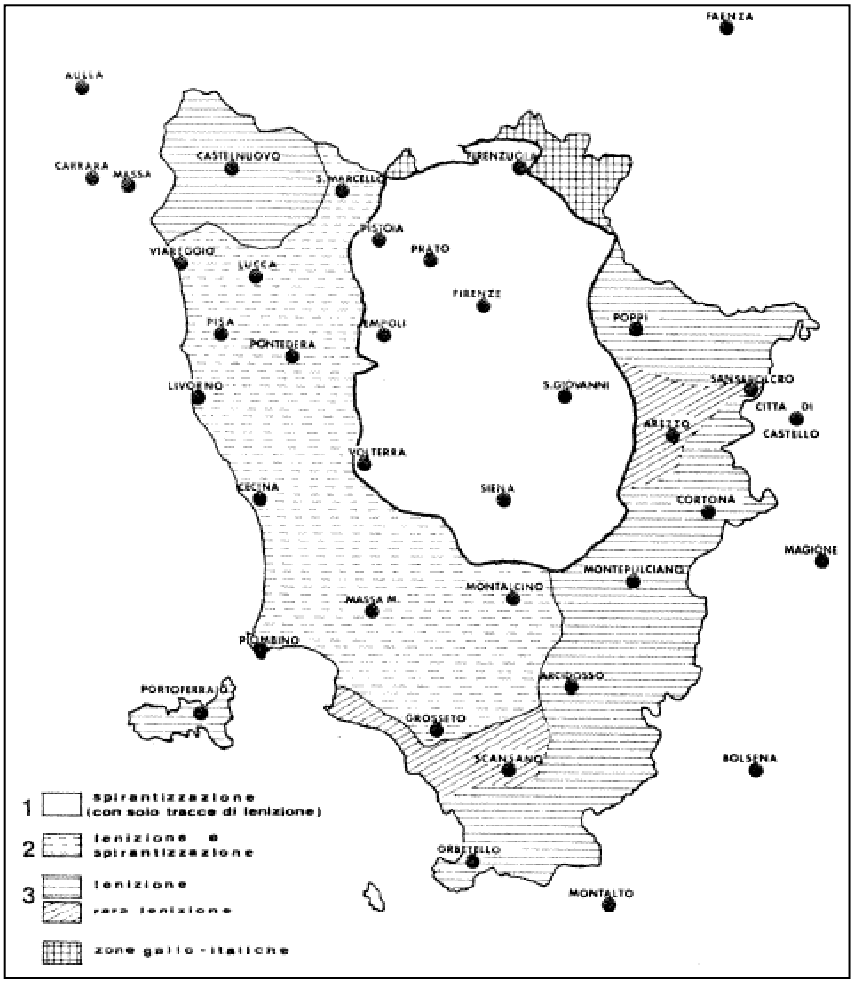


A comparative analysis of the areal distribution of these features reflects the spirantization hierarchy as reported in the literature (see Section 2):

1. voiceless spirantization (maps in the top row) is more widespread than voiced spirantization (maps in the bottom row);

2. the same holds within each row, namely /k/ > /t/ > /p/ and /g/ > /d/ > /b/: i.e. spirantization affects the velar to a greater extent than it does the dental, which in turn is affected more than the bilabial (velar > dental > bilabial).

Figure 4 - Map showing the weakening of intervocalic /k t p/ in Tuscany from Giannelli and Savoia (1978)



This global analysis of the phonetic features underlying the dialectal clusters identified is in line with the primary texts on the topic of Tuscan *gorgia* (see Section 2). Tuscan spirantization originally arose in Florence and spread rapidly in different respects: geographically, by propagating from Florence in all directions, especially southward and westward; and phonologically, by originally involving the velar stop /k/, then /p/ and /t/ up to the voiced stops /b/, /d/ and /g/. This is even more evident by comparing the maps in Figure 2 with the map in Figure 4 from Giannelli and Savoia (1978), visually showing the weakening (carried out in terms of spirantization and/or lenition/voicing) of intervocalic /k t p/ in Tuscany. The white central area corresponds to the spirantization area, whereas the contiguous area on the west side is characterized by the co-occurrence of spirantization and lenition/voicing phenomena suggesting that we are in front of two different patterns of spirantization.

5.3. Old vs. young speakers

Results described in the previous sections show that spirantization in Tuscany is still a native and vital feature, quite resistant to standardization. Giannelli and Savoia (1978, 1980) report that the recent accelerated development and spread of Florentine spirantization throughout Tuscany is increasingly typical of younger generations. This issue can be further investigated on the basis of the ALT corpus, by exploiting one of the main features of this dataset, i.e. the fact that in every location multiple speakers were interviewed (between 4 and 29) and therefore that each PV is anchored to a given location, but also to a specific speaker. To investigate whether and to what extent Tuscan spirantization also spread demographically across generations, we grouped the speakers in an old age group (born in 1930 or earlier – 1930 was the median year of birth) and a young age group (born after 1930). For every age group, we used the phonetic variant testified by the majority of the speakers in the respective group. In this case, we focused on contextualized SCs only.

It turned out that the general clustering pattern is the same across the two age groups with minor differences observed at the level of underlying features. The same typology of features underlies the major clusters for both young and old speakers, albeit with different frequencies associated with different individual features, which is therefore reflected both in the ranking and the importance score assigned to them. In particular, young speakers appear to use the most innovative SCs more than old speakers do, i.e. in the core spirantization cluster, SCs involving voiced stops /g d b/ (as opposed to voiceless ones), and in the external spirantization cluster, SCs involving /p t/ (as opposed to /k/). Different causes may be hypothesized to underly this result. First, according to Giannelli and Savoia (1978, 1980) the spreading of spirantization in young generations is particularly evident in careless, colloquial, fast speech, while ALT data were elicited on the basis of a questionnaire focused on lexico-semantic variation, so that careless, informal and emotive pronunciations are rarely reported in the ALT corpus. In fact, besides specific lexically-oriented questionnaire items, the ALT data do not systematically record phonetic differences among language registers. Second, it should also be considered that the phonetic transcription in ALT is coarse-grained, and finer distinctions between different spirantization degrees are not accounted for.

6. CONCLUSION

In this paper, we showed that the method of spectral partitioning of bipartite graphs applied to synchronic dialectal data can effectively be used to investigate diachronic phonetic processes. This was illustrated through a case study carried out on Tuscan dialects, focusing on the phenomenon of spirantization with a specific view to the so-called Tuscan *gorgia.* It turned out that a careful analysis of the sound correspondences involving voiceless and voiced stops provides truly valuable information for the reconstruction of the diachronic process of spirantization. In particular, we tracked the evolution of the spirantization phenomenon in several respects. First, we tracked spirantization geographically, across Tuscany from the influential center of Florence to the peripheral areas. Second, we tracked it phonologically, from voiceless to voiced stops, and within each voicing class from velars to dentals and then to bilabials. Finally, we tracked it demographically, with young speakers using the most innovative sound correspondences more than old speakers. The fact that these results are in line with the literature on the topic of Tuscan *gorgia* demonstrates the potential of the method of spectral partitioning of bipartite graphs with respect to the reconstruction of diachronic processes starting from diatopically distributed synchronic dialectal data. On the technical side, this study gave us the opportunity to test impact and role of a contextualised representation of SCs which led to a better founded analysis of phonetic patterns of dialectal variation.

Acknowledgements

The research reported in this paper was carried out in the framework of the Short Term Mobility program of international exchanges funded by the National Council of Research (CNR, Italy).

Bibliography

BAFILE 1997 = BAFILE LAURA, *La spirantizzazione toscana nell'ambito della teoria degli elementi*, in AA.VV. (Italian "various authors"), *Studi linguistici offerti a G. Giacomelli dagli amici e dagli allievi*, Firenze, Unipress, 1997, pp. 27–38.

CHURCH − HANKS 1990 = KENNETH CHURCH − PAUL HANKS, *Word association norms, mutual information, and lexicography*, in «Computational Linguistics», 16 (1), 1990, pp. 22-29.

CUCURULLO *et al.* 2006 = SEBASTIANA CUCURULLO − SIMONETTA MONTEMAGNI − MATILDE PAOLI − EUGENIO PICCHI − EVA SASSOLINI, *Dialectal resources on-line: the ALT-Web experience*, in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006), 24-26 May 2006, Genova, Italy, pp. 1846-1851.

LEVENSHTEIN 1965 = VLADIMIR LEVENSHTEIN, *Binary codes capable of correcting deletions, insertions and reversals*, in «Doklady Akademii Nauk SSSR», 163, 1965, pp. 845-848.

GIACOMELLI *et al.* 2000 = GABRIELLA GIACOMELLI − LUCIANO AGOSTINIANI − PATRIZIA BELLUCCI − LUCIANO GIANNELLI − SIMONETTA MONTEMAGNI − ANNALISA NESI − MATILDE PAOLI − EUGENIO PICCHI − TERESA POGGI SALANI (eds.), *Atlante Lessicale Toscano*, Lexis Progetti Editoriali, Roma, 2000.

GIANNELLI 2000 = LUCIANO GIANNELLI, *Toscana*, Pacini Editore, Pisa, 2000 (1976, first edition).

GIANNELLI − SAVOIA 1978 = LUCIANO GIANNELLI − LEONARDO M. SAVOIA, *L'indebolimento consonantico in Toscana (I.)*, in «Rivista Italiana di Dialettologia», Vol. 2, 1978, pp. 23-58.

GIANNELLI − SAVOIA 1980 = LUCIANO GIANNELLI − LEONARDO M. SAVOIA, *L'indebolimento consonantico in Toscana (II.)*, in «Rivista Italiana di Dialettologia», Vol. 4, 1979-80, pp. 38-101.

HAJEK 1996 = JOHN HAJEK, *La Gorgia Toscana and the Problem of Spirantization in Tuscany*, in J.-P. Köster (ed.), *Beiträge zur Phonetik und Linguistik 46*, University of Trier, pp. 5-60.

HALL 1949 = HALL R. A., JR., *A note on gorgia Toscana*, «Italica», 26, 1949, pp. 64–71.

IZZO 1972 = IZZO, H. J. *Tuscan and Etruscan*, University of Toronto Romance Series, 20. Toronto: University of Toronto Press, 1972.

MAROTTA 2001 = MAROTTA GIOVANNA, *Non solo spiranti. La "gorgia toscana" nel parlato di Pisa*, «L'Italia Dialettale», 62 (1), 2001, pp. 27–60.

MONTEMAGNI 2007 = SIMONETTA MONTEMAGNI, *Patterns of phonetic variation in Tuscany: using dialectometric techniques on multi-level representations of dialectal data*, in P. OSENOVA *et al.* (eds.) *Proceedings of the Workshop on Computational Phonology at RANLP-2007*, 26 September 2007, Borovetz, Bulgaria, pp. 49-60.

MONTEMAGNI 2008 = SIMONETTA MONTEMAGNI, *The space of Tuscan dialectal variation. A correlation study*, in «International Journal of Humanities and Arts Computing», Edinburgh University Press, Oct 2008, Vol. 2, No. 1-2, pp. 135-152.

MONTEMAGNI *et al.* forthcoming = SIMONETTA MONTEMAGNI - MARTIJN WIELING - BOB DE JONGE - JOHN NERBONNE, *Patterns of Language Variation and Underlying Linguistic Features: A New Dialectometric Approach*. To appear in N. de Blasi (ed.) *La variazione nell'italiano e nella sua storia. Varietà e varianti linguistiche e testuali. Proceedings volume of Congresso della Società Internazionale di Linguistica e Filologia Italiana*, (XI Congresso SILFI) Naples, Oct. 5-7, 2010.

NERBONNE *et al.* 1999 = JOHN NERBONNE − WILBERT HEERINGA − PETER KLEIWEG, *Edit Distance and Dialect Proximity*, in D. Sankoff, J. Kruskal (eds.), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Stanford, CSLI Press., pp. v-xv.

PELLEGRINI 1977 = GIOVAN BATTISTA PELLEGRINI, *Carta dei Dialetti d'Italia*, Pisa, Pacini Editore, 1977.

ROHLFS 1930 = ROHLFS G., *Vorlateinische Einflüsse in den Mundarten des heutigen Italiens?*, «Germanischromanische Monatsschrift», 18, 1930, pp. 37–56.

SORIANELLO 2001 = SORIANELLO P., *Un'analisi acustica della "gorgia" fiorentina*, «L'Italia Dialettale», 62(1), 2001, pp. 61–94.

SZMRECSANYI (submitted) = BENEDIKT SZMRECSANYI, *Forests, trees, corpora, and dialect grammars*, submitted to Szmrecsanyi and Wälchli (eds)., *Aggregating dialectology and typology: linguistic variation in text and speech, within and across languages* (working title). Berlin: de Gruyter.

WIELING *et al.* 2009 = MARTIJN WIELING − JELENA PROKIC − JOHN NERBONNE, *Evaluating the pairwise string alignment of pronunciations*, in L. BORIN, P. LENDVAI (eds.), *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education* (LaTeCH - SHELT&R 2009), EACL-2009 Workshop, Athens, 30 March 2009, pp. 26-34.

WIELING *et al.* (to appear) = MARTIJN WIELING − ELIZA MARGARETH - JOHN NERBONNE, *Inducing phonetic distances from dialect variation*, in «Computational Linguistics in the Netherlands Journal», 1, pp.109-118.

WIELING − NERBONNE 2010 = MARTIJN WIELING − JOHN NERBONNE, *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*, ACL, Uppsala, Sweden, July 16, 2010, pp. 33-41.

WIELING − NERBONNE 2011 = MARTIJN WIELING − JOHN NERBONNE, *Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features*, in «Computer Speech and Language», 25(3), 2011, pp. 700-715.