BRILL

LANGUAGE DYNAMICS AND CHANGE 14 (2024) 1–33

brill.com/ldc

Extracting Tuscan phonetic correspondences from dialect pronunciations automatically

Arne Rubehn | ORCID: 0009-0001-0888-6411
Multilingual Computational Linguistics, University of Passau, Passau, Germany
arne.rubehn@uni-passau.de

Simonetta Montemagni | ORCID: 0000-0002-2953-8619
Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa, Italy
simonetta.montemagni@ilc.cnr.it

John Nerbonne | ORCID: 0000-0002-3432-675X
Institute of Linguistics, University of Groningen, Groningen, The Netherlands; University of Freiburg, Freiburg, Germany; University of Tübingen, Tübingen, Germany
j.nerbonne@rug.nl

Received 18 March 2023 | Accepted 19 July 2024 | Published online 17 September 2024

Abstract

We present a novel approach to identifying individual pairs of phonetic correspondences in a dataset of dialect pronunciations. This continues work identifying shibboleths (i.e., characteristic features of a given dialect), a category that has interested dialectology and that dialectometrical research has examined mostly in the form of categorical data or entire phonetic transcriptions. This article reaches into segmental sequences (phonetic transcriptions) to identify individual phonetic correspondences. We follow earlier work in examining how distinctive and how representative a given phonetic correspondence is for a selected group of varieties. We proceed from string alignments, and innovate in characterizing the important notions via information theory. Despite minor problems, the method improves on the generality of competing approaches and can be shown to be useful in detecting characteristic phonetic correspondences in Tuscan varieties. We argue that this facilitates

Published with license by Koninklijke Brill BV | DOI:10.1163/22105832-bja10034

deeper investigation into the relation between aggregating approaches to dialectology and approaches proceeding from features.

Keywords

sound correspondence – dialectology – information theory – alignment – Tuscan dialects

1 Introduction

Traditional dialectology emphasized the collection of large amounts of data organized into dialect atlases. The emphasis on size was important to those early researchers because of their conviction that the phenomenon of language variation was broad and deep. To this day, the atlases still provide excellent resources for research into language variation. However, in spite of appreciating large amounts of data, early research was hampered by the absence of tools for analyzing collections involving hundreds of sites and linguistic variables, and thus tens of thousands of observations.

In contrast, published analyses always emphasized the role of *single features*, or small numbers of these. The features may be phonetic, morphological, lexical, syntactic, or otherwise, but the analysis proceeded one feature at a time.

The distribution of the individual variants with respect to geography was the focus of study, and special emphasis was given to features that geographically coincided, leading to "bundles of isoglosses." It is uncommon to see feature-based analyses that justify their conclusions using more than 10 features, and many use much smaller numbers.

An alternative research line that has become popular in recent decades is *dialectometry*, where one proceeds from the same large collection of linguistic material but uses a metric for measuring differences numerically. This step enables aggregate analyses, rising above the level of individual features. For categorical linguistic variables, a simple metric suffices: same vs. different, 0 if the same, 1 if different. For data in the form of phonetic transcriptions, edit distance is often used (Nerbonne, 2003). However the linguistic differences are measured, larger-scale analyses are based on the aggregate differences of a substantial sample.

Intuitively, it is clear that the two approaches ought to be arriving at similar conclusions. Feature-based analyses ought to resemble aggregate analyses as the number of features in the analysis increases, and dialectometrical analysis

should be able to probe the aggregates they are based on to identify elements contributing the most to aggregate differences. It is straightforward to see why this is so in the case of categorical data, where the dialectometric approach simply sums the differences of all the features recorded in a given dataset. For phonetic data, where differences are often measured using edit distance, it is more difficult to see the relation, first since edit distances compare entire words, not simple segments; second, because the comparisons also involve the insertions and deletions of phonetic segments as well as their substitutions; and third, because the operations may be weighted, for example to reflect phonetic similarity (Heeringa, 2004).

The focus of the current paper is the extraction of phonetic correspondences from analyses of dialectal pronunciation using edit distance where segment operation weights are determined using pointwise mutual information (PMI). We hope that this will be a further step toward unifying feature-based and dialectometric analyses of linguistic variation.

We proceed from a dialectometric perspective, exploring techniques for identifying features characteristic of a given region or social group. Such features are intrinsically interesting to a great many lay people, including actors and comedians. They have attracted less attention in work in the dialectometry, even though they are a popular facility in the dialectometrically inspired Gabmap web application (Leinonen et al., 2016). Our primary motivation is setting the stage for understanding the relation between dialectometry and feature-based research in dialectology, but it is also pleasing to satisfy lay curiosity.

In the rest of this introduction, we elaborate on the motivation for this work and discuss prior work, after which we expand on our intended contribution. Following the introduction there are sections on the empirical target of the work—Tuscan dialects—and the dataset we focus on, taken from the *Atlante lessicale toscano*, a lexical atlas focusing on dialectal variation throughout Tuscany, a region in central Italy. We then have sections on our methods, our results and discussion, and a brief concluding section.

1.1 *Motivation*

Dialectometry has advocated the analysis of large aggregates of material, such as all the material in the dialect atlas of a substantial area (Séguy, 1973; Nerbonne, 2009), which may involve hundreds of items sampled at hundreds of sites.[1]

1 Haag (1898) foreshadowed dialectometry in advocating quantifying over individual differences.

Older work in dialectology tended, like dialectometry, to celebrate large collections of data, and by dint of extensive study could identify individual features that characterized different speech areas, such as syllable-final /r/, largely absent in the east of England; the pronunciation of the diphthong /aɪ/, as in *high*, in American speech, largely pronounced as a monophthong /a/ in the American South; or the pronunciation of /p, t, k/ in many German words, which has given way to /p͡f, t͡s/ for /p, t/, respectively, and /x/ and /h/ for /k/, in the southern parts of the German-speaking areas. Of course, lexical and grammatical differences are likewise excellent candidates as potentially distinguishing features, but it is safe to say that pronunciation differences have enjoyed the lion's share of attention. Dialectometry has brought to this study analytical techniques to determine dialect regions exactly and statistical techniques to justify identifying certain features as characteristic.

These two traditions in dialect research—aggregate analyses and feature-based analyses—are best seen as complementary rather than contradictory. Indeed Pickl and Rumpf (2011) point out that the geographic distributions of individual features are often quite different from aggregate distributions, something noted in other quantitative work as well, such as Nerbonne (2009), where the different distributions play a role in the argument for an aggregate view.

In pursuing the goal of seeing how the two approaches might complement one another, we shall proceed from the aggregate view, because the preference for aggregate analyses has been subject to validation studies. Gooskens and Heeringa (2004) showed that aggregate distance measures correlated strongly with Norwegian dialect speakers' perception of similarity ($r = -0.78, p < 0.001$, Mantel test), and Wieling et al. (2014) demonstrated a slightly stronger correlation between native English speakers' perception of "non-nativelikeness" and an aggregate measure of pronunciation difference ($r = 0.81$).

In our work, we proceed from an aggregate analysis of dialect material, to then go on to detect characteristic elements, sometimes referred to as *shibboleths*, or variants of speech that convey information about the origin of a speaker.[2]

2 In case the term is unfamiliar, we note that it stems from a story in the Old Testament where Ephraimites were killed by their enemies when they were caught fleeing from conquered land. The Gileadites, the enemies of the Ephraimites, asked the fugitives to pronounce the word *shibboleth* when they claimed not to be Ephraimites. Unable to pronounce the initial [ʃ], the fugitives revealed themselves to be Ephraimites and thus were killed (Judges 12:5–6).

1.2 *Prior work*

Pickl (2016) analyzed inter alia phonological traits in Bavarian Swabian, proceeding from a site-by-feature matrix and analyzing it using factor analysis. This results in sets of sites sharing the same feature value, but it requires manual coding of the feature values. The factors do group sites, but correspond at best to very low-level dialect areas (Pickl refers to "types"). Sixteen factors accounted for only 62.2 % of the variance.

Wieling and Nerbonne (2011) used bipartite spectral graph partitioning, which automatically associates groups of data collection sites with groups of features, where the authors extracted phonetic correspondences obtained by applying an edit distance algorithm. This paper focused on a comparison between the standard on the one hand and all the dialects on the other. The paper also introduced the notions of *representativeness* and *distinctiveness* to evaluate the success of the attempt to find features characteristic of a given set of varieties. Features are representative when they tend to occur throughout a selected set of the varieties in question and distinctive when they tend to occur only there. Exact definitions are provided below.

Other approaches have separated the determination of groups in data, which relies on aggregate properties (e.g., based on mean differences), from finding characteristic elements in those groups, which can be performed in a separate step.

Montemagni et al. (2012) applied the spectral partitioning method to a matrix of site by contextualized phonetic correspondences in Tuscan dialects, where phonetic correspondences were weighted by their relative frequencies in the data. Representativeness and distinctiveness were used to select the most important correspondences. Montemagni et al. (2013) focused on spirantization phenomena (including the famous *gorgia toscana*) and compared the results that were obtained with and without contextual information. The latter represents the work most similar to our own in tackling the problem of detecting phonetic correspondences, and we generalize on that work by not requiring that researchers specify a phonetic context within which correspondences may be gathered and studied and by separating the determination of the groups from the identification of characteristic elements.

Our procedure will look more widely, admittedly sacrificing sensitivity for coverage. We examine *all* corresponding sound pairs observed in multiple sequence alignments. Since Montemagni et al. (2013) wished to examine the effect of context, the added specification made sense there, but it also restricted the generality of their results.

Prokić and Nerbonne (2013) generalized the search for characteristic elements by using *all* the sound correspondences found in a large set of Bulgarian,

obviating the need to first select the correspondences of interest. They focused on correspondence in alignment positions, which yielded many-to-many pairs of correspondence sets, less precise than our target. They then examined the geographic distributions of the most representative and distinctive correspondences in each of three areas in Bulgaria and showed that these indeed coincide with the dialect areas. We offer an alternative mathematical point of departure in this paper, focusing on corresponding sounds instead of sets of sounds, and move to another language area, Tuscany.

Sound correspondences also play a major part in historical linguistics, where scholars aim to detect regular correspondence patterns in order to detect cognates and infer sound laws. The manual application of the established comparative method, however, is a cumbersome task, which is why substantial advancements have been made to automate different parts of the method (List, 2014, forthcoming). For this purpose, several techniques for calculating phonetic similarity between sounds (and by extension word-forms) have been developed, ranging from parameterized edit distances (Hall and Klein, 2010; Hauer and Kondrak, 2011) over language-specific scoring schemes (List, 2014; Rama, 2016) to adaptations of pointwise mutual information (Jäger, 2013; Jäger et al., 2017; Dellert, 2018)—mostly with the goal of detecting cognates in multilingual word lists.

Identifying regular correspondence patterns is a more challenging task, and only a small number of techniques have been proposed for that. Kondrak (2003, 2009) extracted pairwise correspondences from pairwise alignments, again aiming at improving cognate clustering. List (2019) employed clique-covering techniques to identify multilingual sound correspondence patterns that can enhance reflex prediction (Bodt and List, 2022) and supervised phonological reconstruction (List et al., 2022b). Daneyko (2020) employed probabilistic soft logic (Bach et al., 2017) to subsequently detect sound correspondences and infer sound laws for multiple related languages.

1.3 *Intended contributions*

We aim to detect segment-based shibboleths. Instead of looking for distinctive words, we aim to automatically find *phonetic correspondences* that are characteristic for a given group of varieties. The method will employ the representativeness and distinctiveness of a given phonetic correspondence pair using PMI (Church and Hanks, 1990). In addition to extending the methodology aimed at finding characteristic elements in variationist linguistics to sound segments, we also propose a revision of the treatment of representativeness in an effort to improve performance.

Throughout this work, we use the term "phonetic correspondences" to describe arbitrary sound pairs for which we are able to quantify the characteris-

ticness with respect to a chosen area or dialect cluster. A phonetic correspondence between two sounds *i* and *j* is characteristic if varieties inside the cluster frequently use *i* in positions where varieties outside the cluster use *j*.

The intuitive interpretation of this measurement is that the usage of *i* instead of *j* is very characteristic for speakers of a defined dialect cluster if the calculated characteristicness value is high. However, since we do not make use of underlying representations or notions of standard language, and only operate on surface-level phonetic transcriptions, this metric does not attempt to quantify how certain varieties deviate from an underlying or standard pronunciation. Instead, all surface forms are treated equally, without any notion of directionality. The characteristicness metric should therefore not be interpreted as "speakers of these dialects say *i* instead of *j*," but rather along the lines of "some underlying position tends to surface as *i* among the defined dialects and as *j* among the other dialects."

We would also like to add that the term "correspondence" is used differently in historical linguistics compared with its use here (and elsewhere) in dialectology. Both linguistic subdisciplines seek regular sound correspondence patterns—the dialectologist as (imperfect) indicators of provenance, and the historical linguist as indicators of sound shifts. The dialectologist does not expect correspondences to be perfectly regular, while the historical linguist sees apparent exceptions as reasons to seek refinements in the correspondence, perhaps involving contextual details or the influence of later developments. We do not attempt to discover a set of perfectly regular correspondence patterns that might (nearly) exemplify the Neogrammarian regularity of sound shift. Instead, we draw an aggregate picture of which sounds *tend* to be pronounced in a characteristic way with respect to certain dialects, or which phonetic variations coincide with a bipartition of a dialect continuum.

We test our work by looking for phonetic correspondences in the dialectal corpus of the *Atlante lessicale toscano* (Lexical atlas of Tuscany), and we will regard the work as successful if we detect correspondences that have been noted in previous studies, both traditional and dialectometric.

2 The data

2.1 *Atlante lessicale toscano*

The data that is used in this study comes from the *Atlante lessicale toscano* (Giacomelli et al., 2000), which from now on will be referred to as ALT. As the name suggests, ALT is a linguistic atlas that contains lexical (*lessicale*) data for dialectal varieties spoken in Tuscany (*Toscano*). The corpus was primarily collected

in order to account for lexico-semantic variation, but it still contains valuable information about phonetic variation within the speech varieties (Montemagni et al., 2013). The qualitative adjective *toscano* in the title refers to the political entity of Tuscany rather than to the linguistic classification of Tuscan varieties; the ALT therefore also contains linguistically non-Tuscan varieties which are spoken in peripheral areas of Tuscany.

The ALT data were collected between 1974 and 1986, resulting in about two million individual responses from 2,193 speakers differentiated with respect to age, socioeconomic status, education, and culture who were each asked 745 questions. The entire ALT corpus was compacted into about 380,000 database entries, of which about 350,000 contain georeferenced responses to the questionnaire items, and about 30,000 record additional dialectal items that emerged during the interviews.

In the ALT, each dialectal item is assigned different levels of representation organized in layers of progressively decreasing detail, going from phonetic transcription to the levels of orthographic and normalized representations (Cucurullo et al., 2006). The phonetic representation used in the ALT project was a geographically specialized version of the transcription system from the "Carta dei dialetti italiani" (Grassi et al., 1997). The criteria which guided the definition of the ALT normalized representation level permit one to focus on phonetic/phonological variance (which is the subject of this study), without interference from any other linguistic description level (e.g., morphology) which would produce noise. To illustrate this more concretely, phonetic variants originating from productive phonetic processes are assigned the same normalized form (e.g., [skjaˈtʃːata], [skjaˈtʃːaθa], and [skjaˈtʃːada] are all assigned the same normalized representation, *schiacciata*), whereas, for example, singular vs. plural word-forms are assigned distinct normalized forms (i.e., [skjaˈtʃːata] and [skjaˈtʃːate] are assigned different normalized forms, *schiacciata* and *schiacciate*, respectively).

2.2 *The experimental dataset*

For this study, the data from the ALT was normalized and reduced in some dimensions in order to make the experimental setup more straightforward. In particular, this means:

- Non-Tuscan varieties were excluded from the experimental dataset. The network of locations investigated was restricted to the 213 locations (out of 224) where Tuscan dialects are spoken.
- Phonetically transcribed data were automatically converted to the International Phonetic Alphabet (IPA; Wieling et al. 2016).
- Due to the alignment of the different ALT representation levels (see Section

2.1), the dataset used in this study was formed by automatically extracting, for each attested normalized form (henceforth NF), the set of associated phonetic variants (henceforth PVs).
– When multiple phonetic realizations (PVs) of the same NF were attested in the same location, the most frequent one was chosen.
– Only nouns and adjectives were chosen. Most of them were single words; a few were parts of frozen multi-word expressions.
– The areal coverage of selected NFs was required to be greater than or equal to 100 locations (of the 213 locations).

The resulting experimental dataset built as described above includes 444 NFs, associated with 68,740 different PVs (types) and 448,487 PV tokens. This derived dataset was first used by Montemagni et al. (2012, 2013) in order to study patterns of phonetic variation; in contrast to our study, they retained multiple PVs per site and NF. In order to assess the representativeness of this subset with respect to the whole set of NFs having at least two PVs attested in at least two locations, they measured the correlation between phonetic distances in the overall dataset (Montemagni, 2008) and the selected sample, which turned out to be nearly perfect ($r = 0.994$). Other datasets have been used in a comparable manner in order to obtain data that only contains phonetic variation, without the additional variation caused by morphology, syntax, or the lexicon (Wieling and Nerbonne, 2011; Prokić et al., 2012).

We wish to adhere to the FAIR principles (Wilkinson et al., 2016) for managing scientific data, making it findable, accessible, interoperable, and reusable. We therefore formatted our data according to the CLDF (Cross-Linguistic Data Formats; Forkel et al. 2018) standards and published it as part of Lexibank (List et al., 2022a), a large collection of (retro-)standardized lexical data. By employing those data standards, the published data is conveniently linked to reference catalogs such as Glottolog (Hammarström et al., 2023) and Concepticon (List et al., 2016).

2.3 *Gorgia toscana*

The *gorgia toscana* (literally 'Tuscan throat') refers to a prominent phonological process in Tuscan dialects which may be expected to play a major role in the present study, so we discuss it here in a focused way. In a strict sense, it refers only to the lenition of the unvoiced velar stop /k/ to an [x] or [h] (hence the name), but in a wider sense, analogical lenition processes of the other voiceless plosives (/t/ > [θ]; /p/ > [ɸ]) belong to *gorgia* as well (Hall, 1949). The voiced stops /b/, /d/, and /g/ are also affected by the lenition, producing the surface forms [β], [ð], and [ɣ/ɦ] (Giannelli and Savoia, 1978; Marotta, 2001, 2008; Sorianello, 2001); in addition to that, Binazzi (2019) and Marotta

TABLE 1 Examples for *gorgia*

Spelling	Standard Italian	Tuscan	Gloss
pepe	/ˈpepe/	[ˈpeɸe]	'pepper'
i piedi	/i ˈpjɛdi/	[i ˈɸjɛdi]	'the feet'
schiacciato	/skjaˈtʃaːto/	[skjaˈtʃaːθo], [skjaˈtʃaːo]	'flat'
la torta	/la ˈtorta/	[la ˈθorta]	'the cake'
amico	/aˈmiko/	[aˈmixo], [aˈmiho], [aˈmio]	'friend'
la casa	/la ˈkaza/	[la ˈxaza], [la ˈhaza]	'the house'
abete	/aˈbete/	[aˈβete]	'fir'
e beve	/e ˈbeve/	[e ˈβeve]	'(s)he drinks'
dado	/ˈdado/	[ˈdaðo]	'dice'
e dorme	/e ˈdɔrme/	[e ˈðɔrme]	'(s)he sleeps'
lago	/ˈlago/	[ˈlaɣo]	'lake'
la gamba	/la ˈgamba/	[la ˈɣamba]	'the leg'
ceci	/ˈtʃetʃi/	[ˈtʃeʃi]	'chickpeas'
la cena	/la ˈtʃena/	[la ˈʃena]	'the dinner'
grigio	/ˈgridʒo/	[ˈgriʒo]	'gray'
la gente	/la ˈdʒente/	[la ˈʒente]	'the people'

MONTEMAGNI ET AL., 2013; BINAZZI, 2019

(2008) note that the affricates /t͡ʃ/ and /d͡ʒ/ tend to lose their occlusive parts and become [ʃ] and [ʒ], respectively, a process we will refer to as deaffrication. These lenitions occur only on short stops—long stops remain unaffected (Hall, 1949)—intervocalically or between a vowel and a glide or a liquid. *Gorgia* also applies across word boundaries: /la ˈtorta/ 'the cake' surfaces as [la ˈθorta] in Tuscan (Montemagni et al., 2013).

Table 1 shows examples for each lenition process belonging to Tuscan *gorgia*. It is notable that the lenitions of /k/ and /t/ can even result in complete deletions. This stands in direct relation to the observation that there seems to be a hierarchy among the places of articulation: velar stops are affected the most, while dental stops are less affected than velar stops but more affected than labial stops (Hall, 1949). Another dimension of this hierarchy is that voiceless stops are affected more strongly than their voiced counterparts. The lenition of affricates to fricatives seems to be the least common among the *gorgia* processes discussed.

Research on Tuscan dialectology generally agrees that the *gorgia* phenomenon spread from Florence to the rest of the Tuscan varieties; see Fig. 1 to

FIGURE 1 Physical map of Tuscany with labeled provinces

visualize the spreading of this phenomenon across the Tuscan territory. The areas of Florence, Siena, Prato, and Pistoia are home to the varieties where the lenition has the strongest effect, affecting all voiced and voiceless stops. Further westward, in the areas of Lucca, Pisa, and Grosseto, it is primarily the voiceless stops that are affected by *gorgia*, and in the eastern part of Tuscany, the area of Arezzo, only the /k/ is weakened (Giannelli, 2000; Binazzi, 2019). In their dialectometric study, Montemagni et al. (2013) confirm the picture emerging from traditional dialectology and show that the spreading of Tuscan *gorgia* from Florence to the rest of the Tuscan varieties assumes a wave-like form; moreover, they observe that, for both voiceless and voiced stops, the phenomenon affects the velars to a greater extent than it does the dentals, which in turn are affected more than the bilabials (velar > dental > bilabial).

We shall keep *gorgia* in mind in examining the result of looking for phonetic correspondences. If the work is sound, we predict that correspondences due to *gorgia* will be found. But it is important to note here that other significant correspondences may also exist in the data. We will not interpret the detection of other differences as counterindications, but will rather attempt to interpret them as separate features of genuine dialectal variation. We return to the problem of proposing a rigorous evaluation in the discussion.

3 Methodology

Following the approach of Wieling and Nerbonne (2011), which was also used by Prokić et al. (2012) in order to find shibboleths on a word level, one proceeds from a collection of varieties, for example a dialect area as represented by a subset of the data collection sites that is then compared to all the remaining sites. The goal is then to calculate a given feature's representativeness and distinctiveness with respect to the set of varieties in the bipartite division. More specifically, in this paper we calculate these values for the phonetic correspondences between the two subsets. This is the first of the two innovative contributions we wish to make in the paper.

The entire dataset is defined as U, which contains a set of sites S, a set of normalized forms NF, and a set of phonetic variants $PV_{NF,S}$, which in turn contains a variant $pv_{nf,s}$ for each $nf \in NF$ and each $s \in S$. Additionally, U contains P, a set of IPA symbols that can appear in the elements of $PV_{NF,S}$.

$$U := \{S, NF, PV_{NF,S}, P\}$$

The set of sites constituting the postulated dialect area, which is the basis of the comparison, is symbolized as V, where $V \subset S$. All other sites are contained in $W := S \setminus V$.

We introduced distinctiveness informally above, saying that a feature is to be regarded as distinctive with respect to a subset of sites if it tends to occur only there. Likewise, a feature is representative if it is systematically shared among the sites in question. In order to calculate distinctiveness, a direct comparison between elements of V and W is required, while a comparison of sites within a cluster is necessary for calculating representativeness.

3.1 *Alignments and correspondences*

As the basic unit for sequence comparison, we construct multiple sequence alignments for all phonetic variants $pv \in PV$ that correspond to the same nor-

malized form *nf* ∈ *NF*. Multiple sequence alignments were generated automatically using the LingPy (List and Forkel, 2021) implementation of the Sound Class Algorithm (SCA; List 2012). An example of such a multiple sequence alignment (reduced to unique phonetic variants) for the normalized form *albicocca* 'apricot' is:

a	l	b	i	k	ɔ	kː	a	Anghiari (+71 more sites)
a	l	b	i	h	ɔ	kː	a	Riparbella (+71 more sites)
a	r	b	i	h	ɔ	kː	a	Cecina (+25 more sites)
a	l	b	i	x	ɔ	kː	a	Caprese Michelangelo (+19 more sites)
a	r	b	i	k	ɔ	kː	a	Stia (+5 more sites)
a	-	bː	i	h	ɔ	kː	a	Nusenna (+4 more sites)
a	l	b	i	-	ɔ	kː	a	Antignano (+2 more sites)
a	r	b	i	-	ɔ	kː	a	Rosignano Marittimo (+1 more site)
a	r	b	i	g	ɔ	kː	a	Montemignaio
a	-	b	i	k	ɔ	kː	a	Villa Basilica
a	l	b	i	h	o	kː	a	Vaglia

The sequence alignment thus reflects which sounds of the individual forms correspond to each other. For example, in the second column of the exemplary alignment, one can see how [l] in the Anghiari variety corresponds to [r] and [-] in the varieties spoken in Cecina and Nusenna, respectively. Likewise, the fifth column shows how [k], [x], [h], [g], and [-] correspond to each other in different varieties at this position of the word. Since only full cognate forms are aligned to each other, we can conclude that these sounds are phonetic variants of the same underlying phoneme (or, in a more historical linguistic framing, reflexes of the same proto-sound).

Based on these alignments, we want to count how often certain sounds correspond to each other with respect to the defined clusters. As stated previously, distinctiveness is based on comparing the forms between *V* and *W*, while representativeness is calculated from comparing forms within *V* and *W* respectively. Therefore, we set up three confusion matrices, in which the respective sound pairs are stored: one cross-cluster matrix that represents how often sounds in *V* correspond to sounds in *W*, and two cluster-internal matrices representing how often sounds within *V* and *W* respectively are aligned to each other.

Since we are not interested in full correspondence patterns, but in characteristic sound pairs between the two defined clusters, we require a pairwise representation of the multiple sequence alignment and its sounds. Therefore, we compare each form pair with respect to the full alignment and count the occurring sound pairs. For each aligned form pair, every column represents a sound

pair which is stored in the respective confusion matrix. Naturally, the correct matrix must be chosen with respect to the two varieties the forms belong to: if and only if one of the varieties is defined inside V and the other one is not, then the correspondences are counted in the cross-cluster matrix; if both varieties belong to the same cluster, the correspondences are stored in the respective cluster-internal matrix.

As a result, only the cross-cluster matrix is directed, while the two cluster-internal matrices are symmetric—if a sound i corresponds to another sound j within the cluster, that means that the inverse statement is also true, that is, j also corresponds to i. This is not the case when comparing the two clusters V and W—if i in V corresponds to j in W, that does not imply that j in V corresponds to i in W.

The resulting matrices are indexed by the phonetic symbols defined in P and contain information about how often pairs of sounds correspond to each other, with respect to the user-defined variety clusters. These matrices resemble produced/perceived matrices used in perceptual studies since Miller and Nicely (1955), which have come to be known as "confusion matrices," a convention we follow.

3.2 *Analyzing correspondences*

Based on these three confusion matrices, it is now possible to calculate distinctiveness and representativeness with the help of PMI. Positive PMI values indicate that two events tend to occur together more than one would expect by chance, while a negative value signifies that the two events tend not to. A PMI of 0 indicates that the events are statistically independent.

$$\mathrm{PMI}(x,y) = \ln\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

It should be clear that since PMI is based on the probabilities of occurrences and co-occurrences (naturally estimated based on frequencies), it is a distribution statistic, reflecting the properties of segments and their correspondences across entire samples. The relation to earlier formulations of distinctiveness (Wieling and Nerbonne, 2011) inter alia is fairly simple: if x within the area of study corresponds frequently to y outside that area, it will count as distinctive to the degree that the frequency of its joint occurrence exceeds what one might expect by chance.

$\mathrm{PMI}(x,y)$ can be normalized via dividing it by its joint self information $h(x,y)$ (Bouma, 2009). This yields values in a range between −1 and 1, where 1 signifies complete co-occurrence, and likewise −1 indicates that x and y never occur together.

$$\text{NPMI}(x,y) = \frac{\text{PMI}(x,y)}{h(x,y)}, \text{where} h(x,y) = -\ln(x,y)$$

We note that PMI and NPMI are symmetric measures, since $\text{PMI}(x,y) = \text{PMI}(y,x)$ and likewise $\text{NPMI}(x,y) = \text{NPMI}(y,x)$. We aim to develop an asymmetric measure of distinctiveness, one that, for example, will allow a lenited stop in a central area of *gorgia* to be distinctive when compared to an unlenited stop in a less central area.

We therefore define the distinctiveness (d) of a phonetic correspondence $i : j$ as the difference between $\text{NPMI}(i_V, j_W)$ and the NPMI of the inverted phonetic correspondence $j : i$. It is important to note that we are comparing (i_V, j_W) not to the simple inversion (j_W, i_V) but rather to (j_V, i_W). This yields systematic phonetic correspondences between V and W rather than pairs of sounds that tend often to be confused in both directions, such as a case of free variation between i and j that is statistically independent from the clusters V and W. This might be something like the distinction between [t] and [t̚] in word final position in English. Intuitively, that would mean that i and j would frequently correspond to each other, yielding high PMI scores. But if the distribution of i and j is independent of the samples, both $\text{PMI}(i_V, j_W)$ and $\text{PMI}(j_V, i_W)$ would be similarly high, yielding a $d(i : j)$ around 0—an indicator for statistical independence. We only want to assign high distinctiveness to sound pairs where systematic correspondences between sound $i \in V$ and sound $j \in W$ can be observed, that is, when the investigated correlation occurs frequently while the inverse correlation is rare. This leads us to formalize "distinctive" as follows:

$$d(i : j) = \text{NPMI}(i_V, j_W) - max(0, \text{NPMI}(j_V, i_W))$$

In some cases, an NPMI value for the inverted correspondence cannot be obtained because there are no instances of the inverted relation. That means either (a) that the phonetic correspondence is perfect and exceptionless and therefore only exists in one direction—which is highly unlikely when working with an even moderately large database; or (b) that the phonetic correspondence in general is very rare, because at least one (but maybe even both) of the sounds involved is very rare in the database, which will also lead to a very low representativeness value. In these cases, we define $d(i : j)$ as simply $\text{NPMI}(i_V, j_W)$, since, if the correspondence is unilateral, then there is no need to subtract anything. In the latter (more likely) case, the marginal representativeness will compensate for the high distinctiveness value.

We also note that we calculate $d(i : j)$ only if $\text{NPMI}(i_V, j_W) > 0$. We use this restriction since the difference between two negative NPMI values will yield a positive result in one of the two directions, which is misleading since neg-

ative correlation (i.e., a pair of sounds that only rarely correspond to each other) is not an interesting phonetic correspondence in language variation studies, our focus. Since the correspondences in question are not relevant for this matter, they will be ignored. Clipping negative PMI values at 0 is also known as positive pointwise mutual information (PPMI) and is a common practice that has been shown to work well for various tasks in computational linguistics (Salle and Villavicencio, 2023; Jurafsky and Martin, 2023: § 6.6).

The second important value to be calculated is representativeness (r), which measures how consistently a given sound i is used in the chosen sample of sites V. If i corresponds to itself frequently within the sample of words in V, it is representative; if it corresponds to other sounds more often, the value for r is accordingly lower. Note that this refines earlier formulations by adding a condition. Representativeness is therefore defined as the NPMI of a sound corresponding to itself in V:

$$r(i, V) = \text{NPMI}(i_V, i_V), \text{ where we abbreviate } r(i, V) \text{ as } r_V(i)$$

We need to condition the notion of representativeness to representativeness within a set, so an argument is added. The NPMI function is defined on segments with implicit reference to a set. Given the fact that we proceed from a confusion matrix of corresponding segments, we do not need to explicitly mention that representative items are in corresponding positions in alignments.

Not only should the alignment of i to itself within V be consistent, but the same should also be true for j within W, where $r_W(j)$ is calculated analogically. This is a refinement of formulations in earlier works. The joint representativeness $r_V(i : j)$ is calculated as the harmonic mean between $r_V(i)$ and $r_W(j)$, where $W = U \setminus V$, as usual:

$$r(i : j) = 2 * \frac{r_V(i) * r_W(j)}{r_V(i) + r_W(j)}$$

It might seem unusual to measure the representativeness both inside and outside the cluster, especially given that previous studies have limited this measure to the observed sites. These studies, however, all used an underlying standard form as a reference point for dialectal variation—a piece of information that we disregard on purpose in order to design a method that can measure variation on phonetic surface forms alone. As stated in Section 1.3, our goal is to detect sound pairs that best describe a given bipartition of the dialect spectrum—and therefore, how regularly i_V and j_W correspond to each other.

Since we aim at detecting those characteristic sound pairs, we need to ensure that both parts are being used consistently within their respective clusters likewise.

$d_V(i:j)$ and $r_W(i:j)$ can then be combined to calculate how characteristic $i:j$ is with respect to the subset of varieties under examination. Wieling (2012) notes that representativeness is analogical to recall in information retrieval, and likewise distinctiveness is analogical to precision. Following this analogy, we calculate how characteristic a phonetic correspondence is via an F-score, the harmonic mean of the two component values:

$$c_V(i:j) = 2 * \frac{d_V(i:j) * r_V(i:j)}{d_V(i:j) + r_V(i:j)}$$

$c_V(i:j)$ will be the main value to measure how characteristic a given phonetic correspondence is for the chosen subset of varieties.

3.3 *Attempts at innovation*

Above we have sporadically compared our work to earlier work (Wieling and Nerbonne, 2011; Prokić et al., 2012; Montemagni et al., 2013), because we were concerned that digressions might diminish clarity, but we do wish to emphasize a few points where this paper is innovative in its method.

First, the earlier papers measured the representativeness of a feature in the area being studied by checking the proportion of sites in which it occurs, and they measured distinctiveness by checking the degree to which a feature tended to occur in the given area as opposed to how often it occurred in general (Prokić et al., 2012: p. 73).

Second, while the earlier papers corrected for the overall popularity of the feature within the area under examination, those papers were based on frequencies of single realizations. Prokić and Nerbonne (2013) investigated sounds as characteristic features for defined dialect clusters; however, their metric is calculated only for single structural positions, that is, per column in an alignment. Their method thus detects characteristic sounds with respect to a specific position in a certain word, and not with respect to all the data. We operate instead consistently at the level of correspondences $\langle x, y \rangle$, which is why we also examine the consistency of x appearing in the relevant area as well as the consistency of y appearing outside that area.

Further novelties can be observed with respect to the dialectometric study of Montemagni et al. (2013), where sound correspondences are established between the dialectal variant and the corresponding standard pronunciation, while we compare the surface dialectal forms to each other directly. Moreover, the dialectal areas considered do not result from clustering but rather repre-

sent the starting point, that is, the predefined sets of sites for which underlying features need to be discovered and weighted.

4 Results and discussion

We examined several subsets from the ALT sites to find the most characteristic phonetic correspondences in a proposed subset (dialect area), comparing our detection of phonetic correspondences where possible (where they have been noted elsewhere).

Ideally, one might compare the results obtained here with a consensual gold standard for a relevant set of dialect data, but we know of no such dataset. Lacking that, we compare the results to those of earlier dialectology.

4.1 *Characteristic phonetic correspondences*

We first discuss the subdivision proposed by Giannelli and Savoia (1979), who divide Tuscan dialects into three groups according to the phonetic effects of consonantal weakening: namely, patterns of intervocalic stop spirantization and voicing (which they call "lenition," following the Italian dialectological tradition). Group 1 covers the provinces of Florence, Siena, Prato, and Pistoia, where intervocalic voiceless spirantization is at its most developed stage and voicing appears to be only a marginal phenomenon. In Group 2, covering the northwestern areas of Tuscany (i.e., the provinces of Lucca, Pisa, and Livorno) and the south (Grosseto), there is a significant retention of voiced forms, alongside the frequent use of unlenited and spirantized stops. Here, the spirantization phenomenon is less advanced than in Group 1. Group 3, covering the Tuscan borders from the northern area to the east (i.e., Arezzo province) and from the south, where voicing is rather rare and where voiceless spirantization represents a more recent and still restricted phenomenon. The map provided in Fig. 2 shows the geographic distribution of the three groups. According to Giannelli and Savoia (1979), this situation originates from the historical co-occurrence of two distinct consonantal weakening phenomena, namely spirantization and voicing of voiceless stops, with the former being a more recent and still vital phenomenon showing significant distributional differences throughout the Tuscan territory.

Similar distributional patterns are reported by Pellegrini (1977), where the central area is delimited by isogloss 15 corresponding to voiceless spirantization, and the western area is marked by isogloss 16 where voiceless spirantization is reported to be more recent and still less systematic (see Fig. 3). The other isoglosses delimiting the Tuscan marginal areas correspond to other types of phenomena.

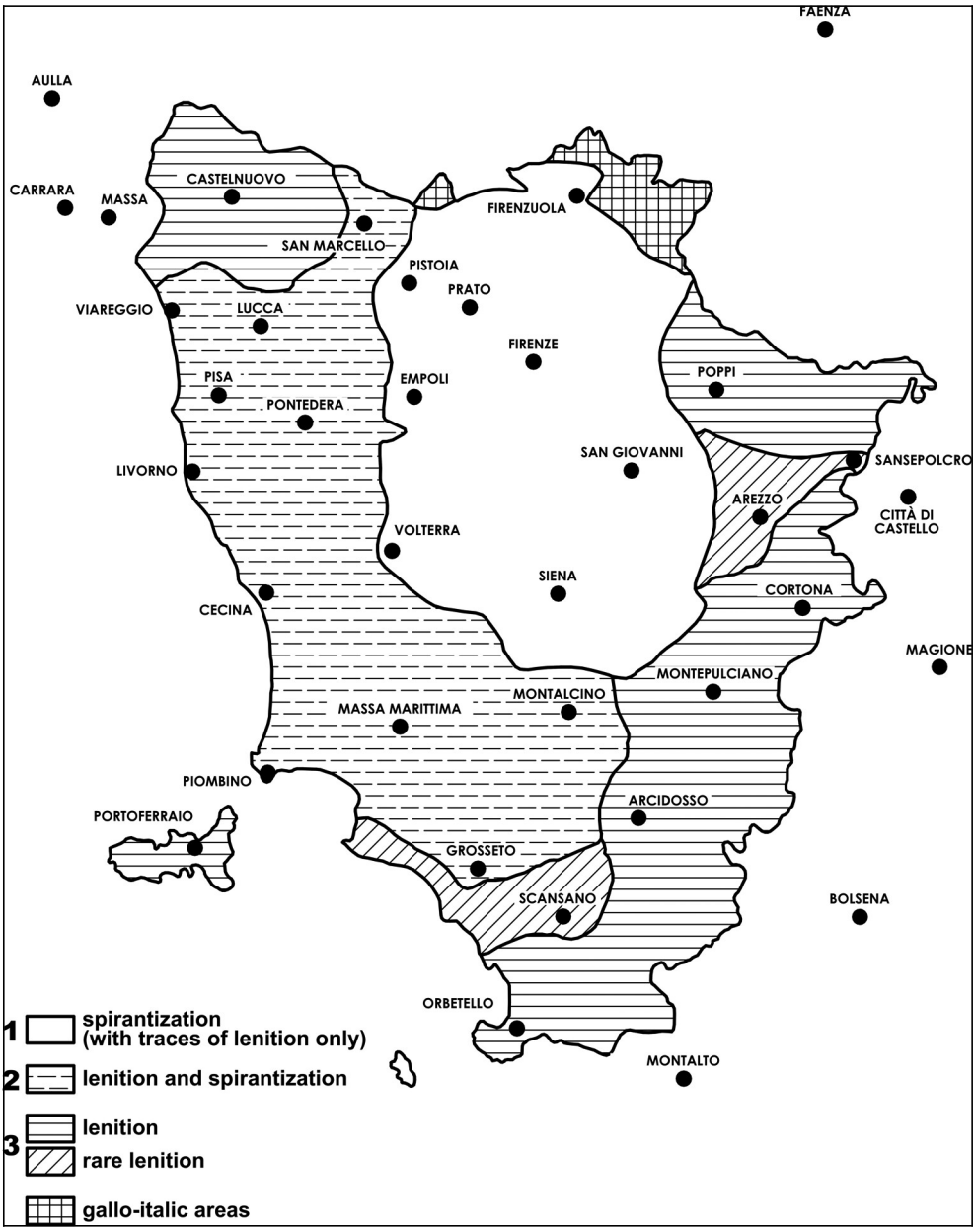

FIGURE 2 Classification of Tuscan dialects according to Giannelli and Savoia (1979) based on the phonetic effects of consonantal weakening (spirantization vs. voicing of voiceless stops)

FIGURE 3 Classification of Tuscan dialects in the map of the dialects of Italy in Pellegrini (1977)

In what follows we illustrate the results we obtained for the dialectal areas proposed by Giannelli and Savoia (1979). Note that the results were filtered in order to reduce the noise represented by too infrequent correspondences: in particular, we only focused on phonetic correspondences involving sounds occurring at least 50 times in the whole dataset. This is an arbitrary and heuristic threshold, which was motivated by the relatively extended areas we decided to focus on (ranging between 60 and 73 locations) on the one hand, and the fact that we do not account for the frequency of the correspondences on the other hand. Since our goal was to identify the characteristic phonetic correspondences for

TABLE 2 Top 10 characteristic phonetic correspondences (Corr.) for Area 1 and the union of Areas 1 & 2 from Fig. 2, with their values for distinctiveness (Dist.), representativeness (Repr.), and characteristicness (Char.)

Area 1				Areas 1 & 2			
Corr.	**Dist.**	**Repr.**	**Char.**	**Corr.**	**Dist.**	**Repr.**	**Char.**
[ɲː]:[ɲ]	0.515	0.753	0.612	[h]:[k]	0.406	0.919	0.563
[θ]:[t]	0.404	0.926	0.563	[θ]:[t]	0.379	0.902	0.534
[sː]:[t͡s]	0.362	0.975	0.528	[ɸ]:[p]	0.305	0.916	0.458
[ɸ]:[p]	0.332	0.947	0.492	[ɲː]:[ɲ]	0.323	0.773	0.456
[h]:[k]	0.298	0.932	0.452	[-]:[h]	0.300	0.739	0.427
[t͡ʃ]:[ʃ]	0.221	0.931	0.357	[ð]:[d]	0.280	0.828	0.419
[d͡ʒ]:[ʒ]	0.203	0.957	0.335	[-]:[x]	0.294	0.685	0.412
[ɣ]:[g]	0.179	0.868	0.297	[ɸ]:[β]	0.271	0.665	0.385
[ð]:[d]	0.178	0.848	0.295	[x]:[k]	0.260	0.687	0.377
[x]:[k]	0.185	0.691	0.292	[d͡z]:[z]	0.228	0.934	0.366

a certain area, we needed to disregard highly distinctive ones whose coverage was restricted to one or very few locations. In other words, we filtered those correspondences which were too infrequent to label them as characteristic.

Let's start by illustrating the results achieved for both the area around Florence (Area 1 in the map in Fig. 2) and the combined areas including Florence and the northwestern coastal areas (Areas 1 & 2 in the map), reported in Table 2. As we explained above, the notion of phonetic correspondence—as we have defined it—is directed, so it is worth noting that when we write [*i*] : [*j*] with respect to a given area, we intend that *i* in the area corresponds to *j* outside it.

The left half of Table 2 shows that spirantization of stops is a commonplace phenomenon in Area 1: among the top 10 correspondences, we observe pairs involving spirantization of all voiceless stops (i.e., /t/, /p/, and /k/). Immediately after, between positions 8 and 14 (not shown in table), we also find the spirantized outcome of all the voiced stops (relatively ordered as follows: [ɣ]:[g], [ð]:[d], and [β]:[b]). Interestingly, in this area deaffrication (spirantization) does not appear to affect affricates: the affricate realization is highly ranked, that is, it is among the top 10 correspondences (see [t͡ʃ]:[ʃ] and [d͡ʒ]:[ʒ]). On the other hand, the reverse correspondence pairs, with deaffrication of postalveolar affricates, appear at the bottom of the list (at positions 81 and 82 of 85), where the initial part of the affricate has been lost.

Among the top 10 correspondences obtained for Area 2 (not listed separately), there are four pairs related to the *gorgia* phenomenon, all showing different spirantized outcomes of /k/, going from [-], through [h], to [x]: Binazzi (2019) reports this as a typical outcome of the velar stop in this area. Interestingly, the spirantized outcomes of voiced and voiceless stops (other than /k/) are ranked lower in the list compared to the central area. This peculiar distribution along the ranked list of correspondence pairs can be seen as a sign of the instability of the area with respect to spirantization, but must mainly be attributed to the fact that it is compared to Area 1 and Area 3 at the same time, lumping together varieties with particularly strong and weak degrees of spirantization respectively. *Gorgia* is therefore not a distinctive feature that would tell a speaker from Pisa apart from both a speaker from Florence and a speaker from Arezzo at the same time. Comparing varieties from Area 2 only to those from Area 3, excluding Area 1 (also not presented in a table), however, reveals the spirantization of voiceless stops as the most characteristic correspondences, while voiced stops are affected to a much lesser degree. This is in line with classical descriptions of the *gorgia* phenomenon. In contrast to Area 1, in Area 2 correspondence pairs involving spirantization of stops (except /k/) are ranked lower in the ordered list (between positions 30 and 46) and are closely followed by their reverse (occurring between positions 56 and 68).

If we move to the right half of Table 2, it immediately catches one's eye that the 10 topmost pairs in the union of Areas 1 & 2 in Fig. 2 (corresponding to central and western Tuscany) confirm the prominence of spirantization phenomena, with particular emphasis on voiceless stops: interestingly, the 5th and 7th pairs ([-]:[h] and [-]:[x]) show the maximum weakening degree, involving the deletion of /k/.

Another phonological process characterizing Areas 1 & 2 is represented by the rhotacism of the lateral consonant in postconsonantal position: the phonetic correspondence [r]:[l] is found in position 14 of the ranking.

Let us now consider the results obtained for Area 3 (not listed in a separate table). It is interesting to note that in the top 10 correspondence pairs, only non-spirantized or less spirantized outcomes of voiceless stops are found: in particular, the top five pairs are represented by [k]:[h], [t]:[θ], [p]:[ɸ], [h]:[-], and [x]:[-]. This area is thus characterized by non-spirantization. Pairs with voiceless stops, which are ranked among the top 10, are closely followed by pairs with voiced stops, namely [d]:[ð], [b]:[β], and [g]:[ɣ]. Similarly to the previous areas, reverse pairs (with spirantized outcomes) are found at the bottom of the list, between positions 90 and 99. On the other hand, deaffrication of postalveolar affricates seems to be a characterizing feature of this area: the pairs [ʒ]:[dʒ] and [ʃ]:[tʃ] are ranked quite high in the list. This distribution sug-

gests that spirantization of affricates tends to occur more frequently in parts of Tuscany where stop spirantization has not otherwise had a strong impact; in particular, in the eastern and southern part of Tuscany. This is supported by the results for all groups discussed above. Marotta (2008) considers the spirantization of affricates (deaffrication) to be a pan-Tuscan phenomenon; on the other hand, Binazzi (2019) notes that it is particularly present in varieties spoken in the area of Arezzo, without considering, however, the phenomenon to be limited to specific varieties of Tuscan.

From the comparison of the results achieved for Areas 1, 1 & 2, and 3, it emerges that, while the first two are generally similar but not identical, the latter represents their complement. This holds for spirantization phenomena but also for other detected phonetic processes such as rhotacism. Further, Area 2, taken alone, is characterized by unstable outcomes as far as spirantization is concerned—which is to be expected when simultaneously comparing it to varieties with both stronger and weaker degrees of spirantization.

4.2 *Comparison to previous studies*

The second approach to evaluation mentioned in the introductory section was trying to replicate previous results. In particular, we refer here to the correspondences obtained by applying the bipartite spectral clustering method reported in Montemagni et al. (2013), where the focus was narrower, namely only on spirantization. Recall that our approach proceeds from a postulated area for which it looks for correspondences more generally, including those not related to spirantization. The map in Fig. 4 visualizes identified clusters, each marked by a different shade, which were identified on the basis of context-free phonetic correspondence pairs.

An obvious difference in the approaches is that the bipartite spectral clustering approach not only identifies sound correspondences, but also clusters the sites they arise from. This will simplify the processing pipeline, but also locks researchers into a single approach for tasks that are normally regarded as separate. An additional difference is that the approach in Montemagni et al. (2013) required that candidate correspondences be identified ahead of time, while the approach in the present paper is applied to *all* available aligned data, making it more general.

To compare results with respect to this work, we selected the white core cluster whose underlying features correspond to spirantization phenomena involving both voiceless and voiced stops. Table 3 reports the results shown by Montemagni et al. (2013) in the first two columns, as well as those obtained using the novel approach proposed in this paper (in the remaining columns), focusing on the area covered by the white cluster. While in Montemagni et al.

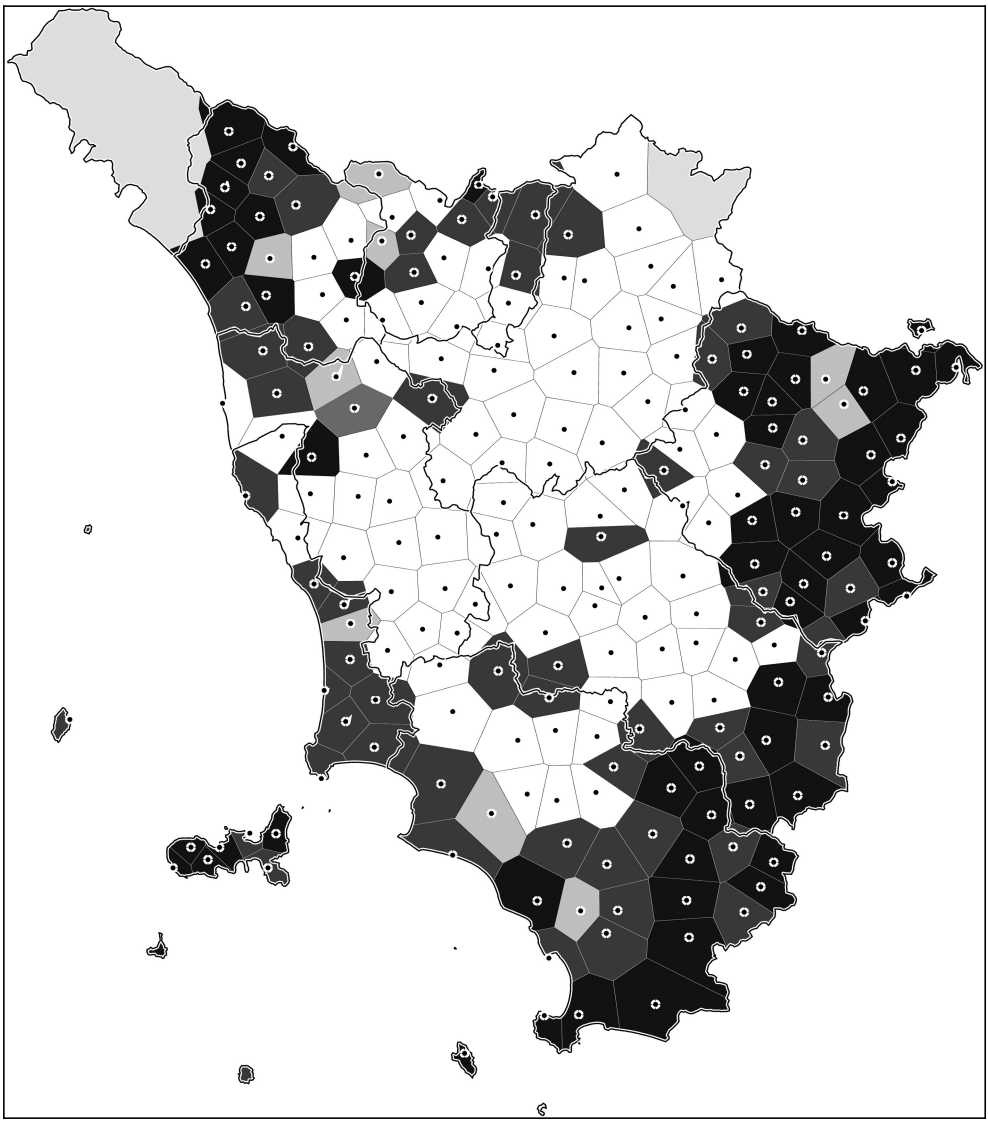

FIGURE 4 Geographic visualization of the clusters obtained with context-free sound correspondence pairs in Montemagni et al. (2013). Different shades indicate different clusters.

(2013) each phonetic correspondence links the dialectal allophone (the right member of the pair) with its realization in standard Italian, correspondence pairs in this study involve phonetic variants on both sides (the variant attested in the area in focus on the left, and the corresponding variant outside of it on the right). It is interesting to note that extracted features overlap significantly. The correspondence pairs identified by applying the hierarchical spectral par-

TABLE 3 Ranked context-free spirantization-related correspondence pairs with associated scores

Context-free correspondence pairs with associated importance score		Spirantization-related correspondence pairs with associated scores				
SpCorr.	**Importance**	**Position**	**Corr.**	**Dist.**	**Repr.**	**Char.**
/t/:[h]	0.500	1	[θ]:[t]	0.492	0.941	0.646
/d/:[ð]	0.484	2	[ɸ]:[p]	0.457	0.957	0.619
/t/:[θ]	0.449	3	[h]:[k]	0.442	0.923	0.598
/p/:[ɸ]	0.421	4	[ð]:[d]	0.324	0.861	0.471
/b/:[β]	0.421	7	[x]:[k]	0.316	0.709	0.437
/g/:[ɣ]	0.405	9	[ɣ]:[g]	0.276	0.868	0.419
/k/:[h]	0.259	12	[h]:[g]	0.102	0.842	0.182
/t/:[-]	0.178	15	[β]:[b]	0.201	0.850	0.325

Left: Context-free spirantization-related correspondences (SpCorr.) underlying the white core cluster in Fig. 4 reported in Montemagni et al. (2013), together with their importance score. *Right*: Spirantization-related phonetic correspondences in the area covered by the white core cluster with associated distinctiveness (Dist.), representativeness (Repr.), and characteristicness (Char.) scores and relative position in the ranked list.

titioning method are highly ranked in the list obtained for the white area: six of them are among the top 10 correspondences (see the "Position" column, which indicates the position of the correspondence in the ranked list) and all of them are among the first third of the list, that is, they represent features characterizing the area. Note that, in contrast to the previous study, in this case the ranked list is not circumscribed to spirantization-related phenomena.

If the results of bipartite spectral clustering and the information-theoretic measures are similar but not identical, we conclude both that they are detecting significant correspondences but also that they are not detecting identical signals. On the other hand, if one compares the importance scores on the left in the table with the characteristic scores, one finds a perfect match in order, suggesting that the information-theoretic characterization is homing in on the same signal or at least a closely related one. Given that we modified the calculation of representativeness (to include correspondences outside the focal area), this could not have been predicted with certainty.

4.3 *Problems and opportunities*

There were problems encountered here that might be addressed in future work, and some new opportunities have arisen through the analysis.

An obvious issue is that, since the correspondences often reflect conditioned sound changes—that is, they are dependent on the phonetic context they occur in—the correspondences are of course also sensitive to the phonetic context in which they occur; but this was ignored in this study. The insensitivity to context arose because we worked only with unigrams in this model, thus completely ignoring the context of sound changes. But we acknowledge that context is important for almost every sound change, including those in Tuscan, such as *gorgia*. Taking context into account (perhaps by working with bigrams or trigrams) could definitely enhance the results, as Montemagni et al. (2013) show, but of course at the combinatorial cost of reducing the amount of data available.

In addition to the challenge of handling context sensitivity, noted above, further opportunities arise from this research. One obvious option is that future work could test the applicability of the proposed measures on other datasets, perhaps in contrast to the alternative calculation of significant correspondences using the definitions of representativeness and distinctiveness proposed in Wieling et al. (2009) and Prokić et al. (2012). It would be interesting to see these compared more rigorously.

The paper further raises the question of whether the proposed measures can be generalized, at least in spirit, to other measures of dialect difference, such as numerical differences in formant frequencies.

Finally, although we have not explicitly discussed the relation between feature-based and aggregating approaches to dialectology, we have shown how to extract characteristic features from the sorts of aggregates normally posited in non-feature-based approaches. The time is ripe for examination of the results of each of the two approaches from the other's perspective: quantitative examination of the characteristic features of a dialect area proposed from a feature-based approach (something like this was done in Wieling et al. 2018), but also a critical appraisal of the somewhat indiscriminate aggregation often used in the alternative approach. If a representative and distinct feature does not arise from calculations like those presented here, that could be due to the way that linguistic prominence has been ignored. We have in mind here that differences in stressed syllables might be preferred by dialect speakers as distinguishing elements, or perhaps emotionally charged words.

5 Conclusions

In this study, we have presented a novel information-theoretic method that can automatically calculate which phonetic correspondences are the most characteristic between an a priori defined group of sites and the rest of the varieties in the dataset. The method reliably detected that the spirantization of unvoiced plosives is a salient phenomenon in central and western Tuscan, whereas the spirantization of postalveolar affricates seems to be prominent in eastern Tuscan varieties. Acknowledging definite room for improvement, the present approach is worthwhile for being more generally applicable to the problem of detecting phonetic correspondences automatically and perhaps for its independent applicability. Finally the work reported on here could serve as a model for investigating the relation between feature-based and aggregating approaches to dialectology.

Data and software availability

Source data

The data used in this study is available at https://github.com/lexibank/ALT.

Software availability

All relevant software for this study, including input data and results, is available at https://github.com/arubehn/TuscanSoundCorrespondences.

Division of labor

We would like to report on our division of labor. Arne Rubehn recast two critical concepts in information theory, developed all the software, and wrote the first draft of the paper. Simonetta Montemagni provided the dataset, consulted on normalizing data, suggested interesting areas, and worked closely on the interpretation of the results. John Nerbonne suggested the topic in the course of a seminar in Tübingen, provided feedback during development and write-up, and wrote a second draft of the paper. All authors contributed equally in revising the manuscript based on the reviewers' suggestions.

Acknowledgments

The authors would like to express their gratitude to Johann-Mattis List for his help on representing and quality-checking the data with CLDF, as well as for helpful comments on constructing and analyzing multiple sequence alignments.

This project was supported by the ERC Consolidator Grant ProduSemy (PI Johann-Mattis List, Grant No. 101044282; see https://doi.org/10.3030/10104428 2). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them.

References

Bach, Stephen, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. Hinge-loss Markov random fields and probabilistic soft logic. *Journal of Machine Learning Research* (*JMLR*) 18(1): 1–67.

Binazzi, Neri. 2019. Toscana, version 2. In Roland Bauer and Thomas Krefeld (eds.), *Lo spazio comunicativo dell'Italia e delle varietà italiane, version 90*, Korpus im Text. http://www.kit.gwi.uni-muenchen.de/?p=12469\&v=2.

Bodt, Timotheus A., and Johann-Mattis List. 2022. Reflex prediction: A case study of Western Kho-Bwa. *Diachronica* 39(1): 1–38.

Bouma, Gerlof. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Gesellschaft für Sprachtechnologie und Computerlinguistik* (*GSCL*) 30: 31–40.

Church, Kenneth, and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1): 22–29.

Cucurullo, Nella, Simonetta Montemagni, Matilde Paoli, Eugenio Picchi, and Eva Sassolini. 2006. Dialectal resources on-line: The ALT-web experience. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk and Daniel Tapias (eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (*LREC'06*), 1846–1851. Genoa: European Language Resources Association (ELRA). https://aclanthology.org/volumes/L06-1/.

Daneyko, Thora. 2020. *Automated Sound Law Inference Using Probabilistic Soft Logic*. MA thesis, University of Tübingen.

Dellert, Johannes. 2018. Combining information-weighted sequence alignment and sound correspondence models for improved cognate detection. In Emily M. Ben-

der, Leon Derczynski and Pierre Isabelle (eds.), *Proceedings of the 27th International Conference on Computational Linguistics*, 3123–3133. Santa Fe, NM: Association for Computational Linguistics. https://aclanthology.org/volumes/C18-1/.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5(1): 1–10.

Giacomelli, Gabriella, Luciano Agostiniani, Patrizia Bellucci, Luciano Giannelli, Simonetta Montemagni, Annalisa Nesi, Matilde Paoli, Eugenio Picchi, and Teresa Poggi Salani (eds.). 2000. *Atlante lessicale toscano*. Rome: Lexis Progetti Editoriali.

Giannelli, Luciano. 2000. *Toscana*. Pisa: Pacini Editore. New updated version of the 1976 edition.

Giannelli, Luciano, and Leonardo M. Savoia. 1978. L'indebolimento consonantico in Toscana. *RID: Rivista italiana di dialettologia* 2(1): 23–58.

Giannelli, Luciano, and Leonardo M. Savoia. 1979. L'indebolimento consonantico in Toscana (II). *RID: Rivista italiana di dialettologia* 3(4): 38–101.

Gooskens, Charlotte, and Wilbert Heeringa. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change* 16(3): 189–207.

Grassi, Corrado, Alberto Sobrero, and Tullio Telmon. 1997. *Fondamenti di dialettologia italiana*. Rome: Laterza.

Haag, Karl. 1898. *Die Mundarten des oberen Neckar- und Donaulandes*. Reutlingen: Hutzler.

Hall, David, and Dan Klein. 2010. Finding cognate groups using phylogenies. In Jan Hajič, Sandra Carberry, Stephen Clark and Joakim Nivre (eds.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1030–1039. Uppsala: Association for Computational Linguistics. https://aclanthology.org/P10-1.

Hall, Robert A. 1949. A note on "gorgia toscana". *Italica* 26(1): 65–71.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. Glottolog 4.8. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://glottolog.org/.

Hauer, Bradley, and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In Haifeng Wang and David Yarowsky (eds.), *Proceedings of 5th International Joint Conference on Natural Language Processing*, 865–873. Chiang Mai: Asian Federation of Natural Language Processing. https://aclanthology.org/I11-1000/.

Heeringa, Wilbert J. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD dissertation, University of Groningen.

Jäger, Gerhard. 2013. Phylogenetic inference from word lists using weighted alignment

with empirically determined weights. *Language Dynamics and Change* 3(2): 245–291.

Jäger, Gerhard, Johann-Mattis List, and Pavel Sofroniev. 2017. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1205–1216. Valencia: Association for Computational Linguistics. https://aclanthology.org/E17-1000/.

Jurafsky, Dan, and James H. Martin. 2023. Speech and language processing (3rd edn. draft). https://web.stanford.edu/~jurafsky/slp3/.

Kondrak, Grzegorz. 2003. Identifying complex sound correspondences in bilingual wordlists. In A. Gelbukh (ed.), *International Conference on Intelligent Text Processing and Computational Linguistics*, 432–443. Berlin: Springer.

Kondrak, Grzegorz. 2009. Identification of cognates and recurrent sound correspondences in word lists. *Traitement automatique des langues et langues anciennes* 50(2): 201–235.

Leinonen, Therese, Çağrı Çöltekin, and John Nerbonne. 2016. Using Gabmap. *Lingua* 178: 71–83.

List, Johann-Mattis. 2012. SCA: Phonetic alignment based on sound classes. In Marija Slavkovik and Dan Lassiter (eds.), *New Directions in Logic, Language, and Computation*, 32–51. Berlin and Heidelberg: Springer. https://doi.org/10.1007/978-3-642-31467-4_3.

List, Johann-Mattis. 2014. *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.

List, Johann-Mattis. 2019. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics* 45(1): 137–161. https://doi.org/10.1162/coli_a_00344.

List, Johann-Mattis. forthcoming. Computational approaches to historical language comparison. In Claire Bowern and Bethwyn Evans (eds.), *Routledge Handbook of Historical Linguistics, vol. 2*, London and New York: Routledge.

List, Johann-Mattis, Michael Cysouw, and Robert Forkel. 2016. Concepticon: A resource for the linking of concept lists. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2393–2400. Portorož, Slovenia: European Language Resources Association. https://aclanthology.org/L16-1.

List, Johann-Mattis, and Robert Forkel. 2021. LingPy: A python library for historical linguistics, version 2.6.9. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022a. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data* 9(1): 1–16.

List, Johann-Mattis, Nathan W. Hill, and Robert Forkel. 2022b. A new framework for fast automated phonological reconstruction using trimmed alignments and sound correspondence patterns. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, 89–96. Dublin: Association for Computational Linguistics. https://aclanthology.org/2022.lchange-1.9.

Marotta, Giovanna. 2001. Non solo spiranti: La "gorgia toscana" nel parlato di Pisa. *L'Italia dialettale* 62: 27–60.

Marotta, Giovanna. 2008. Lenition in Tuscan Italian (Gorgia Toscana). In Joaquim Brandão de Carvalho, Tobias Scheer and Philippe Ségéral (eds.), *Lenition and Fortition*, 235–270. Berlin: Mouton de Gruyter.

Miller, George A., and Patricia E. Nicely. 1955. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America* 27(2): 338–352.

Montemagni, Simonetta. 2008. The space of Tuscan dialectal variation: A correlation study. *International Journal of Humanities and Arts Computing* 2(1–2): 135–152.

Montemagni, Simonetta, Martijn Wieling, Bob De Jonge, and John Nerbonne. 2012. Patterns of language variation and underlying linguistic features: A new dialectometric approach. In *La variazione nell'italiano e nella sua storia: Varietà e varianti linguistiche e testuali. Atti dell'XI Congresso SILFI* (*Società Internazionale di Linguistica e Filologia Italiana*), vol. 2, 879–889. Florence: Cesati.

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

Nerbonne, John. 2003. Linguistic variation and computation. In Ann Copestake and Jan Hajič (eds.), *10th Conference of the European Chapter of the Association for Computational Linguistics*, 3–10. Bergen: ACL. https://aclanthology.org/E03-1088.

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

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

Pickl, Simon. 2016. Fuzzy dialect areas and prototype theory: Discovering latent patterns in geolinguistic variation. In Marie-Hélène Côté, Remco Knooihuizen and John Nerbonne (eds.), *The Future of Dialects: Selected Papers from Methods in Dialectology XV*, 75–98. Berlin: Language Science Press. https://doi.org/10.17169/langsci.b81.78.

Pickl, Simon, and Jonas Rumpf. 2011. Automatische Strukturanalyse von Sprachkarten: Ein neues statistisches Verfahren. In Elvira Glaser, Jürgen Erich Schmidt and

Natascha Frey (eds.), *Dynamik des Dialekts—Wandel und Variation: Akten des 3. Kongresses der Internationalen Gesellschaft für Dialektologie des Deutschen* (IGDD), vol. 3, 267–285. Stuttgart: Franz Steiner.

Prokić, Jelena, Çağrı Çöltekin, and John Nerbonne. 2012. Detecting shibboleths. In Miriam Butt, Sheelagh Carpendale, Gerald Penn, Jelena Prokić and Michael Cysouw (eds.), *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 72–80. Avignon: Association for Computational Linguistics.

Prokić, Jelena, and John Nerbonne. 2013. Analyzing dialects biologically. In Heiner Fangerau, Hans Geisler, Thorsten Halling and William Martin (eds.), *Classification and Evolution in Biology, Linguistics and the History of Science*, 149–161. Suttgart: Franz Steiner.

Rama, Taraka. 2016. Chinese Restaurant Process for cognate clustering: A threshold free approach. ArXiv preprint arXiv:1610.06053. https://arxiv.org/abs/1610.06053.

Salle, Alexandre, and Aline Villavicencio. 2023. Understanding the effects of negative (and positive) pointwise mutual information on word vectors. *Journal of Experimental & Theoretical Artificial Intelligence* 35(8): 1161–1199.

Séguy, Jean. 1973. La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane* 37(145–146): 1–24.

Sorianello, Patrizia. 2001. Un'analisi acustica della "gorgia" fiorentina. *L'Italia dialettale* 62(1): 61–94.

Wieling, Martijn. 2012. *A Quantitative Approach to Social and Geographical Dialect Variation*. PhD dissertation, University of Groningen.

Wieling, Martijn, Jelke Bloem, Kaitlin Mignella, Mona Timmermeister, and John Nerbonne. 2014. Measuring foreign accent strength in English: Validating Levenshtein distance as a measure. *Language Dynamics and Change* 4(2): 253–269.

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

Wieling, Martijn, Jelena Prokić, and John Nerbonne. 2009. Evaluating the pairwise string alignment of pronunciations. In Lars Borin and Piroska Lendvai (eds.), *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education* (*LaTeCH–SHELT&R 2009*), 26–34. Athens: Association for Computational Linguistics. https://aclanthology.org/W09-03.

Wieling, Martijn, Eva Sassolini, Sebastiana Cucurullo, and Simonetta Montemagni. 2016. ALT explored: Integrating an online dialectometric tool and an online dialect atlas. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (*LREC'16*), 3265–3272. Portorož, Slove-

nia: European Language Resources Association (ELRA). https://aclanthology.org/volumes/L16-1/.

Wieling, Martijn, Esteve Valls, R. Harald Baayen, and John Nerbonne. 2018. Border effects among Catalan dialects. In Dirk Speelman, Kris Heylen and Dirk Geeraerts (eds.), *Mixed-effects regression models in linguistics*, 71–97. Berlin: Springer.

Wilkinson, Mark D. et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data* 3(1). http://dx.doi.org/10.1038/sdata.2016.18.