Data, data documentation and analysis scripts for

Border effects among catalan dialects

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

[1]University of Groningen, the Netherlands & [2]Eberhard Karls University, Germany & [3]University of Barcelona, Spain& [4]University of Alberta, Canada

Book: D. Speelman et al. (eds.) **Mixed Effects Regression Models in Linguistics** (subm.)
Preprint: `http://martijnwieling.nl/files/Wieling-etal-2013-LSD.pdf`

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 fitted 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 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. 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 revealed that the pattern observed at the word pronunciation level was 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.

Keywords: Dialectometry, Catalan dialects, border effects, generalized additive modeling, mixed-effects regression.

1 Packages and functions

```
require(parallel)
library(mgcv)

R.Version()$version.string

## [1] "R version 3.0.2 (2013-09-25)"

packageVersion("mgcv")

## [1] '1.7.27'

source('functions/functions.R') # custom functions
```

2 Complete data set

```
load("data/catalan.rda")
```

Legenda tuscan (112608 of 39 variables):

Note that the columns with a suffix of `.c` or `.z` are not described here. These are simply a centered (mean equals zero) or standardized (mean equals zero and standard deviation equals 1) version of the corresponding variables shown below.

1. Word : the word for which pronunciations were obtained
2. Speaker : the speaker whose pronunciations were obtained
3. Location : the location in which speakers were asked for their pronunciations
4. PronDistStdCatalan : Pronunciation distance from standard Catalan
5. Longitude : longitude of the dialect location
6. Latitude : latitude of the dialect location
7. Region : location in Catalonia (C), Andorra (A) or Aragon (L)
8. InAragon : binary value indicating if the location is in Aragon (1) or not (0)
9. IsUrban : binary value indicating if the location is urban (1) or rural (0)
10. CommunitySize.log : number of inhabitants in the location (log-transformed)
11. CommunitySize.log_residGeo : as above, but excluding the influence of geography
12. CommunityAvgAge : average age in the location
13. CommunityAvgAge_residCS_CI_Geo : as above, but excluding the influence of average income, community size and geography
14. CommunityAvgIncome : average income in the location (log-transformed)
15. CommunityAvgIncome_residCS_Geo : as above, but excluding the influence of geography and community size
16. CommunityRelTouristBeds.log : relative number of tourist beds in the location (log-transformed)
17. SpeakerBirthYear : year of birth of the speaker
18. SpeakerBirthYear.z_InAragon : interaction between year of birth of the speaker (z-transformed) and if the location is in aragon or not
19. SpeakerIsMale : binary value indicating if the speaker is male (1) or not (0)

20. SpeakerEduLevel : education level of the speaker from low to high (0: no schooling, 1: primary school, 2: high school, 3: baccalaureate, 4: technical school, 5: university)

21. SpeakerRecordingYear : year when the speaker's pronunciation was recorded

22. WordLength : the number of sounds in the standard pronunciation of the word

23. WordRefVowelRatio : the relative number of vowels in the standard pronunciation of the word

24. WordCategory : the word category: verbs (V), possessives (P), personal pronouns (PP), locatives (L), demonstratives (D), other demonstratives (O), articles (A), clitics (C)

25. WordCategoryIsACD : binary value indicating if the word's category is article (A), clitic (C) or demonstrative (D)

26. IsFieldworkerA : if the fieldworker was Esteve Valls (only for the youngest speakers, recorded around 2010)

3 Analysis and results of complete data set

Mixed-effects regression model

```
cl = makeCluster(4) # 4 cores used in calculating the model

modelCatalan <-
  bam(PronDistStdCatalan.c ~ WordLength.z +
      WordRefVowelRatio.z + WordCategoryIsACD + CommunitySize.log_residGeo.z +
      SpeakerBirthYear.z + InAragon + SpeakerBirthYear.z_InAragon +
      s(Longitude, Latitude) + s(Word, bs = "re") +
      s(Word, CommunityRelTouristBeds.log.z, bs = "re") +
      s(Word, CommunityAvgAge_residCS_CI_Geo.z, bs = "re") +
      s(Word, CommunitySize.log_residGeo.z, bs = "re") +
      s(Word, CommunityAvgIncome_residCS_Geo.z, bs = "re") +
      s(Word, SpeakerBirthYear.z, bs = "re") +
      s(Word, SpeakerEduLevel.z, bs = "re") + s(Word, InAragon, bs = "re") +
      s(Word, SpeakerBirthYear.z_InAragon, bs = "re") + s(Speaker, bs = "re") +
      s(Speaker, WordLength.z, bs = "re") +
      s(Speaker, WordRefVowelRatio.z, bs = "re") +
      s(Speaker, WordCategoryIsACD, bs = "re"), data=catalan,
      gc.level=2, cluster=cl
     ) # duration: 80 minutes

summaryModelCatalan <- summary(modelCatalan) # duration: 60 minutes

save(modelCatalan, file='results/modelCatalan.rda')
save(summaryModelCatalan, file='results/summaryModelCatalan.rda')
```

```
load('results/summaryModelCatalan.rda')
summaryModelCatalan

##
## Family: gaussian
## Link function: identity
##
## Formula:
## PronDistStdCatalan.c ~ WordLength.z + WordRefVowelRatio.z + WordCategoryIsACD +
##     CommunitySize.log_residGeo.z + SpeakerBirthYear.z + InAragon +
##     SpeakerBirthYear.z_InAragon + s(Longitude, Latitude) + s(Word,
##     bs = "re") + s(Word, CommunityRelTouristBeds.log.z, bs = "re") +
##     s(Word, CommunityAvgAge_residCS_CI_Geo.z, bs = "re") + s(Word,
##     CommunitySize.log_residGeo.z, bs = "re") + s(Word, CommunityAvgIncome_residCS...
##     bs = "re") + s(Word, SpeakerBirthYear.z, bs = "re") + s(Word,
##     SpeakerEduLevel.z, bs = "re") + s(Word, InAragon, bs = "re") +
##     s(Word, SpeakerBirthYear.z_InAragon, bs = "re") + s(Speaker,
##     bs = "re") + s(Speaker, WordLength.z, bs = "re") + s(Speaker,
##     WordRefVowelRatio.z, bs = "re") + s(Speaker, WordCategoryIsACD,
##     bs = "re")
##
```

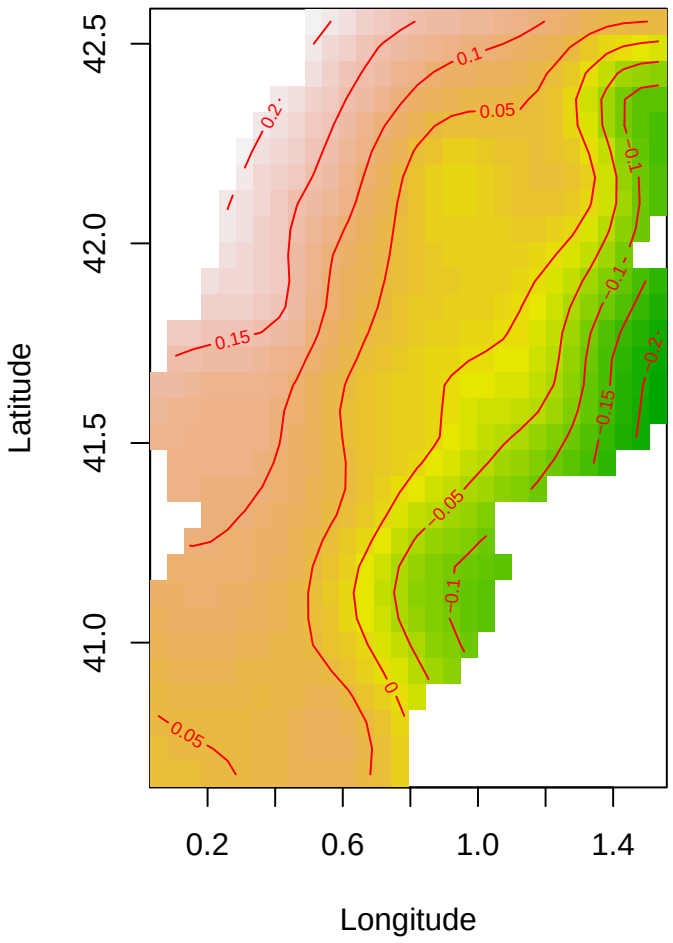
```
## Parametric coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -0.10176    0.02091   -4.87  1.1e-06
## WordLength.z                  0.13030    0.02183    5.97  2.4e-09
## WordRefVowelRatio.z           0.10507    0.01372    7.66  1.9e-14
## WordCategoryIsACD             0.30501    0.04777    6.38  1.7e-10
## CommunitySize.log_residGeo.z -0.00735    0.00377   -1.95  0.05097
## SpeakerBirthYear.z           -0.01140    0.00312   -3.65  0.00026
## InAragon                      0.04608    0.03720    1.24  0.21542
## SpeakerBirthYear.z_InAragon   0.01612    0.00634    2.54  0.01097
##
## Approximate significance of smooth terms:
##                                             edf Ref.df       F p-value
## s(Longitude,Latitude)                      23.9   24.1   24.84  <2e-16
## s(Word)                                   352.3  353.0 6869.46  <2e-16
## s(Word,CommunityRelTouristBeds.log.z)     288.8  357.0   11.98  <2e-16
## s(Word,CommunityAvgAge_residCS_CI_Geo.z) 255.2  357.0    3.57  <2e-16
## s(Word,CommunitySize.log_residGeo.z)      263.6  356.0    3.67  <2e-16
## s(Word,CommunityAvgIncome_residCS_Geo.z) 235.4  357.0   12.69  <2e-16
## s(Word,SpeakerBirthYear.z)                307.1  356.0   53.19  <2e-16
## s(Word,SpeakerEduLevel.z)                 248.5  357.0    5.63  <2e-16
## s(Word,InAragon)                          344.9  356.0 5751.36  <2e-16
## s(Word,SpeakerBirthYear.z_InAragon)       221.9  356.0    6.11  <2e-16
## s(Speaker)                                265.9  313.0   79.80  <2e-16
## s(Speaker,WordLength.z)                   265.9  319.0   66.64  <2e-16
## s(Speaker,WordRefVowelRatio.z)            248.9  319.0    7.17  <2e-16
## s(Speaker,WordCategoryIsACD)              260.4  319.0   88.79  <2e-16
##
## R-sq.(adj) =  0.753   Deviance explained = 76.1%
## fREML score = -31254  Scale est. = 0.03071   n = 112608
```

Visualization of the non-linear effect of geography

```
load('results/modelCatalan.rda')

fixedVals = list(WordLength.z=0, WordRefVowelRatio.z=0, WordCategoryIsACD=0,
                 CommunitySize.log_residGeo.z=0, SpeakerBirthYear.z=0, InAragon=0,
                 SpeakerBirthYear.z_InAragon=0)

vis.gam(modelCatalan, view=c("Longitude","Latitude"), cond=fixedVals,
        plot.type="contour", color="terrain", too.far=0.15, main="")
```



Effect sizes

```
effectSizes = t(data.frame(
    getEffectSize.gam(catalan,summaryModelCatalan,"WordLength.z"),
    getEffectSize.gam(catalan,summaryModelCatalan,"WordRefVowelRatio.z"),
    getEffectSize.gam(catalan,summaryModelCatalan,"WordCategoryIsACD"),
    getEffectSize.gam(catalan,summaryModelCatalan,"CommunitySize.log_residGeo.z"),
    getEffectSize.gam(catalan,summaryModelCatalan,"SpeakerBirthYear.z"),
    getEffectSize.gam(catalan,summaryModelCatalan,"InAragon"),
    getEffectSize.gam(catalan,summaryModelCatalan,"SpeakerBirthYear.z_InAragon"),
    getEffectSizeSpline.gam(modelCatalan,"s(Longitude,Latitude)")
))
```

```
##                               Effect size
## WordLength.z                        0.441
## WordRefVowelRatio.z                 0.649
## WordCategoryisACD                   0.305
## CommunitySize.log_residGeo.z       -0.028
## SpeakerBirthYear.z                 -0.034
## InAragon                            0.046
## SpeakerBirthYear.z_InAragon         0.047
## s(Longitude,Latitude)               0.291
```

Standard deviations of random effects

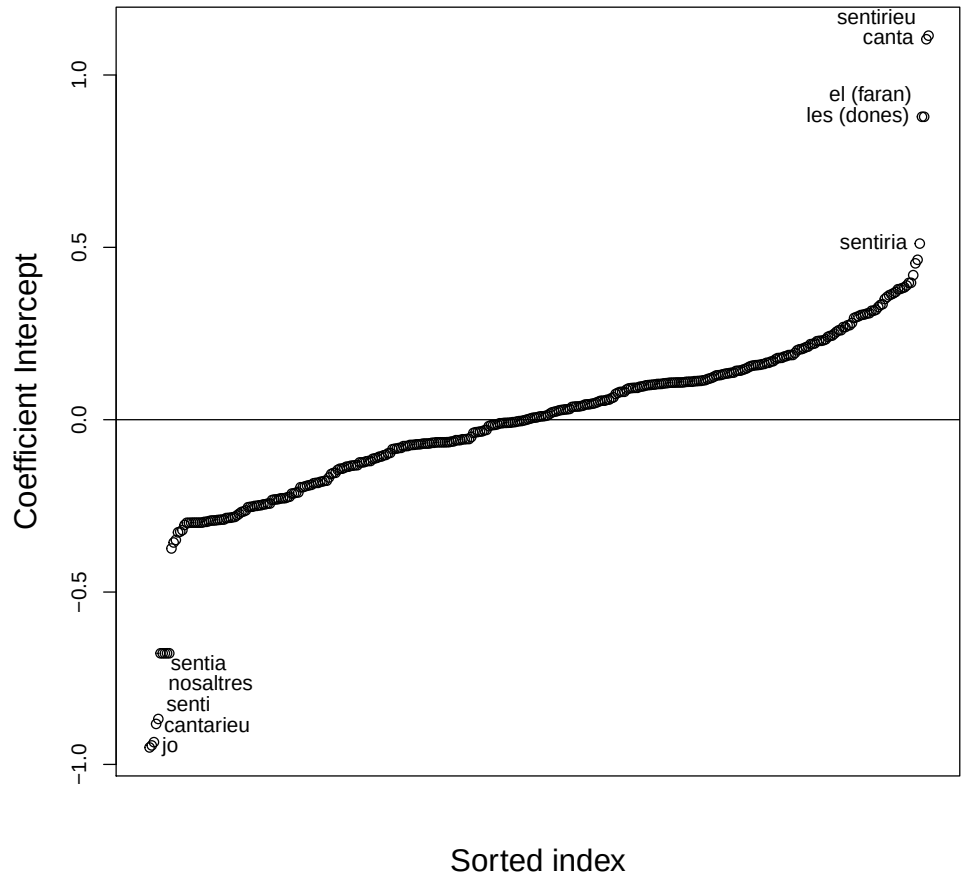
```
coefs = coef(modelCatalan)
stdevs = t(data.frame(
    getSD.gam(coefs, "Word", "Intercept"),
    getSD.gam(coefs, "Word", "CommunityRelTouristBeds.log.z"),
    getSD.gam(coefs, "Word", "CommunityAvgAge_residCS_CI_Geo.z"),
    getSD.gam(coefs, "Word", "CommunitySize.log_residGeo.z"),
    getSD.gam(coefs, "Word", "CommunityAvgIncome_residCS_Geo.z"),
    getSD.gam(coefs, "Word", "SpeakerBirthYear.z"),
    getSD.gam(coefs, "Word", "SpeakerEduLevel.z"),
    getSD.gam(coefs, "Word", "InAragon"),
    getSD.gam(coefs, "Word", "SpeakerBirthYear.z_InAragon"),
    getSD.gam(coefs, "Speaker", "Intercept"),
    getSD.gam(coefs, "Speaker", "WordLength.z"),
    getSD.gam(coefs, "Speaker", "WordRefVowelRatio.z"),
    getSD.gam(coefs, "Speaker", "WordCategoryIsACD")
))
```

```
##                                          Std. dev.
## s(Word)                                      0.255
## s(Word,CommunityRelTouristBeds.log.z)        0.022
## s(Word,CommunityAvgAge_residCS_CI_Geo.z)     0.014
## s(Word,CommunitySize.log_residGeo.z)         0.015
## s(Word,CommunityAvgIncome_residCS_Geo.z)     0.014
## s(Word,SpeakerBirthYear.z)                   0.026
## s(Word,SpeakerEduLevel.z)                    0.014
## s(Word,InAragon)                             0.155
```

```
## s(Word,SpeakerBirthYear.z_InAragon)          0.026
## s(Speaker)                                   0.037
## s(Speaker,WordLength.z)                      0.029
## s(Speaker,WordRefVowelRatio.z)               0.017
## s(Speaker,WordCategoryIsACD)                 0.059
```

By-word random intercepts

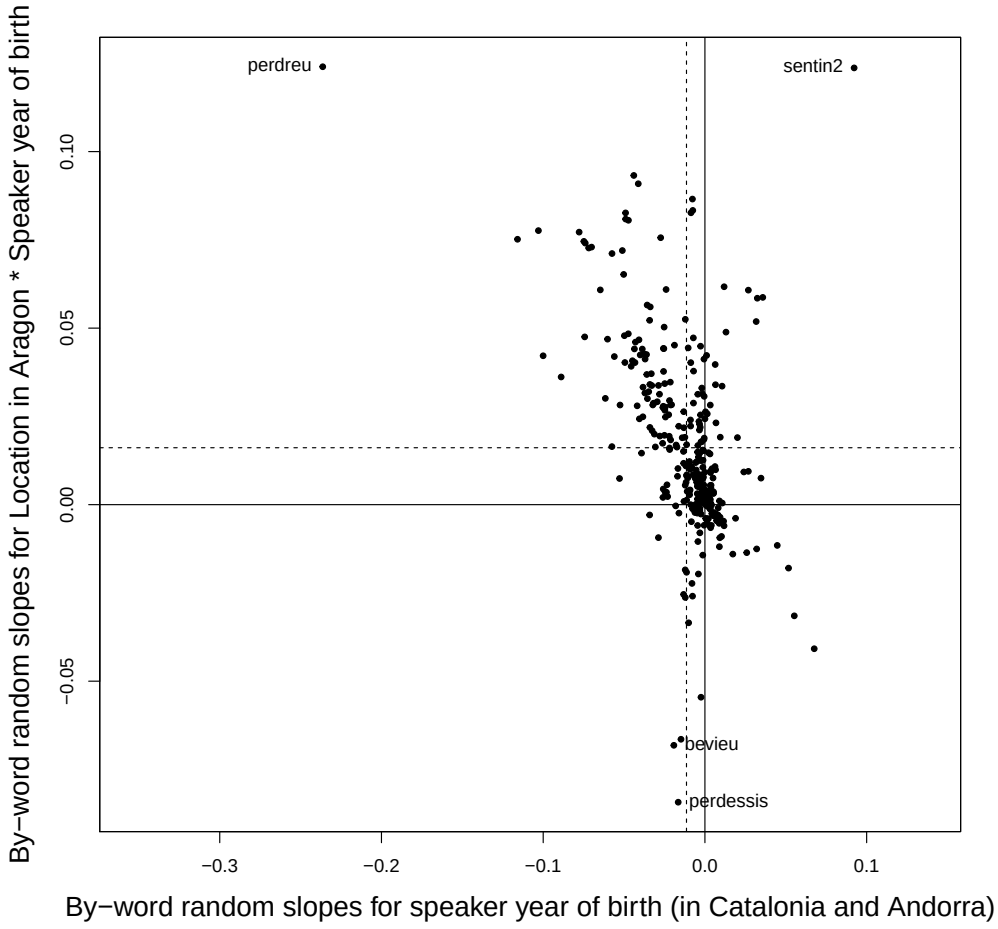
```
plotSlope.gam(modelCatalan,catalan,"Word","Intercept")
```



By-word random slopes

```
ylab = "By-word random slopes for Location in Aragon * Speaker year of birth"

plotSlopes.gam(modelCatalan,catalan,"Word","SpeakerBirthYear.z",
               "SpeakerBirthYear.z_InAragon", xlab, ylab)
```



4 Data set for individual linguistic variables

```
load("data/catalanVars.rda")
```

Legenda tuscanVars (11516 observations of 15 variables):

Note that the columns with a suffix of `.z` are not described here. This is simply a standardized (mean equals zero and standard deviation equals 1) version of the corresponding variables shown below.

1. Word : the word for which responses (with respect to the linguistic variables) were collected
2. Speaker : the speaker whose responses were collected
3. Location : the location where the speaker originates from
4. Var1NonStd : first linguistic variable: replacement of [ʎ] (standard) by [j] (non-standard)
5. Var2NonStd : second linguistic variable: variation in the final morphemes for the Present Subjunctive ([i]: standard, other vowels: non-standard)
6. Var3NonStd : third linguistic variable: use of [β] as opposed to another consonant (mainly [w]) within the possessive adjectives
7. Longitude : longitude of the dialect location
8. Latitude : latitude of the dialect location
9. InAragon : binary value indicating if the location is in Aragon (1) or not (0)
10. Region : location in Catalonia or Andorra (CataloniaAndorra) or Aragon (Aragon)
11. SpeakerIsMale : binary value indicating if the speaker is male (1) or not (0)
12. SpeakerEduLevel : education level of the speaker from low (0) to high (5)
13. SpeakerBirthYear : year of birth of the speaker

5 Analysis and results of individual variables

Model of the first linguistic variable: [ʎ] vs. [j]

```
subsetV1 = droplevels(catalanVars[!is.na(catalanVars$Var1NonStd), ])
dim(subsetV1)

## [1] 3200   15

modelV1 <-
  bam(Var1NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region:SpeakerBirthYear.z +
                   InAragon + s(Longitude, Latitude) + s(Speaker, bs="re") +
                   s(Speaker, InAragon, bs="re"), data=subsetV1,
                   family="binomial"
      ) # duration: 100 seconds

summary(modelV1)

##
## Family: binomial
## Link function: logit
##
## Formula:
## Var1NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region:SpeakerBirthYear.z +
##     InAragon + s(Longitude, Latitude) + s(Speaker, bs = "re") +
##     s(Speaker, InAragon, bs = "re")
##
## Parametric coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                               -5.7363     0.7549   -7.60  3.0e-14
## SpeakerIsMale                              1.1584     0.6282    1.84    0.065
## SpeakerEduLevel.z                          0.0161     0.3821    0.04    0.966
## InAragon                                   3.2121     1.6458    1.95    0.051
## RegionCataloniaAndorra:SpeakerBirthYear.z  3.4241     0.6426    5.33  9.9e-08
## RegionAragon:SpeakerBirthYear.z            5.8840     1.1904    4.94  7.7e-07
##
## Approximate significance of smooth terms:
##                          edf Ref.df Chi.sq p-value
## s(Longitude,Latitude)   8.92   10.1   30.7 0.00072
## s(Speaker)            108.62  312.0  527.7 < 2e-16
## s(Speaker,InAragon)     3.80   62.0   10.1 0.22918
##
## R-sq.(adj) =  0.938   Deviance explained = 92.4%
## fREML score = 3221.7  Scale est. = 1         n = 3200
```

Model of the second linguistic variable: [i] vs. other vowel

```
subsetV2 = droplevels(catalanVars[!is.na(catalanVars$Var2NonStd), ])
dim(subsetV2)

## [1] 6400   15

modelV2 <-
  bam(Var2NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region:SpeakerBirthYear.z +
                   InAragon + s(Longitude, Latitude) + s(Word, bs="re") +
                   s(Speaker, bs="re"), data=subsetV2,
                   family="binomial"
      ) # duration: 30 seconds

summary(modelV2)

##
## Family: binomial
## Link function: logit
##
## Formula:
## Var2NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region:SpeakerBirthYear.z +
##     InAragon + s(Longitude, Latitude) + s(Word, bs = "re") +
##     s(Speaker, bs = "re")
##
## Parametric coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                 0.270      0.633    0.43    0.670
## SpeakerIsMale                              -0.387      0.421   -0.92    0.359
## SpeakerEduLevel.z                          -0.448      0.207   -2.17    0.030
## InAragon                                    5.395      3.267    1.65    0.099
## RegionCataloniaAndorra:SpeakerBirthYear.z  -1.016      0.212   -4.79  1.7e-06
## RegionAragon:SpeakerBirthYear.z             0.331      0.926    0.36    0.721
##
## Approximate significance of smooth terms:
##                         edf Ref.df Chi.sq p-value
## s(Longitude,Latitude)  18.8   20.2    115 3.1e-15
## s(Word)                17.4   19.0    424 < 2e-16
## s(Speaker)            176.1  312.0   1595 < 2e-16
##
## R-sq.(adj) =  0.818   Deviance explained = 79.2%
## fREML score = 7350.8  Scale est. = 1         n = 6400
```

Model of the third linguistic variable: [β] vs. other consonant

```
subsetV3 = droplevels(catalanVars[!is.na(catalanVars$Var3NonStd), ])
dim(subsetV3)

## [1] 1916   15

modelV3 <-
  bam(Var3NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region:SpeakerBirthYear.z +
                   InAragon + s(Longitude, Latitude) + s(Location, bs="re") +
                   s(Word, bs="re") + s(Speaker, bs="re"), data=subsetV3,
                   family="binomial"
      ) # duration: 30 seconds

summary(modelV3)

##
## Family: binomial
## Link function: logit
##
## Formula:
## Var3NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region:SpeakerBirthYear.z +
##     InAragon + s(Longitude, Latitude) + s(Location, bs = "re") +
##     s(Word, bs = "re") + s(Speaker, bs = "re")
##
## Parametric coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                 2.493      0.586    4.26  2.1e-05
## SpeakerIsMale                               0.203      0.492    0.41     0.68
## SpeakerEduLevel.z                          -0.399      0.246   -1.62     0.10
## InAragon                                   13.685    520.963    0.03     0.98
## RegionCataloniaAndorra:SpeakerBirthYear.z  -1.394      0.263   -5.29  1.2e-07
## RegionAragon:SpeakerBirthYear.z             0.300    551.124    0.00     1.00
##
## Approximate significance of smooth terms:
##                          edf Ref.df Chi.sq p-value
## s(Longitude,Latitude)   4.08   4.29  29.83 8.9e-06
## s(Location)            17.72  36.00 242.65  0.0053
## s(Word)                 2.93   5.00   7.31  0.0301
## s(Speaker)            105.09 312.00 428.11 1.5e-08
##
## R-sq.(adj) =  0.895   Deviance explained = 87.1%
## fREML score = 2040.2  Scale est. = 1         n = 1916
```

Visualization of the geographical pattern per linguistic variable

```
fixedVars = list(SpeakerIsMale=0, SpeakerEduLevel.z=0, SpeakerBirthYear.z=0,
                 InAragon=0)

par(mfrow=c(1,3))
vis.gam(modelV1, view=c("Longitude","Latitude"), plot.type="contour",
        color="terrain", too.far=0.15, cond=fixedVars, main="V1")
vis.gam(modelV2, view=c("Longitude","Latitude"), plot.type="contour",
        color="terrain", too.far=0.15, cond=fixedVars, main="V2")
vis.gam(modelV3, view=c("Longitude","Latitude"), plot.type="contour",
        color="terrain", too.far=0.15, cond=fixedVars, main="V3")
```