Revisiting the classification of Gallo-Italic: a dialectometric approach

Marco Tamburelli and Lissander Brasca
School of Linguistics and English Language, Bangor University, Bangor, UK

Abstract

While Gallo-Italic varieties clearly belong to the Romance language family, their subgrouping as either Gallo-Romance or Italo-Romance has been the source of disagreement in the classificatory literature. While earlier analyses tended to classify Gallo-Italic as Gallo-Romance (notably Schmid, 1956; Bec, 1970–1971), later work has either argued for or tacitly assumed a classification of Gallo-Italic as part of the Italo-Romance branch, a view that is both different from as well as irreconcilable with the earlier Gallo-Romance classifications. In this article, we aim to contribute to the development of an empirically based classification of Gallo-Italic through the use of dialectometry applied to atlas corpora, and specifically through the measurement of Levenshtein distance. Using three wordlists (Swadesh 100, Swadesh 200, Leipzig–Jakarta) and comparing twenty-six linguistic varieties across Italy and south-eastern France, we show that Gallo-Italic is best classified as a third subgroup within the Gallo-Romance branch. Our results also clearly identify all the major bundles of isoglosses established through traditional dialectological methods and confirm Gallo-Italic as a relatively homogenous group distinct from Italo-Romance.

Correspondence:
Marco Tamburelli, School of Linguistics and English Language, Bangor University, Bangor, UK.
E-mail:
m.tamburelli@bangor.ac.uk

1 Introduction

Since the work of Bartoli (1936) and Wartburg (1950), it is generally agreed that the Rimini-La Spezia line is an important isogloss for Romance classification in general and for the classification of Italian vernaculars in particular (see Green, 2009; Iacobini, 2009; Repetti, 1996; and the volume edited by Maiden and Parry, 1997, for more recent discussion). There is also broad agreement on the fact that most of the Romance vernaculars traditionally spoken between the Alps and the Rimini-La Spezia line form a generally homogenous group known as 'Gallo-Italic'. As shown in Fig. 1, this group is composed of the Romance varieties historically spoken in the administrative regions of Emilia-Romagna, Liguria, Lombardy, and Piedmont in Italy and the canton of Ticino in Switzerland, as well as in smaller areas in the province of Trento, the Swiss canton of Grisons/Graubünden, and in the northern most part of Tuscany. Gallo-Italic varieties border with Venetian varieties to the east[1] and with Occitan and Franco-Provençal to the west (for an overview, see Harris and Vincent, 2003; and Posner, 1996).

2 The Classification of Gallo-Italic

While the existence of Gallo-Italic as a group is undisputed, the classification of Gallo-Italic within Romance is a point of contention. Specifically, the

Digital Scholarship in the Humanities, Vol. 33, No. 2, 2018.
doi:10.1093/llc/fqx041 Advance Access published on 24 August 2017

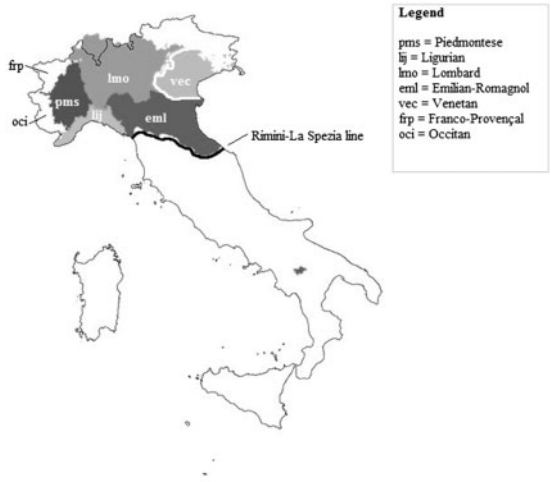

Fig. 1 The Gallo-Italic group and its neighbouring Romance varieties. Each variety is identified via its ISO 639-3 code (Lewis *et al.*, 2014).

scholarly literature on Romance classification proposes two distinct classifications for Gallo-Italic. According to what is perhaps the most influential tradition on the classification of Italian vernaculars, Gallo-Italic varieties belong to the Italo-Romance branch. On this view, Gallo-Italic is part of a dialect group that includes all Romance varieties historically spoken in Italy, Corsica, and Canton Ticino (southern Switzerland), but excluding Occitan, Franco-Provençal/Arpitan (classified as Gallo-Romance), Sardinian (classified as a separate branch), Ladin and Friulian (classified as Rhaeto-Romance).[2] Works that are representative of this tradition include Wartburg (1950), Merlo (1960-1961), Hall (1974), Pellegrini (1973; 1992), Loporcaro (2009), and, in some respects, Lausberg (1956). Within this tradition, Gallo-Italic varieties are either explicitly classified as Italo-Romance (and consequently excluded from the Gallo-Romance group) or indirectly presumed to be Italo-Romance, as is the case with classifications that consider Gallo-Italic as simply 'Northern Italian' or even just 'Italian' (Kabatek and Pusch, 2011).

However, according to a second, less influential tradition, Gallo-Italic is part of the Gallo-Romance branch, separated from Italo-Romance by the Rimini-La Spezia line. Hull (1982) is perhaps the most extensive comparative analysis carried out within this tradition, culminating in a proposal for a genealogical classification of Gallo-Italic as part of a wider historical linguistic branch that he calls 'Padanian', comprising the Gallo-Italic and Rhaetic continua (not unlike the 'Cisalpine' group in the work of Pellegrini, 1992, 1995), and belonging to the larger Gallo-Romance branch within the Western Romance group. Hull concludes that 'the Romance vernaculars of Northern Italy and Rhaetia have conserved, and in many cases have developed further, their original Gallo-Roman structure' (1982: 660), and warns against attaching unwarranted importance to the 'superficial Italic, German and Franco-Occitan influences' which are 'insufficient to warrant a classification of all or part of the Rhaeto-Cisalpine zone as 'Italo-Romance' in the strictly linguistic sense of the term' (1982: 660). Similar conclusions had previously been reached by Ascoli (1890), Schmid (1956), and Bec (1970-1971) who explicitly classified Gallo-Italic varieties as part of the Gallo-Romance branch, a classification that found further support in the work of Pellegrini (1992, 1995) and Kotliarov (2009).

Nevertheless, the classification of Gallo-Italic remains unclear, with most recent work not taking any particular stand on the issue, while at the same time assuming—either implicitly or explicitly—that Gallo-Italic is essentially Italo-Romance,[3] and thus linguistically closer to the varieties south of the Rimini La Spezia than to—say—Provençal or Rumantsch. The assumption that Gallo-Italic belongs to the Italo-Romance branch is rather widespread in the modern sociolinguistic literature (Cerruti, 2011; Dal Negro and Vietti, 2006) but can also be found in the literature on Romance typology (Schmid, 2012), lexicography (Barbato and Varvaro, 2004; Crevatin, 2004), and syntactic analysis (Garzonio and Poletto, 2009).

These two classifications—both based on the family-tree model—are not only distinct from each other, they are also irreconcilable insofar as they take Gallo-Italic to belong to two distinct subgroups that are not in a sisterhood relationship.

As shown by the oval in Fig. 2, the two classifications are irreconcilable since the two nodes under which they classify Gallo-Italic, namely, Gallo-Romance and

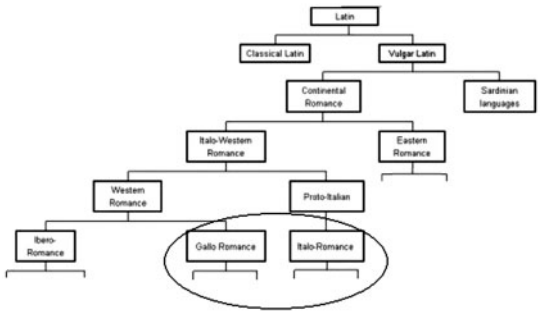

Fig. 2 Partial tree model indicating the position of Gallo-Romance in relation to Italo-Romance.

Italo-Romance, are neither sisters nor in a mother–daughter relationship. In other words, the contention is not only about the classification of Gallo-Italic in itself, but also about its ancestor as either Western-Romance or Proto-Italian, thus making the two classifications considerably different from each other.

2.1 Classificatory criteria

The genetic or genealogical classification of languages is based on the measurement of successive innovations. Each innovation sets a variety apart from its original parent language, and shared innovations among varieties provide evidence for the formation of a subfamily (see Fox, 1995 for a detailed overview). Importantly, innovations are only considered as evidence if they are pervasive. This is to exclude the possibility that some apparently innovative trait resulted from borrowing rather than from systematic change (see Joseph and Janda, 2003 for a complete discussion). For the same reason, the innovations at the basis of classificatory linguistics are mostly phonetic/phonological and occasionally morpho-phonological, since those are the linguistic areas where regular, systematic change can be observed. The more innovations a set of varieties share, the more likely it is that those innovations are due to common ancestry rather than to coincidental, parallel development.

While these principles are generally undisputed, dialectologists have often selected specific linguistic traits that they wished to analyse for signs of innovation and/or archaisms. For example, Pei (1949) compiled a classification of some Romance varieties based solely on the development of stressed vowels (see also Lüdtke, 1956), while Politzer (1947) gave particular importance to the synchronic conservation of plural -s when classifying the Romance varieties of Italy, a position partly endorsed by Pellegrini (1973) and Francescato (1982). Perhaps unsurprisingly, different classifications have emerged depending on the traits selected and/or on the significance that different researchers have associated with particular traits.

However, given that—as we have seen—language grouping in classificatory linguistics is intended to reflect systematic, pervasive change, researchers have increasingly questioned classifications that rely on linguistic traits selected *a priori*. While being a practical necessity in traditional comparative dialectology, the selection of a limited number of specific traits necessarily involves subjective judgements (McMahon and McMahon, 2005; Starostin, 2010; Szmrecsanyi and Wolk, 2011), and may result in erroneous classifications as the preselected traits become overly influential in the final analysis. In keeping with this view, this article aims to contribute to the development of an empirically based classification of Gallo-Italic through the use of dialectometry applied to atlas corpora, and specifically through the measurement of Levenshtein distance. The following research question will therefore be addressed: Should Gallo-Italic be classified as part of the Gallo-Romance branch or the Italo-Romance branch?

Unlike traditional dialectology, dialectometry does not select linguistic traits *a priori*, relying instead on the extraction of patterns from quantitative data. Therefore, dialectometric analyses can offer an insight into unresolved classifications, by largely eliminating the issue of subjective feature selection and enabling the identification of aggregate differences (Nerbonne and Kleiweg, 2007) and 'seemingly hidden structures' (Goebl and Schiltz, 1997: 13) emerging from the combination of individual linguistic variables. Dialectometric measurements in general, and Levenshtein distance in particular, have been successfully applied in the classification of varieties within the Irish Gaelic (Kessler, 1995), Dutch (Heeringa, 2004; Nerbonne, 2005; Nerbonne et al, 1996), and Norwegian (Gooskens and Heeringa, 2004) continua, as well as the Italo-Romance varieties

of Tuscany (Montemagni et al., 2013; Wieling et al., 2014). Moreover, the measurement of linguistic distance has been argued to help evaluate the descriptive power of traditional classifications particularly in cases of disagreement (Tang and Van Heuven, 2009; Wichmann, Holman, Bakker, and Brown, 2010), as is the case for Gallo-Italic.

3 Method

3.1 Wordlist comparison

Wordlist comparison is a quantitative technique aimed at detecting the degree of historical affinity between languages and/or language varieties. This technique relies on the use of one or more lists that contain basic vocabulary items, where each item is a concept (or meaning, hence wordlists are sometimes also called 'meaning lists', McMahon and McMahon, 2005), and associated word forms which are collected for each concept in each of the language variety to be compared. All word forms corresponding to each concept are then compared across language varieties to measure the proportion of sound change correspondences and cognates that are shared, thus yielding a classification of the degrees of phylogenetic relatedness among the linguistic varieties at issue.

Wordlist comparison differs radically from a corpus comparison approach (Heeringa, 2004) as well as from approaches that compare items on word databases (Nerbonne et al., 1996). Specifically, wordlist comparison aims to evaluate only items of basic vocabulary that are known to be relatively resistant to borrowing (Tadmor *et al.*, 2010; Thomason, 2001), such as pronouns, numerals, and body parts. In doing so, wordlists exclude as much as possible items that may be shared across varieties for reasons other than shared genealogy—particularly contact-induced change—thereby leading to more accurate genealogical grouping (Gray and Atkinson, 2003; McMahon et al., 2005; Haspelmath, 2008).

The most widely used wordlists are the ones developed by Morris Swadesh in the 1950s and which came to be known as the 'Swadesh lists'. These lists were compiled by Swadesh himself as part of his pioneering work on quantitative lexical comparison and were based on his own experience of what tended to be 'stable' items in a language's lexicon. The original Swadesh list included 207 items (known as the 'Swadesh 200' list), while a later proposal introduced an alternative version containing 101 items, which came to be known as the 'Swadesh 100' (for a detailed discussion, see Swadesh, 1971). While the Swadesh lists have been widely used in the literature on quantitative comparison, Tadmor (2009) pointed out the need for a more strongly 'empirically-based' alternative that 'makes use of the powers of computational linguistics' (Tadmore, 2009: 72). It is with this aim in mind that the Leipzig–Jakarta list was developed.

Tadmor (2009) and Haspelmath and Tadmor (2009) argue that the Leipzig–Jakarta is probably the most empirically accurate wordlist available, having been developed through quantitative comparison of 57,517 words from a database of forty-one languages representing a broad range of linguistic families and subfamilies from across the world. This process involved an initial list of 1,460 concepts to be associated with their respective word forms in all forty-one languages. However, languages are not always semantically congruent, with some languages having more than one word associated with concepts for which other languages have no word (e.g. they rely on a periphrastic construction). Consequently, the number of word forms collected for the 1,460 concepts varied across languages, ranging between 1,000 and 2,000 word forms depending on language. These word forms and their associated concepts were subsequently weighted for 'borrowability' (Van Hout and Muysken, 1994), namely, the 'relative likelihood that words with particular concepts would be borrowed' (Haspelmath and Tadmor, 2009:1). This process resulted in the compilation of the Leipzig–Jakarta wordlist, containing only the 100 concepts that consistently achieved a low score for borrowability across the corpus. As a result, the Leipzig–Jakarta is probably the most accurately calibrated wordlist available for genealogical classification. Nevertheless, the current study also includes the Swadesh wordlists, since these have been widely used in the literature on quantitative comparison of the major Romance and other Indo-European languages (Forster *et al.*, 1998; McMahon and McMahon, 2003; Rama et al., 2015).

3.2 Materials: wordlists

Three conventional wordlists were compiled for the current study: Swadesh 100 (Swadesh, 1955), Swadesh 200 (Swadesh, 1952), and Leipzig–Jakarta (Tadmor, 2009).

The three wordlists have a high degree of overlap (see Tadmor, 2009 for a detailed analysis). This is chiefly due to two facts. First, all three lists have been developed with the precise intent to minimize the presence of borrowings to increase the accuracy of the resulting genealogical grouping (see discussion in the previous section above). Secondly, there exist only a limited number of concepts to choose from when constructing a wordlist, since only a small subset of word categories and semantic fields tend to be resistant to borrowing (see Tadmor *et al.*, 2010 for a detailed overview).

Specifically, the Leipzig–Jakarta shares 62% of the items with the Swadesh 100 list and 82% of the items with the Swadesh 200 list. The Swadesh 100 list—which was developed from the Swadesh 200 list with the intent to provide a potentially more accurate wordlist (Swadesh, 1955)—resulted from the removal of 113 concepts from the Swadesh 200 list plus the addition of seven concepts. Overall, the three lists contain 223 distinct concepts.

At this point it is important to note that while the Swadesh 100 list is close to being a subset of the Swadesh 200, this does not necessarily mean that the Swadesh 200 will yield a more reliable classification than the Swadesh 100 (Zhang and Gong, 2016). This is due to the fact that wordlists are based on two parameters (McMahon and McMahon, 2005; Swadesh, 1955; Zhang and Gong, 2016). One of these is a quantitative parameter, namely, the number of items that make up the list, and its importance lies in the fact that fewer items necessarily provide fewer measurement points with which to gauge degrees of genealogical relatedness (Embleton, 1986). The other parameter, which is qualitative in nature, is the high resistance to borrowing required of the concepts in the wordlist. While a longer list will include more potential for measurement, it also increases the probability that some of its items might include 'undiagnosed or misdiagnosed loans' that can subsequently 'obscure the familial signal and lead to erroneous classifications' (McMahon et al., 2005:148). As an example, let us take a hypothetical wordlist L which misguidedly includes the concept for 'letter' among its items. Let us then suppose that this list is used to measure the degree of genealogical relatedness between English, French, and German. Due to the fact that the English word for 'letter' is a borrowing from French, our hypothetical list would partially contribute towards the erroneous conclusion that English is genealogically closer to French (*lettre*) than to German (*Brief*). Naturally, a single item would only minimally skew the overall result, but the point remains that the longer the wordlist, the higher the probability that such items may have been unwittingly included (on this point, see also McMahon and McMahon, 2003).

Therefore, a longer wordlist does not necessarily yield more genealogically accurate results, since length is only one of the two relevant parameters, and arguably the less important one as far as establishing degrees of genealogical relatedness is concerned. Indeed, Emory (1963) argued that the Swadesh 100 yields more accurate results than the Swadesh 200 as far as Polynesian languages are concerned (though this is not true of all languages), echoing Swadesh's view that 'quality is at least as important as quantity' (1955: 124).

With these points in mind, we followed current practice in quantitative linguistic comparison (Calude and Pagel, 2014; Rosendal and Mapunda, 2014; Syrjänen *et al.*, 2013) by including all three conventional wordlists in the current study.

3.3 Materials: Atlases

The three wordlists were compiled with the respective word forms taken from two linguistic atlases: the Linguistic Atlas of Italy and southern Switzerland (*Sprach- und Sachatlas Italiens und der Südschweiz*, Jaberg and Jud, 1928–40) and the Atlas Linguistique de la France (Gilliéron and Edmont, 1902). The two atlases share a number of methodological features, not least because the Linguistic Atlas of Italy and southern Switzerland was produced by former pupils of Gilliéron and with the explicit intent to apply the filed techniques of the Atlas Linguistique de la France to the Swiss and Italian areas. Both atlases covered the Romance-speaking areas in their respective countries of interest (i.e. France, Italy, and

Table 1 Summary of the linguistic points from the Linguistic Atlas of Italy and southern Switzerland (AIS) and the Atlas Linguistique de la France (ALF) with respective classifications

Geographical point (Atlas ref.)	Subgrouping	Country (Atlas)
Standard French (N/A)	Oïl, Gallo-Romance	France (N/A)
Haute-Savoie, Chamonix (967)	Franco-Provençal, Gallo-Romance	France (ALF)
Savoie, Chignin (943)	Franco-Provençal, Gallo-Romance	France (ALF)
Aosta Valley, Rhêmes (121)	Franco-Provençal, Gallo-Romance	Italy (AIS)
Lanzo Valley, Stura (143)	Franco-Provençal, Gallo-Romance	Italy (AIS)
Hautes Alpes, Le Monetier (971)	Occitan (Vivaroaupenc), Gallo-Romance	France (ALF)
Susa Valley, Rochemolles (140)	Occitan, Gallo-Romance	Italy (AIS)
Susa Valley, Cesana (150)	Occitan, Gallo-Romance	Italy (AIS)
Var, Aups (886)	Occitan (Provençal), Gallo-Romance	France (ALF)
Turin (155)	Pedemontese, Gallo-Italic	Italy (AIS)
Asti (157)	Pedemontese, Gallo-Italic	Italy (AIS)
Milan (261)	Lombard, Gallo-Italic	Italy (AIS)
Bergamo (246)	Lombard, Gallo-Italic	Italy (AIS)
Nonantola (436)	Emilian, Gallo-Italic	Italy (AIS)
Loiano, (466)	Emilian, Gallo-Italic	Italy (AIS)
Barberino (515)	Central Italian, Italo-Romance	Italy (AIS)
Florence (523)	Tuscan, Italo-Romance	Italy (AIS)
Standard Italian (N/A)	Italian, Italo-Romance	Italy (N/A)
Perugia (565)	Central Italian, Italo-Romance	Italy (AIS)
Lazio, Rieti (624)	Central Italian, Italo-Romance	Italy (AIS)
Naples (721)	Southern Italian, Italo-Romance	Italy (AIS)
Basilicata, Pisticci (735)	Southern Italian, Italo-Romance.	Italy (AIS)
Calabria, Acri (762)	Extreme southern, Italo-Romance	Italy (AIS)
Palermo (803)	Sicilian, extreme southern, Italo-Romance	Italy (AIS)
Macumere (943)	Logudorese, Sardinian	Italy (AIS)
Cagliari (985)	Campidanese, Sardinian	Italy (AIS)

Switzerland), and the authors report that particular attention was paid to collecting data at equidistant intervals. For most locations, data were collected from a single informant, usually male, for 639 localities (Atlas Linguistique de la France) and 604 (Linguistic Atlas of Italy and Southern Switzerland).[4] The informants' age varied widely for both atlases, though the majority of informants were males (due to cultural restrictions of the times) and between the ages of 40 and 70 years. The two atlases were compiled with data elicited by means of a questionnaire that was administered orally by the respective researchers. The questionnaire elicited individual lexical items and simple phrases which were then transcribed by the researchers using a set of symbols and diacritics to achieve accurate phonetic representation.

3.4 Procedure

The three wordlists were applied to a total of twenty-four linguistic points. Twenty points were taken from the Linguistic Atlas of Italy and Southern Switzerland (*Sprach- und Sachatlas Italiens und der Südschweiz*, Jaberg and Jud, 1928–40) and four points from the Atlas Linguistique de la France (Gilliéron and Edmont, 1902). The linguistic points include six points within the Gallo-Italic continuum, as well as six points from Italo-Romance varieties (i.e. varieties south of the Rimini-La Spezia line, including all three traditionally identified subgroups: Central, Southern, and Extreme Southern), two points within Sardinian, and eight points representing Gallo-Romance varieties that are uncontroversially classified as separate from Italo-Romance. Within the varieties spoken in Italy, we included a point immediately north of the Rimini-La Spezia (Loiano, Atlas point 466) and one immediately to the south (Barberino, Atlas point 515) to examine the extent of the linguistic differences across the established bundle of isoglosses. The two selected points are among the closest

Fig. 3 Location of the linguistic points included in the dialectometric comparison

inhabited points across the Rimini-La Spezia line, approximately 25 km apart 'as the crow flies' (40 km by road) and with mostly uninhabited mountainous terrain between them.

Standard Italian and Standard French were also included to provide potential reference points, thus amounting to twenty-six varieties in total. A summary of the varieties included in the comparison is given in Table 1 (North-west to South-east), while Fig. 3 provides the geographical positions of the linguistic points.

As the two atlases use different transcription systems, all items were re-transcribed in the International Phonetic Alphabet by the first author.

3.5 Distance measurements

Distance between the varieties was measured using string edit distance, commonly known as Levenshtein distance. In its simplest form, Levenshtein distance is the sum of the least costly set of operations needed to 'transform one string into another' (Nerbonne and Heeringa, 2010: 553). Figure 4 illustrates how this measurement is achieved when comparing the word for 'liver' between a Lombard variety (Milan, Atlas point 261) and an Occitan variety (Le Monetier, Atlas point 971).

The Levenshtein distance is given by the total number of operations, thus yielding a distance of four for the above example. However, it has been argued that 'more phonetic sensitivity' (Nerbonne and Heeringa, 2010: 553) needs to be incorporated into string distance measures if we are to achieve accurate measurements on linguistic data. In the current research, measurements were therefore carried out according to a more 'linguistically responsible' version of Levenshtein distance measurements (Nerbonne, Colen, Gooskens, Kleiweg, and Leinonen, 2011: 73), which applies parameters that have been developed specifically for linguistic data. These parameters include the incorporation of normalization for word length as well as the requirement that consonants and vowels always be kept distinct. Finally, insertions and substitutions of diacritic marks are weighed 0.5 each (as opposed to the regular weight of 1). This allows to count an oral [a]

and a nasalized [ã] as closer than a pair of distinct vowels such as [a] and [o], thus yielding a more phonetically informed measurement.

Comparison was carried out on all items on the wordlists (i.e. both cognates and non-cognates). As non-cognates necessarily yielded distances approaching 100%, the current comparison reflects lexical distance as well as phonetic distance between the varieties. This, together with the fact that wordlists maximally exclude potential borrowings, should allow for a potentially highly accurate representation of overall linguistic distance between varieties (Bryant *et al.*, 2005; Kessler, 1995; Yang, 2009).

4 Results and Discussion

Figures 5, 6 and 7 show hierarchical clustering of the twenty-six varieties for each of the wordlists using Ward's method, which has been shown to be a highly reliable method for language clustering (Batagelj *et al.*, 1992; Nerbonne et al., 1996; Nerbonne *et al.*, 1999). There was a positive correlation between all three distance matrices, and a Mantel test revealed that all correlations were statistically significant: Swadesh 100 and Swadesh 200 ($r = 0.989$, $P = < 0.001$); Swadesh 100 and Leipzig–Jakarta ($r = 0.992$, $P = < 0.001$); Leipzig–Jakarta and Swadesh 200 ($r = 0.988$, $P = < 0.001$).

While not identical, the three wordlists return similar results and the groups that emerge are relatively similar, with some significant clustering patterns. In particular, we observe clusters that correspond to groups identified by traditional dialectological methods (for an overview, see Harris and Vincent, 2003; and Posner, 1996). Occitan varieties both sides of the French–Italian border form a single cluster in all three cases, as do the Franco-Provençal varieties of Chamonix, Chignign, and Rhêmes. The variety of Stura-Lanzo, however, is consistently clustered with Gallo-Italic rather than with the Franco-Provençal group with which it is traditionally associated. This could be due to the fact that the variety in question has a number of traits (some innovative and some conservative ones) that overlap with Gallo-Italic, such as [j] for Latin -LI-, retention of word-final -l and -s (e.g. Lombard [na:s] and Stura-Lanzo [na:s] but Franco-Provençal [na:] or [no] for 'nose'), retention of postvocalic [ŋ], and consistent use of [y] for Latin Ū (long/u/), which in most Franco-

f	i	d	i	k	
f	ɛ	d	i	k	substitute [ɛ] for [i]
f	ɛ	d		k	delete [i]
f	ɛ	d		ʒ	substitute [ʒ] for [k]
f	ɛ	d		ʒ	e insert [e]

Fig. 4 Example of distance measurement including insertion, deletion and substitution

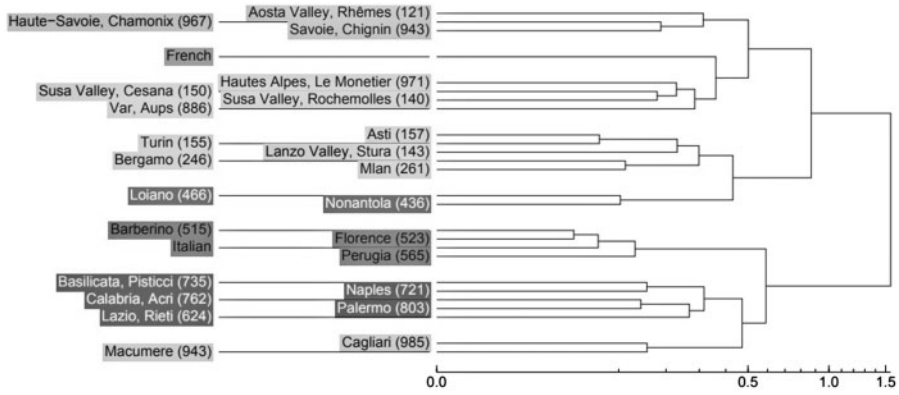

Fig. 5 Hierarchical clustering based on the Swadesh 100 wordlist

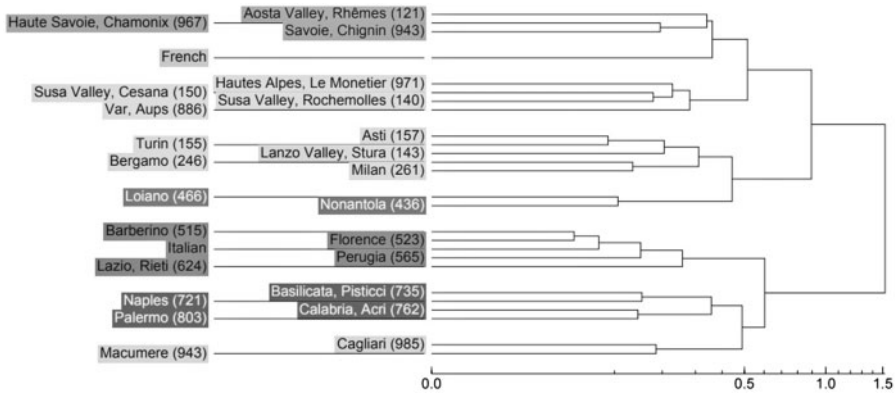

Fig. 6 Hierarchical clustering based on the Swadesh 200 wordlist

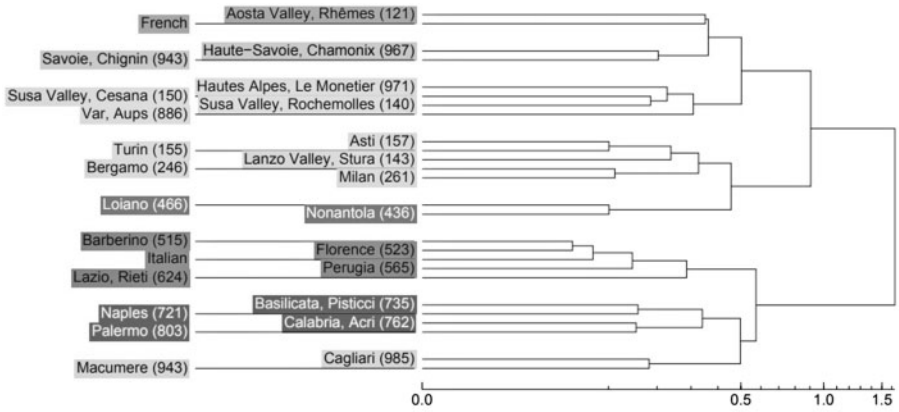

Fig. 7 Hierarchical clustering based on the Leipzig–Jakarta wordlist

Provençal varieties can be realized as [ø] or even undergo deletion.

For two of the wordlists, Franco-Provençal varieties are clustered with French, and thus classed as closer to French than to Occitan/Provençal, in line with Ascoli's (1876) original suggestion and with a more or less established consensus among dialectologists (Posner, 1996). This is not the case for the Swadesh 100 wordlist, however, suggesting that the Swadesh 200 and Leipzig–Jakarta more accurately reflect the traditional distinction between the three Gallo-Romance subgroups. The Rimini-La Spezia line also consistently emerges as a major bundle of isoglosses in all three wordlists, with varieties on each side appearing in separate clusters, showing that the systematic differences originally identified by Bartoli (1936) have been confirmed by the edit distance measurements. In particular, we can observe that Loiano (immediately north of the Rimini-La Spezia bundle of isoglosses) and Barberino (immediately to the south) are embedded in separate clusters, despite being separated by only 25 km of sparsely inhabited territory. Within varieties south of the Rimini-La Spezia line we can observe separate subclusters for varieties each side of the Rome-Ancona bundle of isoglosses originally identified by Rohlfs (1937), separating central varieties from southern varieties and ultimately from Sardinian. The Swadesh 100 list appears to be less sensitive to the isoglottic bundle of the Roma-

Ancona, clustering the variety of Rieti with Southern rather than with Central varieties. Finally, the Tuscan origins of (Standard) Italian emerge in all three wordlists, with the varieties of Florence and Barberino clustering very closely with Italian.

4.1 Classification of Gallo-Italic

The Gallo-Italic group surfaces as a relatively homogenous cluster, with Emilian varieties being slightly removed from the core cluster. As to its classification, comparisons for all three wordlists cluster Gallo-Italic varieties as closer to Gallo-Romance than to varieties south of the Rimini-La Spezia line, or Italo-Romance proper. The confirmed classification is represented in Figure 8.

This classification is consistent with the work of Schmid (1956), Bec (1970–1971), and Hull (1982) but in opposition to the rather widespread stance that takes Gallo-Italic as essentially Italo-Romance (Hall, 1974; Pellegrini, 1973, 1992; Loporcaro, 2009; among others). As dialectometric comparison of wordlists does not select any linguistic feature *a priori*, it might be the case that the analyses that have assumed close affinity between Gallo-Italic and Italo-Romance have been biased towards traits that happen to be specific to Italo-Romance. Alternatively, it may be the case that the sociolinguistic influence supposedly exerted on Gallo-Italic by Tuscan has been previously overestimated (a point originally noted by Hull, 1982). While the distinction between Gallo-Italic and other linguistic groups of Italy (i.e. Tuscan, Central Italian, Southern Italian, and Extreme Southern) is also accepted by traditional analyses, this distinction is often assumed to be an instance of sisterhood, with the different groups as parallel subsections of the Italo-Romance branch (De Mauro, 1963). Our dialectometric analysis shows that this is likely to be inaccurate, as the two subgroups are actually in a hierarchical relationship, with Gallo-Italic being closer to Occitan than to Italian, and thus considerably more distant from Italian than either southern Italian or Sicilian are. In cladistics terms, Gallo-Italic appears as an ingroup of the Gallo-Romance branch rather than in a sisterhood relationship with Italo-Romance. Similarly, the hierarchical clustering also consistently shows Gallo-Italic as being more distant from Italian than Occitan is from French, as the latter pair appear in a sisterhood relationship, while Gallo-Italic and Italian do not. This is in keeping with the view that the Rimini-La Spezia line marks a stronger bundle of isoglosses than the Oc-Oïl line (Wartburg, 1950; Lausberg, 1956). Our

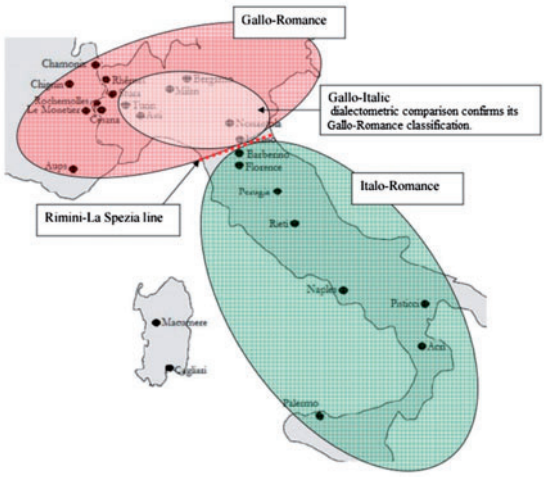

Fig. 8 Map representing the classification of Gallo-Italic as part of Gallo-Romance and separated from Italo-Romance by the Rimini-La Spezia line

results are therefore compatible with the classification of Gallo-Italic as 'a living branch of the Gallo-Romance linguistic tradition' (Hull, 1982: 660).

5 Conclusions

Classification of Gallo-Italic as either Gallo-Romance or Italo-Romance has been a point of contention in the literature. While early studies on Romance classification tended to group Gallo-Italic with Gallo-Romance (notably Schmid, 1956), a later and arguably more influential tradition argued for the classification of Gallo-Italic as Italo-Romance by also relying on *a priori* selected traits, and thus not always keeping with the tenets of the cladistic model. The current dialectmetric study showed that—when relying on all components of wordlist comparison—the relatively large bundle of isoglosses that constitute the Rimini-La Spezia line consistently leads Gallo-Italic to be clustered with Gallo-Romance and as considerably distant from Italo-Romance varieties, lending support to the analyses of Bec (1970-1971), Schmid (1956), and Hull (1982). Specifically, Gallo-Italic forms a relatively homogenous third subgroup within the Gallo-Romance branch, the other two being Occitan and Franco-Burgundian (i.e. Langue d'Oïl and Franco-Provençal), as argued in Hull's (1982) extensive analysis. The Rimini-La Spezia line therefore emerges as the most important isogloss bundle in the North–South dimension of Romance varieties.

Further research may benefit from expanding the set of linguistic properties to also include systematic comparison of grammatical properties and typological traits (Bakker et al., 2009) in combination with the phonetic and lexical ones considered here. This might reveal further similarities between Gallo-Italic and other Gallo-Romance varieties, for example in syllabic structure (Montreuil, 2000) or in the formation of clausal negation (Zanuttini, 1997).

Acknowledgments

The authors would like to thank Wilbert Heeringa for his help and advice during the data analysis. We are also grateful to two anonymous reviewers for comments and suggestions.

References

Ascoli, G. I. (1876). Paul Meyer e il franco-provenzale. *Archivio Glottologico Italiano*, **2**: 385-395.

Bakker, D., Müller, A., Velupillai, V., Wichmann, S., Brown, C. H., Brown, P. and Holman, E. W. (2009). Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology*, **13**(1): 169-181.

Barbato, M. and Varvaro, A. (2004). Dialect dictionaries. *International Journal of Lexicography*, **17**(4): 429-439.

Bartoli, M. (1936). Caratteri Fondamentali delle Lingue Neolatine. *Archivio Glottologico Italiano*, **28**: 97-133.

Batagelj, V., Pisanski, T. and Keržič, D. (1992). Automatic clustering of languages. *Computational Linguistics*, **18**(3): 339-352.

Bec, P. (1970-1971). Manuel pratique de philologie romane. A. and J. Picard, Paris.

Benincà, P. and Haiman, J. (2005). *Rhaeto-Romance Languages*. Routledge, London.

Bryant, D., Filimon, F. and Gray, R. D. (2005). Untangling our past: languages, trees, splits and networks. In Mace, R., Holden, C. J. and Shennan, S. (eds.) *The evolution of cultural diversity: a phylogenetic approach*. Left Coast Press, pp. 67-83.

Calude, A. S. and Pagel, M. (2014). Frequency of use and basic vocabulary. In Filipovic, L. and Pütz, M. (eds.) *Multilingual Cognition and Language Use: Processing and Typological Perspectives*. Amsterdam: John Benjamins Publishing, pp. 45-73.

Cerruti, M. (2011). Regional varieties of Italian in the linguistic repertoire. *International Journal of the Sociology of Language*, **210**: 9-28.

Crevatin, F. (2004). Italo-Romance etymology and dictionaries: A difficult relationship. *International Journal of Lexicography*, **17**(4): 413-428.

Dal Negro, S. and Vietti, A. (2006). The interplay of dialect and the standard in anonymous street dialogues: Patterns of variation in northern Italy. *Language Variation and Change*, **18**(2): 179–192.

De Mauro, T. (1963). *Storia linguistica dell'Italia unita*. Bari: Laterza.

Embleton, S. (1986). *Statistics in historical linguistics*. Bochum: Brockmeyer.

Emory, K. P. (1963). East Polynesian relationships: Settlement pattern and time involved as indicated by

vocabulary agreements. *The Journal of the Polynesian Society*, 78-100.

Fox, A. (1995). Linguistic reconstruction: an introduction to theory and method. Oxford University Press on Demand.

Forster, P., Toth, A. and Bandelt, H. J. (1998). Evolutionary network analysis of word lists: visualising the relationships between Alpine Romance languages. *Journal of Quantitative Linguistics*, **5**(3): 174-187.

Francescato, G. (1982). Rhaeto-Friulian. *Trends in Romance Linguistics and Philology* **3**: 131-169.

Garzonio, J. and Poletto, C. (2009). Quantifiers as negative markers in Italian dialects. *Linguistic Variation Yearbook* **9**(1): 127-151.

Gilliéron, J. and Edmont, E. (1902). Atlas linguistique de la France. Champion.

Goebl, H. and Schiltz, G. (1997). A dialectometrical compilation of CLAE 1 and CLAE 2: Isoglosses and dialect integration. *Computer developed linguistic atlas of England (CLAE)*. Tübingen: Max Niemeyer Verlag, 2.

Gooskens, C. and Heeringa, W. (2004). Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language variation and change*, **16**(03): 189-207.

Gray, R. D. and Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, **426**(6965): 435-439.

Green, J. N. (2009). Romance languages. In Comrie, B. (ed.), *The world's major languages*. London: Routledge, pp. 164–170.

Hall, R. A. (1974). *External history of the Romance languages* (Vol. 1). New York: Elsevier.

Harris, M. and Vincent, N. (eds.). (2003). *The Romance Languages*. London: Routledge.

Haspelmath, M. (2008). Loanword typology: Steps toward a systematic cross-linguistic study of lexical borrowability. In: Stolz, T., Bakker, D. and Salas Palomo, R. (eds.) *Aspects of language contact: New theoretical, methodological and empirical findings with special focus on Romancisation processes*. Berlin: Mouton de Gruyter, pp. 43-62.

Haspelmath, M. and Tadmor, U. (2009). The loanword typology project and the world loanword database. In Haspelmath, M. and Tadmor, U. (eds.) *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter, pp. 55-75.

Heeringa, W. (2004) Measuring dialect pronunciation differences using Levenshtein distance. Ph.D. thesis, University of Groningen.

Hull, G. (1982). The linguistic unity of northern Italy and Rhaetia. Ph.D. thesis, University of Sydney.

Iacobini, C. (2009). The role of dialects in the emergence of Italian phrasal verbs. *Morphology*, **19**(1): 15-44.

Jaberg, K. and Jud, J. (1928-40). *Sprach- und Sachatlas Italiens und der Südschweiz*, Zofingen.

Joseph, B. D. and Janda, R. D. (2003). *The handbook of historical linguistics*. Malden, MA: Blackwell.

Kabatek, J. and Pusch C. D. (2011). The Romance Languages. In Kortmann, B. and Van der Auwera, J. (eds.), *The languages and linguistics of Europe: a comprehensive guide* (Vol. **1**). Berlin: Walter de Gruyter.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. Proceedings of the European ACL. Dublin: ACL. 60– 67.

Kotliarov, Ivan. (2009) A Law of Elision of unstressed Vowels in Western Romance Languages. *Glotta* **85**(1-4): 77-98.

Lausberg, H. (1956). *Romanische Sprachwissenschaft* (Vol. **1**). Berlin: Walter de Gruyter.

Lewis, M. P., Simons, G. F. and Fennig, C. D. (2014). *Ethnologue: Languages of the World*, 17th ed. Dallas, Texas: SIL International. Online version: http://www.ethnologue. com.

Loporcaro, M. (2009). *Profilo linguistico dei dialetti italiani* (Vol. 275). Roma: Laterza.

Lüdtke, H. (1956). Die strukturelle Entwicklung des romanischen Vokalismus. Romanisches Seminar an der Universität Bonn.

Maiden, M. and Parry, M. M. (1997). *The dialects of Italy*. London: Routledge

McMahon, A. and McMahon, R. (2003). Finding families: quantitative methods in language classification. *Transactions of the Philological Society*, **101**(1): 7-55.

McMahon, A. and McMahon R.. (2005). *Language Classification by Numbers*. Oxford: Oxford University Press.

McMahon, A., Heggarty P., McMahon P. and Slaska, N. (2005) Swadesh Sublists and the benefits of borrowing: an Andean case study. *Transactions of the Philological Society* **103**(2): 147-170.

Merlo, C. (1960-1961). I dialetti lombardi. *L'Italia dialettale*, **24**: 1-12.

Montemagni, S., Wieling, M., de Jonge, B. and Nerbonne, J. (2013). Synchronic patterns of Tuscan phonetic variation and diachronic change: Evidence

from a dialectometric study. *Literary and Linguistic Computing*, **28**(1): 157-172.

Montreuil, J. P. (2000). Sonority and derived clusters in Raeto-romance and Gallo-italic. *Amsterdam Studies in the Theory and History of Linguistic Science* **4**: 211-238.

Nerbonne, J. (2005). Computational Contributions to the Humanities. *Literary and linguistic computing*, **20**(1): 25-40.

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P. and Leinonen, T. (2011). Gabmap-a web application for dialectology. *Dialectologia, special issue* **2**: 65-89.

Nerbonne, J. and Heeringa, W. (2010). Measuring dialect differences. In: Auer, P. and Schmidt, J. E. (eds.), *Language and Space. An international Handbook of Linguistic Variation*. Volume **1**: Theories and Methods. Berlin and New York: De Gruyter Mouton, pp. 550–567.

Nerbonne, J., Heeringa, W., Van den Hout, E., Van der Kooi, P., Otten, S. and Van de Vis, W. (1996). Phonetic distance between Dutch dialects: CLIN VI, Proceedings of the Sixth CLIN Meeting, Antwerpen, December 1995.

Nerbonne, J., Heeringa, W. and Kleiweg, P. (1999). Edit distance and dialect proximity. In Sankoff, D. and Kruskal, J. B. (eds.) *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, 2nd edition. Reading: Addison-Wesley Publication.

Nerbonne, J. and Kleiweg, P. (2007). Toward a dialectological yardstick. *Journal of Quantitative Linguistics*, **14**(2-3): 148-166.

Pei, M. A. (1949). A new methodology for Romance classification. *Word*, **5**(2), 135-146.

Pellegrini, G. B. (1973). I cinque sistemi linguistici dell'italo-romanzo. *Revue roumaine de linguistique*, **18**: 105-129.

Pellegrini, G. B. (1992). Il ''cisalpino'' e l'italo-romanzo. *Archivio Glottologico Italiano*, **77**: 272-296.

Pellegrini, G. B. (1995). Il cisalpino e il retoromanzo. Italia settentrionale: crocevia di idiomi romanzi. Atti del Convegno internazionale di studi (Trento 1993) Tübingen, 1-13.

Politzer, R. L. (1947). Final-s in the Romania. *Romanic Review*, **38**(2): 159-166.

Posner, R. (1996) *The Romance Languages*. Cambridge, Cambridge University Press.

Rama, T., Borin, L., Mikros, G. K. and Macutek, J. (2015). Comparative evaluation of string similarity measures for automatic language classification. In Mikros, G. K. and Macutek, J. (eds.) *Sequences in language and text* (Vol. **69**). Berlin/Boston: Walter de Gruyter: 171-200.

Repetti, L. (1996). Teaching about the other Italian languages: Dialectology in the Italian curriculum. *Italica*, 508-515.

Rohlfs, G. (1937). *La struttura linguistica dell'Italia* (Vol. **5**). Keller.

Rosendal, T. and Mapunda, G. (2014). Is the Tanzanian Ngoni language threatened? A survey of lexical borrowing from Swahili. *Journal of Multilingual and Multicultural Development*, **35**(3), 271-288.

Schmid, H. (1956). *Über Randgebiete und Sprachgrenzen. Vox Romanica XV*. Francke, Bern.

Schmid, S. (2012). Phonological typology, rhythm types and the phonetics-phonology interface. A methodological overview and three case studies on Italo-Romance dialects. In Ender A., Leemann, A. A and Wälchli, B. (eds.), *Methods in contemporary linguistics*. A Festschrift in honour of Iwar Werlen. Berlin/New York: Mouton de Gruyter, 45-68.

Starostin, G. (2010). Preliminary lexicostatistics as a basis for language classification: A new approach. *Journal of Language Relationship*, **3**: 79-117.

Swadesh, M. (1952). ''Lexico-statistic dating of prehistoric ethnic contacts.'' *Proceedings of the American philosophical society*, **96**(4): 452-463.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, **21**(2): 121-137.

Swadesh, M. (1971), *Origin and Diversification of Language*, Aldine· Atherton, Inc., Chicago.

Syrjänen, K., Honkola, T., Korhonen, K., Lehtinen, J., Vesakoski, O. and Wahlberg, N. (2013). Shedding more light on language classification using basic vocabularies and phylogenetic methods: a case study of Uralic. *Diachronica*, **30**(3): 323-352.

Szmrecsanyi, B. and Wolk, C. (2011). Holistic corpus-based dialectology. *Revista Brasileira de Linguística Aplicada*, **11**(2), 561-592.

Tadmor, U. (2009). Loanwords in the world's languages: Findings and results. In Haspelmath, M. and Tadmor, U. (eds.) *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter, pp. 55-75.

Tadmor, U., Haspelmath, M. and Taylor, B. (2010). Borrowability and the notion of basic vocabulary. *Diachronica*, **27**(2): 226-246.

Tang, C. and Van Heuven, V. J. (2009). Mutual intelligibility of Chinese dialects experimentally tested. *Lingua*, **119**(5): 709-732.

Temple, R. A. (2000). Old wine into new wineskins. A variationist investigation into patterns of voicing in plosives in the Atlas Linguistique de la France. *Transactions of the Philological Society*, **98**(2): 353-394.

Thomason, S. G. (2001). *Language Contact*. Washington, DC: Georgetown University Press.

Van Hout, R. and Muysken, P. (1994). Modeling lexical borrowability. *Language variation and change*, **6**(01): 39-62.

Wartburg, W. von (1950). *Die Ausgliederung der Romanischen Sprachräume*. Bern: Verlag Francke.

Wichmann, S., Holman, E. W., Bakker, D. and Brown, C. H. (2010). Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, **389**(17): 3632-3639.

Wieling, M., Montemagni, S., Nerbonne, J. and Baayen, R. H. (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language*, **90**(3): 669-692.

Yang, C. (2009). *Nisu Dialect Geography, SIL Electronic Survey Reports, 007*, Dallas: SIL International.

Zanuttini, R. (1997). *Negation and clausal structure*. Oxford University Press.

Zhang, M. and Gong, T. (2016). How Many Is Enough?—Statistical Principles for Lexicostatistics. *Frontiers in Psychology*, **7**.

Notes

1 Some scholars classify Venetian varieties as part of Gallo-Italic (e.g. Lewis *et al.*, 2014) while most do not.

2 Though the classification of Rhaeto-Romance and indeed the existence of Rhaeto-Romance as a separate branch is itself disputed, see Benincà and Haiman (2005) for discussion.

3 On the elusiveness of a linguistic definition for ''Italo-Romance'', see Maiden and Parry, (1997).

4 Though more than one informant was questioned for some areas whenever the fieldworkers thought it necessary. See Temple, 2000 for a detailed discussion of the fieldwork techniques of the Atlas Linguistique de la France.