LEXICAL DIFFERENCES BETWEEN TUSCAN DIALECTS AND STANDARD ITALIAN: ACCOUNTING FOR GEOGRAPHIC AND SOCIODEMOGRAPHIC VARIATION USING GENERALIZED ADDITIVE MIXED MODELING

MARTIJN WIELING

University of Groningen and
University of Tübingen

SIMONETTA MONTEMAGNI

Istituto di Linguistica Computationale
'Antonio Zampolli', CNR

JOHN NERBONNE

University of Groningen and
University of Freiburg

R. HARALD BAAYEN

University of Tübingen and
University of Alberta

This study uses a generalized additive mixed-effects regression model to predict lexical differences in Tuscan dialects with respect to standard Italian. We used lexical information for 170 concepts used by 2,060 speakers in 213 locations in Tuscany. In our model, geographical position was found to be an important predictor, with locations more distant from Florence having lexical forms more likely to differ from standard Italian. In addition, the geographical pattern varied significantly for low- versus high-frequency concepts and older versus younger speakers. Younger speakers generally used variants more likely to match the standard language. Several other factors emerged as significant. Male speakers as well as farmers were more likely to use lexical forms different from standard Italian. In contrast, higher-educated speakers used lexical forms more likely to match the standard. The model also indicates that lexical variants used in smaller communities are more likely to differ from standard Italian. The impact of community size, however, varied from concept to concept. For a majority of concepts, lexical variants used in smaller communities are more likely to differ from the standard Italian form. For a minority of concepts, however, lexical variants used in larger communities are more likely to differ from standard Italian. Similarly, the effect of the other community- and speaker-related predictors varied per concept. These results clearly show that the model succeeds in teasing apart different forces influencing the dialect landscape and helps us to shed light on the complex interaction between the standard Italian language and the Tuscan dialectal varieties. In addition, this study illustrates the potential of generalized additive mixed-effects regression modeling applied to dialect data.*

Keywords: Tuscan dialects, lexical variation, generalized additive modeling, mixed-effects regression modeling, geographical variation

1. INTRODUCTION. In spite of their different origins and histories, it is nowadays a widely acknowledged fact that traditional dialectology (to be understood here as dialect geography) and sociolinguistics (or urban dialectology) can be seen as two streams of a unique and coherent discipline: modern dialectology (Chambers & Trudgill 1998). Chambers and Trudgill (1998:187–88) describe the convergence of these two historically separated disciplines as follows:

> For all their differences, dialectology and sociolinguistics converge at their deepest point. Both are dialectologies, so to speak. They share their essential subject matter. Both fix the attention on language in communities. Prototypically, one has been centrally concerned with rural communities and the other with urban centres, but these are accidental differences, not essential ones and certainly not axiomatic. … A decade or two ago, it might have been possible to think that the common subject matter of dialectology and sociolinguistics counted for next to nothing. Now we know it counts for everything.

* The research reported in this article was carried out in the framework of the Short Term Mobility program of international exchanges funded by the National Research Council (CNR, Italy). This research has also benefited from the Rubicon grant awarded to Martijn Wieling by the Netherlands Organisation for Scientific Research (NWO) and the Alexander von Humboldt Professorship awarded to R. Harald Baayen.

Printed with the permission of Martijn Wieling, Simonetta Montemagni, John Nerbonne, &
R. Harald Baayen. © 2014.

In practice, however, dialectology and sociolinguistics remain separate fields when considering the methods and techniques used for analyzing language variation and change.

Sociolinguistics—whose basic goal consists of identifying the social factors underlying the use of different variants of linguistic variables—has adopted a quantitative approach to data analysis since its inception (e.g. Labov 1966). Over time, different methods for the analysis of linguistic variation were developed, capable of modeling the joint effect of an increasing number of factors related to the social background of speakers (including age, gender, socioeconomic status, etc.) and linguistic features. While early studies focused on simple relationships between the value of a linguistic variable and the value of a social variable (see e.g. Labov 1966, 1972), over time more advanced statistical methods for the analysis of linguistic variation were developed. Since the 1970s, the most common method in sociolinguistic research has been logistic regression (Cedergren & Sankoff 1974), and more recently, mixed-effects regression models have been applied to sociolinguistic data (Johnson 2009, Tagliamonte & Baayen 2012, Wieling et al. 2011).

Traditional dialectology shows a different pattern. Beginning with its origin in the second half of the nineteenth century, it typically relied on the subjective analysis of categorical maps charting the distribution of the different variants of a linguistic variable across a region. Only later, during the last forty years, have quantitative methods been applied to the analysis of dialect variation. This quantitative approach to the study of dialects is known as dialectometry (Goebl 1984, 2006, Nerbonne et al. 1996, Nerbonne 2003, Nerbonne & Kleiweg 2007, Séguy 1973). Dialectometric methods focus mostly on identifying the most important dialectal groups (i.e. in terms of geography) using an aggregate analysis of the linguistic data. The aggregate analysis is based on computing the distance (or similarity) between every pair of locations in the data set based on the complete set of linguistic variables and analyzing the resulting linguistic distance (or similarity) matrix using multivariate statistics to identify aggregate geographical patterns of linguistic variation.

While viewing dialect differences at an aggregate level arguably provides a more comprehensive and objective view than the analysis of a small number of subjectively selected features (Nerbonne 2009), the aggregate approach has never fully convinced linguists of its usefulness because it fails to identify the linguistic basis of the identified groups (see e.g. Loporcaro 2009). By initially aggregating the values of numerous linguistic variables, traditional dialectometric analyses offer no direct method for testing whether and to what extent an individual linguistic variable contributes to observed patterns of variation. Recent developments in dialectometric research have tried to reduce the gap between models of linguistic variation that are based on quantitative analyses and more traditional analyses that are based on specific linguistic features. Wieling and Nerbonne (2010, 2011) proposed a new dialectometric method, the spectral partitioning of bipartite graphs, to cluster linguistic varieties and simultaneously determine the underlying linguistic basis. Originally applied to Dutch dialects, this method was also successfully tested on English (Wieling et al. 2013, Wieling et al. 2014) and Tuscan (Montemagni et al. 2012) dialects. Unfortunately, these methods still disregard social factors, and only take into account the influence of geography.

While some attempts have been made, social and spatial analyses of language are still far from being integrated. Britain (2002) reports that, on the one hand, sociolinguistics fails to incorporate the notion of spatiality in its research. On the other hand, dialectometry mainly focuses on dialect geography and generally disregards social factors. The few exceptions indeed ‘prove’ the proverbial rule. Montemagni and colleagues (2013) and Valls and colleagues (2013) included in their dialectometric analyses social factors concerning the difference between age classes or urban versus rural communities. Unfortunately, the effect of these social factors was evaluated by simply comparing maps visually, as opposed to statistically testing the differences. Another relevant aspect on which the sociolinguistic and dialectometric perspectives do not coincide concerns the role of individual features, which are central in sociolinguistics, but are typically and programmatically disregarded in dialectometry. These issues demonstrate that there is an increasing need for statistical methods capable of accounting for both the geographic and sociodemographic variation, as well as for the impact and role of individual linguistic features.

The present study is methodologically ambitious for its attempt to combine dialectometric and sociolinguistic perspectives along the lines depicted above. The statistical analysis methods we employ enable the incorporation of candidate explanatory variables based on social, geographical, and linguistic factors, making it a good technique to facilitate the intellectual merger of dialectology and sociolinguistics (Wieling 2012). The starting point is the Wieling et al. 2011 study, which proposed a novel method using a generalized additive model in combination with a mixed-effects regression approach to simultaneously account for the effects of geographical, social, and linguistic variables. A basic generalized additive model was used to represent the global geographical pattern, which was employed in a second step as a predictor in a linear mixed-effects regression model. This model predicted word-pronunciation distances from the standard language to 424 Dutch dialects, and it turned out that both the geographical location of the communities, as well as several location-related predictors (i.e. community size and average community age), and word-related factors (i.e. word frequency and category) were significant predictors. While the Wieling et al. 2011 study includes social, lexical, and geographical information, a drawback of that study is that only a single speaker per location was considered, limiting the potential influence of speaker-related variables.

In this article, we present an extended analytical framework that was tested on an interesting case study: Tuscan lexical variation with respect to standard Italian. There are three clear and important differences with respect to the Wieling et al. 2011 study. First, since the software available for generalized additive mixed-effects regression modeling has improved significantly since that time, we are able to advance on that approach by constructing a single generalized additive mixed-effects regression model. This is especially beneficial since we are now in a position to better assess the effect of concept frequency, a variable that has largely been ignored in dialectological studies but is highly relevant as it ‘may affect the rate at which new words arise and become adopted in populations of speakers’ (Pagel et al. 2007:717). Second, in this study we focus on lexical variation rather than variation in pronunciation. We therefore do not try to predict dialect distances, but rather a binary value indicating whether the lexicalization of a concept is different (1) or equal (0) with respect to standard Italian. A benefit of this approach is that it is more in line with standard sociolinguistic practice, which also focuses on binary distinctions. Third, since we take into account multiple speakers per location, we are in an improved position to investigate the contribution of speaker-related variables such as age and gender.

The Tuscan dialect case study we use to investigate the potential of this new method (integrating social, geographical, and lexical factors) is a challenging one. In Italy a complex relationship exists between the standard language and dialects due to the history of this language and the circumstances under which Italy achieved political unification in 1861, much later than most European countries. In Tuscany, a region with a special status among Italian dialects, the situation is even more complex, since standard Italian is based on Tuscan, and in particular on the Florentine variety, which achieved national and international prestige from the fourteenth century onward as a literary language and only later (after the Italian Unification, and mainly in the twentieth century) as a spoken language. Standard Italian, however, has never been identical to genuine Tuscan and is perhaps best described as an 'abstraction' increasingly used for general communication purposes. The aim of this study, therefore, is to investigate this particular relationship between Italian and Tuscan dialects. We focus on lexical variation in Tuscan dialects compared to standard Italian with the goal of defining the impact, role, and interaction of a wide range of factors (i.e. social, lexical, and geographical) in determining lexical choice by Tuscan dialect speakers. The study is based on a large set of dialect data consisting of the lexicalizations of 170 concepts attested by 2,060 speakers in 213 Tuscan varieties drawn from the *Atlante lessicale Toscano* ('Lexical atlas of Tuscany'; Giacomelli et al. 2000).

After discussing the special relationship between standard Italian and the Tuscan dialects in the next section, we describe the Tuscan dialect data set, followed by a more in-depth explanation of the generalized additive modeling procedure, our results, and the implications of our findings.

2. Tuscan dialects and standard italian.

As pointed out by Berruto (2005), Italy's *dialetti* do not correspond to the same type of entity as, for example, the English dialects. Following the Coserian distinction among primary, secondary, and tertiary dialects (Coseriu 1980), the Italian dialects are to be understood as primary dialects (i.e. dialects having their own autonomous linguistic system), whereas the English dialects represent tertiary dialects (i.e. varieties resulting from the social and/or geographical differentiation of the standard language). Italian dialects—or, more technically, Italo-Romance varieties—thus represent not varieties of Italian but independent 'sister' languages arisen from local developments of Latin (Maiden 1995).

A similar 'sisterhood' relationship also exists between the Italian language and Italo-Romance dialects, because, as noted above, Italian has its roots in one of the speech varieties that emerged from spoken Vulgar Latin (Maiden & Parry 1997), namely that of Tuscany, and more precisely the variety of Tuscan spoken in Florence. The importance of the Florentine variety in Italy was mainly determined by the prestige of the Florentine culture, and in particular the establishment of Dante, Petrarch, and Boccaccio, who wrote in Florentine, as the 'three crowns' (*tre corone*) of Italian literature. The fact that standard Italian originated from the Florentine dialect centuries ago changes the type of relationship between standard Italian and Tuscan dialects to a kind of 'parental' relationship instead of a 'sisterhood' relationship. Clearly, this complicates matters with respect to the relationship between the Tuscan dialects and the standard Italian language, and this is the topic of the present study.

Standard Italian is unique among modern European standard languages. Even though it originated in the fourteenth century, it was not consolidated as a spoken national language until the twentieth century. For centuries, Italian was a written literary language, acquired through literacy when one learned to read and write, and was therefore only known to a minority of (literate) people. During this period, people spoke only their local dialect. For a detailed account of the rise of standard Italian the interested reader is referred to Migliorini & Griffith 1984. The particular nature of Italian as a literary language, rather than a spoken language, has been recognized since its origin and has been widely debated from different perspectives (i.e. socioeconomic, political, and cultural) under the general heading of the *questione della lingua* or 'language question'.

At the time of the Italian political unification in 1861, only a very small percentage of the population was able to speak Italian, with estimates ranging from 2.5% (De Mauro 1963) to 10% (Castellani 1982). Only during the second half of the twentieth century did real native speakers of Italian start to appear, as Italian began to be used by Italians as a spoken language in everyday life. Mass media (newspapers, radio, and TV), education, and the introduction of compulsory military service played a central role in the diffusion of the Italian language throughout the country. According to recent statistics by the Italian National Census (*Istituto Nazionale di Statistica*, ISTAT) reported by Lepschy (2002), 98% of the Italian population is able to use their national language. Dialects and standard Italian continue to coexist, however. For example, ISTAT data show that at the end of the twentieth century (1996) 50% of the population used (mainly or exclusively) standard Italian to communicate with friends and colleagues, while this percentage decreased to 34% when communication with relatives was taken into account. More recently, Dal Negro and Vietti (2011) presented a quantitative analysis of the patterns of language choice in present-day Italy on the basis of a national survey carried out by ISTAT in 2006. At the national level, they reported that 45.5% exclusively used Italian in a family setting, whereas 32.5% of the people alternated between dialectal and Italian speech, and 16% exclusively spoke in dialect (with the remaining ones using another language).

The current sociolinguistic situation of Italy is characterized by the presence of regional varieties of Italian (e.g. Berruto 1989, 2005, Cerruti 2011). Following the tripartite Coserian classification of dialects, these can be seen as tertiary dialects (i.e. varieties of the standard language that are spoken in different geographical areas). They differ both from each other and from standard Italian at all levels (phonetic, prosodic, syntactic, and lexical) and represent the Italian actually spoken in contemporary Italy. Common Italian speakers generally speak a regional variety of Italian, referred to as regional Italian. The consequence of this is that there are no real native speakers of standard Italian. Not even a Tuscan or Florentine native speaker could be considered a native speaker of standard Italian, since features that are not part of the standard Italian norm (such as the well-known Tuscan *gorgia*) exist in Tuscan or Florentine Italian.

This clearly raises the question of what we mean by the standard Italian language. Generally speaking, a standard language is a fuzzy notion. Following Ammon (2004), the standard variety of a language can be seen as having a core of undoubtedly standard forms while also having fuzzy boundaries, resulting in a complex gradation between standard and nonstandard. In Italy, a new standard variety, 'neo-standard Italian' (Berruto 1987), is emerging as the result of a restandardization process, which allows for a certain amount of regional differentiation. For the specific concerns of this study, which is aimed at reconstructing the factors governing the lexical choices of Tuscan speakers between dialect and standard language, we refer to the core of undoubtedly standard forms as standard Italian. This is the only way to avoid interference with the regional Italian spoken in Tuscany.

2.1. Previous studies on the relationship between standard italian and tuscan dialects. The specific relationship linking standard Italian and Tuscan dialects has been investigated in numerous studies. Given the goal of our research, we discuss only those studies that focus on the lexical level.

The historical link between the Tuscan dialects and the standard Italian language causes frequent overlap between dialectal and standard lexical forms in Tuscany, and less frequent overlap in other Italian regions (Giacomelli 1978). However, since Tuscan dialects have developed (for several centuries) along their own lines and independently of the (literary) standard Italian language, their vocabulary does not always coincide with that of standard Italian. Following Giacomelli (1975), the types of mismatch between standard Italian and the dialectal forms can be partitioned into three groups. The first group consists of Tuscan words that are used in literature throughout Italy, but are not part of the standard language (i.e. these terms usually appear in Italian dictionaries marked as 'Tuscanisms'). The second group consists of Tuscan words that were part of old Italian and are also attested in the literature throughout Italy, but have fallen into disuse because they are considered old-fashioned (i.e. these terms may appear in Italian dictionaries marked as 'archaisms'). The final group consists of Tuscan dialectal words that have no literary tradition and are not understood outside of Tuscany.

Here our goal is to investigate the complex relationship between standard Italian and the Tuscan dialects from which it originated on the basis of the data collected through fieldwork for the *Atlante lessicale Toscano* (ALT). Previous studies have already explored the ALT data set by investigating the relationship between Tuscan and Italian from the lexical point of view. Giacomelli and Poggi Salani (1984) based their analysis on the dialect data available at the time. More recently, Montemagni (2008a) applied dialectometric techniques to the whole ALT dialectal corpus to investigate the relationship between Tuscan and Italian. In both cases it turned out that the Tuscan dialects overlap most closely with standard Italian in the area around Florence, expanding in different directions and in particular toward the southwest. Obviously, this observed synchronic pattern of lexical variation has the well-known diachronic explanation of the standard Italian language having originated from the Florentine variety of Tuscan.

Montemagni (2008a) also found that the observed patterns varied depending on the speaker's age: only 37% of the dialectal answers of the older speakers (i.e. born in 1920 or before) overlapped with standard Italian, while this increased to 44% for the younger speakers (i.e. born after 1945, when standard Italian started being used progressively more). In addition, words having a larger geographical coverage (i.e. not specific to a small region) were more likely to coincide with the standard language than words attested in smaller areas. These first, basic results illustrate the potential of the ALT data set we use to shed light on the complex relationship between standard Italian and Tuscan dialects.

3. Material.

The material used in this study consists of both lexical and sociolinguistic data and is discussed in detail in the following sections.

3.1. Lexical data.

The lexical data used in this study were taken from the ALT, a specially designed regional atlas in which the dialectal data have a diatopic (geographic), diastratic (social), and diachronic characterization. The diachronic characterization covers only a few generations whose years of birth range from the end of the nineteenth century to the second half of the twentieth century. It is interesting to note that only the younger ALT informants were born in the period when standard Italian started being used as a spoken language. ALT interviews were carried out between 1974 and 1986 in 224 localities of Tuscany. The localities were hierarchically organized according to their size, ranging from medium-sized urban centers (excluding big cities) to small villages and rural areas. In total there were fifty to sixty micro-areas, each placed around an urban center (for more details see Giannelli 1978). In contrast to traditional atlases (which typically rely on elderly and uneducated informants), the ALT includes 2,193 informants who were selected based on a number of parameters, ranging from age and socioeconomic status to education and culture, in order to be representative of the population of each location. The sample size for the individual localities ranges between four and twenty-nine informants, depending on the population size. The temporal window covered by the ALT makes this data set particularly suitable for exploring the complex relationship linking Tuscan dialects to standard Italian along several dimensions (i.e. across space, time, and socially defined groups). The interviews were conducted by a group of trained fieldworkers who employed a questionnaire of 745 target items, designed to elicit variation mainly in vocabulary and semantics.

Because the compilation of the ALT questionnaire was aimed at capturing the specificity of Tuscan dialects and their relationships, concepts whose lexicalizations were identical to Italian (almost) everywhere in Tuscany were programmatically excluded (Giacomelli 1978, Poggi Salani 1978). This makes the ALT data set particularly useful for better understanding the complex relationship linking the standard language and local dialects in the case the two did not coincide.

In this study, we focus on Tuscan dialects only, spoken in 213 out of the 224 investigated locations (see Figure 1; Gallo-Italian dialects spoken in Lunigiana and in small areas of the Apennines were excluded), reducing the number of informants to 2,060. We used the normalized lexical answers to a subset of the ALT onomasiological questions (i.e. those looking for the attested lexicalizations of a given concept).[1] Normalization was meant to abstract away from phonetic variation and in particular away from productive phonetic processes, without removing morphological variation or variation caused by unproductive phonetic processes. Out of 460 onomasiological questions, we selected those that prompted fifty or fewer distinct normalized lexical answers (the maximum in all onomasiological questions was 421 unique lexical answers). We used this threshold to exclude questions having many hapaxes corresponding to nonlexicalized answers: this is the case, for instance, for productive figurative usages (e.g. metaphors such as *cetriolo* ‘cucumber’ and *carciofo* ‘artichoke’ for ‘stupid’) or productive derivational processes (e.g. *scemaccio* and *scemalone* from the lexical root *scemo* ‘stupid’). From the resulting 195-item subset, we excluded a single adjective and twelve verbs (since the remaining concepts were nouns) and all twelve multiword concepts in order to avoid interference from other types of variation. Our final subset, therefore, consisted of 170 concepts and is listed in Table 1.

The representativeness of the selected sample of 170 concepts with respect to the whole set of onomasiological questions was assessed in various ways. First, we measured the correlation between overall lexical distances and lexical distances focusing on the selected sample,[2] which turned out to be very high ($r = 0.94$).[3] Second, we tested the overall distribution of answer types within the selected subset and the whole set of onomasiological questions. In both cases, it turned out that the distribution of answers appears to conform to the asymptotic hyperbolic distribution discussed by Kretzschmar

[1] Although the ALT data set also includes passive vocabulary, for this study we focused on the active vocabulary only.

[2] Lexical distances between each pair of locations (i.e. aggregating over informants) were calculated by measuring the Levenshtein distance between the lexical forms used in both locations per concept and averaging these across all concepts (or only concepts present in the subset). When the Levenshtein distance is used, related lexical forms (which are orthographically similar) do not increase the lexical distance as much as unrelated lexical forms. This approach is in line with the one used by Nerbonne and Kleiweg (2003) and Montemagni (2008b).

[3] A similar approach was taken in Montemagni et al. 2012 with respect to phonetic variation.

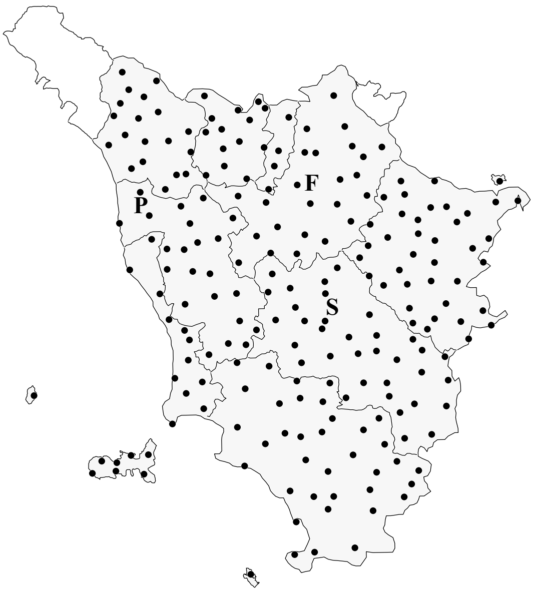

FIGURE 1. Geographical distribution of the 213 locations investigated in this study. ‘F’, ‘S’, and ‘P’ mark the approximate locations of Florence, Siena, and Pisa, respectively.

and Tamasi (2003) as being common to dialect data and known as the ‘A-curve’. In spite of the different sizes of the two data sets, the percentages of hapaxes with respect to the whole set of answer types were comparable, 18.6% in the subset of 170 concepts and 21.3% in the whole set of onomasiological questions. Montemagni (2010) also reports that the A-curve distribution applies to the ALT data set, regardless of the number of answers gathered with respect to a given questionnaire item. We can thus conclude that the selected sample can be usefully exploited for the specific concerns of this study.

The normalized lexical forms in the ALT data set still contained some morphological variation. In order to assess the pure lexical variation we abstracted away from variation originating in, for example, assimilation, dissimilation, or other phonological differences (e.g. the dialectal variants *camomilla* and *capomilla*, meaning ‘chamomile’, were treated as instantiations of the same normalized form), as well as away from both inflectional and derivational morphological variation (e.g. inflectional variants such as singular and plural are grouped together). We compare these more abstract forms to the Italian standard.

The list of standard Italian words denoting the 170 concepts was extracted from the online ALT dialectal resource (ALT-Web, available at http://serverdbt.ilc.cnr.it/altweb). This resource was created as a way for the user to identify the ALT question(s) corresponding to his or her research interests (see Cucurullo et al. 2006). The list of concepts, originally compiled on the basis of lexicographic evidence, was carefully reviewed by members of the *Accademia della Crusca*,[4] the leading institution in the field of research on the Italian language in both Italy and the world, in order to make sure that it contained undoubtedly standard Italian forms and not old-fashioned or literary words originating in Tuscan dialects (see §2.2).

In every location multiple speakers were interviewed (see above) and therefore each normalized answer is anchored to a given location, but also to a specific speaker. Since some speakers provided multiple distinct answers to denote a single concept, the total number of cases (i.e. concept-speaker-answer combinations) was 384,454.

[4] http://www.accademiadellacrusca.it/

abete	‘fir’	faraona	‘guinea fowl’	pigna	‘cone’
acacia	‘acacia’	fiammifero	‘match’	pimpinella	‘pimpernel’
acino	‘grape’	filare	‘spin’	pinolo	‘pine seed’
acquaio	‘sink’	formica	‘ant’	pioppeto	‘poplar grove’
albicocca	‘apricot’	fragola	‘strawberry’	pipistrello	‘bat’
allodola	‘lark’	frangia	‘fringe’	polenta	‘cornmeal mush’
alloro	‘laurel’	frantoio	‘oil mill’	pomeriggio	‘afternoon’
anatra	‘duck’	fregatura	‘cheat’	presine	‘potholders’
angolo	‘ext. angle’	fringuello	‘finch’	prezzemolo	‘parsley’
anguria	‘watermelon’	frinzello	‘badly done darn’	pula	‘chaff’
ape	‘bee’	fronte	‘front’	pulce	‘flea’
arancia	‘orange’	fuliggine	‘soot’	pulcino	‘chick’
aromi	‘aromas’	gazza	‘magpie’	puzzola	‘skunk’
aspide	‘asp’	gelso	‘mulberry’	radice	‘root’
bigoncia	‘vat’	ghiandaia	‘jay’	raganella	‘tree frog’
borraccina	‘moss’	ghiro	‘dormouse’	ramaiolo	‘ladle’
bottiglia	‘bottle’	ginepro	‘juniper’	ramarro	‘green lizard’
brace	‘embers’	gomitolo	‘ball’	rana	‘frog’
braciere	‘brazier’	grandine	‘hail’	ravanelli	‘radishes’
braciola	‘chop’	grappolo	‘cluster’	riccio	‘hedgehog’
bruco	‘caterpillar’	grattugia	‘grater’	riccio (castagna)	‘chestnut husk’
cachi	‘khaki’	grillo	‘cricket’	ricotta	‘ricotta cheese’
caglio	‘rennet’	idraulico	‘plumber’	rosmarino	‘rosemary’
calabrone	‘hornet’	lampo	‘flash’	sagrato	‘churchyard’
calderaio	‘tinker’	lentiggini	‘freckles’	salice	‘willow’
calvo	‘bald’	lucertola	‘lizard’	saliva	‘saliva’
camomilla	‘chamomile’	lumaca	‘snail’	salsiccia	‘sausage’
cantina	‘cellar’	madrina	‘godmother’	scoiattolo	‘squirrel’
capezzolo	‘nipple’	maiale	‘pig’	scorciatoia	‘shortcut’
capocollo	‘Tuscan cold cut from pork shoulder’	maialino	‘piglet’	scrofa	‘sow’
caprone	‘goat’	mammella	‘breast’	seccatoio	‘squeegee’
carbonaio	‘charcoal’	mancia	‘tip’	sedano	‘celery’
cascino	‘cheese mold’	manciata	‘handful’	segale	‘rye’
castagnaccio	‘chestnut cake’	mandorla	‘almond’	sfoglia	‘pastry’
castagneto	‘chestnut’	mangiatoia	‘manger’	siero	‘serum’
cavalletta	‘grasshopper’	matassa	‘hank’	soprassata	‘Tuscan salami made from the pig (offal)’
cetriolo	‘cucumber’	matterello	‘rolling pin’	spazzatura	‘garbage’
ciabatte	‘slippers’	melone	‘melon’	spigolo	‘edge’
ciccioli	‘greaves’	mietitura	‘harvest’	stollo	‘haystack pole’
ciliegia	‘cherry’	mirtillo	‘blueberry’	stoviglie	‘dishes’
cimice	‘bug’	montone	‘ram’	straccivendolo	‘ragman’
cintura (f)	‘belt for woman’	mortadella	‘Italian sausage’	susina	‘plum’
cintura (m)	‘belt for man’	neve	‘snow’	tacchino	‘turkey’
cipresso	‘cypress’	nocciola	‘hazelnut’	tagliere	‘chopping board’
cispa	‘eye gum’	oca	‘goose’	talpa	‘mole’
cocca	‘corner of tissue’	occhiali	‘glasses’	tartaruga	‘tortoise’
coperchio	‘cover’	orcio	‘jar’	trabiccolo (rotondo)	‘dome frame for bed heating’
corbezzolo	‘arbutus’	orecchio	‘ear’	trabiccolo (allungato)	‘elongated frame for bed heating’
corniolo	‘dogwood’	orzaiolo	‘sty’	trogolo	‘trough’
crusca	‘bran’	ovile	‘sheepfold’	truciolo	‘chip’
cuneo	‘wedge’	ovolo	‘royal agaric’	tuono	‘thunder’
dialetto	‘dialect’	padrino	‘godfather’	uncinetto	‘crochet’
ditale	‘thimble’	pancetta	‘bacon’	upupa	‘hoopoe’
donnola	‘weasel’	pancia	‘belly’	verro	‘boar’
duna	‘dune’	panzanella	‘Tuscan bread salad’	vitalba	‘clematis’
edera	‘ivy’	papavero	‘poppy’	volpe	‘fox’
falegname	‘carpenter’	pettirosso	‘robin’		

TABLE 1. List of all 170 lexical items included in this study, including their English translations.

Since Wieling et al. 2011 reported a significant effect of word frequency on dialect distances from standard Dutch pronunciations (with more frequent words having a higher distance from standard Dutch, which was interpreted as a higher resistance to standardization), we obtained the concept frequencies (of the standard Italian lexical form) by extracting the corresponding frequencies from a large corpus of 8.4 million Italian unigrams (Brants & Franz 2009). The corpus-based frequency ranking of these concepts was then compared to the *Grande dizionario italiano dell'uso* ('Comprehensive dictionary of Italian usage'; De Mauro 2000), which represents a standard usage-based reference resource for the Italian language including quantitative information on vocabulary use. In particular, a list of about 7,000 high-frequency concepts highly familiar to native speakers of Italian was identified in this dictionary, representing the so-called BASIC ITALIAN VOCABULARY (BIV). It turned out that 59.4% of the concepts used in our study belonged to the BIV, whereas the remaining concepts refer to an old-fashioned and traditional world (19.4%), denote less common plants and animals (14.7%), or refer to kitchen tools (2.4%). The remaining 4.1% of the concepts represent a miscellaneous group. It is interesting to note that the classification of concepts with respect to this reference dictionary and the frequency data obtained from the large web corpus are aligned, with our most frequent concepts being in the BIV and the less frequent concepts typically corresponding to old-fashioned and traditional notions as well as less common plants and animals.

3.2. SOCIOLINGUISTIC DATA. The speaker information we obtained consisted of the speaker's year of birth, gender, education level (ranging from 1: illiterate or semi-literate, to 6: university degree; for this variable about 1.3% of the values were missing), and employment history (in nine categories: farmer; craftsman; trader or businessman; executive or auxiliary worker; knowledge worker, manager, or nurse; teacher or freelance worker; common laborer or apprentice; skilled or qualified worker; or nonprofessional status such as student, housewife, or retired). Furthermore, we obtained the year of recording for every location, and we extracted demographic information about each of the 213 locations from a website with statistical information about Italian municipalities (Comuni Italiani 2011). We extracted the number of inhabitants (in 1971 or 1981, whichever year was closer to the year when the interviews for that location were conducted), the average income (in 2005, which was the oldest information available), and the average age (in 2007, again the oldest information available) in every location. While the information about the average income and average age was relatively recent and may not precisely reflect the situation at the time when the data set was constructed (between 1974 and 1986), the global pattern will probably be relatively similar.

4. METHODS. Since the statistical method we use, generalized additive mixed-effects regression, is relatively new, the following sections provide a detailed explanation of our approach. To replicate the results published in this article, the data and commands used for the analysis (including results and full-color animated graphs) are available for download from the mind research repository (http://openscience.uni-leipzig.de) and the first author's website (http://www.martijnwieling.nl). In addition, the appendix shows the function call used to fit the complete generalized additive mixed-effects regression model.

4.1. MODELING THE ROLE OF GEOGRAPHY: GENERALIZED ADDITIVE MODELING. In contrast to a linear regression model in which a single predictor is linear in its effect on the dependent variable, in a generalized additive model (GAM) the assumption is relaxed so that the functional relation between a predictor and the response variable need not be linear. Instead, the GAM provides the user with a flexible toolkit for smoothing nonlinear relations in any number of dimensions. Consequently, the GAM is much more flexible than the simple linear regression model. In a GAM, multiple predictors may be combined in a single smooth, yielding essentially a wiggly surface (when two independent variables are combined) or a wiggly hypersurface (when three or more independent variables are combined).

A GAM combines a standard linear model with regression coefficients $\beta_0, \beta_1, \dots, \beta_k$ with smooth functions $s(\dots)$ for one or more predictors: $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + s(X_i) + s(X_j, X_k) + \dots$.

A suitable option for smoothing a single predictor is to use cubic regression splines. These fit piecewise cubic polynomials (functions of the form $y = a + bx + cx^2 + dx^3$) to separate intervals of the predictor values. The transitions between the intervals (located at the knots) are ensured to be smooth since the first and second derivative are forced to be zero. The number of knots determines how smooth the curve is. Determining the appropriate amount of smoothing is part of the parameter estimation process.

To combine predictors that have the same scale (such as longitude and latitude), thin plate regression splines are a suitable choice. These fit a wiggly regression surface as a weighted sum of geometrically regular surfaces. When the predictors do not all have the same scale, tensor products can be used (Wood 2006:162). These define surfaces given marginal basis functions, one for each dimension of the smooth. The basis functions generally are cubic regression splines (but they can be thin plate regression splines as well), and the greater the number of knots for the different basis functions, the more wiggly the fitted regression surface will be. More information about the tensor product bases (which are implemented in the mgcv package for R) is provided by Wood (2006:145–220). A more extended introduction about the use of generalized additive modeling in linguistics can be found in Baayen et al. 2010.

As it turns out, a thin plate regression spline is a highly suitable approach to model the influence of geography in dialectology, since geographically closer varieties tend to be linguistically more similar (e.g. see Nerbonne 2010) and the dialectal landscape is generally quite smooth. Note, however, that the method also can detect steep transitions, as exemplified by the two-dimensional smooths presented by Kryuchkova and colleagues (2012) in their analysis of ERP data associated with auditory lexical processing. Wieling et al. 2011 also used a generalized additive model to represent the global effect of geography, since this measure is more flexible than using, for example, distance from a certain point (Jaeger et al. 2011). In this study, we take a more sophisticated approach, allowing the effect of geography to vary for concept frequency and speaker age. Furthermore, we use a generalized additive LOGISTIC model, since our dependent variable is binary (in line with standard sociolinguistic practice using VARBRUL; Cedergren & Sankoff 1974). Logistic regression does not model the dependent variable directly, but it attempts to model the probability (in terms of logits) associated with the values of the dependent variable. A logit is the natural logarithm of the odds of observing a certain value (in our case, a lexical form different from standard Italian). Consequently, when interpreting the parameter estimates of our regression model, we should realize that these need to be interpreted with respect to the logit scale (i.e. the natural logarithm of the odds of observing a lexical form different from standard Italian). More detailed information about logistic regression is provided by Agresti (2007).

As an illustration of the GAM approach, Figure 2 presents the global effect of geography on lexical differences with respect to standard Italian. The complex wiggly surface shown here was modeled by a thin plate regression spline (Wood 2003), which was also used in Wieling et al. 2011. The (solid) contour lines represent isolines connecting areas that have a similar likelihood of having a lexical form different from standard Italian. Note that the values here represent log-odds values (as we use logistic regression) and should be interpreted with respect to being different from standard Italian. This means that lower values indicate a smaller likelihood of being different (intuitively, it is therefore easiest to view these values as a distance measure from standard Italian). Consequently, the value −0.1 indicates that in those areas the lexical form is more likely to match the Italian standard (the probability is 0.475 that the lexical form is different from the Italian standard form), and the value 0.1 indicates the opposite (the probability is 0.525 that the lexical form is different from the Italian standard form). The colors correspond to the isolines with increasing values indicated by green, yellow, orange, and light gray, respectively. Intuitively, the map can be viewed as a terrain map with a green plane (low) and light gray mountain peaks (high). Thus, the green color indicates a greater likelihood of having lexical forms identical to those in standard Italian, while light gray represents a greater likelihood of having lexical forms different from those in standard Italian. We can clearly see that locations near Florence (indicated by the ‘F’) tend to have lexical variants more likely to be identical to the standard Italian form. This makes sense since Italian originated from the Tuscan dialect spoken in Florence. The 27.5 estimated degrees of freedom invested in this general thin plate regression spline were supported by a chi-square value of 1,580 ($p < 0.001$).

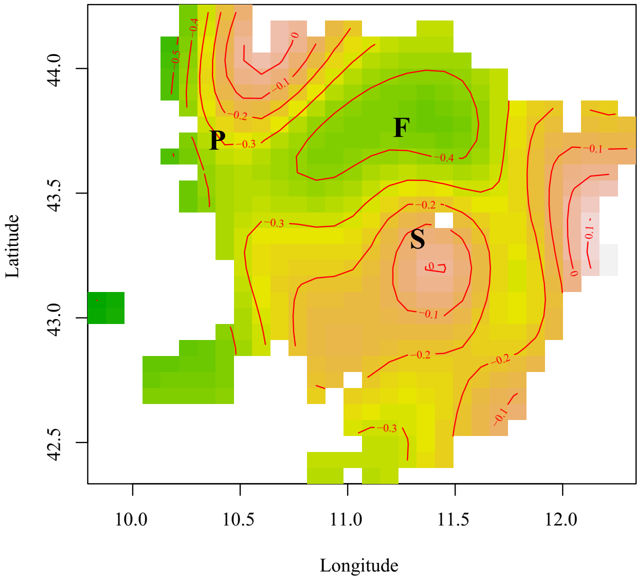

FIGURE 2. Contour plot for the regression surface of predicting lexical differences from standard Italian as a function of longitude and latitude, obtained with a generalized additive model using a thin plate regression spline. The (red) contour lines represent isolines, green and yellow (lower values) indicate a smaller likelihood of having a lexical form different from standard Italian, and orange and light gray (higher values) represent locations with a greater likelihood of having a lexical form different from standard Italian. ‘F’, ‘S’, and ‘P’ mark the approximate locations of Florence, Siena, and Pisa, respectively. The white squares indicate combinations of longitude and latitude for which there are no (nearby) data.

Since Wieling et al. 2011 found that the effect of word frequency on (Dutch) dialect distances varied per location, we initially created a three-dimensional smooth (longitude × latitude × concept frequency), allowing us to assess the concept-frequency-specific geographical pattern of lexical variation with respect to standard Italian. For example, it might be that the geographical pattern presented in Fig. 2 holds for concepts having an average frequency, but might be somewhat different for concepts with a low as opposed to a high frequency. Since our initial analyses revealed that this pattern varied depending on speaker age, we also included the speaker's year of birth in the smooth, resulting in a four-dimensional smooth (longitude × latitude × concept frequency × speaker's year of birth). We model this four-dimensional smooth by a tensor product. In the tensor product, we model both longitude and latitude with a thin plate regression spline (since this is suitable for combining isotropic predictors and also in line with the approach used in Wieling et al. 2011), while the effect of concept frequency and speaker's year of birth are modeled by two separate cubic regression splines.

4.2. Mixed-effects modeling. A generalized additive mixed-effects regression model distinguishes between fixed and random-effect factors. Fixed-effect factors have a small number of levels exhausting all possible levels (e.g. gender is either male or female). Random-effect factors, in contrast, have levels sampled from a much larger population of possible levels. In our study, concepts, speakers, and locations are random-effect factors, since there are many other concepts, speakers, or locations possible. By including random-effect factors, the model can take the systematic variation linked to these factors into account. For example, some concepts will be more likely to be different from standard Italian than others (regardless of location), and some locations (e.g. near Florence) or speakers will be more likely to use lexical variants similar to standard Italian (across all concepts). These adjustments to the population intercept (consequently identified as 'random intercepts') can be used to make the regression formula more precise for every individual location and concept.

It is also possible that there is variability in the effect a certain predictor has. For example, while the general effect of community size might be negative (i.e. larger communities have lexical variants more likely to match the standard Italian form), there may be significant variability for the individual concepts. While most concepts will follow the general pattern, some concepts could even exhibit the opposite pattern (i.e. being more likely to match the standard Italian form in smaller communities). In combination with the by-concept random intercepts, these by-concept random slopes make the regression formula for every individual concept as precise as possible. Furthermore, taking this variability into account prevents type I errors in assessing the significance of the predictors of interest. The significance of random-effect factors in the model was assessed by the Wald test. More information and an introduction to mixed-effects regression models is provided in Baayen et al. 2008.

In our analyses, we considered the three aforementioned random-effect factors (i.e. location, speaker, and concept), as well as several other predictors besides the geographical variation (specific to concept frequency and speaker age). The additional speaker-related variables we included were gender, education level, and employment history (coded in nine binary variables denoting if a speaker has had each specific job or not). The demographic variables we investigated were community size, average community age, average community income, and the year of recording.

To reduce the potentially harmful effect of outliers, several numerical predictors were log-transformed (i.e. community size, average age, average income, education level, and concept frequency). We scaled all numerical predictors by subtracting the mean and dividing by the standard deviation in order to facilitate the interpretation of the fitted parameters of the statistical model.

4.3. Combining mixed-effects regression and generalized additive modeling. In contrast to the approach of Wieling et al. 2011, where first a separate generalized additive model (similar to the one illustrated in Fig. 2) was created and the fitted values of this model were used as a predictor in a mixed-effects regression model, we are now able to create a single generalized additive mixed-effects regression model, which estimates all parameters simultaneously. Since the software to construct a generalized additive model is continuously evolving, this approach was not possible previously. The specification of our generalized additive mixed-effects regression model using the mgcv package for R is shown in the appendix.

5. Results. We fitted a generalized additive mixed-effects logistic regression model, step by step removing predictors that did not contribute significantly to the model. In the following we discuss the specification of the model, including all significant predictors and verified random-effect factors.

Our response variable was binary, with a value of 1 indicating that the lexical form was different from the standard Italian form and a value of 0 indicating that the lexical form was equal to standard Italian. The coefficients and the associated statistics of the significant fixed-effect factors and linear covariates are presented in Table 2. To allow a fair comparison of the effects of all predictors, we included a measure of effect size by specifying the increase or decrease of the likelihood of having a nonstandard Italian lexical form (in terms of logits) when the predictor increased from its minimum to its maximum value. Table 3 presents the significance of the four-dimensional smooth term (modeling the geographical pattern dependent on concept frequency and speaker age), and Table 4 lists the significant random-effects structure of our model.[5]

To evaluate the goodness of fit of the final model (see Tables 2 to 4), we used the index of concordance *C*. This index is also known as the receiver operating characteristic curve area 'C' (see e.g. Harrell 2001). Values of *C* exceeding 0.8 are generally regarded as indicative of a successful classifier. According to this measure, the model performed well with $C = 0.82$.

	ESTIMATE	STD. ERROR	*z*-VALUE	*p*-VALUE	EFF. SIZE
(intercept)	−0.4188	0.1266	−3.31	< 0.001	
Community size (log)	−0.0584	0.0224	−2.60	0.009	−0.3618
Male gender	0.0379	0.0128	2.96	0.003	0.0380
Farmer profession	0.0460	0.0169	2.72	0.006	0.0460
Education level (log)	−0.0686	0.0126	−5.44	< 0.001	−0.2757

Table 2. Significant parametric terms of the final model. A positive estimate indicates that a higher value for this predictor increases the likelihood of having a nonstandard Italian lexical form, while a negative estimate indicates the opposite effect. Effect size indicates the increase or decrease of the likelihood of having a nonstandard Italian lexical form when the predictor value increases from its minimum to its maximum value (i.e. the complete range).

	EST. DOF	CHI. SQ.	*p*-VALUE
Geography × concept frequency × speaker's year of birth	225.9	3,295	< 0.001

Table 3. Significant smooth term of the final model. The estimated degrees of freedom (DOF) of the smooth term is indicated, as well as its significance in the model. Figure 5 below shows the visualization.

[5] The effect of removing the morphological variation (see §3.1) was relatively limited, since the results on the basis of the original data were mainly identical to the results shown in Tables 3, 4, and 5 (where morphological variation was removed). The only difference was that in the data set including the morphological variation, speakers who had a teaching or freelance profession were significantly ($p = 0.03$) more likely to use a standard Italian form than those who had another profession (this variable was not significant in the data set excluding morphological variation: $p = 0.12$).

FACTORS	RANDOM EFFECTS	STD. DEV.	*p*-VALUE
Speaker	(intercept)	0.0100	0.006
Location	(intercept)	0.1874	< 0.001
Concept	(intercept)	1.6205	< 0.001
	Year of recording	0.2828	< 0.001
	Community size (log)	0.1769	< 0.001
	Average community income (log)	0.2657	< 0.001
	Average community age (log)	0.2400	< 0.001
	Farmer profession	0.1033	< 0.001
	Executive or auxiliary worker prof.	0.0650	0.002
	Education level (log)	0.1255	< 0.001
	Male gender	0.0797	< 0.001

TABLE 4. Significant random-effect parameters of the final model. The standard deviation indicates the amount of variation for every random intercept and slope.

5.1. SPEAKER-RELATED PREDICTORS. An inspection of Table 2 shows clearly that the contribution of the speaker-related variables is generally in line with well-established results from sociolinguistics. We see that men were much more likely than women to use nonstandard forms, which is unsurprising since men generally use a higher frequency of nonstandard forms than women (Cheshire 2002), and this is also in line with previously reported gender differences for Tuscany (Cravens & Giannelli 1995, Binazzi 1996). Similarly, farmers were also found to be more likely to use nonstandard forms. A reasonable explanation for this is that people living in rural areas (as farmers tend to do, given the nature of their work) generally favor nonstandard forms and are less exposed to other language varieties (e.g. Chambers & Trudgill 1998). The final significant speaker-related variable was education level. Higher-educated speakers used forms more likely to be identical to the Italian standard. Again, this finding is not unforeseen, as higher-educated people tend to use more standard forms (e.g. Gorman 2010).

As shown in Table 4, the effect of all speaker-related variables varied per concept. For example, Figure 3 visualizes the effect of education level per concept (i.e. the by-concept random slopes for education level). In this graph, each circle represents a concept, and these are sorted (from left to right) by the effect the speaker's education level has on the likelihood of the concept being different from standard Italian. The concept experiencing the strongest negative effect of the speaker's education level (i.e. a greater likelihood for higher-educated speakers to use a lexical form identical to standard Italian) is represented by the first (i.e. left-most) circle, whereas the concept experiencing the strongest positive effect of the speaker's education level (i.e. a greater likelihood for higher-educated speakers to use a lexical form different from standard Italian) is represented by the last (i.e. right-most) circle. Consequently, concepts such as *upupa* 'hoopoe' (a bird species) and *abete* 'fir' follow the general pattern (with higher-educated speakers being more likely to use a standard form; the main, negative effect is indicated by the dashed line), while concepts such as *verro* 'boar' and *cocca* 'corner of a tissue' show the opposite behavior (with less educated people being more likely to use the standard form). As remarked before, taking these by-concept random slopes into account allows us to more reliably assess the general effect of the fixed-effect predictors (i.e. the main effects).

5.2. DEMOGRAPHIC PREDICTORS. Of all demographic predictors (i.e. community size, average community income, and average community age), only the first was significant as a main effect in the model. Larger communities were more likely to have a lexical variant identical to standard Italian (i.e. the estimate in Table 2 is negative). A possible explanation for this finding is that people tend to have weaker social ties in urban com-

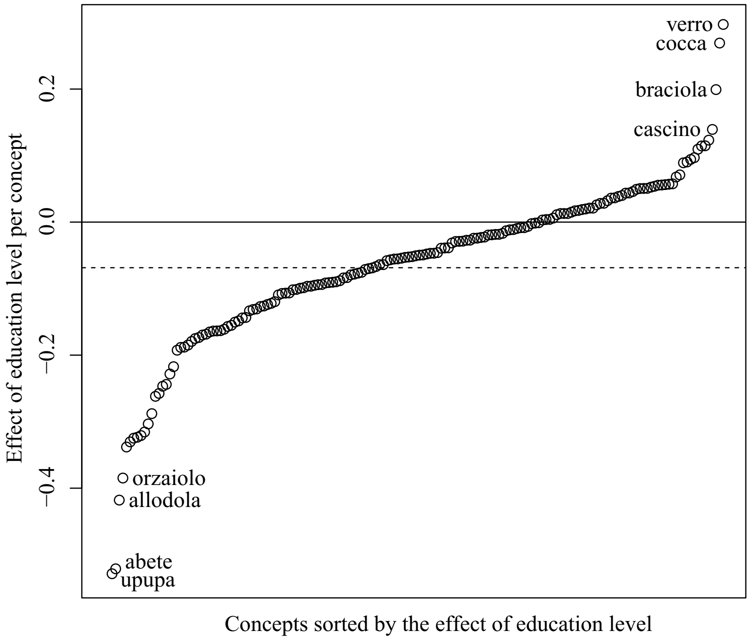

FIGURE 3. By-concept random slopes of education level. The concepts (represented by circles) are sorted by the value of their education-level coefficient (i.e. the effect of education level of the speakers). Strongly negative coefficients (bottom left) are associated with concepts that are more likely to be identical to standard Italian for higher-educated speakers, while positive coefficients (top right) are associated with concepts that are more likely to be different from standard Italian for higher-educated speakers. The estimate of the main effect (see Table 2) is indicated by the dashed line.

munities, which causes dialect leveling (i.e. socially or locally marked variants tend to be leveled in favor of the standard language in conditions of social or geographical mobility and the resulting dialect contact; Milroy 2002).

All demographic variables (i.e. community size, average income, and average age) as well as year of recording showed significant by-concept variation. Similar to Fig. 3 (which showed the effect of education level per concept), Figure 4 visualizes the effect of community size per concept. The graph clearly shows some concepts (e.g. *trabiccolo* ‘elongated frame for bed heating’ and *mirtillo* ‘blueberry’) that are more likely to be identical to standard Italian in larger communities (i.e. consistent with the general pattern; the main effect is indicated by the dashed line), while others behave in completely opposite fashion (i.e. *frinzello* ‘badly done darn’ and *nocciola* ‘hazelnut’) and are more likely to be different from standard Italian in larger communities.

5.3. GEOGRAPHICAL VARIATION AND LEXICAL PREDICTORS. It is clear from the large chi-square value seen in Table 3 that geography is a very strong predictor, in interaction with concept frequency and speaker age. We validated that the geographical pattern was justified by comparing the AIC values (Akaike information criterion; Akaike 1974) for different models. The AIC indicates the relative goodness of fit of the model, with lower values signifying an improved model. Including geography was necessary since the AIC for a model without geography (but including all predictors and random-effect factors shown in Table 2 and Table 4) was 393,242, whereas the AIC for the model including a simple geographical smooth slightly decreased (i.e. improved) to 393,238. Note that the improvement is relatively small (but more than the threshold of two AIC

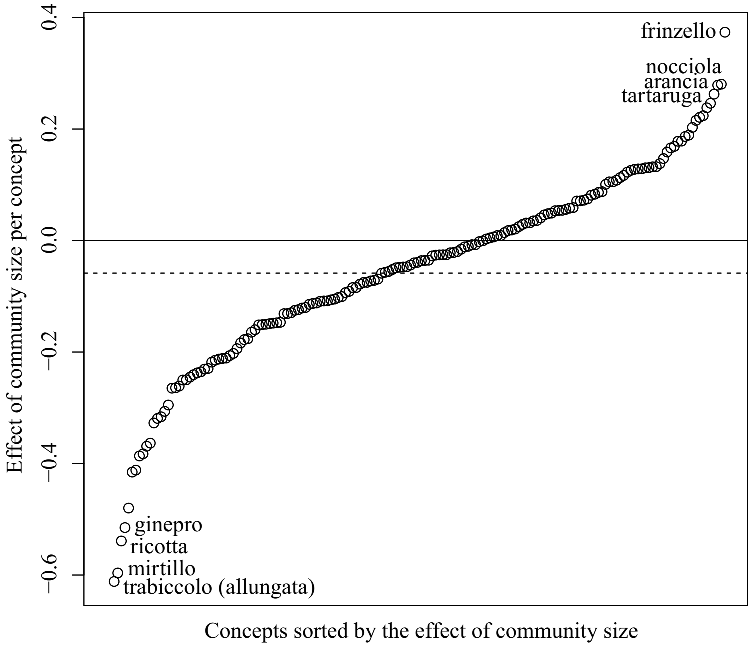

FIGURE 4. By-concept random slopes of community size. The concepts are sorted by the value of their community-size coefficient (i.e. the effect of community size). Strongly negative coefficients (bottom left) are associated with concepts that are more likely to be identical to standard Italian in larger communities, while positive coefficients (top right) are associated with concepts that are more likely to be different from standard Italian in larger communities. The estimate of the main effect (see Table 2) is indicated by the dashed line.

units), since a random intercept for location is included that allows the locations to vary in their likelihood of using a lexical form different from standard Italian, and essentially takes over the role of the geographical smooth when it is not included. When varying the geographical effect by speaker age, the AIC reduced more strongly to 392,727, while varying it by concept frequency resulted in an AIC of 391,041. The best model (with an AIC of 390,479) was obtained when the geographical effect varied depending on concept frequency and speaker age. Figure 5 visualizes the geographical variation related to concept frequency and speaker age. As before, increasing values (i.e. a greater likelihood of having a lexical form different from standard Italian) are indicated by green, yellow, orange, and light gray, respectively.

The three graphs to the left present the geographical patterns for the older speakers, while those to the right present the geographical patterns for the younger speakers. When going from the top to bottom, the graphs show the geographical pattern for increasing concept frequency.

The first observation is that all graphs show the same general trend, according to which speakers from Florence (marked by the ‘F’) or the area immediately surrounding it are more likely to use a standard Italian form than the speakers from the more peripheral areas. Of course, this makes sense since standard Italian originated from Florence. Note, however, that the likelihood of using a standard Italian form varies significantly depending on the age of the speakers and the frequency of concepts.

With respect to the age of the speakers, comparing the left and right graphs yields a straightforward pattern: the right graphs are generally characterized by lower values than the left ones, indicating that the younger speakers are much more likely to use a standard Italian form.

Let us now consider the effect of concept frequency. For the older speakers, we observe that the lexicalizations of high-frequency concepts are less likely to be identical to standard Italian than those of low-frequency concepts (i.e. the graph of the high-frequency concepts is less green than the graph of the low-frequency concepts). For the younger speakers, a slightly different pattern can be observed. While the high-frequency concepts are less likely to be identical to standard Italian than the mean-frequency concepts, the low-frequency concepts are also somewhat less likely to be identical to standard Italian.

In the following section, we discuss these results and offer a possible interpretation for this complex but statistically well-supported geographical pattern.

6. Discussion. In this study we have used a generalized additive mixed-effects regression model to identify the factors influencing the lexical choice of Tuscan speakers between dialect and standard Italian forms. In line with standard results from sociolinguistics, we found clear support for the importance of the speaker-related variables gender, education level, and profession. Men, farmers, and lower-educated speakers were more likely to use a lexical form different from standard Italian.

The only demographic predictor that reached significance in our study was community size. Larger communities were more likely to have a lexical variant identical to standard Italian, and this is in line with results reported in Wieling et al. 2011 for Dutch. Also in agreement with that study is that we did not observe a significant effect of average income. However, in contrast to Wieling et al. 2011, we did not find a significant influence of the average age in a community. The effect of average community age may be less powerful in our study because we also included speaker age (which is obviously much more suitable for detecting the influence of age). The final predictor that did not reach significance in our study was year of recording. This is likely caused by the relatively short time span (with respect to lexical change) in which the data were gathered.

The pattern shown in Fig. 5 revealed that the likelihood of using a lexical form different from standard Italian varied in a geographically coherent way in interaction with speaker age and concept frequency. The interpretation of these results is complicated by the fact that three different types of language variation and change are involved. First, there is dialectal variation within Tuscany, with a history going back to long before the emergence of standard Italian. Second, there is the development of standard Italian from the prestigious dialects of Tuscany, foremost from the literary Florentine variety, but also from the dialects of Tuscany as spoken in and near Florence. Third, with the establishment of a standard language, this standard itself is now affecting the dialects of Tuscany, resulting in dialect leveling.

Our hypothesis is that the pattern of results observed for the older speakers largely reflects dialect differentiation within Tuscany, with relatively little influence from standard Italian. In contrast, the effect of the standard language on the Tuscan dialects is clearly visible when the younger generations are compared with the older generations. With respect to the influence of Tuscan dialects on standard Italian, our results suggest that it is the lower-frequency forms that were borrowed by the standard language, along with the literary vocabulary, from the prestigious Florentine variety.

To see this, consider again the interaction of frequency by geography for the older speakers (Fig. 5, left panels). The older speakers are unlikely to have undergone substantial influence from standard Italian, since many of these speakers grew up when there was no (spoken) standard Italian yet. As a consequence, the changing dialect landscape as a function of increasing concept frequency must reflect, to a considerable extent, original dialect differences within Tuscany. The most striking difference between low-frequency and high-frequency concepts is found for the rural areas to the southwest of Florence. Here, we observe that the higher-frequency concepts are more different from the standard language, while the lower-frequency concepts are more similar to the standard language.

The close similarity of the low-frequency vocabulary to the corresponding vocabulary in the standard language indicates that these concepts must have been borrowed by the standard language from the original Tuscan dialects. Since we are dealing with older speakers, it is unlikely that they would have adopted these low-frequency, often agricultural, concepts from the (at that time) still-emerging standard Italian language.

The greater differences for the high-frequency vocabulary is reminiscent of two independent findings in the literature that suggest that high-frequency words/concepts are more resistant to dialect leveling. Pagel and colleagues (2007), in a study of lexical replacement in Indo-European languages, reported that words denoting frequently used concepts are less prone to be replaced (possibly because they are better entrenched in memory and therefore more resistant to lexical replacement). Wieling et al. 2011 likewise reported a resistance to standardization for high-frequency words in Dutch dialects. We therefore interpret the greater difference from standard Italian for the higher-frequency concepts as reflecting dialect differences within Tuscany that were able to resist leveling toward the emerging norm (rooted in the old Florentine dialect), thanks to better entrenchment in memory (Baayen et al. 1997, Bybee 2003).

Influence in the reverse direction, from standard Italian on Tuscan dialects, is clearly visible for the younger speakers. Compared to the older speakers, the younger speakers have a vocabulary that is much closer to that of standard Italian. Interestingly, our data indicate that younger speakers in the area around Siena are most resistant to dialect leveling for concepts with frequencies in the medium and higher ranges.

The geographical distribution is most similar for older and younger speakers for the lower-frequency concepts. Above, considering the distributional pattern for older speakers, we argued that these lower-frequency concepts must have been, to a considerable extent, incorporated into standard Italian. For the younger speakers, however, the Florentine dialects are closest to standard Italian for the concepts of intermediate frequency. The reason for this is straightforward: the lowest-frequency concepts represent objects that were important in rural agricultural societies, but that have lost importance for modern urban speakers. Many of these concepts have an archaic flavor to the modern ear (such as *stollo* ‘haystack pole’). Since these concepts are hardly used in standard Italian, the standardizing effect of the standard language is limited. In these cases, younger speakers will likely lack the specific words for denoting these concepts and use more general terms instead (mismatching with the standard Italian form).

If our hypothesis is correct, it is important to distinguish between the dynamics of language variation and change between the older and younger speakers. The older speakers show, according to our hypothesis, a dialect landscape in which the higher-frequency concepts resist accommodation to the prestigious Florentine-rooted norm within Tuscany. Conversely, the younger speakers show, again for the higher-frequency concepts, resistance against dialect leveling, but now against standard Italian.

The results that have emerged from our analysis of the ALT corpus thus shed new light on the typology, impact, and role of a wide range of factors underlying the lexical choices by Tuscan speakers. Previous studies, based both on individual words (Giacomelli & Poggi Salani 1984) and on aggregated data (Montemagni 2008a), provided a flat view, according to which Tuscan dialects overlap most closely with standard Italian

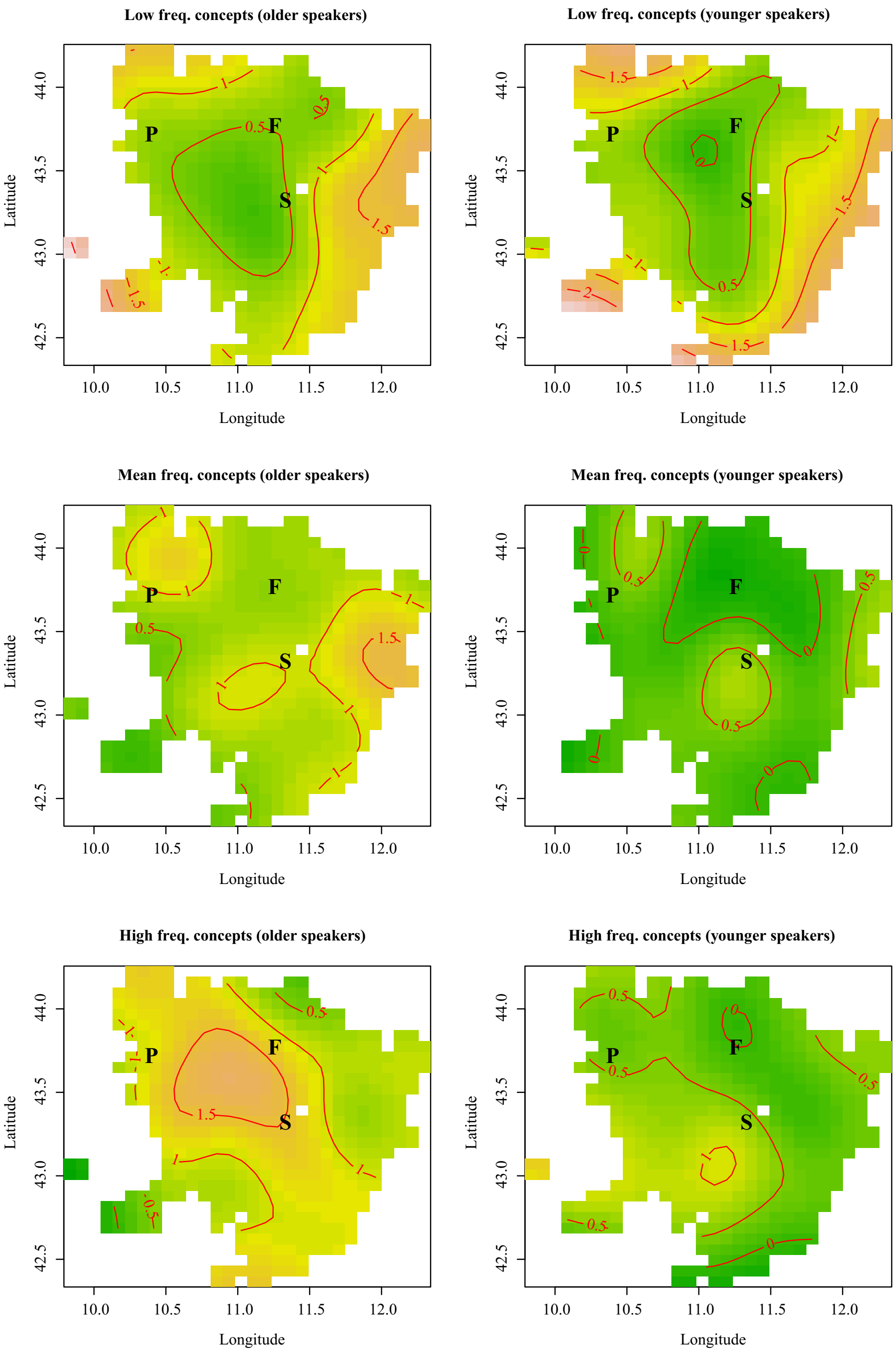
Low freq. concepts (older speakers)
Low freq. concepts (younger speakers)
Mean freq. concepts (older speakers)
Mean freq. concepts (younger speakers)
High freq. concepts (older speakers)
High freq. concepts (younger speakers)
P
F
S
Latitude
Longitude
42.5
43.0
43.5
44.0
10.0
10.5
11.0
11.5
12.0
0
0.5
1
1.5
2

FIGURE 5. Contour plots for the regression surface of predicting lexical differences from standard Italian as a function of longitude, latitude, concept frequency, and speaker age, obtained with a generalized additive model (see also Fig. 2). The left plots visualize the results for older speakers (two standard deviations below the mean year of birth of 1931, i.e. 1888), while the right plots show those for the younger speakers (two standard deviations above the mean year of birth of 1931, i.e. 1974). The top row visualizes the contour plots for low-frequency concepts (two standard deviations below the mean), the middle row for concepts having the mean frequency, and the bottom row for high-frequency concepts (two standard deviations above the mean). See the text for interpretation.

in the area around Florence, with expansions in different directions and in particular toward the southwest. Montemagni's (2008a) aggregate analysis illustrated that a higher likelihood of using standard Italian was connected with speaker age and geographical coverage of words. In the present study, however, a more finely articulated picture emerged about the interplay of Tuscan dialect variation, the transfer of Tuscan vocabulary to standard Italian, and the influence of standard Italian on the modern Tuscan dialect landscape.

Importantly, we would like to stress that the method we applied in this study, generalized additive mixed-effects regression, is able to simultaneously capture the diatopic, diastratic, and diachronic dimensions of language variation. Since the method also allows a focus on individual linguistic features, we think it is an excellent candidate to facilitate the intellectual merger of dialectology and sociolinguistics.

A limitation of this study is that it proceeded from dialect atlas data, which inherently suffers from a sampling bias. Furthermore, in order to keep the analysis tractable and to focus on purely lexical variation we selected a subset of the data from the dialect atlas. While still having a relatively large number of items, our data set consisted only of nouns. Since the influence of word category might also vary geographically (see Wieling et al. 2011), further research is necessary to see if the results of this study extend to other word categories.

Another interesting line of research that might be worth pursuing would be to resort to a more sensitive distance measure with respect to standard Italian, such as the Levenshtein (or edit) distance, rather than the binary lexical difference measure used in this study. In that case, lexical differences that are closely related (i.e. in the case of lexicalized analogical formations) can be distinguished from deeper lexical differences (e.g. due to a different etymon).

In conclusion, thanks to the temporal window covered by the ALT data set, it was possible to keep track of the spreading of standard Italian and its increasing use as a spoken language. Real standardization effects could only be observed with respect to younger speakers, whereas older generations turned out to prefer dialectal variants, especially for the higher-frequency concepts.

Appendix: Function call fitting the generalized additive model

```
library(mgcv) # version 1.7-27

# random intercepts and slopes are denoted by s(...,bs="re")
model = bam (UnequalToStandardItalian ~
CommunitySize.log + MaleGender + FarmerProfession + EducationLevel.log +
te(Longitude,Latitude,ConceptFreqeuncy,SpeakerYearBirth,d=c(2,1,1)) +
s(Speaker,bs="re") + s(Location,bs="re") + s(Concept,bs="re") +
s(Word,YearOfRecording,bs="re") + s(Word,CommunitySize.log,bs="re") +
s(Word,AverageCommunityIncome.log,bs="re") + s(Word,AverageCommunityAge.log,bs="re")
+ s(Word,FarmerProfession,bs="re") + s(Word,ExecutiveOrAuxiliaryWorkerProfession,bs="re")
+ s(Word,EducationLevel.log,bs="re") + s(Word,MaleGender,bs="re"), data=lexdst,
family="binomial")

# show the results of the model
summary(model)
```

REFERENCES

Agresti, Alan. 2007. *An introduction to categorical data analysis*. 2nd edn. Hoboken, NJ: John Wiley & Sons.

Akaike, Hirotugu. 1974. A new look at the statistical identification model. *IEEE Transactions on Automatic Control* 19.716–23.

Ammon, Ulrich. 2004. Standard variety. *Sociolinguistics: An international handbook of the science of language and society*, 2nd edn., ed. by Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier, and Peter Trudgill, vol. 1, 273–83. Berlin: Mouton de Gruyter.

Baayen, R. Harald; Doug J. Davidson; and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59.390–412.

Baayen, R. Harald; Ton Dijkstra; and Robert Schreuder. 1997. Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language* 37.94–117.

Baayen, R. Harald; Victor Kuperman; and Raymond Bertram. 2010. Frequency effects in compound processing. *Compounding*, ed. by Sergio Scalise and Irene Vogel, 257–70. Amsterdam: John Benjamins.

Berruto, Gaetano. 1987. *Sociolinguistica dell'italiano contemporaneo*. Roma: La Nuova Italia Scientifica.

Berruto, Gaetano. 1989. Main topics and findings in Italian sociolinguistics. *International Journal of the Sociology of Language* 76.5–30.

Berruto, Gaetano. 2005. Dialect/standard convergence, mixing, and models of language contact: The case of Italy. *Dialect change: Convergence and divergence in European languages*, ed. by Peter Auer, Frans Hinskens, and Paul Kerswill, 81–97. Cambridge: Cambridge University Press.

Binazzi, Neri. 1996. Giovani uomini e giovani donne di fronte al lessico della tradizione: Risultati di un'analisi sul campo. *Donna e linguaggio: Atti del Convegno Internazionale di Studi*, ed. by Gianna Marcato, 569–79. Padova: Cleup.

Brants, Thorsten, and Alex Franz. 2009. Web 1T 5-gram, 10 European languages, version 1. Philadelphia: Linguistic Data Consortium.

Britain, David. 2002. Space and spatial diffusion. In Chambers et al., 603–37.

Bybee, Joan. 2003. *Phonology and language use*. Cambridge: Cambridge University Press.

Castellani, Arrigo. 1982. Quanti erano gli italofoni nel 1861? *Studi Linguistici Italiani* 8.3–26.

Cedergren, Henrietta J., and David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50.333–55.

Cerruti, Massimo. 2011. Regional varieties of Italian in the linguistic repertoire. *International Journal of the Sociology of Language* 210.9–28.

Chambers, Jack K., and Peter Trudgill. 1998. *Dialectology*. 2nd edn. Cambridge: Cambridge University Press.

Chambers, Jack K.; Peter Trudgill; and Natalie Schilling-Estes (eds.) 2002. *The handbook of language variation and change*. Oxford: Blackwell.

Cheshire, Jenny. 2002. Sex and gender in variationist research. In Chambers et al., 423–43.

Comuni Italiani. 2011. Informazioni e dati statistici sui comuni in Italia, le province e le regioni italiane. Online: http://www.comuni-italiani.it.

Coseriu, Eugenio. 1980. 'Historische Sprache' und 'Dialekt'. *Dialekt und Dialektologie*, ed. by Joachim Göschel, Pavle Ivic, and Kurt Kehr, 106–22. Wiesbaden: Steiner.

Cravens, Thomas D., and Luciano Giannelli. 1995. Relative salience of gender and class in a situation of multiple competing norms. *Language Variation and Change* 7.261–85.

Cucurullo, Nella; Simonetta Montemagni; Matilde Paoli; Eugenio Picchi; and Eva Sassolini. 2006. Dialectal resources on-line: The ALT-Web experience. *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genova, 1846–51.

Dal Negro, Silvia, and Alessandro Vietti. 2011. Italian and Italo-Romance dialects. *International Journal of the Sociology of Language* 210.71–92.

De Mauro, Tullio. 1963. *Storia linguistica dell'Italia unita*. Roma-Bari: Laterza.

De Mauro, Tullio. 2000. *Grande dizionario italiano dell'uso*. Torino: UTET.

Giacomelli, Gabriella. 1975. Dialettologia toscana. *Archivio Glottologico Italiano* 60.179–91.

Giacomelli, Gabriella. 1978. Come e perchè il questionario. In Seminario di Dialettologia Italiana, 19–26.

Giacomelli, Gabriella; Luciano Agostiniani; Patrizia Bellucci; Luciano Giannelli; Simonetta Montemagni; Annalisa Nesi; Matilde Paoli; Eugenio Picchi;

and TERESA POGGI SALANI. 2000. *Atlante lessicale Toscano*. Roma: Lexis Progetti Editoriali.

GIACOMELLI, GABRIELLA, and TERESA POGGI SALANI. 1984. Parole toscane. *Quaderni dell'Atlante Lessicale Toscano* 2.123–229.

GIANNELLI, LUCIANO. 1978. L'indagine come ricerca delle diversità. In Seminario di Dialettologia Italiana, 35–50.

GOEBL, HANS. 1984. *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Tübingen: Niemeyer.

GOEBL, HANS. 2006. Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing* 21.411–35.

GORMAN, KYLE. 2010. The consequences of multicollinearity among socioeconomic predictors of negative concord in Philadelphia. *University of Pennsylvania Working Papers in Linguistics* 16.2.66–75.

HARRELL, FRANK. 2001. *Regression modeling strategies*. Berlin: Springer.

JAEGER, T. FLORIAN; PETER GRAFF; WILLIAM CROFT; and DANIEL PONTILLO. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology: Commentary on Atkinson. *Linguistic Typology* 15.281–319.

JOHNSON, DANIEL EZRA. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3.359–83.

KRETZSCHMAR, WILLIAM A., JR., and SUSAN TAMASI. 2003. Distributional foundations for a theory of language change. *World Englishes* 22.377–401.

KRYUCHKOVA, TATIANA; BENJAMIN V. TUCKER; LEE H. WURM; and R. HARALD BAAYEN. 2012. Danger and usefulness in auditory lexical processing: Evidence from electroencephalography. *Brain and Language* 122.81–91.

LABOV, WILLIAM. 1966. *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

LABOV, WILLIAM. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

LEPSCHY, GIULIO. 2002. *Mother tongues & other reflections on the Italian language*. Toronto: University of Toronto Press.

LOPORCARO, MICHELE. 2009. *Profilo linguistico dei dialetti italiani*. Roma-Bari: Laterza.

MAIDEN, MARTIN. 1995. *A linguistic history of Italian*. London: Longman.

MAIDEN, MARTIN, and MAIR PARRY. 1997. *The dialects of Italy*. London: Routledge.

MIGLIORINI, BRUNO, and T. GWYNFOR GRIFFITH. 1984. *The Italian language*. London: Faber and Faber.

MILROY, LESLEY. 2002. Social networks. In Chambers et al., 549–72.

MONTEMAGNI, SIMONETTA. 2008a. Analisi linguistico-computazionali del corpus dialettale dell'Atlante Lessicale Toscano: Primi risultati sul rapporto toscano-italiano. *Discorsi di lingua e letteratura italiana per Teresa Poggi Salani*, ed. by Annalisa Nesi and Nicoletta Maraschio, 247–60. Pisa: Pacini.

MONTEMAGNI, SIMONETTA. 2008b. The space of Tuscan dialectal variation: A correlation study. *International Journal of Humanities and Arts Computing* 2.135–52.

MONTEMAGNI, SIMONETTA. 2010. Esplorazioni computazionali nello spazio della variazione lessicale in Toscana. *Atti del Convegno Parole: Il lessico come strumento per organizzare e trasmettere gli etnosaperi*, ed. by Nadia Pratera, Antonio Mendicino, and Cinzia Citraro, 619–44. Rende: Centro Editoriale e Librario dell'Università della Calabria.

MONTEMAGNI, SIMONETTA; MARTIJN WIELING; BOB DE JONGE; and JOHN NERBONNE. 2012. Patterns of language variation and underlying linguistic features: A new dialectometric approach. *La variazione nell'italiano e nella sua storia: Varietà e varianti linguistiche e testuali: Atti dell'XI Congresso Società Internazionale di Linguistica e Filologia Italiana*, ed. by Patricia Bianchi, Nicola De Blasi, Chiara De Caprio, and Francesco Montuori, 879–89. Firenze: Franco Cesati Editore.

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* 28.157–72.

NERBONNE, JOHN. 2003. Linguistic variation and computation. *Proceedings of the 10th meeting of the European Chapter of the Association for Computational Linguistics*, Budapest, 3–10.

NERBONNE, JOHN. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3.175–98.

Nerbonne, John. 2010. Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.3821–28.

Nerbonne, John; Wilbert Heeringa; Erik van den Hout; Peter van de Kooi; Simone Otten; and Willem van de Vis. 1996. Phonetic distance between Dutch dialects. *Papers from the sixth CLIN Meeting*, ed. by Gert Durieux, Walter Daelemans, and Steven Gillis, 185–202. Antwerp: Centre for Dutch Language and Speech.

Nerbonne, John, and Peter Kleiweg. 2003. Lexical distance in LAMSAS. *Computers and the Humanities* 37.339–57.

Nerbonne, John, and Peter Kleiweg. 2007. Toward a dialectological yardstick. *Journal of Quantitative Linguistics* 14.148–67.

Pagel, Mark; Quentin D. Atkinson; and Andrew Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449.717–20.

Poggi Salani, Teresa. 1978. Dialetto e lingua a confronto. In Seminario di Dialettologia Italiana, 51–65.

Séguy, Jean. 1973. La dialectométrie dans l'atlas linguistique de Gascogne. *Revue de Linguistique Romane* 37.1–24.

Seminario di Dialettologia Italiana. 1978. *Atlante lessicale toscano—Note sul questionario*. Firenze: Facoltà di Lettere e Filosofia.

Tagliamonte, Sali A., and R. Harald Baayen. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24.135–78.

Valls, Esteve; Martijn Wieling; and John Nerbonne. 2013. Linguistic advergence and divergence in northwestern Catalan: A dialectometric investigation of dialect leveling and border effects. *Literary and Linguistic Computing* 28.119–46.

Wieling, Martijn. 2012. *A quantitative approach to social and geographical dialect variation*. Groningen: University of Groningen dissertation.

Wieling, Martijn, and John Nerbonne. 2010. Hierarchical spectral partitioning of bipartite graphs to cluster dialects and identify distinguishing features. *Proceedings of the 2010 ACL Workshop on Graph-based Methods for Natural Language Processing*, Uppsala, 33–41.

Wieling, Martijn, and John Nerbonne. 2011. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language* 25.700–715.

Wieling, Martijn; John Nerbonne; and R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation socially and geographically. *PLoS ONE* 6.9.e23613. Online: http://dx.plos.org/10.1371/journal.pone.0023613.

Wieling, Martijn; Robert G. Shackleton, Jr.; and John Nerbonne. 2013. Analyzing phonetic variation in the traditional English dialects: Simultaneously clustering dialects and phonetic features. *Literary and Linguistic Computing* 28.31–41.

Wieling, Martijn; Clive Upton; and Ann Thompson. 2014. Analyzing the BBC *Voices* data: Contemporary English dialect areas and their characteristic lexical variants. *Literary and Linguistic Computing* 29.107–17.

Wood, Simon. 2003. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B* 65.95–114.

Wood, Simon. 2006. *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.

Wieling
University of Groningen
Department of Humanities Computing
P.O. Box 716
9700 AS Groningen, The Netherlands
[wieling@gmail.com]
[simonetta.montemagni@ilc.cnr.it]
[j.nerbonne@rug.nl]
[harald.baayen@uni-tuebingen.de]

[Received 9 January 2012;
revision invited 24 September 2012;
revision received 25 April 2013;
accepted 29 October 2013]