CNR Short Term Mobility Program – 2011

Report of the research activity carried out
from May 16th to May 28th 2011
at ILC-CNR (Istituto di Linguistica Computazionale "Antonio Zampolli")
by Martijn Wieling (Università di Groningen, The Netherlands)

Title of Research Program
Definizione di un modello computazionale della variazione dialettale basato sull'integrazione di fattori socio-demografici e geografici
Computational models of dialectal variation based on the combination of geographic and socio-demographic factors

Sintesi delle attività svolte

Lo studio della variazione linguistica può essere affrontato da prospettive diverse: può riguardare la variazione nello spazio (variazione diatopica), oppure rispetto al profilo socio-demografico del parlante (variazione diastratica). Generalmente questo tipo di studi sono trattati nell'ambito di discipline diverse: se da un lato la dialettologia tradizionale si è primariamente concentrata sul rapporto tra lingua e geografia e sulla differenziazione spaziale della lingua, la sociolinguistica o dialettologia urbana ha riguardato piuttosto le relazioni tra la lingua e le caratteristiche socio-demografiche dei parlanti. Anche sul versante dei modelli computazionali della variazione linguistica si è osservata una suddivisione analoga: fino dalle origini, la tradizione di studi dialettometrici si è focalizzata sul ruolo rivestito dalla distanza geografica nel definire il quadro complessivo della variazione dialettale.

L'obiettivo di ricerca che ci siamo proposti nell'ambito del Programma per la Mobilità di breve durata (STM-2011) di cui Martijn Wieling dell'Università di Groningen (The Netherlands) è stato il fruitore si colloca all'interno della tradizione di studi dialettometrici. In particolare, il fine ultimo della ricerca proposta era quello di applicare una tecnica innovativa appena messa a punto da Martijn Wieling in collaborazione con la University of Alberta (Edmonton, Canada) in grado di rispondere a un interrogativo aperto degli studi dialettometrici riguardante l'interazione di fattori geografici e socio-demografici nella definizione del quadro della variazione dialettale. Tale tecnica è stata testata con risultati promettenti su dati riguardanti la variazione fonetica per l'olandese e il catalano.

Uno studio di questo tipo richiedeva un corpus di materiali dialettali che permettesse l'osservazione congiunta di variabili geografiche e socio-demografiche, ovvero in cui ciascuna attestazione dialettale fosse geo-referenziata ma anche socio-demograficamente referenziata. Il corpus dialettale dell'*Atlante Lessicale Toscano* (ALT, accessibile dal sito di ALT-Web http://serverdbt.ilc.cnr.it/altweb) rappresentava una risorsa ideale a tal fine in quanto specificamente concepita per lo studio di dinamiche linguistiche sia a livello areale sia a livello socio-culturale. Lo studio si è focalizzato sul livello di variazione lessicale e ha riguardato la combinazione di medodi e tecniche di analisi quantitativa sviluppati in ambito dialettometrico e socio-linguistico.

I risultati raggiunti dell'ambito delle attività di ricerca svolte sono particolarmente promettenti sia sul versante metodologico (è la prima volta che questo tipo di tecnica viene applicato allo studio della variazione lessicale) sia sul versante della dialettologia toscana (è la prima volta che il corpus dei materiali dell'ALT viene ispezionato nel suo complesso da questa duplice prospettiva), e sono stati illustrati in un articolo, allegato nella sua versione pre-finale, che intendiamo sottoporre per la pubblicazione alla maggiore rivista internazionale di studi dedicata al tema della variazione linguistica, ovvero *Language Variation and Change* della Cambridge University Press.

LEXICAL DIFFERENCES BETWEEN TUSCAN DIALECTS AND STANDARD ITALIAN: A SOCIOLINGUISTIC ANALYSIS USING MIXED-EFFECTS REGRESSION MODELLING

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

[a]Department of Humanities Computing, University of Groningen, [b]Istituto di Linguistica Computationale "Antonio Zampolli", CNR, Italy, [c]Department of Linguistics, University of Alberta

m.b.wieling@rug.nl, simonetta.montemagni@ilc.cnr.it, baayen@ualberta.edu, j.nerbonne@rug.nl

Abstract

In this study, we used a mixed-effects logistic regression model in combination with generalized additive logistic modeling to predict lexical differences in Tuscan dialects with respect to standard Italian. We used lexical information for 170 concepts in 213 locations in Tuscany. Although geographical position is an important predictor with locations distant from Florence having lexical forms more likely to differ from standard Italian, several other factors emerged as significant. The model predicts that lexical variants used by older speakers and in smaller as well as poorer communities are more likely to differ from standard Italian. The impact of the demographic variables, however, varied from concept to concept. For a majority of concepts, smaller and poorer communities have lexical forms different from standard Italian. For a smaller minority of concepts, however, larger and richer communities have lexical forms different from standard Italian. Similarly, the effect of speaker age and the average community age also varied per concept. While not significant as a fixed effect, the concept frequency showed significant geographical variation. These results clearly identify important factors involved in dialect variation at the lexical level. In addition, this study illustrates the usefulness of mixed-effects regression techniques together with generalized additive modeling for analyzing lexical dialect data.

Key words

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

1. Introduction

In this study we investigate a Tuscan lexical dialect dataset using advanced regression techniques in order to identify sociolinguistic and word-related factors which play an important role in predicting lexical differences with respect to standard Italian. We use 170 concepts for which we have lexical information for 213 Tuscan dialects (for both young and old speakers) and standard Italian.

Tuscany is a region with a special status in the complex puzzle of Italian dialects. According to the main scholars of Tuscan dialectology (Giacomelli, 1975; Giannelli, 2000), Tuscan dialects are neither northern nor southern dialects; this follows from their status as the source of Italian as well as representing a compromise between northern and central-southern dialects.

Tuscany is a region with a special status in the complex puzzle of Italian dialects. According to the main scholars of Tuscan dialectology (Giacomelli, 1975; Giannelli, 2000), Tuscan dialects are neither northern nor southern dialects, but rather represent a compromise between them. Standard Italian is based on Tuscan, and in particular on the Florentine variety, which achieved national and international prestige from the fourteenth century onwards as a literary language and later (after the Italian Unification, and mainly in the twentieth century) as a spoken language. However, standard Italian has never been identical to genuine Tuscan; the standard language is perhaps best described as an "abstraction" increasingly used for general communication purposes.

The relationship between standard Italian and the dialects spoken in Italy has been widely debated since the origin of the Italian language (i.e. *the questione della lingua*, or language question). Considering this situation, a study investigating the relationship between standard Italian and Tuscan dialects from which it originated on the basis of the data collected through fieldwork for a regional linguistic atlas, the *Atlante Lessicale Toscano* (ALT; Giacomelli et al., 2000), can help shed new light on the widely debated Italian *questione della lingua*. 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, words that were identical to Italian (almost) everywhere in Tuscany were programmatically excluded (Giacomelli, 1978; Poggi Salani, 1978).

Previous studies have already explored this dataset with a specific view to investigating the relationship between Tuscan and Italian. 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 in the area around Florence, expanding in different directions and in particular towards southwest. Montemagni (2008) also showed 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.

Our study is methodologically ambitious. On the one hand, we take a dialectometric perspective by using a large set of dialect data, 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 replicable study (Nerbonne, 2009). On the other hand, we explicitly investigate sociolinguistic as well as word-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).

Using a mixed-effects regression approach to combine the dialectometric and dialectological approach is a recent development, but has already proven to be successful. Wieling, Nerbonne and Baayen (submitted) 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, population size, average population age, and word frequency as significant factors in explaining the distance from standard Dutch for the pronunciation of individual words in different dialects. Our study differs somewhat from this study, as we do not try to predict dialect distances, but (using logistic regression) a binary value indicating if the lexical form of a concept with respect to standard Italian is different (1) or equal (0).

The 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 for young speakers than old speakers, 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, 2008).

In the next section, we will discuss the Tuscan dialect dataset, followed by a more in-depth explanation of the mixed-effects regression 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* ('Lexical Atlas of Tuscany', henceforth ALT; Giacomelli et al., 2000) which is a specially designed linguistic atlas of Tuscany in which dialectal data have both a diatopic (geographic) and diastratic (social) characterization. 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. 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 only used the normalized lexical forms in the ALT data source, which abstract away from phonetic variation. Unfortunately, morphological variation (both derivational and inflectional) was not excluded from these normalized forms. Therefore our lexical differences with respect to standard Italian will also necessarily contain some morphological differences. However, the effect of these morphological differences on our resulting model is limited, as we did not find distinct differences when analyzing an adapted dataset where much inflectional morphology was automatically removed. Our automatic procedure simply ignored differences at the end of the lexical form and therefore treated (e.g.) *mirtillo* and *mirtilli*, the singular and plural form of blueberry, as identical lexical forms. In addition, when we did not predict binary differences with respect to the lexical form in standard Italian, but more gradual differences (using the normalized edit distance, counting the average number of insertions, deletions and substitutions to transform a dialectal lexical form into the standard Italian lexical form), we also did not find distinct differences with respect to the original model. Using this gradual approach, morphological differences are likely to result in a lower distance than lexical differences, as a larger part of the lexical form will differ in the latter case. We therefore believe the normalized lexical forms in the ALT can be suitably used to evaluate lexical differences in Tuscan dialects, as the effect of morphology appears to be limited.

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 ranging up to 50 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 (since the remaining concepts were nouns) and all multi-word concepts (in order to be able to evaluate the effect of frequency fairly; multi-words have a relatively low frequency and relatively much variation). Our final subset, therefore, consisted of 170 concepts and is listed in Table 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 was created for query purposes. This list, originally compiled on the basis of lexicographic evidence, was carefully revised 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 from Tuscan dialects.

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 much slower than normal regression). Consequently, we grouped the speakers in an old age group (born in 1930 or earlier – 1930 was the median year of birth) and a young age group (born after 1930). For every age group, we used the 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 69,266.

As Wieling et al. (submitted) reported a significant effect of word frequency on dialect distances from standard Dutch, we obtained the word 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.

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

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

Figure 1. Geographical distribution of the locations.

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 locations (Comuni Italiano, 2011). We extracted the number of inhabitants in every location 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 in every location (in 2005; which was the oldest information available) and the average age 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), it is unlikely that the pattern would have been considerably different.

3. Methods

3.1. Modeling the role of geography: generalized additive modeling

An important factor in dialectometry is geography, as geographically closer varieties tend to be linguistically more similar (e.g., see Nerbonne, 2010). A problem in standard regression analysis is that geographic location (as determined by longitude

and latitude) cannot be modeled in a flexible way. The standard regression model can only include longitude and latitude as separate predictors (spanning a plane), or as a multiplicative interaction between longitude and latitude (representing a hyperbolic plane). This clearly is not flexible enough to adequately model the influence of geography.

In line with Wieling et al. (submitted), we use a generalized additive model (GAM) to more adequately model geography. Generalized additive models are an extension of multiple regression and provide flexible tools to model complex surfaces. In agreement with Wieling et al. (submitted) we turn to thin plate regression splines (Wood, 2003), which are able to model a complex wiggly surface (i.e. the influence of geography) as a weighted sum of geometrically simpler, analytically well defined, surfaces (see Wood, 2006 for a detailed discussion). The significance of a thin plate regression spline is evaluated with an *F*-test evaluating whether the estimated degrees of freedom invested in the spline yield an improved fit of the model to the data. Besides predicting linguistic distances on the basis of geography (Wieling et al., submitted), generalized additive models have been used successfully in modeling experimental data in psycholinguistics (Tremblay and Baayen, 2010; Baayen, Kuperman and Bertram, 2010; Baayen, 2010) and biology (e.g., see Schmidt, Kiviste and von Gadow, 2011 for spatial explicit modeling in ecology).

In this study, we use a generalized additive model to generate a two-dimensional surface estimator (based on the combination of longitude and latitude) which estimates lexical differences using thin-plate regression splines as implemented in the `mgcv` package for R (Wood, 2006). Figure 2 shows the resulting regression surface for the complete area under study using a contour plot. The (solid) contour lines represent isoglosses connecting areas which have a similar likelihood of having a lexical form different from standard Italian. Note that the values represent log-odds values and should be interpreted with respect to being different from standard Italian, which means that lower values indicate a smaller likelihood of being different (intuitively it is easiest to view these values as a distance measure from standard Italian). Consequently, the value 0 indicates that in those regions the lexical form is more likely to match with the Italian standard (the probability is 0.45 that the lexical form is *different* from the Italian standard form) and the value 1 indicates the opposite

(the probability is approximately 0.73 that the lexical form is different from the Italian standard form). Correspondingly, darker shades of gray indicate a greater likelihood of having a lexical form equal to standard Italian, while lighter shades of gray represent a greater likelihood of having a lexical form different from standard Italian. We can clearly see that locations near Florence (the dark circle near the top-right) tend to have lexical variants more likely to be equal to the standard Italian form. This makes sense as Italian originated from the region around Florence.

The 26.9 estimated degrees of freedom invested in the thin plate regression spline were supported by a *z*-value of 50 ($p < 0.0001$).

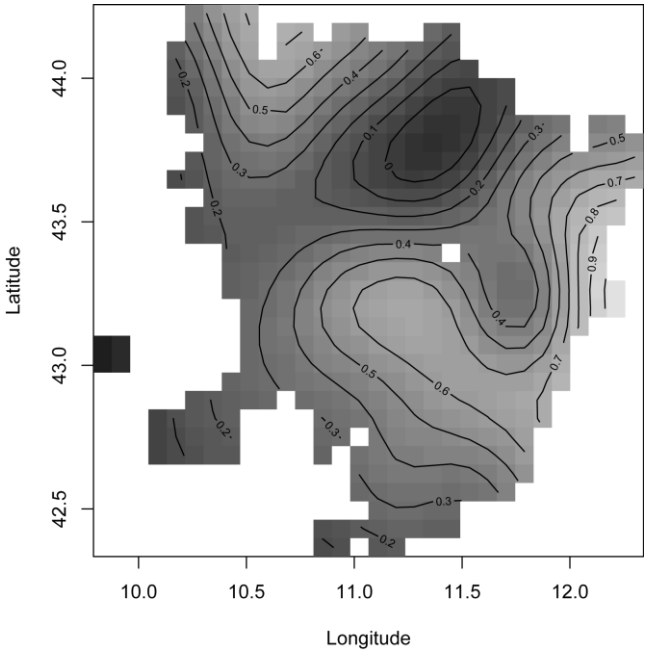

Figure 2. Contour plot for the regression surface of predicting lexical differences from standard Italian as a function of longitude and latitude obtained with a generalized additive model using a thin plate regression spline. The (black) contour lines represent probability isoglosses, darker shades 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.

3.2. Mixed-effects modeling

In this study we apply mixed-effects regression modeling (for introductions, see, e.g., Baayen, 2008, Ch. 7 and Baayen, Davidson and Bates, 2008), comparable to the approach taken by Wieling et al. (submitted). In mixed-effects regression modeling a distinction is made between fixed-effect and random-effect factors. Fixed-effect

factors are factors with a small number of levels that exhaust all possible levels (e.g., the age group is either young or old). Random-effect factors, in contrast, have levels sampled from a much larger population of possible levels.

In our data, there are two random-effect factors that are likely to introduce systematic variation, namely location, and concept. Our observations (i.e. having a lexical form equal or different from standard Italian for a certain concept) are specific to 213 locations. As these locations are a sample of a much larger set of possible locations where Tuscan dialects are spoken, location is a random effect. Each of the 170 concepts was attested in almost all 213 locations. As our concepts are also sampled from a much larger population of possible concepts, concept is our second random effect.

In mixed-effects modeling, random-effect factors are viewed as sources of random noise that can be linked to specific observational units, in our case, locations and concepts. In its simplest form, the variability associated with a random-effect factor is restricted to adjustments to the population intercept. For example, some concepts might be more likely to have a lexical form different from standard Italian, while other concepts might show the opposite pattern. These adjustments are assumed to follow a normal distribution with mean zero and a standard deviation to be estimated from the data. When these adjustments have been estimated, it is possible to adjust the population intercept such that it becomes precise for each individual concept. In this case, these adjusted intercepts are referred to as by-concept random intercepts.

It is also possible that the variation associated with a random-effect factor affects the slopes of other predictors. For example, the slope of population size might vary with concept, indicating that some concepts might be more strongly affected by the population size than others (the mean influence of population size over all concepts equals the general model coefficient of population size). A mixed-effects model will estimate the by-concept biases in the slope of population size, and by adding these estimated adjustments to the general population size slope, by-concept random slopes are obtained that make the estimated effect of population size as precise as possible for each concept. The justification of random intercepts and random slopes is verified

with likelihood ratio tests, which evaluate whether the increase in the number of parameters is justified given the increase in goodness of fit.

Statistical models combining mixed-effects regression and generalized additive modeling are still being developed and are as a consequence not completely stable (e.g., see Schmidt et al., 2011). We therefore followed the approach taken by Wieling et al. (submitted) and Schmidt et al. (2011), by using the generalized additive model to predict if the lexical form is different from standard Italian solely on the basis of longitude and latitude. We then use the fitted values of this simple model (see Figure 2), as a predictor (GAM Distance) representing geography in our model.

In our analyses, we considered the two aforementioned random-effect factors (i.e. location and concept) as well as several other predictors besides the GAM Distance. The only lexical variable we included was concept frequency (based on the frequency of the standard Italian lexical form). The location-related variables we investigated were population size, average population age, average population income and 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 recurrent problem in large-scale regression studies is collinearity of the predictors. 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).

In order to reduce the potentially harmful effect of outliers, a number of numerical predictors were log-transformed (i.e. population 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 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 effects. 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 fixed-effect factors and covariates are shown in Table 2. The included random-effect structure is shown in Table 3. The model converged in about 3 CPU hours.

The inclusion of the fixed-effect and random-effect factors shown in Tables 2 and 3 was warranted, as every additional factor increased the goodness of fit of the model. Tables 4 and 5 show the increase in goodness of fit for the fixed-effect and random-effect factors measured by the increase of the log-likelihood and the decrease of the Akaike Information Criterion (Akaike, 1974). Log-likelihood ratio tests for the fixed-effects were carried out with maximum likelihood estimation (for fixed-effect factors) or relativized maximum likelihood estimation (for random-effect factors), as recommended by Pinheiro and Bates (2000).

	Estimate	Std. Error	*z*-value	*p*-value
Intercept	-2.9943	0.3455	-8.668	< 0.001
Old instead of young speakers	0.5782	0.0313	18.460	< 0.001
Geography (GAM Distance)	5.8018	0.5269	11.011	< 0.001
Population size (log)	-0.1028	0.0289	-3.557	< 0.001
Population average income (log)	-0.0771	0.0344	-2.243	0.025

Table 2. Fixed-effect factors of the final model. A positive estimate indicates that that a higher value for this predictor increases the likelihood of having a different lexical form than the standard Italian one. A negative estimate reduces the likelihood of having a different lexical form than the standard Italian one.

Factors	Random effects	Std. Dev.	Correlation
Location	Intercept	0.32549	
	Concept frequency (log)	0.16976	0.213
Concept	Intercept	1.85494	
	Old instead of young speakers	0.30304	
	Population size (log)	0.17240	-0.287
	Average population income (log)	0.35244	
	Average population age (log)	0.28309	-0.794

Table 3. Random-effect factors of the final model. The standard deviation indicates the amount of variation for every random intercept and slope. The correlation parameters indicate how related the random slopes and intercepts are (the correlation value involves the random effect on the same line and the line directly above).

To assess the influence of each additional fixed-effect factor, the random effects were held constant and consisted of the complete random-effect structure shown in Table 3. The baseline model, to which the inclusion of the first fixed-effect factor (the contrast between the old and young speakers) was compared, only consisted of the complete random effect structure (without fixed-effect factors). Subsequently, the second model (including geography in addition to the age group contrast), was compared to the model including the age group contrast only. The results are shown in Table 4.

Similarly, the importance of additional random-effect factors was assessed (in Table 5) by fixing the fixed-effect factors to those included in the final model (shown in Table 2). The baseline model only included the fixed-effect factors. The model including a random intercept for concept was compared to this baseline model. The model to which the inclusion of the first random slope (the concept frequency; fourth line in Table 5) as well as the correlation between this random slope and the random intercept of location (0.213; see Table 3) was compared only consisted of the random intercepts for concept and location together with the complete fixed-effect structure of the final model.

	Log-likelihood increase	AIC decrease	Likelihood ratio test
Random-effect factors			
+ Old instead of young speakers	83.6	165.2	$p < 0.0001$
+ Geography (GAM Distance)	54.4	106.7	$p < 0.0001$
+ Population size (log)	6.1	10.3	$p = 0.0005$
+ Population average income (log)	3.0	4.0	$p = 0.0143$

Table 4. Goodness of fit of the fixed-effect factors of the model. Every row specifies the significant increase in goodness of fit obtained by adding the current predictor to the model including all preceding predictors. All models include the random-effect factors listed in Table 3.

	Log-likelihood increase	AIC decrease	Likelihood ratio test
Fixed-effect factors			
+ Random intercept concept	10779.1	21556.3	$p < 0.0001$
+ Random intercept location	362.4	722.9	$p < 0.0001$
+ Concept frequency (location)	52.2	100.4	$p < 0.0001$
+ Old instead of young speakers (concept)	51.5	101.1	$p < 0.0001$
+ Population size (concept)	99.9	195.8	$p < 0.0001$
+ Average population income (concept)	335.6	669.1	$p < 0.0001$
+ Average population age (concept)	278.2	552.4	$p < 0.0001$

Table 5. Goodness of fit of the random-effect factors. Each row specifies the significant increase in goodness of fit of the model resulting from inclusion of the specified random slopes to the preceding model. All models include the fixed-effect factors listed in Table 2.

4.1. Demographic predictors

When inspecting the *z*-values in Table 2, it is clear that GAM Distance representing geography (see Figure 2) is a very strong predictor. Because GAM Distance represents the fitted values of a generalized additive model with respect to pronunciation distance from the standard (adjusted $R^2 = 0.01$) the strong support for this predictor is not surprising. Note that the log-odds estimate of 5.83 corresponds with a probability of 0.997, indicating that the fitted values of the GAM model (GAM Distance) are kept almost unchanged. It is also reassuring that GAM Distance remains significant in a model in which location is included as a random-effect factor. In the following we will investigate to what extent other predictors help in explaining lexical differences with respect to the standard Italian form.

The most important predictor (having the highest *z*-value) was the contrast between the age groups, with older speakers being more likely to have a lexical form different from standard Italian. In addition, the effect of this contrast showed significant variation between concepts. While for the great majority of concepts, old speakers were more likely to have a lexical form different from standard Italian, for a few concepts the pattern was reversed. Figure 3 shows the by-concept random slopes for the contrast between old and young speakers. A higher value for a concept indicates that this concept is more likely to differ from the standard Italian form for old speakers as opposed to young speakers. The dashed line indicates the model parameter for the contrast (see Table 2), indicating that the older speaker group is more likely to have a lexical form different from the standard Italian form than the young speaker group (with concepts *abete*, 'fir' and *edera*, 'ivy' being most extreme). This result is not surprising as younger speakers tend to converge to standard Italian.

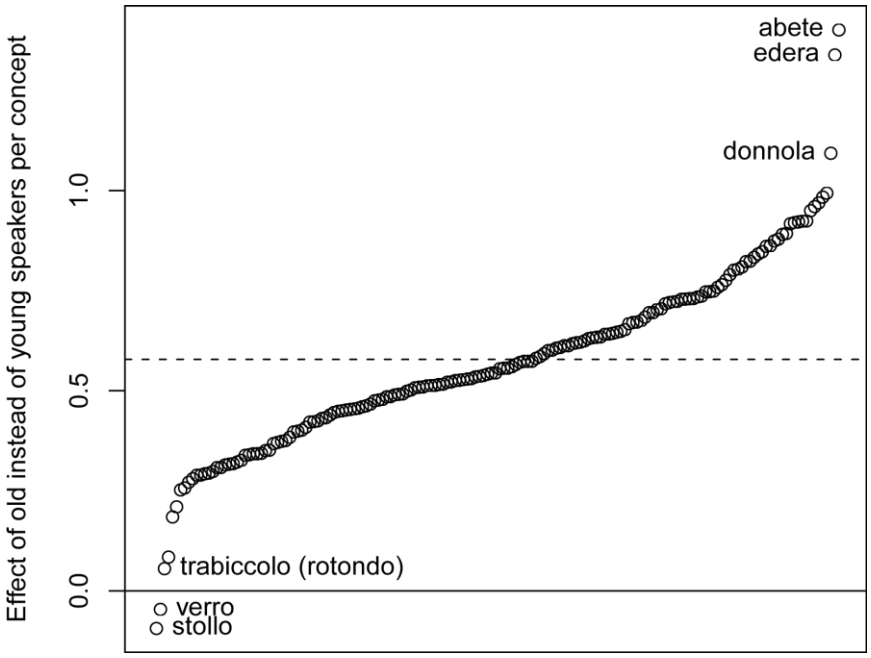

Figure 3. By-concept random slopes of the contrast between old and young speakers. The concepts are sorted by their coefficient for the contrast between old and young speakers. The model estimate (see Table 2) is indicated by the dashed line.

Interestingly, the two concepts (*verro*, 'boar' and *stollo*, 'haystack pole') show an opposite pattern, with younger speakers being more likely to have a lexical form different from standard Italian than older speakers. These two concepts involve very old-fashioned concepts for which old speakers in many cases know a specific term, but the young speakers have no word in their vocabulary to indicate the distinction from the more general term (e.g., they use the more general words 'pig' and 'pole' or resort to a multi-word expression).

Of all location-based predictors (i.e. the population size, the average population income and the average population age in a location) only the first two were significant predictors in the general model. Larger populations and populations with a higher average income were more likely to have a lexical variant close to standard Italian (i.e. the estimates in Table 2 are negative). A possible explanation for these findings is that people tend to have weaker social ties in urban populations, which causes dialect leveling (Milroy, 2002). As the standard Italian language is more prestigious than dialectal forms (Danesi, 1974), conversations will be normally held in standard Italian and consequently leveling will proceed in the direction of standard Italian. The higher prestige of standard Italian might also be an explanation why communities with a higher average income tend to be closer to standard Italian. The other location-based predictor, average age, was not significant in the general model, but this might be due to having two age groups per location (which are much more suitable to detect age differences). 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.

The location-related variables (i.e. population size, average income and average age) showed significant by-concept variation. Figure 4 (illustrating population size) shows some concepts following the general pattern (the model estimate is indicated by the dashed line), with bigger populations being more likely to have a lexical form equal to standard Italian (i.e. *ovile*, 'sheepfold', *scoiattolo*, 'squirrel' and *donnola*, 'weasel'), while others behave in completely opposite fashion (*stollo*, 'haystack pole', *castagnaccio*, 'chestnut cake' and *melone*, 'melon'). Similar to the by-concept random slopes for the age group contrast, we see (with the exception of *melone*) old-fashioned

concepts which behave in the opposite direction (i.e. other concepts which are not shown in the graph, but are also included in this set, include *verro*, 'boar' and *ditale*, 'thimble'). It might be that in the larger and richer towns, people are less likely to remember these old forms, as they are used less frequently. Note that the (slight) similarity between the by-concept random slopes for the age contrast and population size, are expressed by the correlation coefficient of -0.287 shown in Table 3.

Figure 5 illustrates the by-concept random slopes for average age and average income. While the model estimate of average income (indicated by the dashed vertical line) indicates that richer populations are more likely to have a lexical form equal to standard Italian (with concepts *ovolo,*'royal agaric' (a mushroom) and *arancia*, 'orange' being close to the extreme), the concepts in the lower-left quadrant (e.g., *riccio* and 'hedgehog' and *castagnaccio*, 'chestnut cake') show the opposite pattern, with richer populations more likely to have a lexical form different from standard Italian.

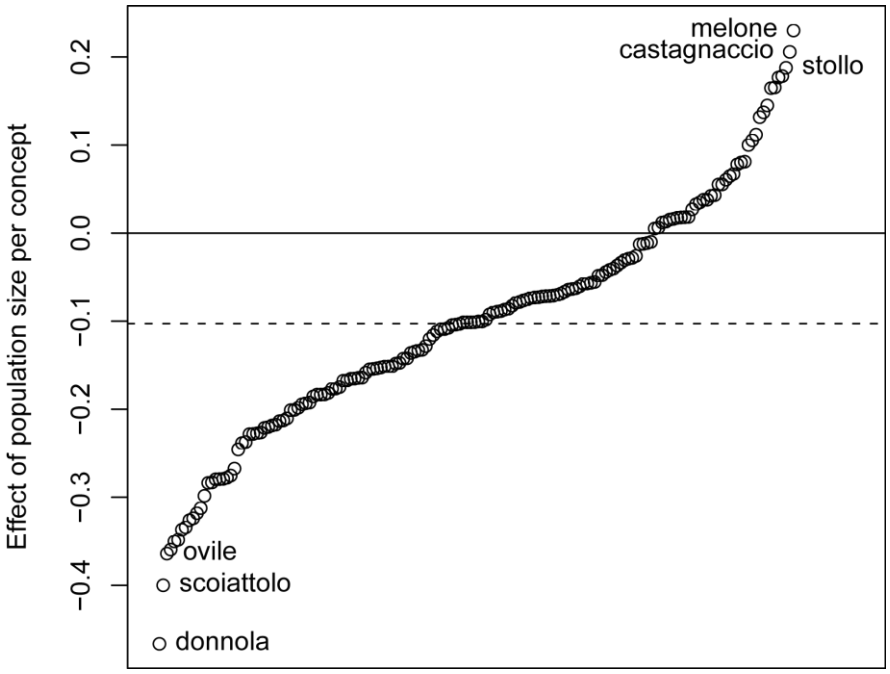

Figure 4. By-concept random slopes of population size. The concepts are sorted by the value of their population size coefficient. The model estimate (see Table 2) is indicated by the dashed line.

In addition, the strong correlation (-0.794; see Table 3) between the by-concept random slopes of average age and average income indicate that the effect average age and average income have on distinct concepts is closely linked. Concepts that are more likely to differ from the Italian lexical form in poorer locations, are also more likely to differ from the Italian lexical form in locations with a higher average age (e.g., *ovolo* and *arancia*). Similarly, concepts which follow the opposite pattern and are more likely to differ from standard Italian for richer populations, are also more likely to differ from the Italian lexical form in younger populations. A similar result was also reported by Wieling et al. (submitted) where there they found that by-word random slopes of average age, average income as well as population size were closely linked. However, in our case there was no support for a link between by-concept random slopes for population size and the other by-concept random slopes as well. As we again see some old-fashioned concepts in the bottom-left corner, it might be that in richer, younger towns, people are less likely to remember these old forms and therefore are less likely to be identical to standard Italian.

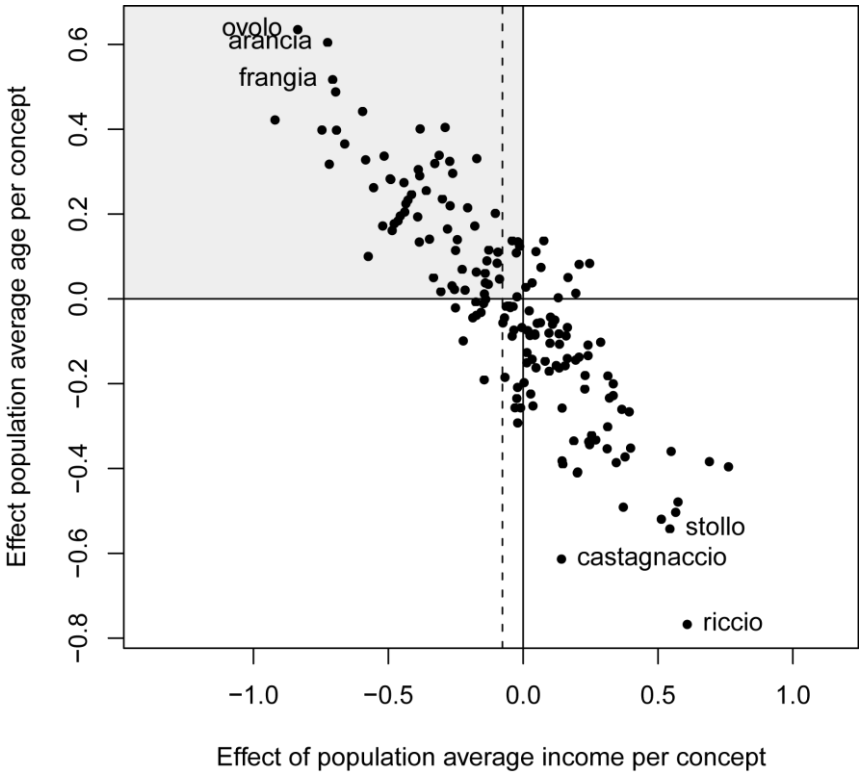

Figure 5. By-concept random slopes of average population age and income. The dashed vertical line marks the model estimate (see Table 2) of average income (average age was not included in the model as it was not significant). The grey quadrant indicates where most dots (concepts) are located.

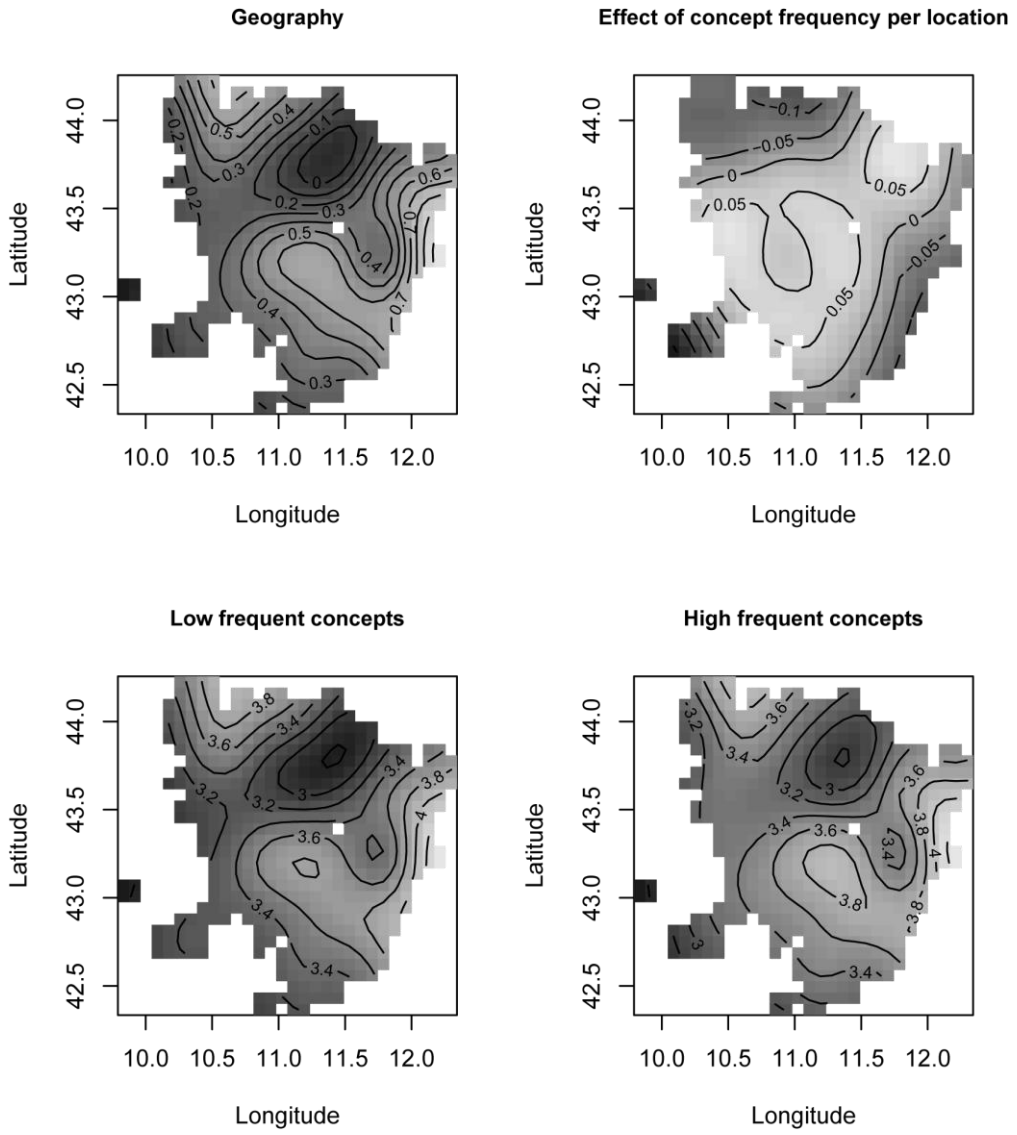

Figure 6. Concept frequency and lexical differences from standard Italian. Upper-left: distance only predicted from longitude and latitude. Upper-right: the geographical distribution of random slopes for concept frequency. Bottom panels: the combined effect of geography and low-frequent words (bottom-left) versus high-frequent words (bottom-right). Lighter shades of gray denote a greater likelihood of having a lexical form different from standard Italian.

4.2. Concept frequency

Concept frequency was not significant in our general model. However, we did find significant geographical variation, shown in Figure 6. The top-left graph is equal to Figure 2 and shows the general effect of geography on predicting lexical differences with respect to standard Italian (note that the values in this graph cannot be compared to the values in the other three graphs, as these values are based on a logistic GAM, while the others are not). The top-right graph shows that the effect of concept

frequency differs per location with more frequent concepts being more likely to differ from standard Italian, especially in the region around Florence. We see the opposite pattern in the more peripheral northern and south-eastern areas, where more frequent concepts are more likely to be similar to the standard Italian form. A possible explanation for this pattern might be that areas relatively distinct from standard Italian (the north and the southeast) are more likely to adopt standard Italian lexical forms for the more frequent concepts, whereas in the region around Florence these high-frequent concepts might already be equal to standard Italian.

4. Discussion and conclusions

In this study we have shown that the lexical variation in Tuscan dialects with respect to the standard Italian lexical form can adequately be modeled by a logistic mixed-effects regression model in combination with generalized additive modeling. We clearly showed support for the importance of geography, speaker age and several location-related variables. In addition we showed that mixed-effects regression also enable a detailed investigation of the precise effect of different location-related (or concept-related) variables for individual concepts (or locations).

As we remarked in the introduction, our dataset unfortunately contained some morphological variation. While we have verified by additional tests that this likely did not influence our results significantly, it would be good to verify that this indeed is the case by manually removing the morphological variation and investigating if the results remain the same.

Instead of using 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 can be distinguished from more rigorous lexical differences. As this would not require 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.

When keeping lexical differences binary, however, it would also be interesting to investigate the importance of other speaker characteristics (e.g., education) by creating groups based on the selected characteristic.

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) for the year 2011.

References

Akaike, Hirotugu (1974). A new look at the statistical identification model. *IEEE transactions on Automatic Control*, 19(6): 716-723.

Baayen, R. Harald (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baayen, R. Harald (2010). The directed compound graph of English. An exploration of lexical connectivity and its processing consequences. In S.Olson (ed.), *New impulses in word-formation* (Linguistische Berichte Sonderheft 17). Buske, Hamburg, 383-402.

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.

Baayen, R. Harald, Victor Kuperman and Raymond Bertram (2010). Frequency effects in compound processing. In S. Scalise and I. Vogel (eds.), *Compounding*. Benjamins, Amsterdam / Philadelphia, 257-270.

Brants, Thorsten & Alex Franz (2009). Web 1T 5-gram, 10 European Languages Version 1. Linguistic Data Consortium, Philadelphia.

Communi Italiano (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-italiono.it. Last accessed: 2011-05-23.

Giacomelli, G. (1975). Aree lessicali toscane. *La ricerca dialettale*, 1: 115-152.

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, T. Poggi Salani (2000). *Atlante Lessicale Toscano*. Lexis Progetti Editoriali, Roma.

Giacomelli, G. and T. Poggi Salani (1984). Parole toscane. In *Quaderni dell'Atlante Lessicale Toscano*, 2(3): 123-229.

Giannelli, Luciano (2000). *Toscana*. Second edition. Pisa, Pacini.

Johnson, Daniel Ezra (2008). Getting off the GoldVarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Language and Linguistic Compass*, 3(1): 359–383.

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

Nerbonne, John (2010). Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365: 3821-3828.

Pinheiro, José and Douglas Bates (2000). *Mixed-effects models in S and S-PLUS. Statistics and Computing*. Springer, New York.

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.

Schmidt, Matthias, Andres Kiviste and Klaus von Gadow (2011). A spatially explicit height-diameter model for Scots pine in Estonia. *European Journal of Forest Research*, 130: 303-315.

Tremblay, Antoine and R. Harald Baayen (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (ed.), *Perspectives on formulaic language: Acquisition and communication*. The Continuum International Publishing Group, London, 151-173.

Wieling, Martijn, John Nerbonne and R. Harald Baayen (submitted). Quantitative Social Dialectology: Explaining Linguistic Variation Socially and Geographically.

Wood, Simon (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B* (Statistical Methodology), 65(1): 95-114.

Wood, Simon (2006). *Generalized additive models: an introduction with R*. Chapman & Hall/CRC.