BORDER EFFECTS AMONG CATALAN DIALECTS

Martijn Wieling[1], Esteve Valls[2], R. Harald Baayen[3,4] and John Nerbonne[1,5]

[1]Department of Humanities Computing, University of Groningen, The Netherlands

[2]Department of Catalan Philology, University of Barcelona, Spain

[3]Department of Quantitative Linguistics, University of Tübingen, Germany

[4]Department of Linguistics, University of Alberta, Edmonton, Canada

[5]Freiburg Institute for Advanced Studies, University of Freiburg, Germany

m.b.wieling@rug.nl, e.valls@ub.edu, harald.baayen@uni-tuebingen.de, j.nerbonne@rug.nl

Abstract In this study, we investigate which factors influence the linguistic distance of Catalan dialectal pronunciations from standard Catalan. We use pronunciations from three regions where the northwestern variety of the Catalan language is spoken (Catalonia, Aragon and Andorra). In contrast to Aragon, Catalan has an official status in both Catalonia and Andorra, which likely influences standardization. Because we are interested in the potentially large range of differences that standardization might promote, we examine 357 words in Catalan varieties and in particular their pronunciation distances with respect to the standard. In order to be sensitive to differences among the words, we fit a generalized additive mixed-effects regression model to this data. This allows us to examine simultaneously the general (i.e. aggregate) patterns in pronunciation distance and to detect those words that diverge substantially from the general pattern. The results reveal higher pronunciation distances from standard Catalan in Aragon than in the other regions. Furthermore, speakers in Catalonia and Andorra, but not in Aragon, show a clear standardization pattern, with younger speakers having dialectal pronunciations closer to the standard than older speakers. This clearly indicates the presence of a border effect within a single country with respect to word pronunciation distances. Since a great deal of scholarship focuses on single segment changes, we compare our analysis to the analysis of three segment changes that have been discussed in the literature on Catalan. This comparison shows that the pattern observed at the word pronunciation level is supported by two of the three cases examined. As not all individual cases conform to the general pattern, the aggregate approach is necessary to detect global standardization patterns.

1 Introduction

In this study we investigate a Catalan dialect data set in order to identify social and linguistic factors which play an important role in predicting the distance between dialectal pronunciations and the Catalan standard language (which is a formal variety of Catalan mainly based on the dialects of the eastern counties of Catalonia, including those of the Barcelona area). We use Catalan dialect pronunciations of 320 speakers of varying age in 40 places located in three regions where the northwestern variety of the Catalan language is spoken (the autonomous communities Catalonia and Aragon in Spain, and the state of Andorra). Our approach allows us to investigate border effects caused by different policies with respect to the Catalan language. As the Catalan language has been the native and official language (i.e. used in school and in public media) of both Andorra and Catalonia, but not in Aragon,[1] we will contrast these two regions in our analysis.

We show that the speakers of Catalan in Catalonia and Andorra use a variety of Catalan closer to the standard that those in Aragon. Because this tendency is particularly strong among younger speakers, we argue that it is at least in part due to the introduction of Catalan as an official language in the 1980's in Catalonia and Andorra but not in Aragon. Naturally the differences we find may have existed before the language became official in Catalonia, but this cannot explain the larger differences among the young.

Since we suspect that the changes associated with standardization will be far-ranging, we deliberately conduct our analysis in a way that is likely to detect a wide range of differences, effectively aggregating over all differences with respect to the standard in each variety we examine. By taking into account many variables, we deliberately deviate from common sociolinguistic practice which typically focuses on only a small number of variables. We cast a wider net in an effort to obtain a more comprehensive (i.e. aggregate) view, and avoid selecting only those variables that behave as predicted. In a second step, we will investigate whether the aggregate pattern observed at the word pronunciation level also holds when focusing on the more commonly investigated sound (phonemic) level.

---

[1] In Andorra, Catalan is the only official language. In Catalonia, where Spanish and Aranese (a variety of Occitan) are also official, Catalan was the vehicular language of education during the 1920s and the 1930s and achieved this status again after Franco's dictatorship in the early 1980s (Woolard and Gahng, 2008). That means that all subjects except second and third languages are taught in Catalan in the public schools of Catalonia and Andorra. In Aragon, Catalan has only been a voluntary subject in schools in the eastern counties (where Catalan is spoken) since 1984 (Huguet, Vila and Llurda, 2000). The standard variety used at all schools in these areas is the one sanctioned by the *Institut d'Estudis Catalans* (Fabra, 1918).

1.1 Border effects

Border effects in European dialectology have been studied intensively (see Woolhiser, 2005 for an overview). In most of these studies, border effects have been identified on the basis of a qualitative analysis of a sample of linguistic features. In contrast, Goebl (2000) used a dialectometric approach and calculated aggregate dialect distances based on a large number of features to show the presence of a clear border effect at the Italian-French and Swiss-Italian borders, but only a minimal effect at the French-Swiss border. This approach is arguably less subjective than current practice in social dialectology (focusing on a pre-selected small set of items), as many features are taken into account simultaneously and the measurements are very explicit. However, Woolhiser (2005) is very critical of this study, as Goebl does not discuss the features he used and also does not consider the sociolinguistic dynamics as well as ongoing dialect changes (i.e. he uses static dialect atlas data).

Border effects have generally been studied with respect to national borders. In the present paper, we focus on one language border within a single nation state, and on a second border between two states. The former kind of border has been scarcely studied at all (Woolhiser, 2005).

Several researchers have offered hypotheses about the presence and evolution of border effects in Catalan. For example, Pradilla (2008a, 2008b) indicates that the border effect between Catalonia and Valencia might increase, as the two regions recognize different varieties of Catalan as standard (i.e. the unitary Catalan standard in Catalonia and the Valencian Catalan substandard in Valencia). In a similar vein, Bibiloni (2002, p. 5) discusses the increase of the border effect between Catalan dialects spoken on either side of the Spanish-French border in the Pyrenees during the last three centuries. More recently, Valls, Wieling and Nerbonne (2013) conducted a dialectometric analysis of Catalan dialects and found, on the basis of aggregate dialect distances (average distances based on hundreds of words), a clear border effect contrasting Aragon with Catalonia and Andorra. This dialectometric approach is an improvement over Goebl's (2000) approach, since they measure dialect change by including pronunciations for four different age groups (measuring dialect evolution by the apparent-time construct; Bailey, 1991). However, it ignores other sociolinguistic variables due to its purely dialectometric nature.

1.2 Combining dialectometry and social dialectology

The methodology used in the present study essentially follows dialectometry, which has generally focused on determining aggregate pronunciation distances, and the geographical pattern of aggregate variation (Wieling, 2012, Ch. 1). In contrast, many dialectologists have focused on the influence of specific social factors on the realization of (individual) linguistic variables. Instead of examining a large

set of items simultaneously, however, social dialectologists have generally investigated smaller sets of pre-selected linguistic variables.

We grant the essential correctness of Woolhiser's (2005) critique that dialectometry has at times been blind to the potential importance of non-geographic conditioning factors. Therefore, in this study, we combine perspectives from two approaches, dialectometry and social dialectology. Following dialectometry, we will measure distances for a large set of dialectal pronunciation data, preventing in this way biased choices in the selection of material (Nerbonne, 2009). (Of course, as we work with a pre-existing pronunciation data set our analysis will be biased as well towards the material included in this set.) In line with social dialectology, however, in analyzing these distances, we will also take several social factors into account. We have not conducted surveys to determine how the differences we measure are perceived socially. In this sense, we are not in a position to gauge the social meaning of the changes we examine, as sociolinguists often expect. We nonetheless explore the hypothesis that linguistic changes are being brought about by a social change, namely the change to using standard Catalan in schools and public media in part of the Catalan-speaking area. In this sense we are conducting a sociolinguistic study.

In addition, we aim to clarify the relationship between aggregate (dialectometric) analyses, which often ignore the linguistic details most responsible for aggregate relations, and analyses based on selected linguistic features (most non-dialectometric analyses). While dialectometric analyses have aimed at establishing the relations among varieties, analyses based on selected linguistic features such as rhoticization, the raising of front vowels or varying verbal inflections are often motivated both by the wish to establish the social affinities of variation, but also by the wish to adduce linguistic structure in the variation.[2]

1.3 Hypotheses

In our analysis we will contrast the area where Catalan is recognized as an official language (Catalonia and Andorra) with the area where it is not (Aragon). This contrast allows us to investigate the influence of an internal border within the same country (i.e. Aragon versus Catalonia) as opposed to a national border (Andorra-Spain). Based on the results of Valls et al. (2013), we expect to observe larger pronunciation distances from standard Catalan in Aragon than in the other two regions[3]. More importantly, however, we expect that the models will differ

[2] Wieling and Nerbonne (2011, 2015) summarize several earlier attempts to ascertain the linguistic foundations of aggregate dialectometric differences, so we shall not review those here.

[3] It might be argued that this pattern is due to the fact that the Catalan standard language is mainly based on the eastern dialects of Catalonia. Although it is true that the northwestern varieties of Catalonia and Andorra have historically converged towards the (closer and more prestigious) eastern varieties during the

with respect to the importance of the sociolinguistic factors. Mainly, we expect to see a clear effect of speaker age (i.e. with younger speakers having pronunciations closer to standard Catalan) in the area where Catalan has the status of an official language, while we do not expect this for Aragon, as there is no official language policy which might 'attract' the dialect pronunciations to the standard. In contrast to the exploratory visualization-based analysis of Valls et al. (2013), our (regression) analysis allows us to assess the significance of these differences. For example, while Valls et al. (2013) state that urban communities have pronunciations more similar to standard Catalan than rural communities, this pattern might be non-significant (as they reach this conclusion on the basis of visualization only).

In addition we shall examine a methodological hypothesis, namely that the standardization we are interested in will be more insightfully investigated from an aggregate, dialectometric perspective rather than from the perspective of a small number of sound changes. In defense of the plausibility of this view we note that standardization efforts are unlikely to be undertaken if only a small number of linguistic items is at stake. Standardization normally involves a large number of changes, certainly when viewed from the perspective of all the different varieties affected. However, while we do intend to examine this hypothesis, we do not propose to test it rigorously in this study.

2 Material

2.1 Pronunciation data

The Catalan dialect data set contains basilectal phonetic transcriptions (using the International Phonetic Alphabet) of 357 words in 40 dialectal varieties and the Catalan standard language. The locations are spread out over the state of Andorra (2 locations) and two autonomous communities in Spain (Catalonia with 30 locations and Aragon with 8 locations). In all locations, Catalan has traditionally been the dominant language. Figure 1 shows the geographical distribution of these locations. The locations were selected from 20 counties, and for each county the (urban) capital as well as a rural village was chosen as a data collection site. In every location eight speakers were interviewed, two per age group (F1: born between 1991 and 1996; F2: born between 1974 and 1982; F3: born between 1946 and 1960; F4: born between 1917 and 1930). All data was transcribed by a single transcriber (Esteve Valls), who also did the fieldwork for the youngest (F1) age-group between 2008 and 2011. The fieldwork for the other age groups was conducted by another fieldworker (Mar Massanell) between 1995 and 1996. The complete data set we use contains 357 items, consisting of 16 articles, 81 clitic pronouns, 8

20th century, Valls et al. (2013, section 4.2) have shown that the standardization process has been much more effective in the diffusion of the prestigious features westwards.

demonstrative adjectives, 2 neuter pronouns, 2 locative adverbs, 220 inflected forms of 5 verbs, 20 possessive adjectives and 8 personal pronouns. The complete item list and a more detailed description of the data set are given by Valls et al. (2013). Note that the data set did not contain any nouns and only contained a limited number of verbs. The fact that over 60% of the words studied are forms of only five verbs means that the sample is biased toward these words. A follow-up study using different material would be worthwhile. However, these five verbs are representative of the five regular paradigms in Catalan and allow us to take into account all the regular inflections of the Catalan verbs.

The standard Catalan pronunciations were transcribed by the second author and are based on the *Gramàtica Catalana* (Fabra, 1918) and the proposal of the *Institut d'Estudis Catalans* for an oral Standard Catalan language (*Institut d'Estudis Catalans*, 1999a, 1999b).

2.2 Sociolinguistic data

Besides the information about the speakers present in the corpus (i.e. gender, age and education level of the speaker), we extracted additional demographic information about each of the 40 locations from the governmental statistics department of Catalonia (*Institut d'Estadística de Catalunya*, 2008, 2010), Aragon (*Instituto Aragonés de Estadística*, 2007, 2009, 2010) and Andorra (*Departament d'Estadística del Govern d'Andorra*, 2010). The information we extracted for each location was the number of inhabitants (i.e. community size), the average community age, the average community income, and the relative number of tourist beds (i.e. per inhabitant; used to estimate the influence of tourism) in the most recent year available (ranging between 2007 and 2010). There was no location-specific income information available for Andorra, so for these two locations we used the average income of the country (*Cambra de Comerç – Indústria i Serveis d'Andorra*, 2008).

As the data for the older speakers (age groups F2, F3 and F4) was collected in 1995, the large time span between the recordings and measurement of demographic variables might be problematic. We therefore obtained information on the average community age, average community income and community size for most locations in 2000 (which was the oldest data available online). Based on the high correlations between the data from the year 2000 and the most recent data for each of the separate measures (in all cases $r > 0.9$, $p < 0.001$), we decided to use the most recent demographic information in this study. No historical information about the number of tourist beds was available for Catalonia and Aragon, but we do not have reason to believe that this correlation strength should be lower than for the other variables (and thus we can use the most recent data).

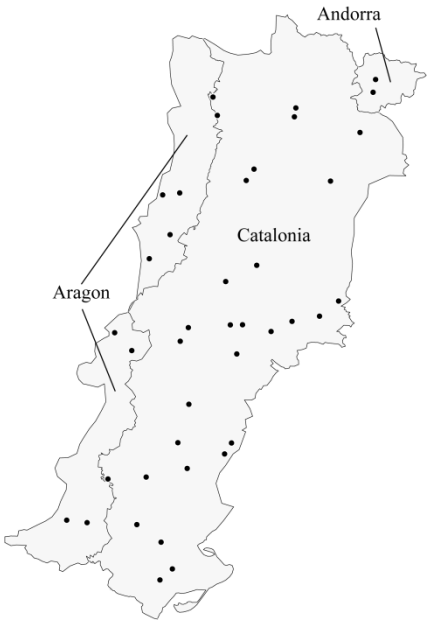

Fig. 1. Geographical distribution of the locations. Two locations are found in Andorra, eight in Aragon and the remaining thirty locations are found in Catalonia.

3 Methods

3.1 Obtaining pronunciation distances

For all 320 speakers, we calculated the pronunciation distance between the standard Catalan pronunciations and their dialectal counterparts by using a modified version of the Levenshtein distance (Levenshtein, 1965). The Levenshtein distance transforms one string into the other by minimizing the number of insertions, deletions and substitutions. For example, the Levenshtein distance between two Catalan variants of the word 'if I drank', [beɣésa] and [bejɣés] is 3:

be ɣésa	insert j	1
bejɣésa	subst. ɛ́ for é	1
bejɣɛ́sa	delete a	1
bejɣɛ́s		
		3

This sequence corresponds with the following alignment:

b	e		ɣ	é	s	a	
b	e	j	ɣ	ɛ́	s		
		1		1		1	

The standard Levenshtein distance does not distinguish vowels from consonants and therefore could align these together. In order to prevent these (linguistically) undesirable alignments, a syllabicity constraint is normally added, allowing only alignments of vowels with vowels, consonants with consonants, and /j/ and /w/ with both consonants and vowels. It prevents alignments of other sounds, as these are assigned a very large (arbitrary) distance (Heeringa, 2004; Heeringa et al., 2006).

It is clear that these Levenshtein pronunciation distances are very crude as the Levenshtein algorithm does not distinguish (e.g.,) substitutions involving similar sound segments, such as /e/ and /ɛ/, from more different sound segments, such as /e/ and /u/. Wieling, Prokić and Nerbonne (2009) proposed a method to automatically obtain more sensitive sound segment distances on the basis of how frequent they align according to the Levenshtein distance algorithm. Sound segments aligning relatively frequently obtain a low distance, while sound segments aligning relatively infrequently are assigned a high distance. The sound distances are based on calculating the Pointwise Mutual Information score (PMI; Church and Hanks, 1990) for every pair of sound segments. The automatically obtained sound segment distances were found to be phonetically sensible (based on six independent dialect data sets; Wieling, Margaretha and Nerbonne, 2012) and also improved pronunciation alignments when these sound segment distances were integrated in the Levenshtein distance algorithm (Wieling et al., 2009). A detailed description of the PMI-based approach can be found in Wieling et al. (2012). Similar to the study of Wieling et al. (2011) on pronunciation differences between Dutch dialects and standard Dutch, our pronunciation distances are not based on the Levenshtein distance (with syllabicity constraint), but rather on the PMI-based Levenshtein distance. Using this phonetically more sensitive measure, the difference of the example alignment shown above is 0.107. The calculation is illustrated below:

b	e		ɣ	é	s	a	
b	e	j	ɣ	ɛ́	s		
		0.0339		0.0345		0.0388	

On average, longer words will have a greater pronunciation distance (i.e. more sounds may change) than shorter words. Therefore we normalize the PMI-based word pronunciation distances by dividing by the alignment length. Since the distribution of the Levenshtein distances was skewed, we log-transformed these distances (after adding a small value, 0.01, to prevent taking the log of 0). Note that log-transforming the PMI-based Levenshtein distances has been previously re-

ported to increase the match with perceptual distances (for native-likeness; Wieling et al., 2014). After log-transformation, we centered the Levenshtein distances (i.e. subtracted the mean value). Consequently, a Levenshtein distance of 0 indicates the average Levenshtein distance, whereas negative and positive values are indicative of Levenshtein distances lower or higher than the average, respectively.

3.2 Mixed-effects regression modeling

The usefulness of a generalized linear mixed-effects regression model (GLMM) in language variation research has already been argued for extensively by Tagliamonte and Baayen (2012). In summary, a generalized linear mixed-effects regression model allows the researcher to determine which variables (i.e. predictors) are important in language variation, while also taking into account that the interviewed informants as well as the specific linguistic items included are a source of variation. While the GLMM is suitable to determine the preference for a certain form over another (e.g., *was* versus *were* in the study of Tagliamonte and Baayen, 2012), the dependent variable may also be numerical instead of binary. In our case, the numerical dependent variable will be the pronunciation distance from standard Catalan on the basis of the log-transformed and centered PMI-based Levenshtein distance.

As explained by Tagliamonte and Baayen (2012), a mixed-effects regression model distinguishes fixed-effect factors from random-effect factors. Fixed-effect factors have a small (fixed) number of levels that exhaust all possible levels (e.g., gender is either male or female), while random-effect factors have levels sampled from a large population of possible levels (e.g., we use 357 words, but could have included other words). A mixed-effects regression analysis allows us to take the systematic variability linked to our speakers, locations and words (i.e. our random-effect factors) into account. For example, some words might (generally) be more similar to standard Catalan than other words. By estimating how much more similar these words are, the general regression formula can be adapted for every individual word. These adjustments to the general model's intercept are called 'random intercepts'. For example, Figure 2 shows the effect of the (standardized) year of birth of the speakers on the (log-transformed and centered) linguistic distance from standard Catalan for two different words, *meves* 'my' (feminine plural possessive), and *ell* 'he'. In these graphs, each circle corresponds to the pronunciation of *meves* (left graph) or *ell* (right graph) of a single speaker. The dashed line (which is the same in both graphs) indicates the general effect (across all words) of the year of birth of the speaker on the linguistic distance from standard Catalan (i.e. the fixed effect). It shows a slightly negative slope, with the intercept (i.e. the height at where the standardized year of birth of the speaker equals zero; the reason for standardizing the predictors is explained below) being close to zero. The solid line in each graph shows the word-specific effect of year of birth of the speaker on the linguistic distance from standard Catalan (i.e. the fixed effect plus

the random intercept and random slope; see below). Clearly, the solid line belonging to the word *meves* has an intercept which is higher than the dashed line (i.e. *meves* generally has a higher linguistic distance from standard Catalan than the average word), while the solid line of *ell* is positioned much lower (and thus *ell* is, on average, more similar to standard Catalan).

Similarly, the effect of a certain predictor may also vary per word. For example, while in general younger speakers may have pronunciations closer to standard Catalan than older speakers (shown by the dashed line in Figure 2 whose slope is slightly negative) the precise effect could vary per word. Some words may even show a completely opposite pattern, with older speakers having pronunciations closer to standard Catalan. These (by-word) random slopes, in combination with the random intercepts, allow the regression formula to be adapted for every individual word (or other random-effect factor). For example, the solid lines in Figure 2 show that the effect (i.e. slope) of the year of birth of the speaker for the word *meves* is slightly more negative than the general pattern (i.e. younger speakers use a pronunciation closer to standard Catalan), while the effect for the word *ell* shows the opposite pattern with a positive slope. For the word *ell*, younger speakers have adopted a slightly different pronunciation ([éj]) than the one used in standard Catalan and by older speakers ([éʎ]), as the sound [ʎ] is disappearing from most young phonetic inventories.

In order to prevent type-I errors, it is important to consider both random intercepts as well as random slopes (Jaeger, 2008; Baayen et al., 2008; Tagliamonte and Baayen, 2012; Barr et al., 2013; Bates et al., 2015). A more detailed introduction about mixed models applied to language data is given by Baayen (2008, Ch. 7) and Baayen et al. (2008). While Barr et al. (2013) advocate an approach where the random-effects structure is maximally complex, we do not favor this approach given the large size of our dataset. Furthermore, Bates et al. (2015) show that the approach of Barr et al. (2013) may result in overfitting and convergence errors. Consequently, we will only fit the random-effects structure supported by the data.

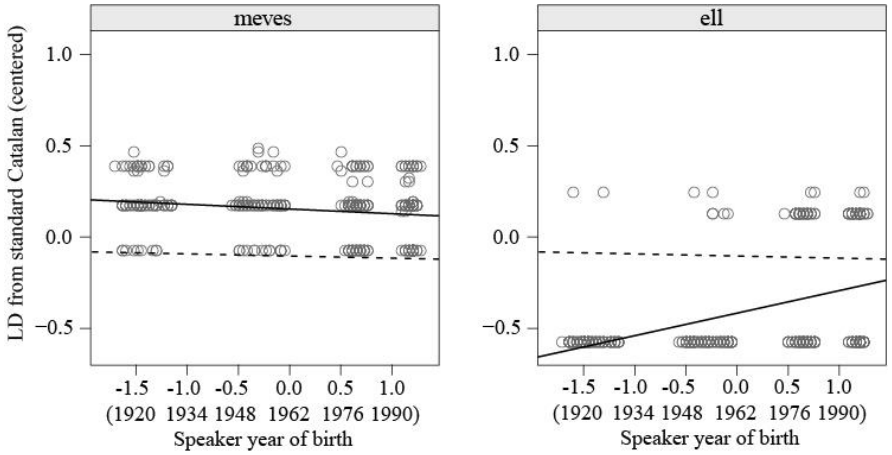

Fig. 2. Example of random slopes and intercepts for the standardized year of birth of the speaker per word. For ease of interpretation, the actual year of birth values have been added below the standardized values. The dashed line indicates the general model estimate (the intercept and the coefficient for speaker year of birth) for all words, while the solid lines indicate the estimates of the intercept and the slope for the two words (i.e. the total effect: fixed-effect intercept and slope plus random intercept and slope). The circles represent the distances for individual variants of the words *meves* (left) and *ell* (right). The dependent variable was centered, so an LD of 0 indicates the mean distance from standard Catalan.

3.3 Generalized additive mixed-effects regression modeling

The difference between a generalized additive model (GAM; Hastie and Tibshirani, 1986) and the generalized linear regression model explained earlier is that the former allows the explicit inclusion of non-linear relationships via so-called smooths. While non-linearities can be included in a generalized linear regression model, in that case the specific form (e.g., a parabola) needs to be specified in advance. A generalized additive mixed-effects regression model does not require a predefined form, but rather determines the shape of the relationship (i.e. modeled by so-called smooths) itself. Furthermore, a smooth can contain multiple numerical variables and thus represent a (potential) non-linear surface. Importantly, if a pattern is linear rather than non-linear, the GAM smooth will reflect this as well. Consequently, it is more flexible than (generalized) linear mixed-effects regression.

There are several choices to make regarding the smooths. First of all, the researcher has to choose the basis functions for each smooth. For example, smooths may consist of a series of cubic polynomials (i.e. a cubic regression spline). Another type of basis function is the thin plate regression spline, which is a combination of several simpler functions (such as a linear function, a quadratic function, a logarithmic function, etc.). Furthermore, a limit needs to be specified for the com-

plexity of each smooth. For a cubic regression spline, this limit is specified as the number of knots, which are the points at which the cubic polynomials are connected. The higher this number, the more cubic polynomials may be used to model the smooth. For the thin plate regression spline, which is the basis function we use (as it is the best approximation of the optimal fit; Wood, 2006), the complexity is limited by the number of simpler functions used to model the smooth. The actual complexity of the smooth is indicated by estimated degrees of freedom (edf). If the edf value is equal to 1, the smooth models a linear pattern, whereas an edf value higher than 1 indicates a non-linear pattern. Importantly, visualization is essential to investigate the specific shape of the smooth.

Crucially, overfitting is prevented internally by using cross-validation. Furthermore, The GAM implementation we use (i.e. the *mgcv* R package, version 1.8.8; Wood, 2003, 2011) allows that random intercepts and slopes are included as well. In this generalized additive modeling framework, random intercepts and slopes are represented by smooths with an associated *p*-value, indicating if their inclusion is necessary or not. Consequently, model comparison is not required to assess if random intercepts and slopes are necessary to include.

An important focus of dialectometry is the relationship between dialect distance and geographic location (e.g., see Nerbonne, 2010). While it has become standard practice to analyze the influence of geography on language variation by using geographic distance as an independent variable (Nerbonne & Heeringa, 2007), this approach necessarily assumes that locations having the same distance from some reference point are relatively similar (irrespective of their absolute position). This is obviously not very flexible, and does not allow for distinct, irregularly shaped dialect areas (as the effect of distance is assumed to be the same in every direction). Instead of using distance, we fit a more flexible two-dimensional non-linear surface to the dialect data, with as geographical predictors the longitude and latitude of the locations for which dialect data is available. In this way, geography is modeled by a two-dimensional surface, rather than a set of distances. Of course, the random-effect factor location (i.e. the random intercept for location) would also be able to model the effect of geography (if the geographical smooth were absent). However, such an approach would not take advantage of the fact that people living in nearby locations generally have a more similar pronunciation than those living far apart.

Instead of using a generalized linear mixed-effects regression model, we therefore use a generalized additive mixed-effects regression model where geography is modeled by a non-linear interaction (represented by a two-dimensional thin plate regression spline) of longitude and latitude. (Note that location is included as a random-effect factor as well, to capture location-based effects not present in the non-linear interaction of longitude and latitude.) A similar approach was taken by Wieling et al., 2011 to model the effect of geography on Dutch dialect distances (compared to standard Dutch).

Figure 3 shows the resulting surface for the complete area under study using a contour plot (note that the effects of social and lexical variables are also taken into

account in the model from which this surface is extracted; see Section 4). The (red) contour lines represent distance isoglosses connecting areas which have a similar pronunciation distance from standard Catalan. Wherever the contour lines are not regular circles, the treatment of geography is more sophisticated than in models which examined linguistic variation as a function of geographic distance alone (Nerbonne and Heeringa, 2007, inter alia). A green color indicates the use of pronunciations closest to the standard language, while yellow, orange, pink and light gray indicate increasingly greater pronunciation distances (on average, considering all words) from standard Catalan, respectively. The measurement points are identified by a single character corresponding to the region (A: Aragon, C: Catalonia, D: Andorra). We can clearly identify the separation between the dialects spoken in the east of Catalonia compared to the Aragonese varieties in the west. The local cohesion in Figure 3 is sensible, as nearby communities tend to speak dialectal varieties which are relatively similar.

The complexity of the surface shown in Figure 3 is reflected by the estimated degrees of freedom of the spline, in this case 12. The thin plate regression spline was highly significant as the 12.0 estimated degrees of freedom invested in it were supported by an F-value of 17 ($p < 0.0001$). This indicates that the non-linear surface is clearly warranted.

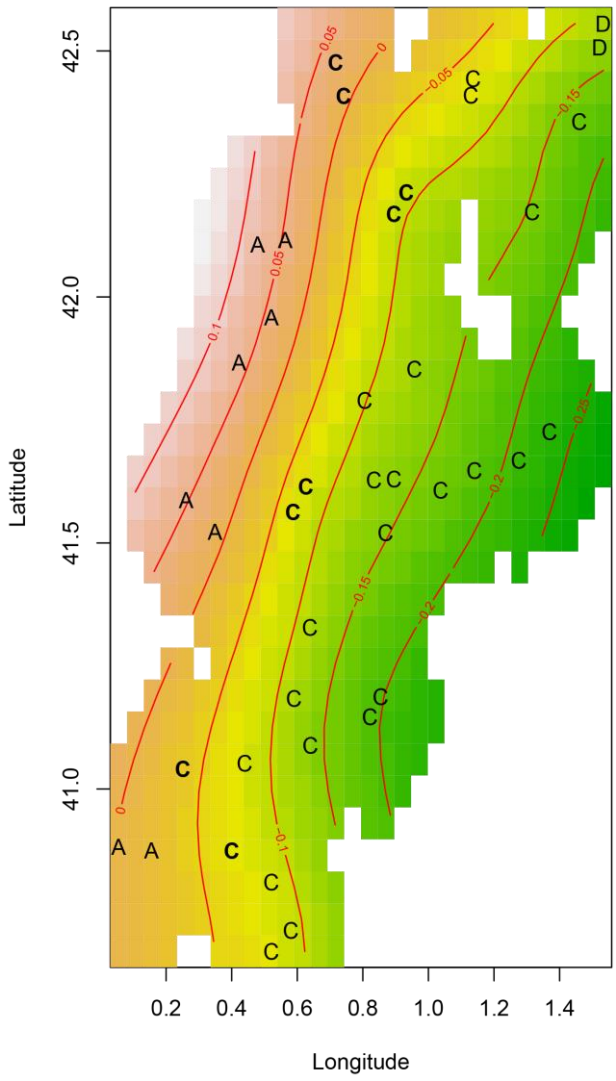

Fig. 3. Contour plot for the regression surface of pronunciation distance as a function of longitude and latitude obtained with a generalized additive model using a thin plate regression spline. The (red) contour lines represent (log-transformed Levenshtein) distance isoglosses, a green color (lower values, negative in the east) indicate smaller distances from the standard language, while a yellow, orange, pink and light gray color (i.e. increasingly higher values) represents greater distances. The characters indicate the region of the measurement points (A: Aragon, C: Catalonia, D: Andorra). The C characters in boldface indicate eight sites in Catalonia, later compared to the eight sites in Aragon, discussed in Section 4.1.

Social and lexical variables

In addition to the random-effect factors for word, speaker and location, and the smooth combining longitude and latitude representing geography, we considered several other predictors. Based on our initial analyses which showed that the pronunciations of articles, clitic pronouns and demonstrative adjectives (i.e. words such as 'this' and 'that') differed significantly more from the corresponding standard Catalan pronunciations than the other word categories, we included a factor to distinguish these two word groups (i.e. articles, clitic pronouns and demonstrative adjectives versus verbs, neuter and personal pronouns, possessive adjectives and locative adverbs). Other word-specific variables we included were the length of the word (i.e. the number of sound segments in the standard Catalan pronunciation) and the relative frequency of vowels in the standard Catalan pronunciation of each word. In addition, we included several location-specific social variables: community size, the average community age, the average community income and the relative number of tourist beds (as a proxy for the amount of tourism). The speaker-related variables we took into account were the year of birth, the gender, and the education level of the speaker. Finally, we used a factor to distinguish speakers from Catalonia and Andorra as opposed to Aragon.

Collinearity of predictors (i.e. predictors which are highly correlated with each other) is a general problem in large-scale regression studies. In our data set, communities with a larger population tend to have a higher average income and lower average age (all $|r|$'s > 0.65). Furthermore, the articles, clitic pronouns and demonstrative adjectives were much shorter than the other words, and thus the word category factor distinguishing these types from the other words is strongly related to word length ($|r| = 0.77$) . While the residualization of predictors which are highly correlated has been a popular approach, Wurm and Fisicaro (2014) convincingly argued that it is not a useful remedy for collinearity. Consequently, we only included the strongest predictor from each of the two groups of related predictors.

A few numerical predictors (i.e. community size and the relative number of tourist beds) were log-transformed (i.e. instead of the original value, the logarithm of that value was used) in order to reduce the potentially harmful effect of outliers. To facilitate the interpretation of the fitted parameters of our model, we scaled all numerical predictors by subtracting the mean and dividing by the standard deviation. As indicated above, we log-transformed and centered our dependent variable (i.e. the pronunciation distance per word from standard Catalan, averaged by dividing by the alignment length). Consequently, the value 0 represents the mean log-distance, negative values a smaller distance, and positive values a larger distance from the standard Catalan pronunciation. The significance of the fixed-effect factors, covariates, and smooths was extracted from the GAM model summary.

4 Results[4]

As not all words in our data set are pronounced by every speaker, the total number of cases (i.e. word-speaker combinations) in this study is 112,608.

We fitted a generalized additive mixed-effects regression model, step by step removing predictors that did not contribute significantly to the model. Predictors which correlated highly (indicated above) were not included at the same time (i.e. population average age, population average income and population size; and word length and word category), but only the strongest predictor was included for each of the two sets of predictors (if significant). With respect to the random effects, we assessed the significance of all possible random slopes and intercepts for the random-effect factors *location*, *speaker* and *word*. We only retained random intercepts and slopes when they were associated with a significant *p*-value (< 0.05) in the model summary. We will discuss the specification of the model including all significant predictors and random effects. The model explained 73.5% of the variation in pronunciation distances from standard Catalan. This value also incorporates the variability linked to the random-effect factors. This indicates that the model is highly capable of predicting the individual distances (for specific speaker and word combinations), providing support for our approach of integrating geographical, social and lexical variables. The main contributor (62.8%) for this good fit was the variability associated with the words (i.e. the random intercepts for word). Without random-effect factors, the fixed-effect factors explained 16% of the variation. To compare the relative influence of each of these (fixed-effect) predictors, we included a measure of effect size by specifying the increase or decrease of the dependent variable when the predictor increased from its minimum to its maximum value. The effect size of the geographical smooth was calculated by subtracting the minimum from the maximum fitted value (see Figure 3). Of course, the estimates of the standardized predictors may also be used as a measure of effect size, but there is no such estimate for the effect of geography, and not all numerical predictors are normally distributed. On the basis of our measure of effect size, we clearly observe that geography and the word-related predictors have the greatest influence on the pronunciation distance from standard Catalan.

The coefficients and the associated statistics of the fixed-effect factors and covariates included in the final model are shown in Table 1. The random-effect factors included are shown in Table 2. The fact that a random intercept for location was necessary indicates that there is variability associated with the locations which is not captured by the geographical smooth. As an example of the random-effect structure, Figure 4 shows the by-word random intercepts. In general, the words *cantaríeu*, *jo* and *nosaltres* are more likely to be similar to the standard Catalan pronunciations than *sentiríeu*, *canta* and *el (faran).*

[4] The paper package associated with this paper and available at the Mind Research Repository contains all data, methods and results for reproducibility. It can be found at: http://openscience.uni-leipzig.de/index.php/mr2/article/view/46.

4.1 Demographic predictors

None of the location-based predictors (i.e. the relative number of tourist beds, community size, average community income and average community age was significant as a main effect in our general model (see Table 1). All location-based predictors, however, showed significant word-related variation (see Table 2). For example, while there is no main effect of average community income, the pronunciation of some words will be closer to the standard in richer communities, while for some other words this pattern will be reversed.

The non-linear interaction of longitude and latitude (see Figure 3) shows that the Aragonese varieties have a higher distance from standard Catalan than the other varieties. In fact, if the non-linear interaction is replaced by a contrast between the Aragonese varieties versus the other varieties (also including location as a random-effect factor), the contrast is highly significant, $p < 0.0001$, and indicates that the Aragonese speakers have a larger pronunciation distance from standard Catalan than the other speakers. The same result is found when the dataset is restricted to the eight Aragonese sites and a subset of eight Catalan sites located close to the border (indicated by boldface C's in Figure 3).

With respect to the speaker-related predictors, only year of birth for Catalonia and Andorra was a significant predictor, indicating that younger speakers in those two regions use pronunciations which are more similar to standard Catalan than older speakers. The effect of year of birth was not significant for Aragon, and significantly different from the effect in Catalonia and Andorra ($p = 0.02$). This result confirms the existence of a clear border effect between Aragon on the one hand, and Catalonia and Andorra on the other. We interpret this difference as the effect of the Catalan language becoming official again in the 1980s in Catalonia.

We did not find an effect of gender despite this being reported in the literature frequently (see Cheshire, 2002 for an overview). Similarly, Wieling et al. (2011) also did not find a gender effect with respect to the pronunciation distance from the standard language (Dutch) in their study. We also did not find gender differences when investigating individual linguistic variables (see Section 4.3, below).

We did not find support for the inclusion of education level as a fixed-effect predictor in our model. The education measure alone (without any other social status measures) might have too little power to discover social class effects (Labov, 2001: Ch. 5; but see Gorman, 2010 for a new analysis of Labov's data suggesting that education does have sufficient power). Furthermore, when investigating individual linguistic variables (see Section 4.3), education only appeared once as a significant predictor.

Table 1. Fixed-effect factors and covariates of the final model. Negative estimates indicate more standard-like pronunciations (for increasing values of the predictors), and positive estimates less standard ones. Effect size indicates the increase or decrease of the dependent variable when the predictor value increases from its minimum to its maximum value (i.e. the complete range). The geographical smooth (Figure 3; 12 estimated degrees of freedom) is represented by the final row. Its effect size equals the minimum value subtracted from the maximum value of the fitted smooth.

	Estimate	Std. error	*p*-value	Effect size
Intercept	-0.033	0.018	0.061	
Vowel ratio per word	0.109	0.014	< 0.001	0.674
Word category is A/D/C	0.101	0.034	0.003	0.101
Speaker year of birth (Aragon)	0.005	0.004	0.282	0.014
Speaker year of birth (Catalonia and Andorra)	-0.012	0.005	0.028	-0.034
s(longitude,latitude) [12.0 edf]			< 0.001	0.310

Table 2. Significant random-effect parameters of the final model.

Factors	Random effects	Std. dev.	*p*-value
Word	Intercept	0.258	< 0.0001
	Relative nr. of tourist beds	0.025	< 0.0001
	Average community age	0.031	< 0.0001
	Community size (log)	0.020	< 0.0001
	Average community income	0.032	< 0.0001
	Speaker education level	0.009	< 0.0001
	Speaker year of birth (Cat. + And.)	0.029	< 0.0001
	Speaker year of birth (Aragon)	0.019	< 0.0001
Speaker	Intercept	0.025	0.0004
	Vowel ratio per word	0.009	< 0.0001
	Word category is A/D/C	0.018	< 0.0001
	Word length	0.013	< 0.0001
Location	Intercept	0.026	< 0.0001
	Speaker year of birth (Cat. + And.)	0.021	< 0.0001
	Vowel ratio per word	0.015	< 0.0001
	Word category is A/D/C	0.071	< 0.0001
	Word length	0.037	< 0.0001

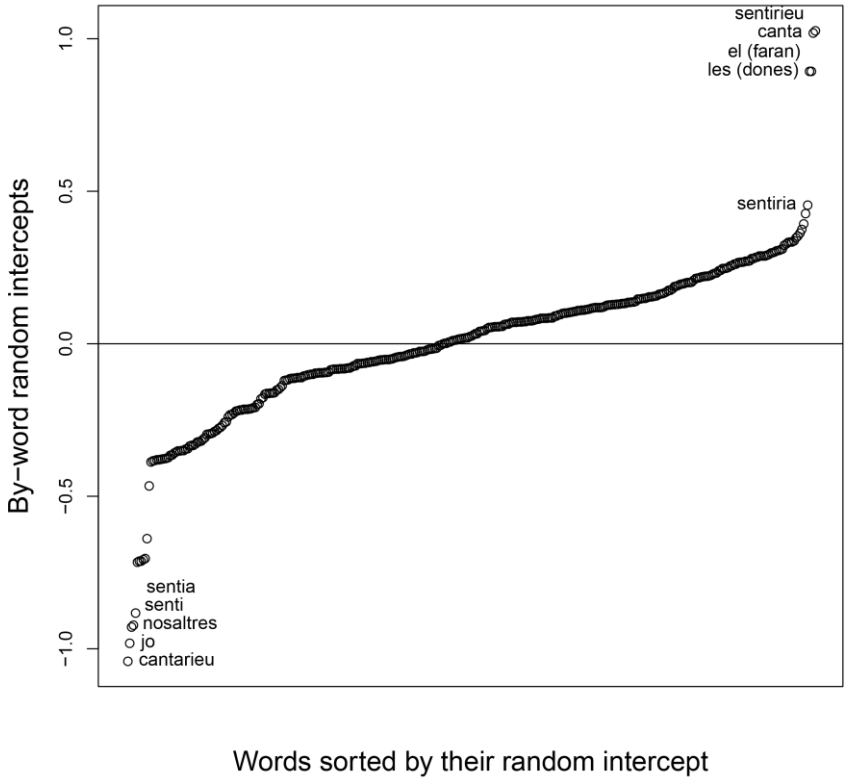

Fig. 4. By-word random intercepts. The words are sorted by the value of their intercept. Negative values (bottom-left) are associated with words which are generally (across all varieties) more similar to the standard, while positive values (top-right) are associated with words which are generally more different from the standard language. The dashed line shows the population intercept (see Table 1).

4.2 Predictors specific to lexical identity

Two variables specific to lexical identity we tested appeared to be significant predictors of the pronunciation distance from standard Catalan. It is not surprising that the binary predictor distinguishing articles, clitic pronouns and demonstratives from the other word types was highly significant, since we grouped these word categories on the basis of their higher distance from the standard language (according to our initial analyses). Articles and clitic pronouns are relatively short (in many cases only having a length of one or two sounds), and when they are different from the standard, their relative distance will be very high. While the demonstratives are not as short, they tend to be either completely identical to the standard pronunciation, or almost completely different from the standard pronunciation,

which might explain their larger distances. As word length correlated highly ($|r|$ = 0.77) with the binary group distinction, we only included the better predictor of the two. Given that word length was not significant, we included the binary group distinction between articles, clitic pronouns and demonstratives versus the other word types.

Finally, the number of vowels compared to the total number of sounds in the reference pronunciation was a highly significant predictor. This is not surprising (and similar to the result reported by Wieling et al., 2011 for Dutch) as vowels are much more variable than consonants (e.g., Keating et al., 1994). Similarly to word length, including this predictor allows us to more reliably assess the effect of the more interesting predictors.

With respect to the random effects, all lexical variables showed significant variation in their strength for individual speakers and locations. This reflects that, for example, some speakers will pronounce words with a large number of vowels closer to the standard Catalan pronunciation than others.

4.3 Comparison to individual linguistic variables

This paper proceeds from an aggregate, dialectometric perspective and applies a novel statistical technique, generalized additive mixed-effects regression modeling to a large collection of Catalan dialect variation data with the goal of understanding the (quite effective) standardization policies now in place in Catalonia and Andorra. The advantage of the aggregate perspective is its bird's eye view of language variation, which, in this case has meant a view encompassing over 100,000 pronunciations, 357 words (though note the lack of nouns, and the limited number of distinct verbs) as pronounced by eight speakers in each of the 40 different northwestern Catalan varieties. The aggregate perspective clearly runs the risk of losing sight of important details of language variation, but we have shown that mixed-effects regression modeling, in which words are individually modeled, can effectively detect very different levels of influence among individual words, thus protecting us against the risk of missing details, at least to some extent.

Standard sociolinguistic practice is rather different. With the goal of identifying individual phonemic changes in progress, and in particular, their social motivation, sociolinguists ignore aggregate tendencies in favor of detailed studies on the influence of social and structural factors on linguistic variation (Chambers, 2009). This low-level focus has certainly proven effective in understanding individual sound changes and in isolating the social dynamics that may underlie them, but it clearly runs the risk of selectively focusing on non-representative material and myopically losing sight of global tendencies.

With respect to the present study on the effects of a policy of language standardization, we might expect there to be global effects, and, in fact, this is just what we have shown. Age was shown to be significant, where the young, who have mandatorily been exposed to standard Catalan in school (and via public media), speak varieties of Catalan that are more standard like. Might we have reached sim-

ilar conclusions by examining individual linguistic variables? After all, individual phoneme effects will also be reflected directly in pronunciation distances.

To answer this question, we have examined three different linguistic variables reported in the literature, to see if the effect observed at the aggregate level also could be found when focusing on a lower level. In each case we examine examples of the variables in our own data, taking care that only examples in the relevant phonetic contexts are used. Naturally we study each of them on the basis of the pronunciations of the eight speakers per site at the 40 sites described above.

The first linguistic variable (V1) we investigated was the replacement of [ʎ] (standard) by [j] (non-standard). This change has been reported by Recasens (1996, p. 324) and is caused by the influence of the Spanish language, from which [ʎ] has almost completely disappeared. The following 10 words present in our data set were used to examine this phenomenon: *aquell*, *aquella*, *aquells*, *aquelles*, *ell*, *ella*, *ells*, *elles*, *allò*, and *allí*.

The second linguistic variable (V2) is the variation in the final morphemes for the present subjunctive. The standard uses [i] as its subjunctive theme vowel, while other vowels indicate a non-standard pronunciation. This difference is described by Massanell (2001). We examined this variable by focusing on the following 20 items: *canti* (1[-PLU]), *cantis*, *canti* (3[-PLU]), *cantin*, *perdi* (1[-PLU), *perdis*, *perdi* (3[-PLU]), *perdin*, *begui* (1[-PLU]), *beguis*, *begui* (3[-PLU]), *beguin*, *senti* (1[-PLU]), *sentis*, *senti* (3[-PLU]), *sentin*, *serveixi* (1[-PLU]), *serveixis*, *serveixi* (3[-PLU]), and *serveixin*.

The final linguistic variable (V3) is the use of [β] as opposed to another consonant (mainly [w]) within the feminine possessive adjectives. The progressive substitution of [w] for the standard [β] in the Tremp area is discussed by Romero (2001). To investigate this pattern, we investigated the following six items: *meva*, *meves*, *teva*, *teves*, *seva*, and *seves*.

Table 3 shows the significance of the social variables (gender, education level and age – the latter separated for the two areas) in addition to the influence of geography (visualized in Figure 5). The estimates were obtained by creating three separate generalized additive mixed-effects logistic regression models (one for each linguistic variable). This approach is similar to the approach outlined in Section 3, except that we now use logistic regression, since in each of the three models, the dependent variable has only two values: 1 (the variant of a speaker differs from the standard language) and 0 (the variant of a speaker is equal to the standard language). In logistic regression the estimates need to be interpreted with respect to the logit scale (i.e. the log of the odds of observing a non-standard as opposed to a standard Catalan form). A positive estimate therefore indicates that an increase in the predictor results in a higher likelihood of using a non-standard variant, while a negative estimate indicates the opposite (thus the signs of the estimates can be compared to those in Table 1). This logistic regression approach corresponds with standard sociolinguistic practice (Labov 2001).

The geographical pattern (visualized in Figure 5) varies for each variable, but in general shows that the Aragonese varieties (in the west) are more likely to have a

non-standard variant than the varieties in Catalonia and Andorra. Again, excluding the geographical smooth and replacing it by a binary predictor distinguishing Aragon from the other regions, reveals that the Aragonese speakers are significantly more likely to use a non-standard form than the speakers from Catalonia or Andorra. The same holds when focusing on the eight Aragonese sites compared to the eight sites in Catalonia close to the border with Aragon.

With respect to the social variables, both V2 and V3 show a pattern consistent with the result presented in Table 1 (i.e. younger speakers are more likely to conform to the standard in Catalonia and Andorra, but not in Aragon). V1 shows that younger speakers in Catalonia and Andorra are more likely to differ from the standard language than the older speakers (caused by the move towards Spanish, as mentioned earlier), but that this effect is even stronger in Aragon (where the influence of standard Spanish is stronger). Only V2 showed a significant influence of the education level of the speaker (with more highly educated people being more likely to use the standard variant). In summary, the aggregate result with respect to year of birth is supported by two of the three individual variables.[5]

Of course, the aggregate result is not always reflected by the behavior of individual variables, and there are two reasons for this. First, the aggregate analysis shows the general pattern when taking into account the complete set of words, and it is unlikely that all individual linguistic variables exhibit this exact same pattern. The second reason is that the aggregate analysis involves pronunciation distances, which also include pronunciation differences that are outside of the focus of the specifically selected linguistic variables.

By way of illustration that individual words do not all have to adhere to the aggregate pattern, Figure 6 shows the by-word random slopes for the speaker's year of birth for Aragon (*x*-axis) and Catalonia and Andorra (*y*-axis). Consequently, words (i.e. dots) to the right of the *y*-axis (the vertical dashed line indicates the non-significant positive effect of speaker's year of birth for Aragon; see Table 1) and below the *x*-axis (the horizontal dashed line indicates the negative effect of speaker's year of birth for Catalonia and Andorra; see Table 1) roughly adhere to the general pattern. For words in that area, younger speakers (i.e. having a higher year of birth) in Catalonia and Andorra have a pronunciation closer to standard Catalan than older speakers, while the effect is opposite (but non-significant) in Aragon. Whereas many words follow the aggregate pattern, some words even show opposite patterns, such as *perdi3*, 'waste' (3[-PLU]). These words differ *more* from the standard for younger speakers in Catalonia and Andorra as opposed

[5] While the precise effect of speaker's year of birth is different for both regions (Aragon, and Catalonia and Andorra) across all three variables, the difference in the effect of this predictor on Aragon as opposed to Catalonia and Andorra was never significant (all p's > 0.07) due to the small number of locations in Aragon (i.e. eight) and the limited number of words. Therefore, strictly speaking, none of the variables completely adheres to the aggregate pattern (where this difference was significant).

to older speakers, and differ *less* from the standard for younger people as opposed to older people in Aragon. Consequently, a linguistic variable consisting of such words would show a completely different pattern (such as V1, illustrated earlier). The aggregate approach, however, is necessary to draw more general conclusions.

5 Discussion and conclusions

In this study we have used a generalized additive mixed-effects regression model to provide support for the existence of a border effect between Aragon (where the Catalan language does not have an official status) and Catalonia and Andorra (where Catalan is an official language). Our analysis clearly indicated a greater distance from standard Catalan for speakers in Aragon as opposed to those in Catalonia and Andorra. Furthermore, our analysis identified a significant effect of speaker age (with younger speakers having pronunciations closer to standard Catalan) for Catalonia and Andorra, but not for Aragon. This provides strong evidence for the existence of a border effect in these regions caused by different language policies and is in line with the results of Valls et al. (2013). Also, our analysis revealed the importance of several word-related factors in predicting the pronunciation distance from standard Catalan and confirms the utility of using generalized additive mixed-effects regression modeling to analyze dialect distances, with respect to traditional dialectometric analyses.

Table 3. Significance of social predictors (rows) for each of the three models corresponding each to a single linguistic variable (columns). Only if an estimate was significantly different from zero (or close to significance) is its estimate printed. A positive estimate indicates a greater likelihood of having a non-standard variant for increasing values of the predictor, while a negative estimate indicates the opposite. In all cases, geography shows a significant non-linear pattern (visualized in Figure 5) as the edf values are greater than 1. Note that the estimates for the year of birth do not differ significantly for the two regions. Significance: *p < 0.05; **p < 0.001.

	V1: [ʎ] vs. [j]	V2: [i] vs. other vowel	V3: [β] vs. other consonant
Speaker is male	1.1 (p = 0.08)	n.s.	n.s.
Speaker education level	n.s.	-0.4*	-0.4 (p = 0.1)
Speaker year of birth (Catalonia and Andorra)	3.1**	-1.0**	-1.4**
Speaker year of birth (Aragon)	6.4**	n.s.	n.s.
Geography	[9.4 edf]**	[20.5 edf]**	[3.8 edf]**

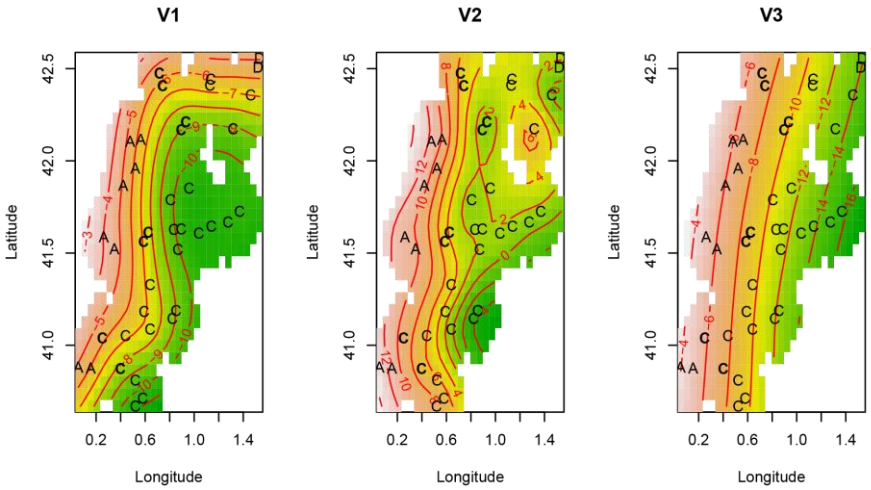

Fig. 5. Contour plot for the regression surfaces for each of three linguistic variables as a function of longitude and latitude obtained with a generalized additive model using a thin plate regression spline. The (red) contour lines represent isoglosses reflecting the probability (in terms of logits) of using a non-standard Catalan form, a green color (lower values in the east) indicates a smaller likelihood of using a non-standard variant, while a yellow, orange, pink and light gray color (i.e. increasingly higher values) represent a greater likelihood of using a non-standard variant. The characters indicate the region of the measurement points (A: Aragon, C: Catalonia, D: Andorra). The C characters in boldface indicate eight sites in Catalonia, later compared to the eight sites in Aragon.

Methodologically, we have attempted on the one hand to include candidate social variables as well as geography in a single aggregate (dialectometric) analysis. We wished to include both sorts of variables in an effort to meet objections such as Woolhiser's (2005) that dialectometry systematically ignores social variables. However, note that our analysis retains the aggregate perspective of dialectometry, despite the limitations caused by the data set (i.e. no nouns and only five distinct verbs). On the other hand, we have also included structural, linguistic factors in the analysis, such as the varying degree to which different words are influenced by geographic and social factors, as well as (e.g.,) the relative number of vowels in a word. Of course these linguistic techniques may seem insensitive when compared to studies in other variationist traditions (i.e. where individual sound changes are investigated), but they enable analyses to be more comprehensive, i.e. based on large amounts of data including many variables, and it has also been our point here to introduce the methodology.

With regard to the comparison to single-variable analyses, standard in sociolinguistics, we presented additional analyses at the level of three individual linguistic variables that have been discussed in the literature, and we showed that two of the three variables supported the general pattern. These analyses also illustrated that

an aggregate approach is needed, as individual linguistic variables may not be representative of the global pattern.

In contrast to the (exploratory visualization-based) conclusion of Valls et al. (2013) that the older speakers in urban communities use pronunciations closer to standard Catalan than the older speakers in rural communities, we did not find a significant effect of community size (nor a significant interaction between speaker age and community size). In fact when using the binary distinction Valls and colleagues based their conclusion on (i.e. distinguishing urban and rural communities in twenty different counties), the results are not at all significant (p = 0.3). This clearly illustrates the need for adequate statistical models, to prevent reaching statistically unsupported conclusions.

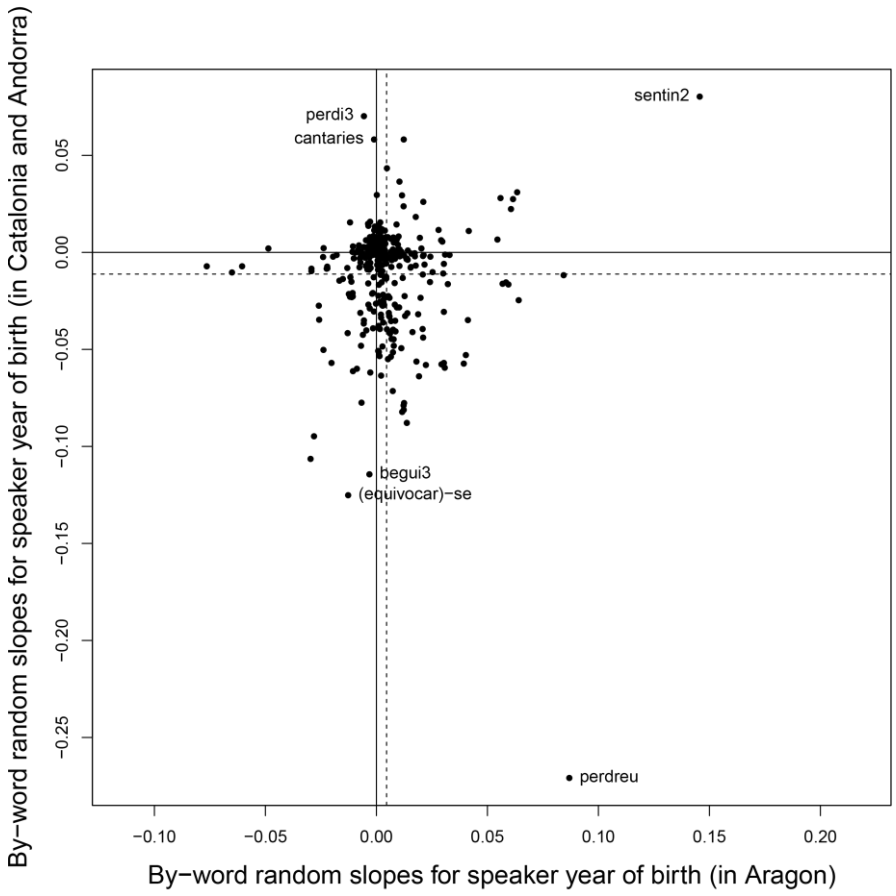

Fig. 6. By-word random slopes for the speaker's year of birth in Aragon (*x*-axis) and Catalonia and Andorra (*y*-axis). The dashed lines indicate the model estimates (see Table 1).

We did not find support for the importance of education level of the speaker. This might seem surprising given that one of the main reasons for the border effect is the official status of the Catalan language in both Catalonia and Andorra (and therefore its use in education), but not in Aragon. However, this education effect might be partly captured by year of birth, as there is a positive correlation between education level and the year of birth of the speaker ($r = 0.3$). Furthermore, the influence of mass media or the speaker's job might mask the potential standardizing effect of education on the speaker's pronunciation.

We also did not find support for the general influence of any of the demographic variables. This contrasts with the study of Wieling et al. (2011) on Dutch dialects, who found a significant effect of community size (larger communities use pronunciations closer to the standard) and average community age (older communities use pronunciations closer to the standard language). However, the number of locations in the present study was small and might have limited our power to detect these effects – in the study of Wieling et al. (2011) more than ten times as many locations were included.

It should be clear that we think that the standardization policy has led to pronunciation change. We have asked ourselves whether our reasoning commits the fallacy known as *post hoc, ergo propter hoc* – i.e. whether we might be mistaking a mere correlation between standardization policy and pronunciation change for a causal relation between the two. The temporal order is indeed as it should be, i.e. the behavioral change followed the policy change with younger people in Catalonia (where Catalan was used in schools and public media again after Franco's dictatorship) speaking a more standard-like dialect. Nonetheless, the relation might also be indirect, i.e. the policy change might have influenced attitudes which in turn influence phonetic behavior. And it is also possible that the policy change was motivated by linguistic ideology, but it would take us too far afield to explore those issues here. We admit therefore that we cannot claim to have proven that the policy change caused the pronunciation change, even if that is our interpretation.

We see three promising extensions of this study. First, replicating this study using new material (i.e. using a random set of words) would be useful to see if the results on the basis of our study (with a biased set of items) are valid in general.

Second, it would be interesting to investigate standardization towards Spanish, by comparing the dialectal pronunciations to the Spanish standard language instead of the Catalan standard language. In our data set there are clear examples of the usage of a dialectal form closer to the standard Spanish pronunciation than to the standard Catalan pronunciation, and it would be rewarding to investigate which word- and speaker-related factors are related to this.

The third extension involves focusing on the individual sound correspondences between Catalan dialect pronunciations and pronunciations in standard Catalan. These sound correspondences can easily be extracted from the alignments generated by the Levenshtein distance algorithm. When focusing on a specific set of locations (e.g., the Aragonese locations), it would be computationally feasible to create a generalized additive mixed-effects regression model to investigate which

factors determine when a sound in a certain dialectal pronunciation is different from the corresponding sound in the standard Catalan pronunciation.

Acknowledgements.

We thank the two anonymous reviewers for their extensive comments which have helped to improve this manuscript. This research was partly funded by the project *Descripción e interpretación de la variación dialectal: aspectos fonológicos y morfológicos del catalán* (FFI2010-22181-C03-02), financed by MICINN and FEDER.

References

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

Baayen, R.H., Davidson, D.J., Bates, D.M.: Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59(4), 390–412 (2008)

Bailey, G., Wikle, T., Tillery, J., Sand, L.: The apparent time construct. Language Variation and Change 3, 241–264 (1991)

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J.: Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278 (2013)

Bates, D., Kliegl, R., Vasishth, S. and Baayen, R. H: Parsimonious mixed models. http://arxiv.org/abs/1506.04967 (2015)

Bibiloni, G.: Un estàndard nacional o tres estàndards regionals? In: Joan, B. (ed.), Perspectives sociolingüístiques a les Illes Balears. Res Publica, Eivissa (2002)

Cambra de Comerç - Indústria i Serveis d'Andorra: Informe econòmic (2008)

Chambers, J.: Sociolinguistic Theory: Linguistic Variation and its Social Significance. Third ed. Wiley-Blackwell (2009)

Cheshire, J.: Sex and gender in variationist research. In: Chambers, J.K., Trudgill, P., Schilling-Estes, N. (eds.). The Handbook of Language Variation and Change, pp. 423–443. Blackwell Publishing Ltd. (2002)

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

Departament d'Estadística del Govern d'Andorra: Societat i població: http://www.estadistica.ad. Last accessed: February 28, 2011 (2010)

Fabra, P.: Gramàtica catalana. Institut d'Estudis Catalans, Barcelona (1918)

Goebl, H.: Langues standards et dialectes locaux dans la France du Sud-Est et l'Italie septentrionale sous le coup de l'effet-frontière: une approche dialectomètrique. International journal of the sociology of language 145, 181–215 (2000)

Gorman, K.: The consequences of multicollinearity among socioeconomic predictors of negative concord in Philadelphia. In: Lerner, M. (ed.), University of Pennsylvania Working Papers in Linguistics, 16(2), pp. 66–75 (2010)

Hastie, T. J., Tibshirani, R. J: Generalized additive models (Vol. 43). CRC Press (1990).
Heeringa, W.: Measuring Dialect Pronunciation Distances using Levenshtein Distance. PhD thesis, Rijksuniversiteit Groningen (2004)
Heeringa, W., Kleiweg, P., Gooskens, C., Nerbonne, J.: Evaluation of String Distance Algorithms for Dialectology. In: Nerbonne, J., Hinrichs, E. (eds.) Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, pp. 51–62 (2006)
Huguet, A., Vila, I., Llurda, E.: Minority language education in unbalanced bilingual situations: A case for the linguistic interdependence hypothesis. Journal of Psycholinguistic Research 3, 313–333 (2000)
Instituto Aragonés de Estadística: Población y Territorio. http://www.aragon.es. Last accessed: February 28, 2011 (2007, 2009, 2010)
Institut d'Estadística de Catalunya: Territori. http://www.idescat.cat. Last accessed: February 28, 2011 (2008, 2010)
Institut d'Estudis Catalans: Proposta per a un estàndard oral de la llengua catalana I. Fonètica. Barcelona: Institut d'Estudis Catalans (1999a)
Institut d'Estudis Catalans: Proposta per a un estàndard oral de la llengua catalana II. Morfologia. Barcelona: Institut d'Estudis Catalans (1999b)
Jaeger, F.: Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of memory and language 59(4), 434–446 (2008)
Keating, P., Lindblom, B., Lubker, J., Kreiman, J.: Variability in jaw height for segments in English and Swedish VCVs. Journal of Phonetics 22, 407–422 (1994)
Labov, W.: Principles of Linguistic Change, Volume 2. Social Factors. Blackwell Publishers Inc (2001)
Levenshtein, V.: Binary codes capable of correcting deletions, insertions and reversals (in Russian). Doklady Akademii Nauk SSSR 163, 845–848 (1965)
Massanell, M.: Morfologia flexiva actual de la Seu d'Urgell i Coll de Nargó: estadis en el procés d'orientalització del català nord-occidental. Zeitschrift für Katalanistik 14, 128–150 (2001)
Nerbonne, J.: Data-driven dialectology. Language and Linguistics Compass 3(1), 175–198 (2009)
Nerbonne, J.: Measuring the diffusion of linguistic change. Philosophical Transactions of the Royal Society B: Biological Sciences 365, 3821–3828 (2010)
Nerbonne, J., Heeringa, W.: Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation. In: Featherston, S., Sternefeld, W. (eds.) Roots: Linguistics in Search of its Evidential Base, pp. 267–297. Berlin, Mouton De Gruyter (2007)
Pradilla, M.-À.: Sociolingüística de la variació i llengua catalana. Institut d'Estudis Catalans, Barcelona (2008a)
Pradilla, M.-À.: La tribu valenciana. Reflexions sobre la desestructuració de la comunitat lingüística. Onada, Benicarló (2008b)
Recasens, D.: Fonètica descriptiva del català (assaig de caracterització de la pronúncia del vocalisme i consonantisme del català al segle XX). Barcelona: Institut d'Estudis Catalans (1996)
Romero, S.: Canvi lingüístic en morfologia nominal a la Conca de Tremp. Universitat de Barcelona. PhD thesis. Available at http://hdl.handle.net/10803/2082 (2001)
Tagliamonte, S., Baayen, R.H.: Models, forests and trees of York English: Was/were variation as a case study for statistical practice. Language Variation and Change 24(2), 135–178 (2012)

Valls, E., Wieling, M., Nerbonne, J.: Linguistic advergence and divergence in north-western Catalan: A dialectometric investigation of dialect leveling and border effects. LLC. The Journal of Digital Scholarship in the Humanities, 28(1) (2013)

Wieling, M.: A Quantitative Approach to Social and Geographical Dialect Variation. PhD thesis, Rijksuniversiteit Groningen (2012)

Wieling, M., Bloem, J., Mignella, K., Timmermeister, M., Nerbonne, J.: Measuring foreign accent strength in English: Validating Levenshtein distance as a measure. Language Dynamics and Change 4(2), 253-269 (2014)

Wieling, M., Margaretha, E., Nerbonne, J.: Inducing a measure of phonetic similarity from pronunciation variation. Journal of Phonetics 40(2), 307–314. (2012)

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

Wieling, M., Nerbonne, J.: Advances in dialectometry. Annual Review of Linguistics 1, 243-264 (2015)

Wieling, M., Nerbonne, J., Baayen, R.H.: Quantitative Social Dialectology: Explaining Linguistic Variation Socially and Geographically. PLoS ONE, 6(9): e23613 (2011)

Wieling, M., Prokić, J., Nerbonne, J.: Evaluating the pairwise string alignment of pronunciations. In: Borin, L, Lendvai, P. (eds.) Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education, Workshop at the 12th Meeting of the European Chapter of the Association for Computational Linguistics. Athens, 30 March 2009, pp. 26–34 (2009)

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

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

Wood, S.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B 73 (1), 3–36 (2011)

Woolard, K. and Gahng, T.-J.: Changing language policies and attitudes in autonomous Catalonia. Language in Society 19, 311–330 (2008)

Woolhiser, C.: Political borders and dialect divergence/convergence in Europe. In: Auer, P., Hinskens, F., Kerswill, P. (eds.) *Dialect Change. Convergence and divergence in European languages*, pp. 236–262. Cambridge University Press, New York (2005).

Wurm, L.H. and Fisicaro, S.A.: What residualizing predictors in regression analysis does (and what it does *not* do). *Journal of Memory and Language* 72, 37-48 (2014).