

**LEXICAL DIFFERENCES BETWEEN TUSCAN DIALECTS AND
STANDARD ITALIAN: A SOCIOLINGUISTIC ANALYSIS USING
GENERALIZED ADDITIVE MIXED MODELING**

Martijn Wieling^a, Simonetta Montemagni^b, John Nerbonne^a and R. Harald Baayen^c

^aDepartment of Humanities Computing, University of Groningen, ^bIstituto di Linguistica
Computazionale “Antonio Zampolli”, CNR, Italy, ^cDepartment of General Linguistics, University of
Tübingen and Department of Linguistics, University of Alberta
m.b.wieling@rug.nl, simonetta.montemagni@ilc.cnr.it, j.nerbonne@rug.nl, baayen@ualberta.edu

Abstract

This study is the first to use 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 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 old versus young speakers. Several other factors emerged as significant. The model predicts that lexical variants used by older speakers and 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, smaller communities have lexical forms different from standard Italian. For a smaller minority of concepts, however, larger communities tend to have lexical forms different from standard Italian (more than smaller communities). Similarly, the effect of average community income and average community age also varied per concept. These results clearly identify important factors involved in dialect variation at the lexical level in Tuscany. In addition, this study illustrates the potential of generalized additive mixed modeling applied to dialect data.

Key words

Tuscan dialects, Lexical variation, Generalized additive modeling, Mixed-effects regression modeling

1. Introduction

In this study we investigate a Tuscan lexical dialect dataset using a generalized additive model (GAM) in order to identify sociolinguistic and concept-related factors which play an important role in predicting lexical differences with respect to standard Italian.

1.1. *The relationship between standard Italian and Tuscan*

Standard Italian is unique among modern European standard languages. Although Italian 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. A good account of the rise of standard Italian is provided by Migliorini and Griffith (1984). The particular nature of Italian as a literary language, rather than a spoken language was recognized since its origin and widely debated from different (i.e. socio-economic, political and cultural) perspectives as the *questione della lingua* or ‘language question’.

At the time of the Italian political unification in 1860 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 real native speakers of Italian started to appear, as Italian started to be used by Italians as a spoken language in everyday life. Mass media (newspapers, radio and TV) and education played a central role in the diffusion of the Italian language throughout the country. According to the most recent statistics of ISTAT (*Istituto per le ricerche statistiche*) reported by Lepschy (2002), 98% of the Italian population is able to use their national language. However, dialects and standard Italian appear to coexist. 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.

To see the reason for the coexistence of standard Italian and local dialects, the origin of the standard language has to be taken into account. Italian has its roots in one of the speech varieties that emerged from spoken Vulgar Latin (Maiden and 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 the Italian literature. Consequently, Italian dialects do not represent varieties of the Italian language, but they are simply ‘sisters’ of the Italian language (Maiden, 1995).

In contrast to other Italian regions where a sort of ‘sisterhood’ relationship holds between the standard language and local dialects, in Tuscany this relationship is complicated by the fact that standard Italian originated from the Florentine dialect centuries ago. This also causes the frequent overlap between dialectal and standard Italian forms in Tuscany, which occurs much less frequently in other Italian regions (Giacomelli, 1978). However, since the Florentine dialect has developed (for several centuries) along its own lines and independently of the (literary) standard Italian language, its vocabulary does not always coincide with 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 which 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 which *were* part of old Italian and are also attested in the literature throughout Italy, but have fallen into disuse as they are considered old-fashioned (i.e. these terms may appear in Italian dictionaries glossed as ‘archaisms’). The final group consists of Tuscan dialectal words which have no literary tradition and are not understood outside of Tuscany.

This study investigates the relationship between standard Italian and the Tuscan dialects from which it originated on the basis of the data collected through fieldwork for a regional linguistic atlas, the *Atlante Lessicale Toscano* (‘Lexical Atlas of Tuscany’, henceforth ALT; Giacomelli et al., 2000). We undertook the study to shed new light on the widely debated Italian *questione della lingua*. The ALT is a specially designed lexical atlas in which the dialectal data both have a diatopic (geographic)

and diastratic (social) characterization. In particular, the advanced regression techniques we apply make it possible to keep track of the sociolinguistic and lexical factors at play in the complex relationship linking the Tuscan dialects with standard Italian. The ALT data appear to be particularly suitable to explore the Italian language question from the Tuscan point of view. Since 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 means that the ALT dataset was collected with the main purpose of better understanding the complex relationship linking the standard language and local dialects in the case the two did not coincide (see above).

Previous studies have already explored the ALT dataset 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 that time. Montemagni (2008) more recently 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 towards the southwest. Obviously, this observed synchronic pattern of lexical variation has the well-known diachronic explanation that the standard Italian language originated from the Florentine variety of Tuscan.

Montemagni (2008) also found that the observed patterns varied depending on the speaker's age: only 37 percent of the dialectal answers of the old speakers overlapped with standard Italian, while this percentage increased to 44 for the young speakers. 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 dataset, which will be discussed in the following section, to shed light on the widely debated *questione della lingua* from the point of view of Tuscan dialects.

1.2. Regression models applied to dialect variation

The present study is methodologically ambitious. On the one hand, we take a dialectometric perspective by using a large set of dialect data (i.e. 170 concepts in 213 Tuscan varieties), seeking in this way to strengthen the signals in the data and to prevent potentially biased choices among linguistic features, and subsequently to obtain a study whose replicability does not depend on the choice of a small number of features (Nerbonne, 2009). On the other hand, we explicitly investigate sociolinguistic as well as concept-related features, generally ignored in the dialectometric approach. In this study, therefore, we attempt to combine perspectives from dialect geography (dialectometry) and social dialectology (sociolinguistics). The statistical analysis we employ (i.e. generalized additive modeling) enables the incorporation of candidate explanatory variables based on both social and geographical factors, making it a good technique to facilitate the intellectual merger of dialectology and sociolinguistics (Chambers and Trudgill, 1998).

Using a generalized additive model in combination with a mixed-effects regression approach to combine the dialectometric and dialectological approach is a very recent development, and has already proven to be successful. Wieling, Nerbonne and Baayen (2011) showed in a study on Dutch dialects that the distance from standard Dutch could be predicted by both the geographical location of the communities, as well as several location- and word-related factors. They identified, among others, community size, average community age, and word frequency as significant factors in explaining the pronunciation distance of individual words in different dialects from standard Dutch. Wieling et al. (2011) used a basic generalized additive model to represent the global geographical pattern, but they took a two-step approach by adding the global geographical pattern as a predictor in their linear mixed-effects regression model. Since the software available for generalized additive mixed modeling has improved significantly since the study of Wieling et al. (2011), our study is able to advance on their approach by constructing a single generalized additive mixed-effects regression model. In this model, we allow the effect of geography to vary with concept frequency and speaker age and also take additional socio-linguistic factors into account. Furthermore, our study differs from Wieling et al.'s (2011) approach in two ways. First, here we focus on lexical variation rather than variation in pronunciation.

Second, we do not try to predict dialect distances, but a binary value indicating if the lexical form of a concept with respect to standard Italian is different (1) or equal (0).

Using a mixed-effects regression approach has clear advantages over conventional regression analyses. First, it has a lower chance of incorrectly judging a predictor as significant (Baayen, 2008: Ch. 7). Second, it allows us to make specific predictions for individual concepts and locations. For example, while most concepts will be more likely to have a lexical variant equal to standard Italian in large communities as opposed to small communities, some concepts might show an opposite pattern (as we will observe later, this is indeed the case). These advantages of mixed-effects regression have already resulted in clear recommendations for researchers in sociolinguistics to embrace mixed-effects regression (Johnson, 2009).

In the next sections, we will discuss the Tuscan dialect dataset, followed by a more in-depth explanation of the generalized additive modeling procedure, the results and the discussion.

2. Material

2.1. *Lexical data*

The lexical data used in this study were taken from the *Atlante Lessicale Toscano* (ALT). ALT interviews were carried out between 1974 and 1986 in 224 localities of Tuscany, with 2193 informants selected with respect to a number of parameters ranging from age and socio-economic status to education and culture. It is interesting to note that only the younger informants of ALT were born in the period when standard Italian was used a spoken language. 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.

In this study, we focused 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). 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). Out of 460 onomasiological questions, we selected

only those which prompted 50 or fewer normalized lexical answers (the maximum in all onomasiological questions was 421 unique lexical answers). We used this threshold to exclude questions having many hapaxes which did not appear to be lexical (a similar approach was taken in Montemagni, 2007); for instance, the questionnaire item looking for denominations of ‘stupid’ included 372 different normalized answers, 122 of which are hapaxes representing productive figurative usages (e.g., metaphors such as *cetriolo* ‘cucumber’ and *carciofo* ‘artichoke’), or originating from productive derivational processes (e.g., *scemaccio* and *scemalone* from the lexical root *scemo* ‘stupid’), or multi-word expressions (e.g., *mezzo scemo* ‘half stupid’, *puro locco* ‘pure stupid’ and the like). From the resulting 195-item subset, we excluded a single adjective and twelve verbs (as the remaining concepts were nouns) and all twelve multi-word concepts. Our final subset, therefore, consisted of 170 concepts and is listed in Table 1.

The normalized lexical forms in the ALT data source still contained some morphological variation. In order to assess the pure lexical variation we abstracted away from variation originating in, e.g., assimilation, dissimilation, or other phonological differences (e.g., the dialectal variants *camomilla* and *capomilla*, meaning ‘chamomile’, have been treated as instantiations of the same normalized form) as well as 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.¹

The list of standard Italian forms for the 170 concepts was extracted from the online ALT corpus (ALT-Web; available at <http://serverdbt.ilc.cnr.it/altweb>) within which it had been created for query purposes. This list, originally compiled on the basis of lexicographic evidence, was carefully reviewed by members of the *Accademia della Crusca*, 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 real Italian and not old-fashioned or literary words originating in Tuscan dialects.

¹ The effect of the morphological variation was relatively limited, as the results using the unaltered ALT normalized lexical forms were highly similar to the results based on the lexical forms where morphological variation was filtered out.

In every location multiple speakers were interviewed (between 4 and 29) and therefore each normalized answer is anchored to a given location, but also to a specific speaker. While we could have included all speakers separately (a total of 2081), we decided against this, as this would be computationally infeasible (logistic regression is computationally quite slow). Consequently, we grouped the speakers in an older age group (born in 1930 or earlier – 1930 was the median year of birth) and a younger age group (born after 1930). For both age groups, we used the lexical form pronounced by the majority of the speakers in the respective group. As not all concepts were attested in every location, the total number of cases (i.e. concept-speaker group combinations) was 69,259.

As 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 against standardization), we obtained the concept frequencies (of the Italian lexical form) by extracting the corresponding frequencies from a large corpus of 8.4 million Italian unigrams (Brants and Franz, 2009). While the frequencies of other lexical forms are likely to be different, these frequencies should give a good idea about the relative frequencies of different concepts.

Since concrete concepts may be easier to remember (Wattenmaker and Shoben, 1987) and stored differently in the brain than abstract concepts (Crutch and Warrington, 2005), we also investigated if the concreteness of a concept played a role in lexical differences with respect to standard Italian. We obtained concreteness scores (for the English translations of the Italian concepts) from the MRC Psycholinguistic Database (Coltheart, 1981). The concreteness scores in this database ranged from 100 (abstract) to 700 (concrete). Our most abstract concept ('cheat') had a score of 329, while our most concrete concepts (e.g., 'cucumber' and 'grasshopper') had a score of about 660.

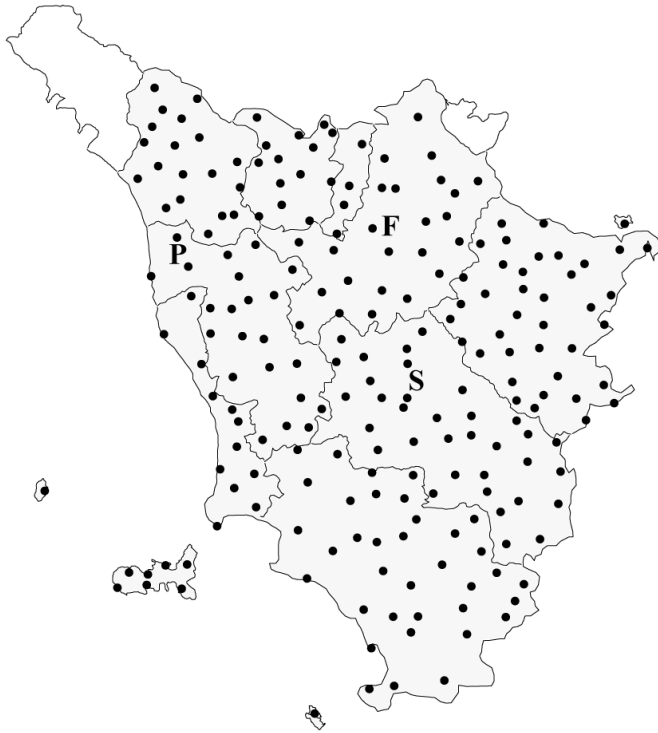


Figure 1. Geographical distribution of the 213 locations investigated in this study. The ‘F’, ‘S’ and ‘P’ mark the approximate locations of Florence, Sienna and Pisa, respectively.

2.2. *Sociolinguistic data*

Besides the age information about the speaker group (old and young) and the year of recording for every location, we extracted additional 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 (ranging from 9 to 97,114) in all 213 locations in 1971 or 1981 (whichever year was closer to the year when the interviews for that location were conducted). In addition, we extracted the average income per inhabitant (ranging from 6,461 to 17,039 euro) in every location (in 2005; which was the oldest information available) and the average age (ranging from 41 to 54) in every location (in 2007; again the oldest information available). 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 dataset was constructed (between 1974 and 1986), the global pattern will probably be relatively similar.

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

Table 1. List of all 170 lexical items included in this study including their English translation

3. Methods

3.1. *Modeling the role of geography: generalized additive modeling*

In contrast to a linear regression model in which a single predictor is simply linear in its effect on the dependent variable, in a generalized additive model (GAM) the assumption is relaxed that the functional relation between a predictor and the response variable is linear. Instead, the GAM provides the user with a flexible toolkit for smoothing nonlinear relations (i.e. a ‘wiggly curve’) in any number of dimensions. Consequently, the GAM is much more flexible than the simple linear regression model. In addition, multiple predictors may be combined in a single smooth function, yielding a wiggly (hyper)surface. For example, longitude and latitude may be combined in a single smooth (function) to model how these two variables interact. As it turns out, this is a highly suitable approach to model the influence of geography in dialectology, as geographically closer varieties tend to be linguistically more similar (e.g., see Nerbonne, 2010). Wieling et al. (2011) also used a generalized additive model to represent the global effect of geography. In this study, we will take a more sophisticated approach, allowing the effect of geography to vary for concept frequency and also to differ between young and old speakers. Furthermore, we will use a generalized additive *logistic* model, as our dependent variable is binary. 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 (Agresti, 1996). 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).

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 by Wieling et al. (2011). The (solid) contour lines represent aggregate isoglosses connecting areas which 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 -0.2 indicates that in those areas the lexical form is more likely to match the Italian standard (the probability is 0.45 that the lexical form is *different* from the Italian standard form) and the value 0.2 indicates the opposite (the probability is approximately 0.55 that the lexical form is different from the Italian standard form). Correspondingly, darker shades of gray indicate a greater likelihood of having lexical forms identical to those in standard Italian, while lighter shades of gray represent a greater likelihood of having lexical forms different from those in standard Italian. We can clearly see that locations near Florence (indicated by the white star) tend to have lexical variants more likely to be the same as the standard Italian form. This makes sense as Italian originated from the Tuscan dialect spoken in Florence. The 27.02 estimated degrees of freedom invested in this general thin plate regression spline were supported by a z -value of -24 ($p < 0.001$).

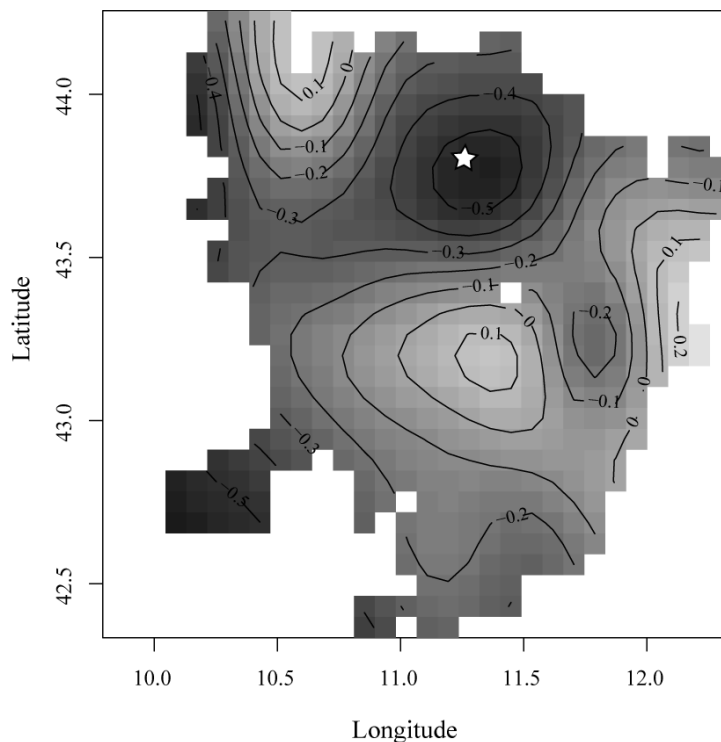


Figure 2. Contour plot for the regression surface of predicting lexical differences from standard Italian as a function of longitude and latitude obtained with a generalized additive model using a thin plate regression spline. The (black) contour lines represent aggregate isoglosses, darker shades of gray (lower values) indicate a smaller likelihood of having a lexical form different from standard Italian, while lighter shades of gray (higher values) represent locations with a greater likelihood of having a lexical form different from standard Italian. The star marks the approximate location of Florence.

As Wieling et al. (2011) found that the effect of word frequency on (Dutch) dialect distances varied per location, we created a three-dimensional smooth (longitude x latitude x concept frequency), allowing us to assess the concept frequency-specific geographical pattern of linguistic variation with respect to standard Italian. For example, it might be that the geographical pattern presented in Figure 2, may hold for concepts having an average frequency, but might be radically different for concepts with a low as opposed to a high frequency. In addition, we will investigate whether these patterns also vary for old speakers as opposed to young speakers (i.e. we create two three-dimensional smooths, one for old speakers and one for young speakers). We model these three-dimensional smooths by a tensor product which allows combinations of non-isotropic predictors (i.e. measurements of the predictors are not on the same scale: e.g., longitudinal degrees versus frequency; Wood, 2006, p. 162). In the tensor product, we model both longitude and latitude with a thin plate regression spline (as this is suitable for combining isotropic predictors and also in line with the approach of Wieling et al., 2011), while the concept frequency effect is modeled by a cubic regression spline, which is computationally more efficient than the thin plate regression spline. More information about these tensor product bases (which are implemented in the `mgcv` package for R) is provided by Wood (2006; Ch. 4).

3.2. *Mixed-effects modeling*

A generalized additive *mixed* model distinguishes between fixed and random-effect factors. Fixed-effect factors have a small number of levels exhausting all possible levels (e.g., our age group is either young or old). Random-effect factors, in contrast, have levels sampled from a much large population of possible levels. In our study, concepts and locations are random-effect factors, as we could have included many other concepts and locations. 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) will be more likely to be 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 (in general, larger communities have lexical variants closer to standard Italian), 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 be similar to standard Italian in small 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 models is given by Baayen (2008, Ch. 7) and Baayen et al. (2008).

In our analyses, we considered the two aforementioned random-effect factors (i.e. location and concept) as well as several other predictors besides the (concept frequency and speaker age group-specific) geographical variation. The only lexical variables we included were concept frequency (based on the frequency of the standard Italian lexical form) and the concreteness rating of each concept. The location-related variables we investigated were community size, average community age, average community income and the year of recording. The only speaker-related variable we took into account was age group (old: born in 1930 or earlier; young: born after 1930).

A common problem in large-scale regression studies is predictor collinearity. In our dataset, communities with a higher average age tend to have a lower average income. To be able to assess the pure effect of each predictor, we decorrelated average age from average income by using as predictor the residuals of a linear model regressing average age on average income (instead of the original average age values). Since the new predictor correlated highly ($r = 0.9$) with the original predictor, we can still interpret the new predictor as representative of average age (but now excluding the effect of average income).

To reduce the potentially harmful effect of outliers, several numerical predictors were log-transformed (i.e. community size, average age, average income 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. The significance of fixed-effect factors was evaluated by means of the Wald test (reporting a z -value) for the coefficients in a logistic regression model.

4. 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 will discuss the specification of the model including all significant predictors and verified random-effect factors.

Our dependent value 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. Table 3 presents the significance of the three-dimensional smooth terms (modeling the concept-frequency related geographical pattern for both the old and young speaker group)² and Table 4 lists the significant random-effect structure of our model.

To evaluate goodness of fit of the final model (see Tables 2 to 4), we used the index of concordance C . The index of concordance C 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 $C = 0.85$.

	Estimate	Std. error	z -value	p -value
Intercept	-0.4129	0.13950	-2.960	0.003
Old instead of young speakers	0.4407	0.01925	22.896	< 0.001
Community size (log)	-0.1154	0.02658	-4.343	< 0.001

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 non-standard Italian lexical form. A negative estimate reduces the likelihood of having a different lexical form than the standard Italian form.

² We verified the necessity of concept frequency and the contrast between old and young speakers in the geographical smooth (all p 's < 0.001).

	Est. d.o.f.	Chi. sq.	<i>p</i> -value
Geography x concept frequency (old)	53.88	221.5	< 0.001
Geography x concept frequency (young)	59.48	326.6	< 0.001

Table 3. Significant smooth terms of the final model. For every smooth the estimated degrees of freedom is indicated as well as its significance in the model. See Figure 3 for the geographical patterns.

Factors	Random effects	Std. dev.	<i>p</i> -value
Location	Intercept	0.2410	< 0.001
Concept	Intercept	1.7748	< 0.001
	Average community income (log)	0.3127	< 0.001
	Average community age (log)	0.2482	< 0.001
	Community size (log)	0.1166	0.006

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

4.1. Geographical variation and lexical predictors

Inspecting Table 3, it is clear that the geographical pattern is a very strong predictor, and it varied significantly with concept frequency (which was not significant by itself in our general model) and speaker age group. Figure 3 visualizes the geographical variation related to concept frequency and speaker age. Lighter shades of gray indicate a greater likelihood of having a lexical form different from standard Italian.

The three graphs to the left present the geographical patterns for old speakers, while those to the right present the geographical patterns for young speakers. In general, the graphs for younger speakers are somewhat darker than those for the older speakers, supporting the finding (discussed in Section 4.2 below) that older speakers have a greater likelihood of using a lexical form different from standard Italian than the younger speakers.

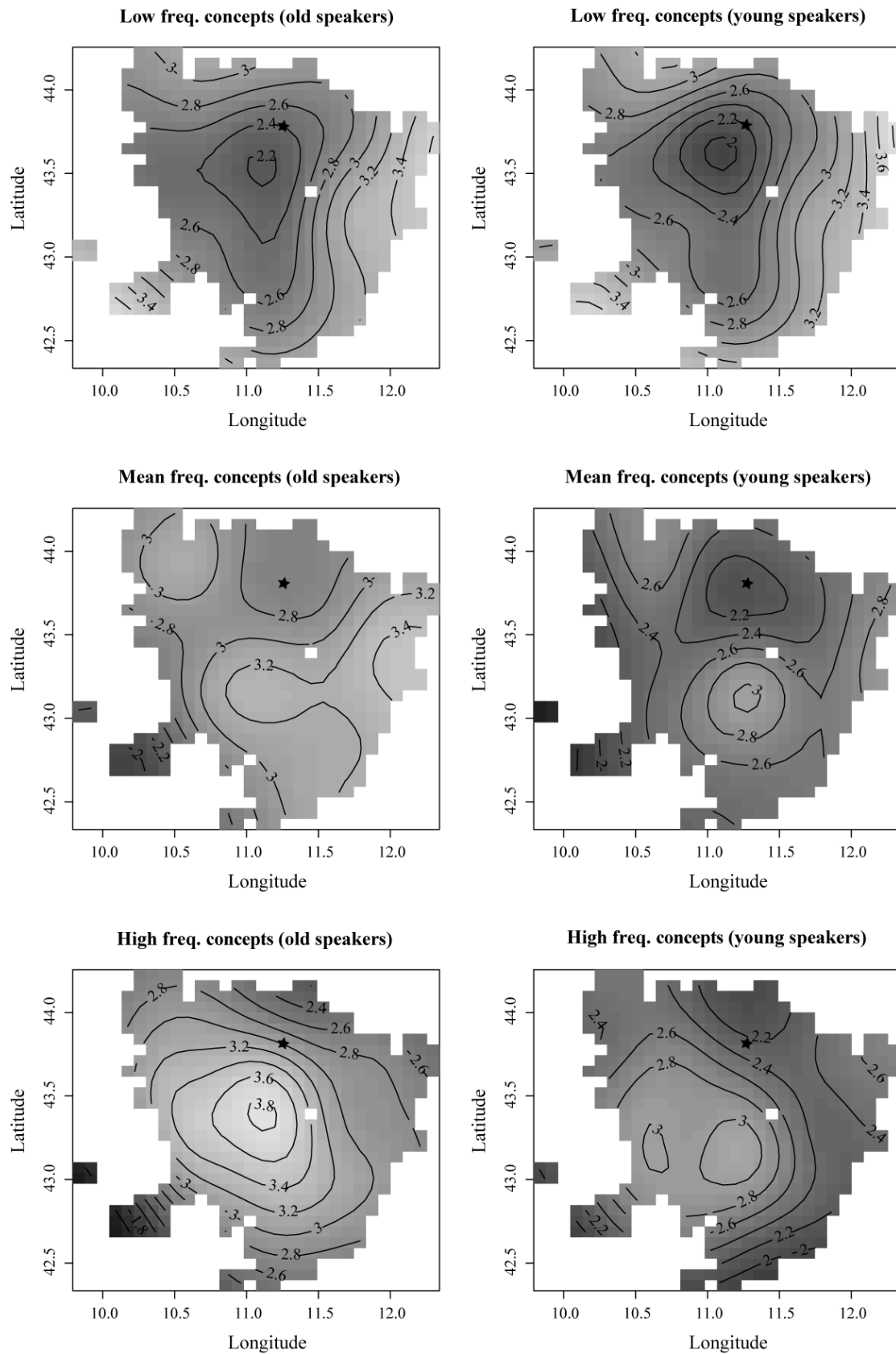


Figure 3. Contour plots for the regression surface of predicting lexical differences from standard Italian as a function of longitude, latitude, concept frequency and speaker age group obtained with a generalized additive model. The (black) contour lines represent aggregate isoglosses, darker shades of

gray (lower values) indicate a smaller lexical ‘distance’ from standard Italian (i.e. a smaller likelihood of having a lexical form different from standard Italian), while lighter shades of gray (higher values) represent locations with a larger lexical ‘distance’ from standard Italian. The star marks the approximate location of Florence. The left plots visualize the results for old speakers, while the right plots those for young speakers. 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). Note that the contour lines drop towards a pit (with a smaller lexical ‘distance’) in the center of the low frequency plots, and rise towards a peak (with a higher lexical ‘distance’) in the center of the high frequency plots (which also resemble the mean-frequency plots).

The first thing to note is that in the top graphs Florence (indicated by the star) is located in (approximately) the area with the smallest likelihood of having a non-standard Italian lexical form. This clearly makes sense as standard Italian originates from the Florentine variety.

The second observation is that, going from the top to the bottom graphs, we see a strong effect of concept frequency, both for older speakers and (to a slightly reduced extent) for younger speakers.³ More frequent concepts in the central Tuscan area, including Florence, are more likely to differ from standard Italian than lower frequent concepts (i.e. the values in the bottom maps in the central area are higher than the values in the central and top maps).

Third, we observe a reverse pattern in the more peripheral areas (in the Tuscan archipelago in the west, but also in the north and east), with a greater likelihood of having a non-standard Italian lexical form for low frequency concepts than for high frequency concepts.

When looking in more detail at the data, high frequency concepts typically include cases for which standard Italian and Tuscan dialects diverge (e.g., standard Italian *angolo* ‘angle’, in Tuscany *canto*, *cantonata* or *cantone*; or standard Italian *pomeriggio* ‘afternoon’, in Tuscany *sera* ‘evening’ or multi-word expressions such as *dopo mangiato/pranzo/desinare* meaning ‘after lunch’, but also *dopo mezzogiorno*

³ While the general effect of concept frequency appeared to be stronger in old speakers as opposed to young speakers, this interaction was not significant.

‘after noon’). For mean frequency concepts, the standard Italian and dialectal words share the same etymology, with the latter frequently (but not always) representing analogical variants of the former (e.g., for ‘ivy’, the standard Italian form is *edera*, whereas the set of dialectal forms includes *ellera*, *ellora*, *lellera*, *lallera*, etc.).⁴ Finally, the low frequency concepts belong to an obsolete, progressively disappearing rural world and include concepts such as *bigoncia* ‘vat’, *seccatoio* ‘squeegee’, and *stollo* ‘haystack pole’.

To understand the pattern of results, we need to distinguish between three dimensions of change. First, as one moves out from the heartland of Tuscany, it is more likely that different words are used for a certain concept. This is the well-known effect of (increasing) geographical distance (e.g., see Nerbonne and Kleiweg, 2007). We see this effect most clearly for the low frequency concepts, and reversed, for the high frequency concepts.

Second, the standard literary Italian language was ill-equipped for use in everyday discourse (very likely involving high frequency concrete concepts), and consequently the lexical gaps of the standard Italian (literary) language were filled with dialectal forms whose origin was not necessarily from Tuscany. Furthermore, during the evolution of the standard Italian language in the past centuries, alternative cognitively well-entrenched forms of high frequency concepts might have been preferred over the Florentine forms, and thus became part of the standard language instead.

Third, with the relatively recent emergence of the standard spoken language, a new wave of change is affecting Tuscany, causing Tuscan speakers to adopt the new standard Italian norm. This process is clearly documented in Figure 3, which shows that younger speakers (right panels) have moved closer to standard Italian than the older speakers (left panels).

Considering these three dimensions, the high frequency concepts in central Tuscany are more different (than lower frequency concepts) from standard Italian for two reasons. First, the high frequency Florentine forms likely did not contribute as

⁴ Note that the normalization process we have employed in this study still distinguishes these forms which represent lexical variants in their own right, in spite of their origin from the same etymon.

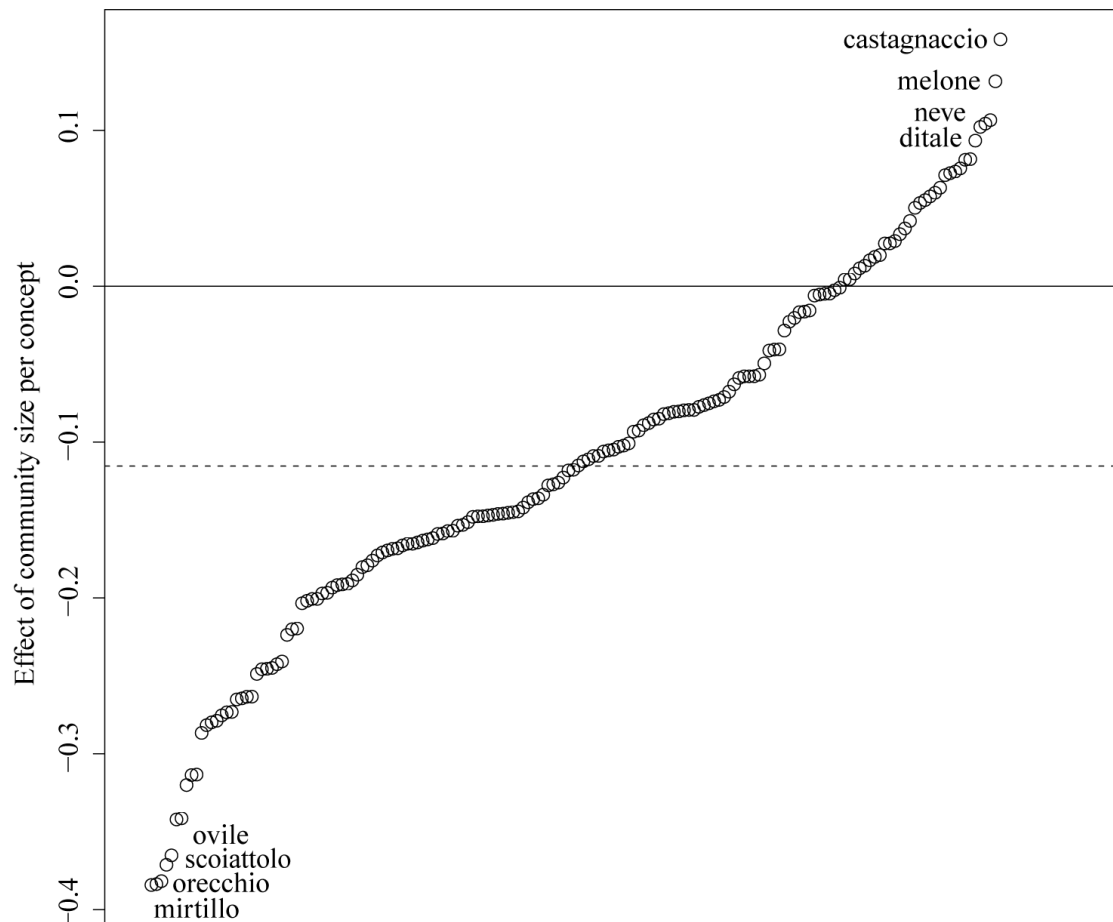
prominently to the standard Italian language because of the competition of alternative non-Florentine (well-entrenched) high frequency forms. Second, high frequency lexical forms are most resistant to replacement by the current equivalents in standard Italian because they are cognitively well entrenched in the central Tuscan (high prestige) speakers' mental lexicons. This explanation is in line with the finding by Wieling et al. (2011) where they reported that high frequency words were more resistant to standardization than low frequency words (see also Pagel et al., 2007).

With respect to the peripheral areas, the low frequency concepts (mainly belonging to an obsolete, progressively disappearing rural world) differ most from standard Italian as these are either represented by original Tuscan forms that were different from the Florentine norm due to geographical distance (or separation from the mainland), or represented by non-Tuscan dialectal forms (especially in the north and east, which border to other dialect areas). For medium and high frequency concepts in these peripheral areas, the pattern reverses, and the lexical forms are more likely to match the standard Italian form. With no close cultural ties to central Tuscany, and no prestige of its own, these dialects have been more open to accepting the standard Italian forms (spread via education and mass media).

We also investigated the effect of concept concreteness, but we did not find support for the significance of this predictor. As our most abstract concepts were only mildly abstract (according to the categorization of Crutch and Warrington, 2005), this might have limited our ability to investigate the effect of abstract versus concrete concepts on lexical differences with respect to standard Italian.

4.2. Demographic predictors

When inspecting Table 2, it is clear that the contrast between the age groups was highly important, judging by its high z -value. Old speakers were much more likely to have a lexical form different from standard Italian. This result is not surprising as younger speakers tend to converge to standard Italian.



Concepts sorted by the effect of community size

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). The strongly negative coefficients (bottom left) are associated with concepts that are more likely to be identical to standard Italian in larger communities, while the positive coefficients (top right) are associated with concepts that are more likely to be different from standard Italian in larger communities. The model estimate (see Table 2) is indicated by the dashed line.

Of all location-based predictors (i.e. the community size, the average community income and the average community age) only the first was a significant predictor in the general model. Larger communities were more likely to have a lexical variant close to standard Italian (i.e. the estimate in Table 2 is negative). A possible explanation for these findings is that people tend to have weaker social ties in urban communities, which causes dialect leveling (Milroy, 2002). As the standard Italian language is more prestigious than dialectal forms (Danesi, 1974), conversations will normally be held in standard Italian and leveling will proceed in the direction of standard Italian.

The other location-based predictors, average age and average income, were not significant in the general model. In the study of Wieling et al. (2011) on Dutch dialects, average age was identified as a significant predictor of pronunciation distance from standard Dutch, while average income was not. The effect of average community age may be less powerful in our study, because we have two age groups per location (which are much more suitable to detect age differences). In line with Wieling et al. (2011), the effect of average income pointed to a negative influence (with richer communities having lexical variants closer to the standard), but not significantly so ($p = 0.3$). Also note that year of recording was not a significant predictor, which is likely due to the relatively short time span (with respect to lexical change) in which the data was gathered.

All location-related variables (i.e. community size, average income and average age) showed significant by-concept variation. Figure 4, illustrating the effect of community size, shows some concepts (i.e. *ovile* ‘sheepfold’, *scoiattolo* ‘squirrel’, *orecchio* ‘ear’, and *mirtillo* ‘blueberry’) which are more likely to be identical to standard Italian in larger communities (i.e. consistent with the general pattern; the model estimate is indicated by the dashed line), while others behave in completely opposite fashion (i.e. *castagnaccio* ‘chestnut cake’, *melone* ‘melon’, *neve* ‘snow’, and *ditale* ‘thimble’) and are more likely to be different from standard Italian in larger communities. Many of these latter concepts (e.g., *castagnaccio*, but also *verro* ‘boar, male swine’, and *stollo* ‘haystack pole’, which are not marked in the graph) involve very old-fashioned rural concepts which may have fallen into disuse in larger cities, but not in smaller, more traditional, villages. As a consequence, people in larger cities may have forgotten the (old-fashioned) standard Italian lexical form, and use multi-word phrases or more general terms instead (e.g., ‘pig’ instead of ‘boar’). It is interesting to note that the set of latter concepts also includes *melone* ‘melon’ and *ditale* ‘thimble’, which represent two of the few well-known cases in which all Tuscan dialects diverge from standard Italian.

Figure 5 illustrates the by-concept random slopes for average age and average income (both were not significant as a fixed-effect factor). In the left part of the graph we see concepts which are more likely to have a standard Italian lexical form in richer communities (with concepts *caprone* ‘goat’, *cocca* ‘corner’, e.g., of a handkerchief,

and *grattugia* ‘grater’ being close to the extreme), while the concepts in the bottom-right quadrant (e.g., *pimpinella* ‘pimpernel’, *stollo* ‘haystack pole’, and *ditale* ‘thimble’) demonstrate the opposite pattern, being more likely to have a lexical form different from standard Italian in richer communities.

The by-concept random slopes of average age and average income are closely linked (their correlation is $r = -0.623$, $p < 0.001$), which is reflected by the general negative trend in the scatter plot (see Figure 5). Concepts that are more likely to differ from the Italian form in poorer locations are also more likely to differ from the Italian form in locations with a higher average age (e.g., *cocca* and *grattugia* in the top-left). Similarly, concepts which follow the opposite pattern and are more likely to differ from standard Italian for richer communities, are also more likely to differ from the Italian lexical form in younger communities (e.g., *pimpinella* and *stollo* in the bottom-right). A similar finding was also reported by Wieling et al. (2011) where they found that by-word random slopes of average age, average income, as well as community size were closely linked. In our case, however, there was no support for a link between by-concept random slopes for community size and the other by-concept random slopes.

Somewhat comparable to the by-concept random slopes for community size, the concepts in the bottom-right quadrant (i.e. showing a higher likelihood of having a non-standard Italian lexical form for richer as well as younger communities) include both old-fashioned rural concepts (e.g., *castagnaccio*, and *stollo*) and concepts for which Tuscan dialects diverge from standard Italian (e.g., *ditale*, and *pigna* ‘cone’). The reasons for this pattern are different, however. On the one hand, it might be that in richer, younger towns, people are less likely to remember these old-fashioned forms, but on the other hand, for those few cases in which standard Italian and Tuscan dialects diverge they prefer the local form (which is, for those cases, widely used throughout the whole of the Tuscan area) in favor of the standard Italian form.

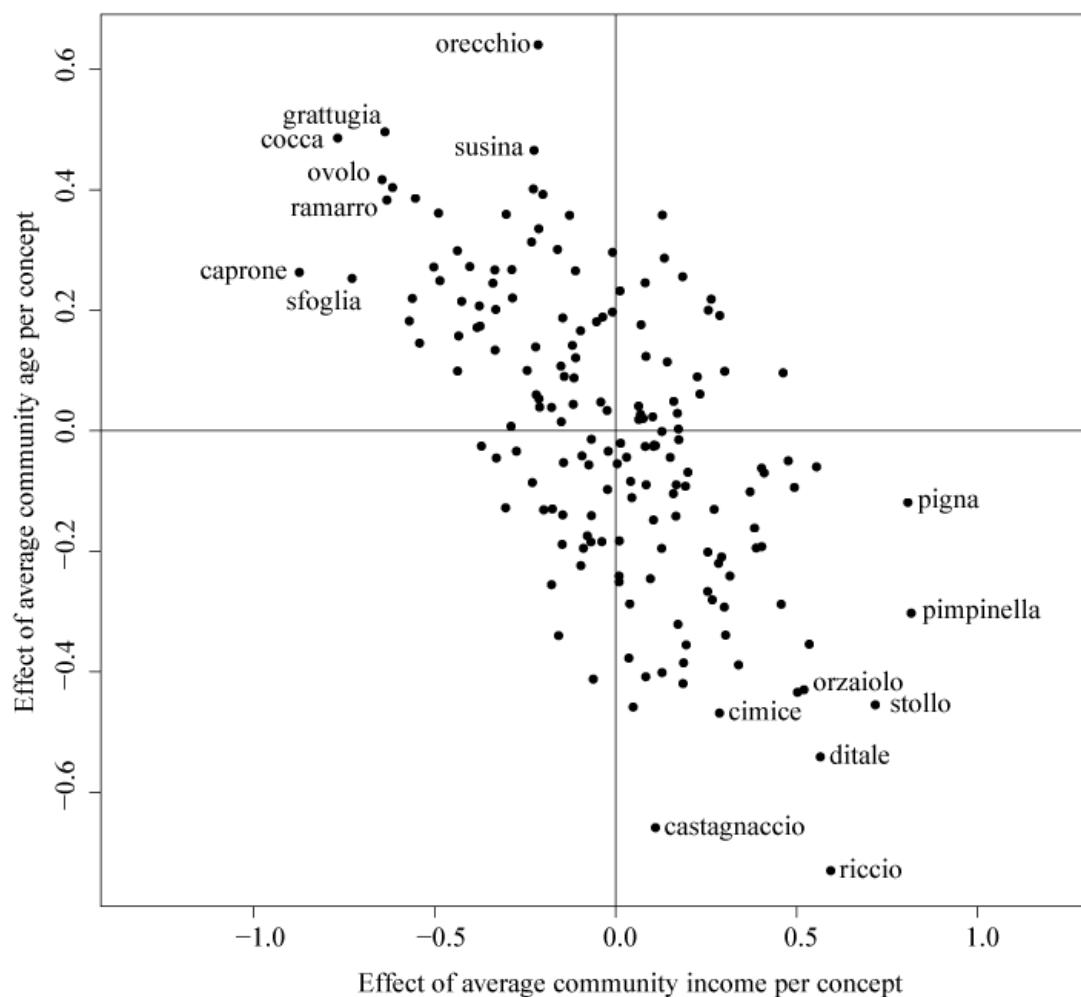


Figure 5. By-concept random slopes of average community age and income.

5. Discussion

In this study we have shown that the lexical variation in Tuscan dialects with respect to the standard Italian lexical form can be adequately modeled by a generalized additive mixed model. We found clear support for the importance of speaker age, community size, as well as geography, which varied significantly depending on concept frequency and speaker age. In addition, we illustrated that the mixed-effects regression approach also enabled a detailed investigation of the precise effect of different location-related variables for individual concepts.

The results which have emerged from our analysis of the ALT corpus also shed new light on the widely debated *questione della lingua* from the point of view of Tuscan

dialects. Previous studies, based both on individual words (Giacomelli and Poggi Salani, 1984) and on aggregated data (Montemagni, 2008), provided a flat view according to which Tuscan dialects overlap most closely with standard Italian in the area around Florence, with expansions in different directions and in particular towards the southwest. The aggregated analysis of the ALT lexical data (Montemagni, 2008) also illustrated that a higher likelihood of using standard Italian is connected with speaker age and geographical coverage of words. The results of the analysis introduced in this paper, however, provide a much more finely articulated picture in which new factors are shown to play a significant role and also allowed us to speculate about the spread of standard Italian with a particular emphasis on its relationship to the Florentine variety from which it originated.

For example, we observed that in the central Tuscan area, including Florence, more frequent concepts are more likely to differ from standard Italian than lower frequency concepts, whereas in the marginal areas in the north, east and west a reverse pattern was observed. There, infrequent concepts are highly different from standard Italian and frequent concepts are more similar to the standard. Frequency of concepts thus shows markedly different effects based on the history of the Italian language (originating from Florence) and the status attributed to the dialect in the specific area: the standard Italian language diverged more from the Florentine variety it originated from for high frequency concepts than for low frequency concepts, and also dialects from the central Tuscan area, including Florence, are accorded higher prestige than the dialects spoken in marginal areas of Tuscany, and are therefore able to counterbalance the rising of standard Italian.

On the demographic side, besides finding a significant effect of speaker age (with younger speakers using lexical forms more likely to be equal to standard Italian), we observed that larger communities are more likely to use standard Italian vocabulary than smaller communities. In addition, the effect of community size, but also average community age and income (even though these two were not significant as main effects), shows significant by-concept variation: concepts belonging to an obsolete disappearing rural world, such as ‘haystack pole’ and ‘boar’, are more likely to differ from the standard in bigger, richer and younger communities, due to the fact that they are no longer part of everyday life.

Given the general features of the ALT dataset, it would also be feasible to investigate other speaker-related characteristics by creating a different grouping (than the present age-based split). For example, it would be interesting to investigate the importance of speaker education or profession in this way.

In all experiments reported so far we used a binary lexical difference measure with respect to standard Italian. It would also be possible to use a more sensitive distance measure such as the Levenshtein (or edit) distance. In that case, lexical differences which are closely related (i.e. in the case of lexicalized analogical formations) can be distinguished from more rigorous lexical differences. As this would not require the time-consuming logistic regression analysis, it would be possible to analyze all individual speakers (and incorporating their speaker-specific characteristics in the model specification) instead of simply grouping them.

Acknowledgements

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

References

- Agresti, A. (2007). *An introduction to categorical data analysis*. John Wiley & Sons, Hoboken, NJ, 2nd edition.
- Baayen, R.H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.
- Baayen, R.H., D.J. Davidson and D.M. Bates (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4): 390-412.
- Brants, T. and A. Franz (2009). Web 1T 5-gram, 10 European Languages Version 1. Linguistic Data Consortium, Philadelphia.
- Castellani, A. (1982). Quanti erano gli italofoeni nel 1861? *Studi Linguistici Italiani*, 8/1, Roma, Salerno Editrice: 3-26.
- Comuni Italiani (2011). Informazioni e dati statistici sui comuni in Italia, le province e le regioni italiane. Sito ufficiale, CAP, numero abitanti, utili link. <http://www.comuni-italiano.it>. Last accessed: 2011-05-23.

- Chambers, J.K. and P. Trudgill (1998). *Dialectology*. Second edition. Cambridge University Press, Cambridge.
- Coltheart, M. (1981). The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4): 497-505.
- Crutch, S.J. and E.K. Warrington (2005). Abstract and concrete concepts have structurally different representational frameworks. *Brain*, 128(3): 615-627.
- Danesi, M. (1974). Teaching Standard Italian to dialect speakers: a pedagogical perspective of linguistic systems in contact. *Italica*, 51(3): 295-304.
- De Mauro, T. (1963). *Storia linguistica dell'Italia unita*. Bari-Roma, Laterza.
- Giacomelli, G. (1975). Dialettologia toscana. *Archivio glottologico italiano*, 60: 179-191.
- Giacomelli, G. (1978). Come e perchè il questionario. In G. Giacomelli et al. (eds.), *Atlante lessicale toscano - Note al questionario*, Firenze, Facoltà di Lettere e Filosofia, 19-26.
- Giacomelli, G., L. Agostiniani, P. Bellucci, L. Giannelli, S. Montemagni, A. Nesi, M. Paoli, E. Picchi and T. Poggi Salani (2000). *Atlante Lessicale Toscano*. Lexis Progetti Editoriali, Roma.
- Giacomelli, G. and T. Poggi Salani (1984). Parole toscane. *Quaderni dell'Atlante Lessicale Toscano*, 2(3): 123-229.
- Harrell, F (2001). *Regression modeling strategies*. Springer, Berlin.
- Johnson, D.E. (2009). Getting off the GoldVarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Language and Linguistic Compass*, 3(1): 359–383.
- Lepschy, G. (2002). *Mother Tongues & other Reflections on the Italian Language*. University of Toronto Press, Toronto.
- Maiden, M. (1995). *A Linguistic History of Italian*. Longman, London.
- Maiden, M. and M. Parry (1997). *The Dialects of Italy*. Routledge, London.
- Migliorini, B. and T.G. Griffith (1984). *The Italian language*. Faber and Faber, London.
- Milroy, L. (2002). Social Networks. In J. Chambers, P. Trudgill and N. Schilling-Estes (eds.), *The Handbook of Language Variation and Change*. Blackwell Publishing Ltd., 549-572.
- Montemagni, S. (2007). Patterns of phonetic variation in Tuscany: using dialectometric techniques on multi-level representations of dialectal data. In P.

- Osenova et al. (eds.), *Proceedings of the Workshop on Computational Phonology at RANLP-2007*, 49-60.
- Montemagni, S. (2008) Analisi linguistico-computazionali del corpus dialettale dell'Atlante Lessicale Toscano. Primi risultati sul rapporto toscano-italiano. In A. Nesi and N. Maraschio (eds.), *Discorsi di lingua e letteratura italiana per Teresa Poggi Salani* (Strumenti di filologia e critica, vol. 3), Pisa, Pacini, 247-260.
- Nerbonne, J. (2009). Data-driven dialectology. *Language and Linguistics Compass*, 3(1): 175-198.
- Nerbonne, J. (2010). Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365: 3821-3828.
- Pagel, M., Q. Atkinson and A. Meade (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449: 717-720.
- Poggi Salani, T. (1978). Dialetto e lingua a confronto. In G. Giacomelli et al., *Atlante lessicale toscano - Note al questionario*, Firenze, Facoltà di Lettere e Filosofia, 51-65.
- Wattenmaker, W.D. and E.J. Shoben (1987). Context and the recallability of concrete and abstract sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(1): 140-150.
- Wieling, M., J. Nerbonne and R.H. Baayen (2011). Quantitative Social Dialectology: Explaining Linguistic Variation Socially and Geographically. *PLoS ONE*, 6(9): e23613.
- Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B* (Statistical Methodology), 65(1): 95-114.
- Wood, S. (2006). *Generalized additive models: an introduction with R*. Chapman & Hall/CRC.