

**THE EFFECTS OF LANGUAGE POLICIES ON STANDARDIZATION OF
CATALAN DIALECTS: A SOCIOLINGUISTIC ANALYSIS USING
GENERALIZED ADDITIVE MIXED-EFFECTS REGRESSION MODELLING**

Martijn Wieling^a, Esteve Valls^b, R. Harald Baayen^c and John Nerbonne^a

^aDepartment of Humanities Computing, University of Groningen ^bDepartment of Catalan Philology,
University of Barcelona, ^cDepartment of General Linguistics, University of Tübingen and Department
of Linguistics, University of Alberta

m.b.wieling@rug.nl, e.valls@ub.edu, baayen@ualberta.edu, j.nerbonne@rug.nl

Martijn Wieling, ^aDepartment of Humanities Computing, University of Groningen, P.O. Box 716, 9700
AS Groningen, The Netherlands, +31614108622 / +31503635977

Short title: Language policies influence Catalan dialects

**THE EFFECTS OF LANGUAGE POLICIES ON STANDARDIZATION OF
CATALAN DIALECTS: A SOCIOLINGUISTIC ANALYSIS USING
GENERALIZED ADDITIVE MIXED-EFFECTS REGRESSION MODELLING**

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 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. We fitted a generalized additive mixed-effects regression model to the pronunciation distances of 357 words from standard Catalan. The results revealed higher pronunciation distances from standard Catalan in Aragon than in the other regions. Furthermore, speakers in Catalonia and Andorra, but not in Aragon, showed a clear standardization pattern, with younger speakers having dialectal pronunciations closer to the standard than older speakers. These results clearly show the effect of language policies on standardization patterns and border effects in dialects of a single language. In addition, this study illustrates the usefulness of generalized additive modeling for analyzing dialect data.

Acknowledgements

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.

1. Material

In this study we investigate a Catalan dialect data set using a generalized additive mixed-effects regression model in order to identify sociolinguistic and word-related factors which play an important role in predicting the distance between dialectal pronunciations and the Catalan standard language. We use Catalan dialect pronunciations of 320 speakers of varying age in 40 places located in three regions where 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 official status in Andorra (as the only language) and Catalonia (where both Catalan and Spanish are the official languages; Woolard and Gahng, 2008), but not in Aragon (Huguet, Vila and Llorca, 2000), we will contrast these two regions in our analysis.

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, Goebel (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 the traditional approach in dialectology, as many features are taken into account simultaneously and the measurements are very explicit. However, Woolhiser (2005) is very critical of this study, as Goebel does not discuss the included features and also does not consider the

sociolinguistic dynamics as well as the ongoing dialect change (i.e. he uses dialect atlas data).

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 both regions recognize a different variety of Catalan as the standard language (i.e. the unitary Catalan standard in Catalonia and the Valencian Catalan substandard in Valencia). In a similar vein, Bibiloni (2002: 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 (submitted) 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 Goebel'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

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). In line with social dialectology,

however, in analyzing these distances, we will also take several social factors into account.

In addition, we aim to be innovative in exploring 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 verbal inflections are often motivated both by the wish to clarify the social affinities of variation, but also by the wish to adduce linguistic structure in the variation. Wieling and Nerbonne (2011) summarize several earlier attempts to ascertain the linguistic foundations of aggregate dialectometric differences, so we shall not review those here, but the attempt here, using a generalized additive model, is notable for the opportunity it offers to quantify the relative importance of different (word-related) features. We cannot propose canonical techniques for combining the aggregating (dialectometric) and selected-features approaches, but this paper does claim to present a novel and potentially very useful manner of approaching the two sorts of questions simultaneously.

1.3. Regression models to study pronunciation variation

In this study, we use the same Catalan dialect data set as studied by Valls et al. (submitted). We measure the pronunciation distances of a large number of words (following dialectometry; see Heeringa, 2004), and we will use a generalized additive mixed-effects regression model to predict these distances *per word* for all individual

speakers (8 per location) on the basis of geography, word-related features and several sociolinguistic determinants.

In our regression analysis we will contrast the area where Catalan is an official language (Catalonia and Andorra) with the area where this is not the case (Aragon). Based on the results of Valls et al. (submitted), we expect to observe larger pronunciation distances from standard Catalan in Aragon than in the other two regions. Furthermore, we expect that the models will differ 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. (submitted), the regression analysis allows us to assess the significance of these differences. For example, while Valls et al. (submitted) state that urban communities have pronunciations more similar to standard Catalan than rural communities, this pattern might not be significant.

This is one of the first studies using generalized additive mixed-effects regression modeling in language variation research. Wieling et al. (2011) successfully used linear mixed-effects regression modeling to show that the distance from standard Dutch could be predicted based on geography and several word- and location-related factors. For example, they identified word frequency, community size and average community age as significant predictors. In their study, Wieling et al. (2011) used a basic generalized additive model to represent the non-linear effect of geography and subsequently used its fitted values as a predictor in a linear mixed-effects regression model. Due to improvements in the software available for generalized additive

modeling (we use the *mgcv* package in *R*; Wood, 2006), we were able to advance on their approach by creating a generalized additive mixed-effects regression model directly, instead of this two-step approach.

The advantage of using a mixed-effects regression approach is that it takes the random-effects structure of the data into account (i.e. the variability linked with the set of speakers, locations and words), and consequently lowers the chance of incorrectly judging a predictor as significant (Baayen, 2008: Ch. 7). A more detailed explanation of applying mixed-effects modeling in predicting pronunciation distances is given by Wieling et al. (2011).

2. Material

2.1. *Pronunciation data*

The Catalan dialect data set contains 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, 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 between 1995 and 1996. The complete data set we use contains 357 words, consisting of 16 articles, 81 clitic pronouns, 8 demonstrative pronouns, 2 neuter pronouns, 2 locative adverbs, 220 verbs (inflected forms of 5 verbs), 20 possessive pronouns and 8 personal pronouns. The original data set consisted of 363 words, but 6 words were excluded as they did not have a pronunciation in the standard Catalan language. A more detailed description of the data set is given by Valls et al. (submitted).

As Wieling et al. (2011) reported a significant effect of word frequency on dialect distances from standard Dutch, we also investigated if we could obtain Catalan word frequencies. While we were able to find a dictionary with frequency information (Rafel, 1996-1998), it did not contain contextual information which was necessary for clitics and articles (representing almost one third of our data). Consequently, we did not include frequency information.

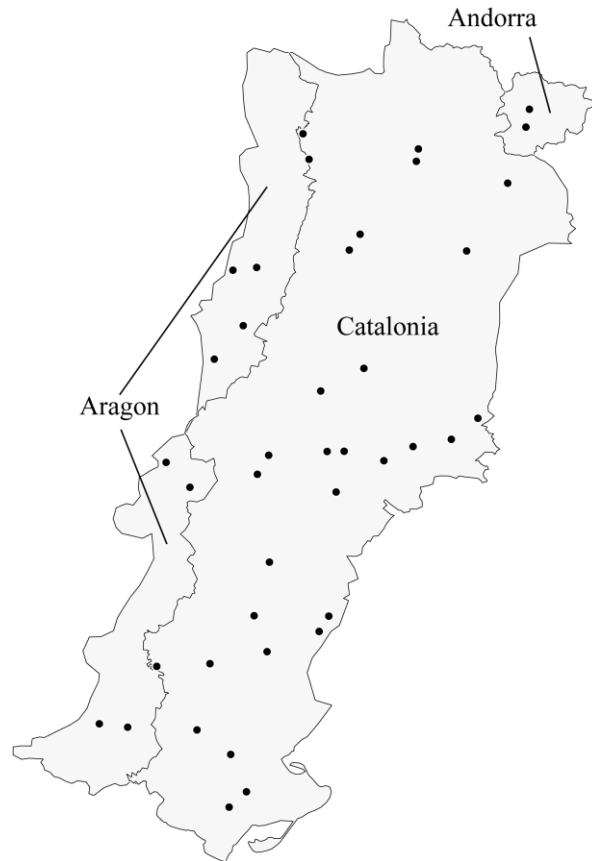


Figure 1. Geographical distribution of the locations. Two locations are found in Andorra, eight in Aragon and the remaining thirty locations are found in Catalonia.

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 as a proxy to measure 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 (in all cases $r > 0.9$, $p < 0.001$), we decided to use the most recent demographic information in this study. No less recent 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.

3. Methods

3.1. Obtaining pronunciation distances

For all 320 speakers, we calculated the pronunciation distance between standard Catalan and their dialectal pronunciations by using 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', [beyesa] and [bejyes] is 3:

beyesa	insert j	1
bejyesa	subst. ε/e	1
bejyεsa	delete a	1
bejyεs		
<hr/>		3

This sequence corresponds with the following alignment:

b	e		γ	e	s	a
b	e	j	γ	ε	s	
<hr/>						
			1	1	1	

The regular 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 such that these alignments do not occur.

It is clear that the regular Levenshtein pronunciation distances are very crude as the Levenshtein algorithm does not distinguish (e.g.,) substitutions involving similar sound segments, such as /e/ and /ε/, from more different sound segments such as versus /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 acoustically sensible (based on six independent dialect data sets; Wieling, Margaretha and Nerbonne, accepted) 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. (accepted). Similar to the study of Wieling et al. (2011), our pronunciation distances will be based on the PMI-based Levenshtein distance.

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 the alignment length.

3.2. Modeling the role of geography: generalized additive modeling

An important focus of dialectometry is the relationship between dialect distance and geographic location (e.g., see Nerbonne, 2010). A linear regression model is clearly not suitable to model the complex interaction between longitude and latitude which is characteristic of geographical dialect patterns. We therefore follow Wieling et al. (2011) and turn to a generalized additive model (GAM) instead. A GAM is more flexible than a linear regression model, as it does not restrict the functional relation between a predictor and the response variable to be linear. Instead, smoothing functions are used to estimate these functional relationships. A single smooth function may include multiple predictors simultaneously, resulting in the estimation of a complex surface. In this study, we will combine longitude and latitude in a single smooth function (i.e. a thin plate regression spline; Wood, 2003; see also Wieling et al., 2011) to model their interaction on dialect distance from standard Catalan. This

approach is similar to that of Wieling et al. (2011), but due to software improvements we were able to include all other factors and covariates, as well as the random-effects structure in our generalized additive model (i.e. as opposed to creating a second linear mixed-effects regression model including the fitted values of a simple GAM).

Figure 2 shows the resulting regression surface for the complete area under study using a contour plot. The thin plate regression spline was highly significant as the invested 23.9 estimated degrees of freedom were supported by an F -value of 24.9 ($p < 0.001$). The (solid) contour lines represent distance isoglosses connecting areas which have a similar distance from standard Catalan. Darker shades of gray indicate smaller distances, lighter shades of gray represent greater distances from the standard Catalan language. 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 2 is sensible, as nearby communities tend to speak dialectal varieties which are relatively similar (Nerbonne and Kleiweg, 2007).

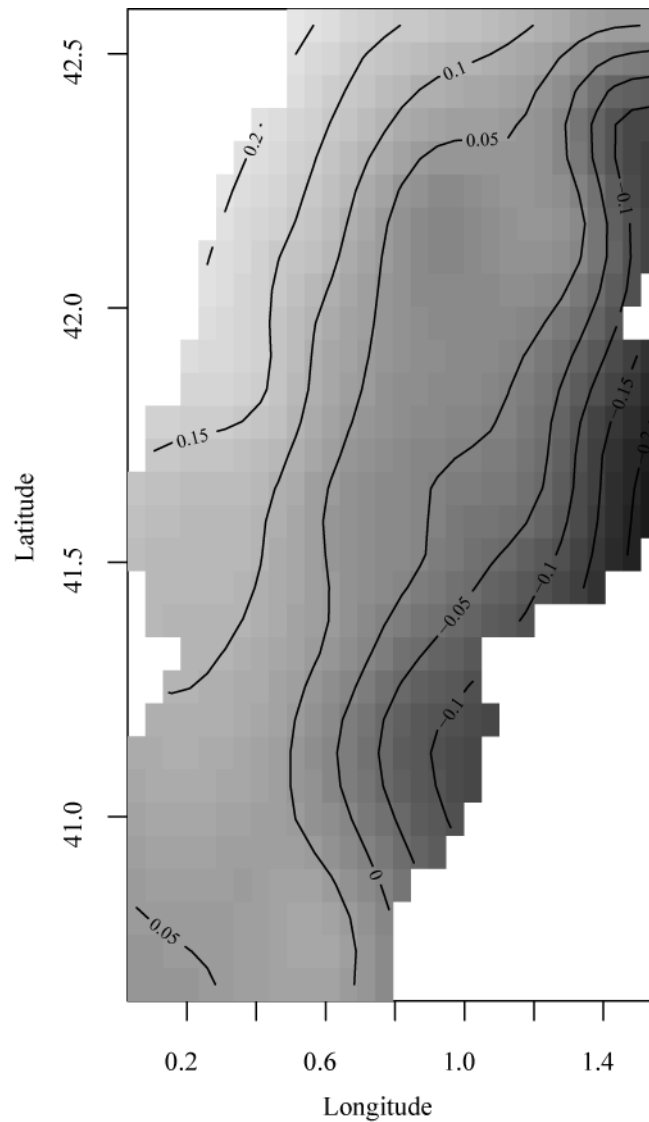


Figure 2: Contour plot for the regression surface of pronunciation distance as a function of longitude and latitude obtained with a generalized additive model using a thin plate regression spline. The (black) contour lines represent distance isoglosses, darker shades of gray (lower values, negative in the east) indicate smaller distances from the standard language, while lighter shades of gray (higher values) represent greater distances.

3.3. *Mixed-effects modeling*

Our generalized additive *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 (intercept of the) general regression formula can be adapted for every individual word to make it as precise as possible. These adjustments to the general model's intercept are called 'random intercepts'. For example, Figure 3 shows the random intercepts for four different words, *ho (són)* '(they are) ...' (plural neuter pronoun denoting anything that cannot be expressed by a noun), *meves* 'my' (feminine plural possessive), *aquest* 'this' (masculine singular demonstrative), and *ell* 'he'. Clearly the words *meves* and *aquest* have a higher linguistic distance from standard Catalan than the average word (indicated by the dashed line), while an opposite pattern can be observed for the words *ho (són)* and *ell* (i.e. these words are more similar to standard Catalan).

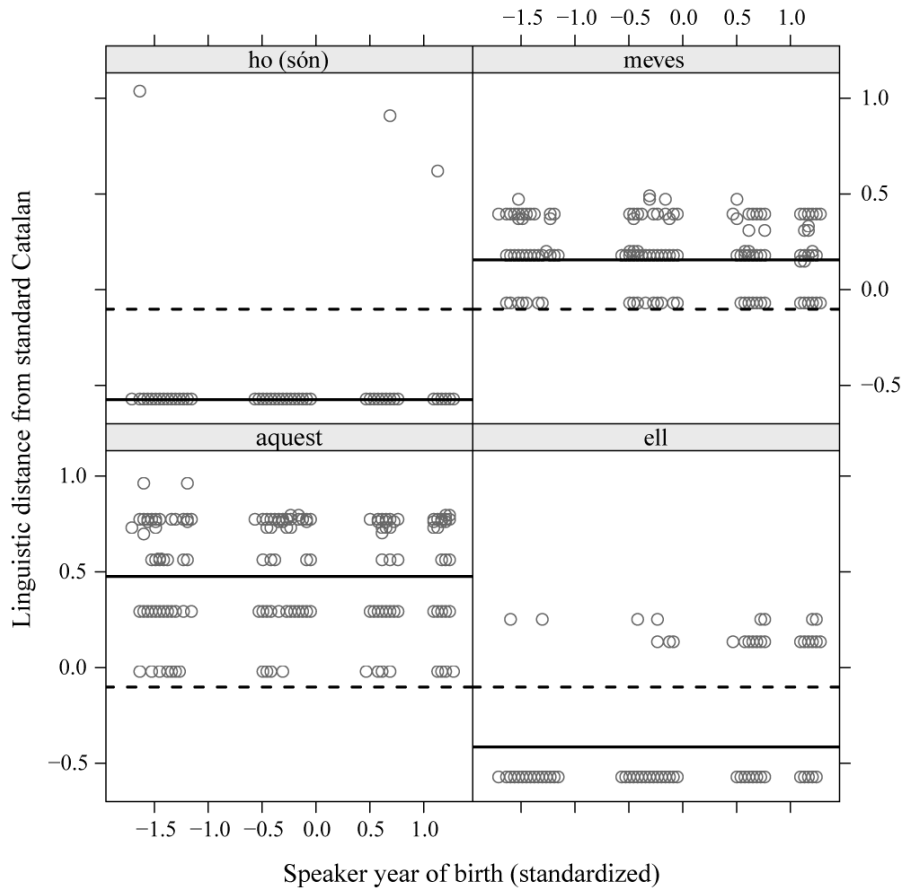


Figure 3: Example of random intercepts per word. The dashed line indicates the general model estimate, while the solid lines indicate the estimates of the intercept for each individual word.

Similarly, the effect of each predictor may also vary. For example, while in general younger speakers may have pronunciations closer to standard Catalan than older speakers, 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, make the regression formula as precise as possible for each individual word (or other random-effect factor). As an illustration, Figure 4 shows the random slopes for speaker age (combined with the random intercepts) for the same four words discussed earlier. The general effect of speaker year of birth is negative (i.e. younger speakers have a pronunciation closer to standard Catalan than older speakers) and is

indicated by the dashed line. While the words *ho* (*són*) and *aquest* do not show a clear effect of speaker year of birth, the word *meves* behaves in accordance with the general effect (i.e. slightly negative). Surprisingly, the word *ell* shows an opposite pattern, with younger speakers having a pronunciation more distant from standard Catalan than older speakers. In this case, younger speakers have adopted a slightly different pronunciation ([éj]) than used in standard Catalan and by older speakers ([éλ]), as the sound [λ] is disappearing from the phonetic inventories of most young speakers.

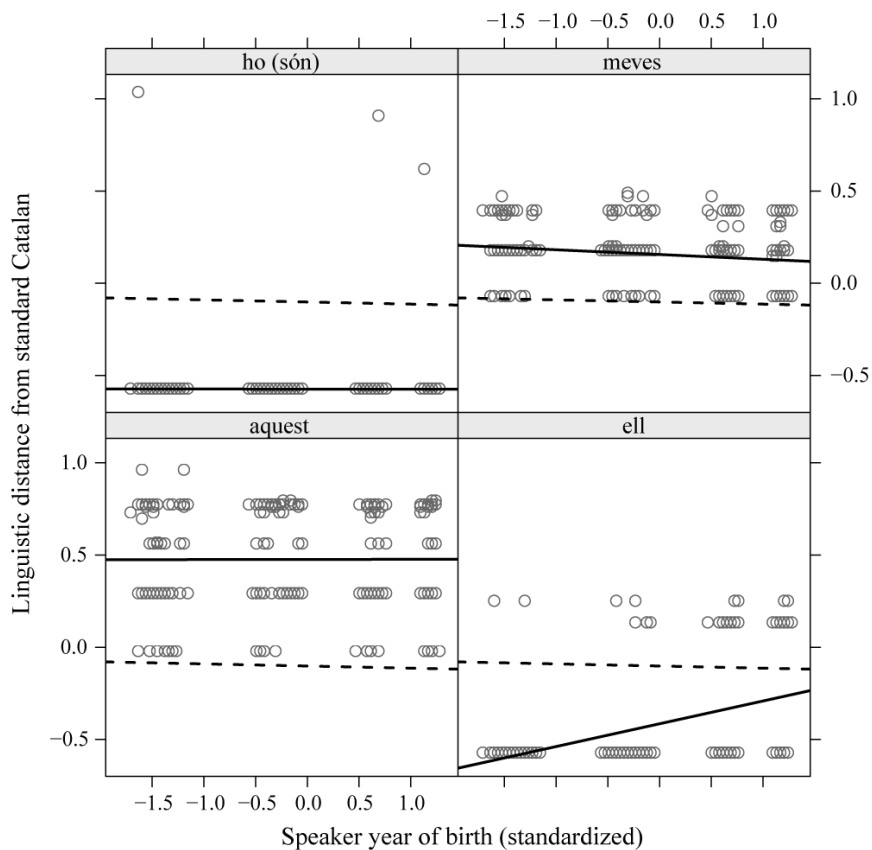


Figure 4: Example of random slopes for speaker year of birth per word. The dashed line indicates the general model estimate (the intercept and the coefficient for speaker year of birth), while the solid lines indicate the estimates of the intercept and the slope for each individual word.

By taking the systematic variability of these random-effect factors into account, the coefficients of the general model can be more precisely determined and type-I errors

are prevented. The significance of random-effect factors in the model was assessed by the Wald test. A more detailed introduction about mixed models applied to language data is given by Baayen (2008, Ch. 7) and Baayen et al. (2008).

Besides 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 articles, clitic pronouns and demonstrative pronouns had a significantly larger distance from the corresponding standard Catalan pronunciations than the other word categories, we included a factor to distinguish these two groups. Other lexical variables we included were the length of the word (i.e. the number of sounds 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 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 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 and also show a specific geographical distribution, somewhat similar to Figure 2 (e.g., the largest communities appear mainly in the east). To be able to assess the pure effect of each predictor, we took out the effect of other correlated variables, by instead using as predictor the residuals of a linear model regressing that predictor on the correlated variables (i.e. one way only, so we took out the effect of community size from average income, but not the other way around). In

this context, geography was represented by the fitted values of a GAM predicting the pronunciation distance from standard Catalan only based on longitude and latitude. As the new predictors all correlated positively with the original predictors, they can still be interpreted in the same way as the original predictors.

A few numerical predictors (i.e. community size and the relative number of tourist beds) were log-transformed 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. In addition, we log-transformed and centered our dependent variable (i.e. the pronunciation distance per word from standard Catalan, averaged by alignment length). Consequently, the value 0 represents the mean distance, negative values a smaller distance, and positive values a larger distance from the standard Catalan pronunciation. The significance of each fixed-effect factor was evaluated by means of the Wald test (reporting an *F*-value).

4. Results

For the purpose of another study, a multiple alignment of the sounds in every pronunciation was made. This multiple alignment did not reveal any transcription errors and therefore signifies the high quality of the data set. For this reason, we did not remove pronunciations with large distances from standard Catalan, as these are genuine distances instead of noise. As not all words are pronounced by every speaker, the total number of cases (i.e. word-speaker combinations) 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. We will discuss the specification of the model including all significant predictors and verified random

effects. The model explained 75% of the variation in pronunciation distances from standard Catalan. 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 (63%) 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 20% of the variation.

As our initial analyses (investigating the random intercepts of word, location and speaker) revealed that the inclusion of a random intercept for location was not warranted given its limited improvement in goodness of fit, we excluded location as a random-effect factor.

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.

4.1. Demographic predictors

Of all location-based predictors (i.e. the relative number of tourist beds, community size, average community income and average community age), only community size was close to significance ($p = 0.051$) as a main effect in our general model (see Table 1). All location-based predictors, however, showed significant word-related variation. 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.

	Estimate	Std. Error	<i>p</i> -value
Intercept	-0.10175	0.02091	< 0.001
Word length	0.13023	0.02183	< 0.001
Vowel ratio per word	0.10501	0.01372	< 0.001
Word category is A/D/C	0.30515	0.04777	< 0.001
Community size (log)	-0.00736	0.00377	0.051
Speaker year of birth	-0.01144	0.00311	< 0.001
Location is in Aragon	0.04702	0.03720	0.206
Location is in Aragon * Speaker year of birth	0.01682	0.00630	0.008
s(longitude,latitude) [23.9 edf]			< 0.001

Table 1: Fixed-effect factors and covariates of the final model. Note that community size was included as it was close to significance. The factor distinguishing locations in Aragon from those in Catalonia and Andorra was included as the interaction with year of birth of speaker was significant. The geographical smooth (Figure 2; 23.9 estimated degrees of freedom) is represented by the final row.

It might seem strange that the factor distinguishing the locations in Aragon from those in Catalonia and Andorra was not significant, but the smooth function representing geography (see Figure 2) already shows that the Aragonese varieties have a higher distance from standard Catalan than the other varieties. Note that the contour lines in Figure 2 all run roughly north-south, and the distances increase monotonically as one looks further west. In fact, when we exclude the smooth function the factor is highly significant ($p < 0.001$) and assigns higher distances from standard Catalan to the Aragonese varieties.

Factors	Random effects	Std. Dev.	<i>p</i> -value
Word	Intercept	0.25470	< 0.0001
	Relative nr. of tourist beds	0.02162	< 0.0001
	Average community age	0.01377	< 0.0001
	Community size (log)	0.01496	< 0.0001
	Average community income	0.01404	< 0.0001
	Speaker year of birth	0.02595	< 0.0001
	Speaker education level	0.01370	< 0.0001
	Location is in Aragon	0.15592	< 0.0001
	Loc. is in Aragon * Sp. year of birth	0.02448	< 0.0001
Speaker	Intercept	0.03691	< 0.0001
	Word length	0.02911	< 0.0001
	Vowel ratio per word	0.01678	< 0.0001
	Word category is A/D/C	0.05901	< 0.0001
Residual		0.17249	

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

With respect to the speaker-related predictors, only year of birth was a significant predictor indicating that younger speakers use pronunciations which are more similar to standard Catalan than older speakers. However, the significant interaction in Table 1 (i.e. Location is in Aragon * Speaker year of birth) indicates that this pattern does not hold for speakers from Aragon. In line with our hypothesis, there is no effect of speaker age for the Aragonese speakers. This result suggests the existence of a border effect between Aragon on the one hand, and Catalonia and Andorra on the other.

We did not find an effect of gender (in both the fixed-effect and the random-effect structure), despite this being reported in the literature frequently (see Cheshire, 2002 for an overview). However, as Wieling et al. (2011) also did not find a gender effect with respect to the pronunciation distance from the standard language in their study, it might be that this phenomenon is more strongly linked to individual sounds (e.g., see Chambers and Trudgill, 1998: Ch. 5.3) than to pronunciation distances between complete words.

We also did not find support for the inclusion of education level as a covariate in our model. The reason for this might be similar to the reason for the absence of a gender effect, but the education measure alone (without any other social status measures) might simply have too little power to discover social class effects (Labov, 2001: Ch. 5.9). Interestingly, we do see that the effect of (some of) these speaker-related variables varies significantly *per word* (see Table 2).

4.2. Lexical predictors

All lexical variables we tested were significant predictors of the pronunciation distance from standard Catalan and also showed significant by-speaker random slopes.

It is not surprising that the factor distinguishing articles, clitic pronouns and demonstratives from the other words 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 essentially semi-words as they attach to nouns or verbs, respectively. Because of this they are relatively short (in many cases only having a length of one or two sounds), and when they are different from the standard, the relative distance will be very high. While the demonstratives are not so short, they tend to be either completely identical to the

standard pronunciation, or almost completely different from the standard pronunciation, explaining their larger distances.

We were somewhat surprised that the number of sounds in the reference pronunciation contributed significantly to the distance from the standard, as we normalized dialect distances by dividing them by the alignment length (which correlates highly, $r > 0.95$, with the number of sounds in the reference pronunciation). This result, however, indicates that longer words have a higher average distance from the standard pronunciation than shorter words. A possible reason for this might be that longer words potentially allow for more variation in their pronunciation without affecting the understandability (e.g., changing one third of the sounds of a word consisting of only three sounds will likely change the meaning, while this is less probable for a word consisting of six sounds).

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) as vowels are much more variable than consonants (e.g., Keating et al., 1994).

Besides playing a significant role as fixed-effect factors, all lexical predictors show significant variation in their strength for individual speakers. This reflects that, for example, some speakers will pronounce words with a large number of vowels closer to the standard Catalan pronunciation than others.

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. (submitted). 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.

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. Note that our analysis retains the aggregate perspective of dialectometry. 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 effects may seem crude when compared to studies in other variationist traditions, but our point has been to introduce the methodology.

In contrast to the conclusion of Valls et al. (submitted) 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 clear significant effect of community size (nor a significant interaction between speaker age and community size). In fact

when using the binary distinction they based their conclusion on (i.e. distinguishing urban and rural communities in twenty different counties), the results do not even approach significance ($p = 0.18$). This clearly illustrates the need for adequate statistical tests, to prevent reaching statistically unsupported conclusions.

We did not find support for the general influence of the other demographic variables. This contrasts with the study of Wieling et al. (2011), 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 this study was only small and might have limited our power to detect these effects (i.e. in the study of Wieling et al., 2011 more than ten times as many locations were included).

We see two promising extensions of this study. First, it would be interesting to compare 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 second extension involves focusing on the individual sound correspondences between Catalan dialect pronunciations and pronunciations in standard Catalan (or in another language, such as Latin for a more historically motivated reference point). 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 (or Latin) pronunciation. Of course, this approach is also possible for all locations, but due to the larger size, much more patience will be required before all parameters are successfully estimated.

References

- Baayen, R. Harald (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.
- Baayen, R. Harald, Doug J. Davidson and Douglas M. Bates (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4): 390-412.
- Bailey, Guy, Tom Wikle, Jan Tillery and Lori Sand (1991). The apparent time construct. *Language Variation and Change*, 3, 241-264.
- Bibiloni, Gabriel (2002). Un estàndard nacional o tres estàndards regionals? In Joan, Bernat (ed.), *Perspectives sociolingüístiques a les Illes Balears*. Res Publica, Eivissa.
- Cambra de Comerç - Indústria i Serveis d'Andorra (2008). *Informe economic 2008*.
- Chambers, Jack and Peter Trudgill (1998). *Dialectology*. Second edition. Cambridge University Press.
- Cheshire, Jenny (2002). Sex and gender in variationist research. In J. Chambers, P. Trudgill and N. Schilling-Estes (eds.). *The Handbook of Language Variation and Change*. Blackwell Publishing Ltd., 423-443.
- Church, Kenneth & Patrick Hanks (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1): 22-29.
- Departament d'Estadística del Govern d'Andorra (2010). Societat i població. <http://www.estadistica.ad>. Last accessed: February 28, 2011.

- Goebel, Hans (2000). Langues standards et dialectes locaux dans la France du Sud-Est et l'Italie septentrionale sous le coup de l'effet-frontière: une approche dialectométrique. *International journal of the sociology of language*, 145: 181-215.
- Heeringa, Wilbert (2004). *Measuring Dialect Pronunciation Distances using Levenshtein Distance*. PhD thesis, Rijksuniversiteit Groningen.
- Huguet, Angel, Ignasi Vila and Enric Llorca (2000). Minority language education in unbalanced bilingual situations: A case for the linguistic interdependence hypothesis. *Journal of Psycholinguistic Research*, 3: 313-333.
- Instituto Aragonés de Estadística (2007, 2009, 2010). Población i Territorio. <http://www.aragon.es>. Last accessed: February 28, 2011.
- Institut d'Estadística de Catalunya (2008, 2010). Territori. <http://www.idescat.cat>. Last accessed: February 28, 2011.
- Keating, Patricia, Björn Lindblom, James Lubker and Jody Kreiman (1994). Variability in jaw height for segments in English and Swedish VCVs. *Journal of Phonetics*, 22: 407-422.
- Labov, William (2001). *Principles of Linguistic Change, Volume 2. Social Factors*. Blackwell Publishers Inc.
- Levenshtein, Vladimir (1965). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163: 845-848.
- 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.
- Nerbonne, John and Peter Kleiweg (2007). Toward a dialectological yardstick. *Quantitative Linguistics*, 14: 148-167.

- Pradilla, Miquel-Àngel (2008a). *Sociolingüística de la variació i llengua catalana*. Institut d'Estudis Catalans, Barcelona.
- Pradilla, Miquel-Àngel (2008b). *La tribu valenciana. Reflexions sobre la desestructuració de la comunitat lingüística*. Onada, Benicarló.
- Rafel, Joaquim (1996-1998). *Diccionari de freqüències. Corpus textual informatitzat de la llengua catalana*. Barcelona: Institut d'Estudis Catalans.
- Valls, Esteve, Martijn Wieling and John Nerbonne (submitted). Linguistic advergence and divergence in north-western Catalan: A dialectometric investigation of dialect leveling and border effects.
- Wieling, Martijn, Eliza Margaretha and John Nerbonne (accepted). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*.
- Wieling, Martijn and John Nerbonne (2011). Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language*, 25(3), 700-715.
- Wieling, Martijn, John Nerbonne and R. Harald Baayen (2011). Quantitative Social Dialectology: Explaining Linguistic Variation Socially and Geographically. *PLoS ONE*, 6(9): e23613.
- Wieling, Martijn, Jelena Prokić & John Nerbonne (2009). Evaluating the pairwise string alignment of pronunciations. In Lars Borin and Piroska Lendvai (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, 26-34.
- 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.

Woolard, Kathryn and Tae-Joong Gahng (2008). Changing language policies and attitudes in autonomous Catalonia. *Language in Society*, 19: 311-330.

Woolhiser, Curt (2005). Political borders and dialect divergence/convergence in Europe. In Peter Aurer, Frans Hinskens and Paul Kerswill (eds.), *Dialect Change. Convergence and divergence in European languages*. Cambridge University Press, New York, 236-262.