Data, data documentation and analysis scripts for

English accents and their determinants.

Martijn Wieling[(1)] & Jelke Bloem[(2)] & Kaitlin Mignella[(1)] & Mona Timmermeister[(3)] & R. Harald Baayen[(4,5)] & John Nerbonne[(1,6)]

[1]University of Groningen, the Netherlands & [2]University of Amsterdam, the Netherlands & [3]Utrecht University, the Netherlands & [4]Eberhard Karls University, Germany & [5]University of Alberta, Canada & [6]University of Freiburg, Germany

Status: Submitted
Preprint: `http://www.martijnwieling.nl/files/WielingEtAl-accents-submitted.pdf`

Abstract

In this study we investigate determinants of the strength of foreign accents in English pronunciation. We use pronunciation data from more than 800 speakers with a variety of language backgrounds and analyze the data with an eye to assessing the presence of a critical period in second language learning. In our dataset, speakers with a non-Indo-European native language had a clear breakpoint at the age of 6, whereas speakers with an Indo-European background had only a minor breakpoint around the age of 16. However, resampling the data in an attempt to verify the results showed that both language groups showed a mirrored bimodal pattern. In sum, our study does not support the existence of a stable critical period within which a second language can be learned with a high degree of proficiency, but rather a complex interaction between various social, educational and maturational factors.

Keywords: Second language learning, Critical period hypothesis, Piecewise regression, Mixed-effects regression, English pronunciation.

1 Packages and functions

```
library(lme4)
library(mgcv)
library(car)

set.seed(100) # seed for random number generator

R.Version()$version.string

## [1] "R version 3.1.1 (2014-07-10)"

packageVersion('lme4')

## [1] '1.1.7'

packageVersion('car')

## [1] '2.0.21'
```

2 English accents data set

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

Legenda `accents` (806 observations of 17 variables) and `nat` (272 observations of 17 variables):

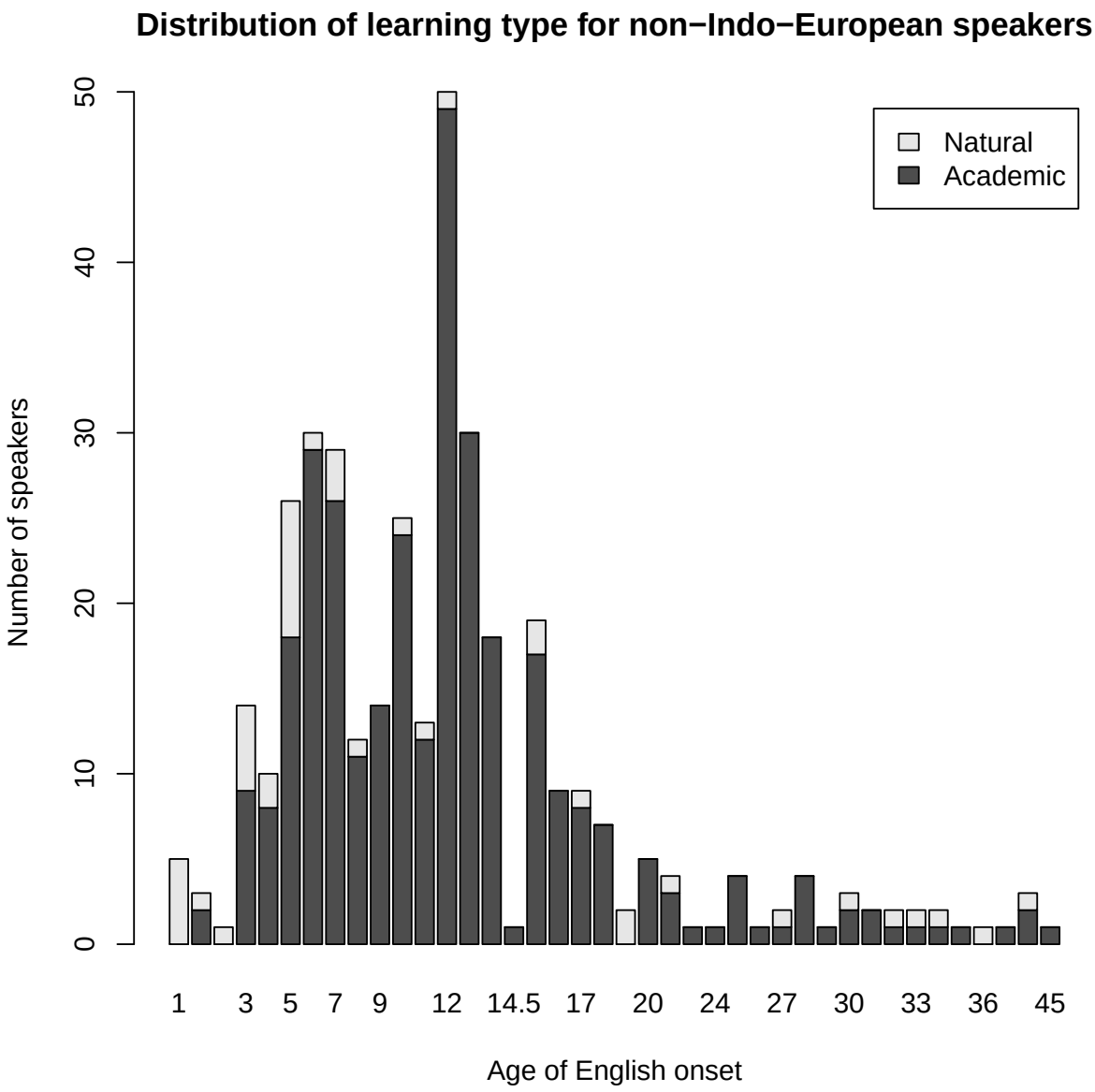
1. Speaker : the speaker
2. Nativelikeness : the human-rated nativelikeness of the speaker (only for the 272 non-native speakers in `nat`)
3. LD.log : log-transformed average Levenshtein distance with respect to average native American English (see Wieling, Bloem, Mignella et al., forthcoming, *Language Dynamics and Change*)
4. Country : the country of birth of the speaker
5. Language : the native language of the speaker
6. IsIESpeaker : if the speaker has an Indo-European Language (TRUE) or not (FALSE)
7. Age : the age of the speaker
8. IsMale : the gender of the speaker (1: male, 0: female)
9. AEO : the age of English onset of the speaker
10. LR : the cumulative length of residence in an English-speaking country of a speaker
11. NrLang : the number of additional languages the speaker speaks (besides English)
12. NrLangNIE : the number of additional non-Indo-European languages the speaker speaks (besides English)
13. NrLangIE : the number of additional Indo-European languages the speaker speaks (besides English)
14. IsNaturalLearner : if the speaker is a natural learner of English (TRUE) or an academic learner (FALSE)
15. PopSize.log : the log-transformed population size of the speaker's country of birth (in 2011)
16. GNI.log : the log-transformed gross national income of the speaker's country of birth (in 2011)
17. CountryEduYrs : the average number of years of education in the speaker's country (in 2011)

3 Analysis and results: Levenshtein distances

3.1 Descriptives

```
# gender distribution of participants
table(accents$IsMale)

##
##   0   1
## 367 439

round( 100*table(accents$IsMale)/nrow(accents), 1 )

##
##    0    1
## 45.5 54.5

# language background distribution of participants
table(accents$IsIESpeaker)

##
## FALSE  TRUE
##   368   438

round( 100*table(accents$IsIESpeaker)/nrow(accents), 1 )

##
## FALSE  TRUE
##  45.7  54.3

# number of unique non-English languages in the dataset
length(unique(accents$Language))

## [1] 171

# average age of participants (and standard deviation)
mean(accents$Age)

## [1] 32.69

sd(accents$Age)

## [1] 12.34

# average age of English onset (and standard deviation)
mean(accents$AEO)

## [1] 12.32

sd(accents$AEO)

## [1] 7.367
```

```r
# average length of residence in an English-speaking country (and standard deviation)
mean(accents$LR)

## [1] 7.736

sd(accents$LR)

## [1] 11.66

# learning style distribution of participants
round( 100*table(accents$IsNaturalLearner)/nrow(accents), 1 )

##
## FALSE  TRUE
##  88.3  11.7

# correlation between the average number of education years per country and the
# gross national income
cor(accents$CountryEduYrs, accents$GNI.log)

## [1] 0.8345

# correlation between Nativelikeness ratings and Levenshtein distances
# (note the correlation might be slightly different from the one reported
#  by Wieling, Bloem, Mignella et al. (forthcoming) as only the
#  non-native speakers are considered here)
cor(accents$Nativelikeness, accents$LD.log, use='pairwise')

## [1] -0.804

# relation of natural vs. academic learners dependent on the
# age of English onset in the non-IE speakers
# (note the greater proportion of natural learners before an
#  age of English onset < 7)
accentsNIE = accents[accents$IsIESpeaker==F,]
tab = table(accentsNIE$IsNaturalLearner,accentsNIE$AEO)
barplot(tab, legend = c('Academic','Natural'),
        xlab='Age of English onset', ylab = 'Number of speakers',
        main='Distribution of learning type for non-Indo-European speakers')
```



3.2 Assessing which random-effect factors are needed

```
# testing random intercepts
m0 = lmer(LD.log ~ AEO + (1|Country), data=accents)
m = lmer(LD.log ~ AEO + (1|Country) + (1|Language), data=accents)

# language is not needed as a random intercept
AIC(m0) - AIC(m)

## [1] -0.4039
```

```
anova(m0, m, refit=F)

## Data: accents
## Models:
## m0: LD.log ~ AEO + (1 | Country)
## m: LD.log ~ AEO + (1 | Country) + (1 | Language)
##    Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m0  4 -421 -403    215     -429
## m   5 -421 -397    216     -431   1.6      1       0.21
```

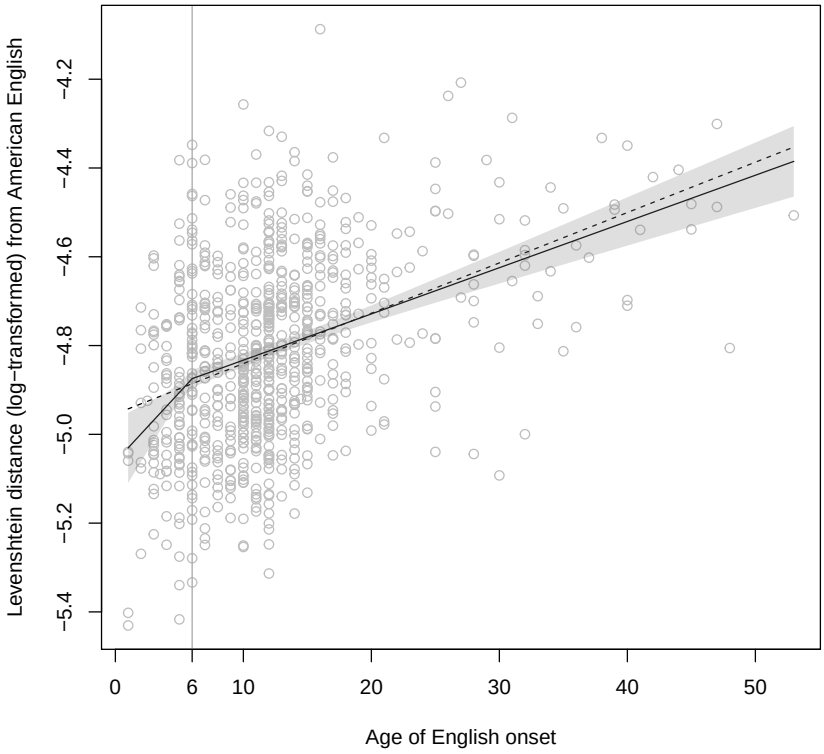
3.3 Breakpoint analysis: single breakpoint

```
# determine breakpoint by iterating over possible breakpoints between an AEO of 1 and 30
deviances = rep(Inf, 30)

for (i in (min(accents$AEO)+1):min(30,max(accents$AEO)-1)) {
    breakpoint = i
    accents$ShiftedAEO = accents$AEO - breakpoint;
    accents$PastBreakPoint = as.factor(accents$ShiftedAEO > 0)
    m = lmer(LD.log ~ ShiftedAEO:PastBreakPoint + (1|Country),
             data=accents, REML=F)
    deviances[i] = deviance(m)
}

breakpoint = which(deviances == min(deviances))
breakpoint

## [1] 6
```

```
# model parameters of model with breakpoint
accents$ShiftedAEO = accents$AEO - breakpoint
accents$PastBreakPoint = as.factor(accents$ShiftedAEO > 0)
m1 = lmer(LD.log ~ ShiftedAEO:PastBreakPoint + (1|Country), data=accents,
          REML=F)
summary(m1)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: LD.log ~ ShiftedAEO:PastBreakPoint + (1 | Country)
##    Data: accents
##
##      AIC      BIC   logLik deviance df.resid
##   -451.8   -428.4    230.9   -461.8      801
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -2.702 -0.670 -0.029  0.642  4.374
##
## Random effects:
```

```
##  Groups   Name        Variance Std.Dev.
##  Country  (Intercept) 0.00869  0.0932
##  Residual             0.02855  0.1690
## Number of obs: 806, groups:  Country, 139
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                    -4.877505   0.013329    -366
## ShiftedAEO:PastBreakPointFALSE  0.041802   0.008389       5
## ShiftedAEO:PastBreakPointTRUE   0.010192   0.000959      11
##
## Correlation of Fixed Effects:
##             (Intr) SAEO:PBPF
## SAEO:PBPFAL  0.288
## SAEO:PBPTRU -0.537 -0.243
```

```
# approximate visualization of breakpoint (using linear model)
my = lm(LD.log ~ ShiftedAEO:PastBreakPoint,data=accents)
mydf = as.data.frame(accents$AEO)
mydf$simple = predict(lm(LD.log ~ AEO, data=accents)) # line without breakpoint
mydf$fit = predict(my,interval="confidence")[,1]
mydf$low = predict(my,interval="confidence")[,2]
mydf$high = predict(my,interval="confidence")[,3]
colnames(mydf)[1] = "AEO"
mydf = mydf[order(mydf$AEO),]

plot(accents$AEO,accents$LD.log,col="gray",xlab="Age of English onset",
     ylab="Levenshtein distance (log-transformed) from American English",axes=F)
ticks=c(0,breakpoint,10,20,30,40,50,60)
axis(side = 1, at = ticks)
axis(side = 2)
abline(v=breakpoint,col="darkgray")
lines(mydf$AEO,mydf$fit,lty=1)
lines(mydf$AEO,mydf$simple,lty=2)
polygon(c(mydf$AEO,rev(mydf$AEO)),c(mydf$high,rev(mydf$low)),
        col=rgb(.5,.5,.5,alpha=.25),border=NA) # shaded confidence interval
box()
```




```
# compare models with and without breakpoint (the AIC reduction should be >= 2)
m0 = lmer(LD.log ~ AEO + (1|Country), data=accents, REML=F)
AIC(m0) - AIC(m1)

## [1] 11.15

anova(m0,m1)

## Data: accents
## Models:
## m0: LD.log ~ AEO + (1 | Country)
## m1: LD.log ~ ShiftedAEO:PastBreakPoint + (1 | Country)
##    Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m0  4 -441 -422    224     -449
## m1  5 -452 -428    231     -462  13.2      1    0.00029
```

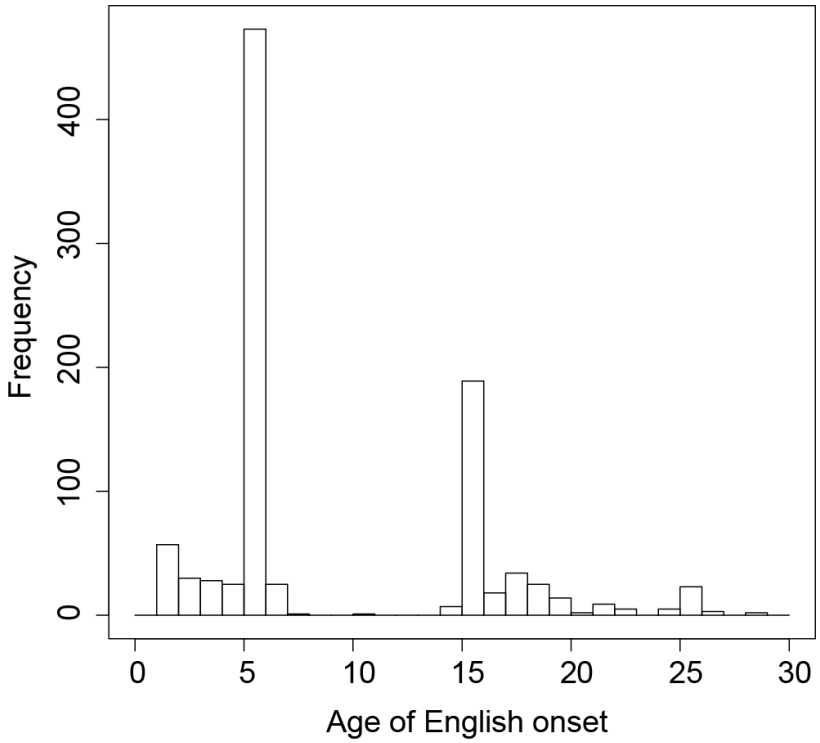
3.3.1 Breakpoint validation: bootstrapping

```
# validating the breakpoint by bootstrapping (1000 iterations)
breakpoints = rep(0, 1000)
AICvals = rep(NA,1000)
```

```
pvals = rep(NA,1000)
for (j in 1:1000) { # 1000 iterations
    dat = accents[c(sample(nrow(accents), nrow(accents), replace=T)), ]

    deviances = rep(Inf, 30)
    for (i in (min(dat$AEO)+1):min(30,max(dat$AEO)-1)) {
        breakpoint = i
        dat$ShiftedAEO = dat$AEO - breakpoint
        dat$PastBreakPoint = as.factor(dat$ShiftedAEO > 0)
        m = lmer(LD.log ~ ShiftedAEO:PastBreakPoint + (1|Country),
                 data=dat, REML=F)
        deviances[i] = deviance(m)
    }
    breakpoints[j] = which(deviances == min(deviances))

    # evaluate significance of the best model for this iteration
    breakpoint = breakpoints[j]
    dat$ShiftedAEO = dat$AEO - breakpoint
    dat$PastBreakPoint = as.factor(dat$ShiftedAEO > 0)
    m = lmer(LD.log ~ ShiftedAEO:PastBreakPoint + (1|Country),
             data=dat, REML=F)
    m0 = lmer(LD.log ~ AEO + (1|Country), data=dat, REML=F)
    AICvals[j] = AIC(m0) - AIC(m) # AIC comparison
    pvals[j] = anova(m0,m)$"Pr(>Chisq)"[2] # p-value
}

# saved as computation takes 30 mins.
save(breakpoints,file='results/breakpoints.rda')
save(AICvals,file='results/AICvals.rda')
save(pvals,file='results/pvals.rda')
```

```
# breakpoints are not stable
load('results/breakpoints.rda')
load('results/AICvals.rda')
load('results/pvals.rda')
table(breakpoints)

## breakpoints
##   2   3   4   5   6   7   8  11  15  16  17  18  19  20  21  22  23  25  26  27  29
##  57  32  30  27 478  25   2   1   8 192  18  38  26  14   2   9   5   5  26   3   2

# number of significant breakpoints
sum(AICvals >= 2)

## [1] 976

sum(pvals < 0.05)

## [1] 976
```

```
# distribution of non-significant breakpoints
table(breakpoints[which(AICvals<2)])

##
##  3  4  5  6  8 15 16 18 19 26
##  2  2  2  5  1  1  3  4  1  3

# visualization of significant breakpoints
sigbp = breakpoints[which(AICvals>=2)]
hist(sigbp,main='',xlab='Age of English onset',breaks=seq(0,30)); box()
```



3.4 Breakpoint analysis: two breakpoints

```
# determine separate breakpoints for IE as opposed to non-IE speakers
# by iterating over possible breakpoints between an AEO of 1 and 30

# IE speakers
deviances = rep(Inf, 30)
accentsIE = accents[accents$IsIESpeaker==T,]
accentsIE = droplevels(accentsIE)

for (i in (min(accentsIE$AEO)+1):min(30,max(accentsIE$AEO)-1)) {
    breakpoint = i
    accentsIE$ShiftedAEO = accentsIE$AEO - breakpoint;
    accentsIE$PastBreakPoint = as.factor(accentsIE$ShiftedAEO > 0)
    m = lmer(LD.log ~ ShiftedAEO:PastBreakPoint + (1|Country),
```

```
            data=accentsIE, REML=F)
    deviances[i] = deviance(m)
}

breakpointIE = which(deviances == min(deviances))
breakpointIE

## [1] 17

# non-IE speakers
deviances = rep(Inf, 30)
accentsNIE = accents[accents$IsIESpeaker==F,]
accentsNIE = droplevels(accentsNIE)

for (i in (min(accentsNIE$AEO)+1):min(30,max(accentsNIE$AEO)-1)) {
    breakpoint = i
    accentsNIE$ShiftedAEO = accentsNIE$AEO - breakpoint;
    accentsNIE$PastBreakPoint = as.factor(accentsNIE$ShiftedAEO > 0)
    m = lmer(LD.log ~ ShiftedAEO:PastBreakPoint + (1|Country),
             data=accentsNIE, REML=F)
    deviances[i] = deviance(m)
}

breakpointNIE = which(deviances == min(deviances))
breakpointNIE

## [1] 6
```

```
# model parameters of two separate models with breakpoint
accentsIE$ShiftedAEO = accentsIE$AEO - breakpointIE
accentsIE$PastBreakPoint = as.factor(accentsIE$ShiftedAEO > 0)
mIE = lmer(LD.log ~ ShiftedAEO:PastBreakPoint + (1|Country), data=accentsIE,
           REML=F)
summary(mIE)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: LD.log ~ ShiftedAEO:PastBreakPoint + (1 | Country)
##    Data: accentsIE
##
##      AIC      BIC   logLik deviance df.resid
##   -273.9   -253.5    142.0   -283.9      433
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -2.606 -0.662 -0.056  0.654  4.252
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Country  (Intercept) 0.00364  0.0603
##  Residual             0.02812  0.1677
```

```
## Number of obs: 438, groups:  Country, 76
##
## Fixed effects:
##                                Estimate Std. Error t value
## (Intercept)                    -4.77491    0.01817  -262.8
## ShiftedAEO:PastBreakPointFALSE  0.01753    0.00236     7.4
## ShiftedAEO:PastBreakPointTRUE   0.00899    0.00188     4.8
##
## Correlation of Fixed Effects:
##             (Intr) SAEO:PBPF
## SAEO:PBPFAL  0.758
## SAEO:PBPTRU -0.388 -0.356
```

```
# model parameters of model with breakpoint
accentsNIE$ShiftedAEO = accentsNIE$AEO - breakpointNIE
accentsNIE$PastBreakPoint = as.factor(accentsNIE$ShiftedAEO > 0)
mNIE = lmer(LD.log ~ ShiftedAEO:PastBreakPoint + (1|Country), data=accentsNIE,
            REML=F)
summary(mNIE)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: LD.log ~ ShiftedAEO:PastBreakPoint + (1 | Country)
##    Data: accentsNIE
##
##      AIC      BIC   logLik deviance df.resid
##   -202.1   -182.6    106.1   -212.1      363
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5541 -0.6867 -0.0154  0.6452  2.5442
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Country  (Intercept) 0.00929  0.0964
##  Residual             0.02803  0.1674
## Number of obs: 368, groups:  Country, 85
##
## Fixed effects:
##                                Estimate Std. Error t value
## (Intercept)                    -4.80471    0.01929  -249.1
## ShiftedAEO:PastBreakPointFALSE  0.05057    0.01039     4.9
## ShiftedAEO:PastBreakPointTRUE   0.00794    0.00144     5.5
##
## Correlation of Fixed Effects:
##             (Intr) SAEO:PBPF
## SAEO:PBPFAL  0.329
## SAEO:PBPTRU -0.553 -0.287
```

```
# comparable results when determining two separate breakpoints in a single model
deviances = rep(Inf, 30^2)
```

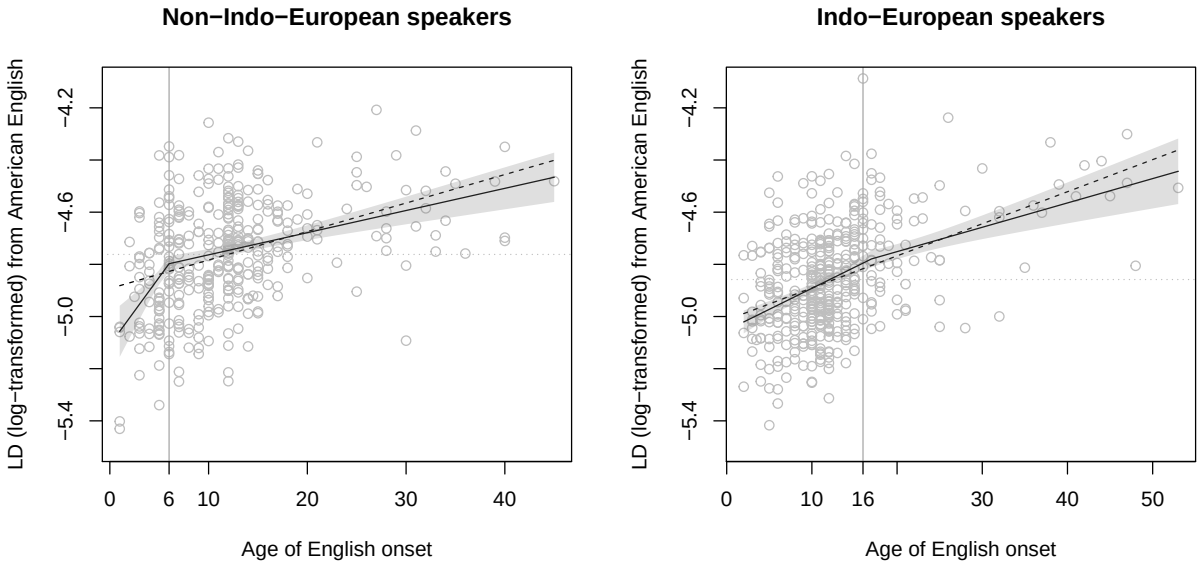
```
dim(deviances) = c(30,30)
for (i in (min(accentsIE$AEO)+1):min(30,max(accentsIE$AEO)-1)) {
    for (j in (min(accentsNIE$AEO)+1):min(30,max(accentsNIE$AEO)-1)) {
        breakpointIE = i
        breakpointNIE = j
        accents$ShiftedAEO = NA
        accents[accents$IsIESpeaker==T,]$ShiftedAEO =
                accents[accents$IsIESpeaker==T,]$AEO - breakpointIE
        accents[accents$IsIESpeaker==F,]$ShiftedAEO =
                accents[accents$IsIESpeaker==F,]$AEO - breakpointNIE
        accents$PastBreakPoint = as.factor(accents$ShiftedAEO > 0)

        m = lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + (1|Country),
                 data=accents, REML=F)
        deviances[i,j] = deviance(m)
    }
}

breakpointIE = which(deviances == min(deviances),arr.ind=T)[1]
breakpointNIE = which(deviances == min(deviances),arr.ind=T)[2]

# breakpoint for Indo-European speakers
breakpointIE

## [1] 16

# breakpoint for non-Indo-European speakers
breakpointNIE

## [1] 6
```

```
# model parameters of the single model with two breakpoints
accents$ShiftedAEO = NA
accents[accents$IsIESpeaker==T,]$ShiftedAEO =
        accents[accents$IsIESpeaker==T,]$AEO - breakpointIE
accents[accents$IsIESpeaker==F,]$ShiftedAEO =
        accents[accents$IsIESpeaker==F,]$AEO - breakpointNIE
accents$PastBreakPoint = as.factor(accents$ShiftedAEO > 0)
m2 = lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + (1|Country),
          data=accents, REML=F)
summary(m2)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + (1 | Country)
##    Data: accents
##
##      AIC      BIC   logLik deviance df.resid
##   -482.4   -449.6    248.2   -496.4      799
##
## Scaled residuals:
```

```
##    Min     1Q Median     3Q    Max
## -2.576 -0.646 -0.038  0.649  4.277
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Country  (Intercept) 0.00619  0.0787
##  Residual             0.02809  0.1676
## Number of obs: 806, groups:  Country, 139
##
## Fixed effects:
##                                                Estimate Std. Error t value
## (Intercept)                                    -4.79886    0.01298    -370
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE  0.05339    0.00991       5
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE   0.01694    0.00212       8
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE   0.00762    0.00127       6
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE    0.00969    0.00174       6
##
## Correlation of Fixed Effects:
##                      (Intr) SAEO:IIESFALSE:PBPF SAEO:IIESTRUE:PBPF SAEO:IIESFALSE:...
## SAEO:IIESFALSE:PBPF  0.267                                                        ...
## SAEO:IIESTRUE:PBPF   0.533  0.178                                                 ...
## SAEO:IIESFALSE:PBPT -0.445 -0.218              -0.265                             ...
## SAEO:IIESTRUE:PBPT  -0.254 -0.074              -0.240              0.127
```

```r
# approximate visualization of both breakpoints (using linear model)
my = lm(LD.log ~ ShiftedAEO:PastBreakPoint,data=accentsNIE)
mydf = as.data.frame(accentsNIE$AEO)
mydf$simple = predict(lm(LD.log ~ AEO, data=accentsNIE))
mydf$fit = predict(my,interval="confidence")[,1]
mydf$low = predict(my,interval="confidence")[,2]
mydf$high = predict(my,interval="confidence")[,3]
colnames(mydf)[1] = "AEO"
mydf = mydf[order(mydf$AEO),]

par(mfrow=c(1,2))
plot(accentsNIE$AEO,accentsNIE$LD.log,col="gray",main="Non-Indo-European speakers",
     xlab="Age of English onset",
     ylab="LD (log-transformed) from American English",
     axes=F, ylim=c(-5.5,-4.1))
ticks=c(0,breakpointNIE,10,20,30,40,50,60)
axis(side = 1, at = ticks)
axis(side = 2)
abline(v=breakpointNIE,col="darkgray")
abline(h=mean(accentsNIE$LD.log),lty=3,col='gray')
lines(mydf$AEO,mydf$fit,lty=1)
lines(mydf$AEO,mydf$simple,lty=2)
polygon(c(mydf$AEO,rev(mydf$AEO)),c(mydf$high,rev(mydf$low)),
        col=rgb(.5,.5,.5,alpha=.25),border=NA)
box()
```

```
my = lm(LD.log ~ ShiftedAEO:PastBreakPoint,data=accentsIE)
mydf = as.data.frame(accentsIE$AEO)
mydf$simple = predict(lm(LD.log ~ AEO, data=accentsIE))
mydf$fit = predict(my,interval="confidence")[,1]
mydf$low = predict(my,interval="confidence")[,2]
mydf$high = predict(my,interval="confidence")[,3]
colnames(mydf)[1] = "AEO"
mydf = mydf[order(mydf$AEO),]

plot(accentsIE$AEO,accentsIE$LD.log,col="gray",main="Indo-European speakers",
     xlab="Age of English onset",
     ylab="LD (log-transformed) from American English",
     axes=F, ylim=c(-5.5,-4.1))
ticks=c(0,10,breakpointIE,20,30,40,50,60)
axis(side = 1, at = ticks)
axis(side = 2)
abline(v=breakpointIE,col="darkgray")
abline(h=mean(accentsIE$LD.log),lty=3,col='gray')
lines(mydf$AEO,mydf$fit,lty=1)
lines(mydf$AEO,mydf$simple,lty=2)
polygon(c(mydf$AEO,rev(mydf$AEO)),c(mydf$high,rev(mydf$low)),
        col=rgb(.5,.5,.5,alpha=.25),border=NA)
box()
```




```
# compare model with two breakpoints (m2) to the model with one breakpoint (m1)
AIC(m1) - AIC(m2)

## [1] 30.63

anova(m1,m2)
```

```
## Data: accents
## Models:
## m1: LD.log ~ ShiftedAEO:PastBreakPoint + (1 | Country)
## m2: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + (1 | Country)
##    Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m1  5 -452 -428    231     -462
## m2  7 -482 -450    248     -496  34.6      2      3e-08

# compare model with two breakpoints to a model with only a different effect of
# AEO for Indo-European as compared to non-Indo-European speakers
m0alt = lmer(LD.log ~ AEO*IsIESpeaker + (1|Country),
             data=accents, REML=F)
AIC(m0alt) - AIC(m2)

## [1] 19.26

anova(m0alt,m2)

## Data: accents
## Models:
## m0alt: LD.log ~ AEO * IsIESpeaker + (1 | Country)
## m2: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + (1 | Country)
##       Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m0alt  6 -463 -435    238     -475
## m2     7 -482 -450    248     -496  21.3      1      4e-06

# test if the individual breakpoint (stored in model mNIE) is needed
# for the non-Indo-European speakers
m0nonIE = lmer(LD.log ~ AEO + (1|Country),
               data=accentsNIE, REML=F)
AIC(m0nonIE) - AIC(mNIE)

## [1] 13

anova(m0nonIE,mNIE)

## Data: accentsNIE
## Models:
## m0nonIE: LD.log ~ AEO + (1 | Country)
## mNIE: LD.log ~ ShiftedAEO:PastBreakPoint + (1 | Country)
##         Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m0nonIE  4 -189 -174   98.6     -197
## mNIE     5 -202 -183  106.1     -212    15      1    0.00011

# test if the individual breakpoint (stored in model mIE) is also needed
# for the Indo-European speakers
m0IE = lmer(LD.log ~ AEO + (1|Country),
            data=accentsIE, REML=F)
AIC(m0IE) - AIC(mIE)

## [1] 3.898
```

```
anova(m0IE,mIE)

## Data: accentsIE
## Models:
## m0IE: LD.log ~ AEO + (1 | Country)
## mIE: LD.log ~ ShiftedAEO:PastBreakPoint + (1 | Country)
##      Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m0IE  4 -270 -254    139     -278
## mIE   5 -274 -254    142     -284   5.9      1      0.015
```

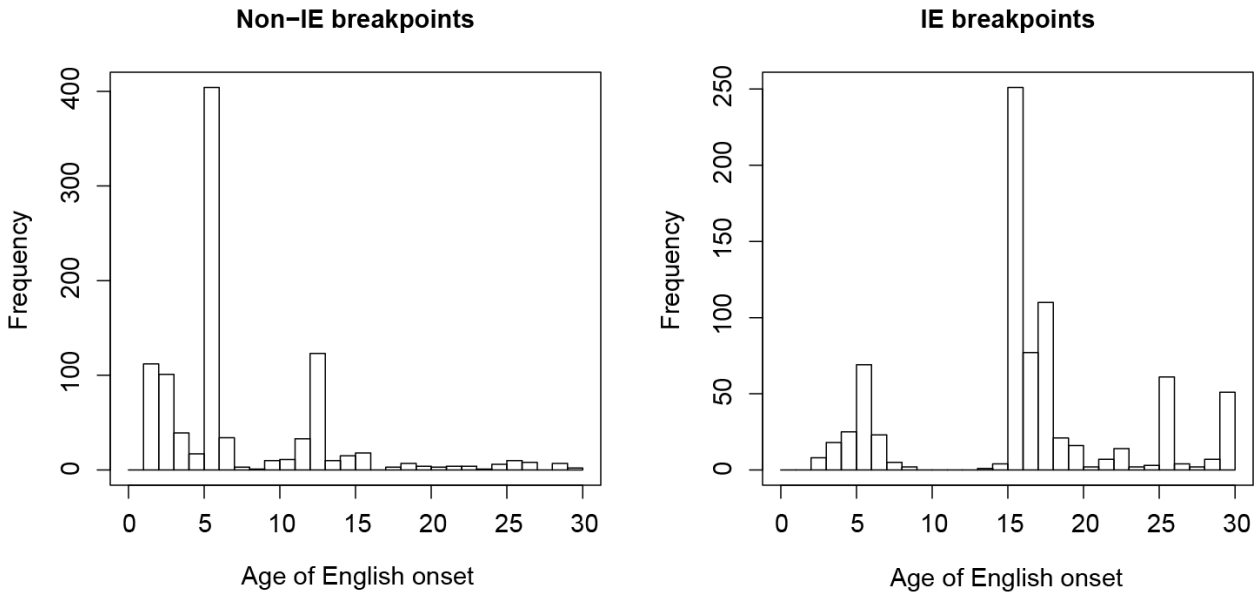
3.4.1 Breakpoint validation: bootstrapping

```
# validating the breakpoint for IE speakers by bootstrapping (1000 iterations)
breakpointsIE = rep(0, 1000)
AICvalsIE = rep(NA,1000)
pvalsIE = rep(NA,1000)
for (j in 1:1000) { # 1000 iterations
    dat = accentsIE[c(sample(nrow(accentsIE), nrow(accentsIE), replace=T)), ]

    deviances = rep(Inf, 30)
    for (i in (min(dat$AEO)+1):min(30,max(dat$AEO)-1)) {
        breakpointIE = i
        dat$ShiftedAEO = dat$AEO - breakpointIE
        dat$PastBreakPoint = as.factor(dat$ShiftedAEO > 0)
        m = lmer(LD.log ~ ShiftedAEO:PastBreakPoint + (1|Country),
                 data=dat, REML=F)
        deviances[i] = deviance(m)
    }
    breakpointsIE[j] = which(deviances == min(deviances))

    # evaluate significance of the best model for this iteration
    breakpointIE = breakpointsIE[j]
    dat$ShiftedAEO = dat$AEO - breakpointIE
    dat$PastBreakPoint = as.factor(dat$ShiftedAEO > 0)
    m = lmer(LD.log ~ ShiftedAEO:PastBreakPoint + (1|Country),
             data=dat, REML=F)
    m0 = lmer(LD.log ~ AEO + (1|Country), data=dat, REML=F)
    AICvalsIE[j] = AIC(m0) - AIC(m) # AIC comparison
    pvalsIE[j] = anova(m0,m)$"Pr(>Chisq)"[2]
}

# saved as computation takes 30 mins.
save(breakpointsIE,file='results/breakpointsIE.rda')
save(AICvalsIE,file='results/AICvalsIE.rda')
save(pvalsIE,file='results/pvalsIE.rda')


# validating the breakpoint for NIE speakers by bootstrapping (1000 iterations)
breakpointsNIE = rep(0, 1000)
```

```
AICvalsNIE = rep(NA,1000)
pvalsNIE = rep(NA,1000)
for (j in 1:1000) { # 1000 iterations
    dat = accentsNIE[c(sample(nrow(accentsNIE), nrow(accentsNIE), replace=T)), ]

    deviances = rep(Inf, 30)
    for (i in (min(dat$AEO)+1):min(30,max(dat$AEO)-1)) {
        breakpointNIE = i
        dat$ShiftedAEO = dat$AEO - breakpointNIE
        dat$PastBreakPoint = as.factor(dat$ShiftedAEO > 0)
        m = lmer(LD.log ~ ShiftedAEO:PastBreakPoint + (1|Country),
                 data=dat, REML=F)
        deviances[i] = deviance(m)
    }
    breakpointsNIE[j] = which(deviances == min(deviances))

    # evaluate significance of the best model for this iteration
    breakpointNIE = breakpointsNIE[j]
    dat$ShiftedAEO = dat$AEO - breakpointNIE
    dat$PastBreakPoint = as.factor(dat$ShiftedAEO > 0)
    m = lmer(LD.log ~ ShiftedAEO:PastBreakPoint + (1|Country),
             data=dat, REML=F)
    m0 = lmer(LD.log ~ AEO + (1|Country), data=dat, REML=F)
    AICvalsNIE[j] = AIC(m0) - AIC(m) # AIC comparison
    pvalsNIE[j] = anova(m0,m)$"Pr(>Chisq)"[2] # ML comparison
}

# saved as computation takes 30 mins.
save(breakpointsNIE,file='results/breakpointsNIE.rda')
save(AICvalsNIE,file='results/AICvalsNIE.rda')
save(pvalsNIE,file='results/pvalsNIE.rda')
```

```
# breakpoints are (again) not stable
load('results/breakpointsNIE.rda')
load('results/AICvalsNIE.rda')
load('results/pvalsNIE.rda')
load('results/breakpointsIE.rda')
load('results/AICvalsIE.rda')
load('results/pvalsIE.rda')

table(breakpointsNIE)

## breakpointsNIE
##   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  18  19  20  21  22  23
## 113 102  39  19 406  34   3   1  10  11  33 124  10  15  18   3   7   4   4   4   4
##  24  25  26  27  29  30
##   1   6  10   9   7   3

# number of significant breakpoints
sum(AICvalsNIE >= 2)
```

```
## [1] 990

sum(pvalsNIE < 0.05)

## [1] 992

# distribution of non significant breakpoints
table(breakpointsNIE[which(AICvalsNIE<2)])

##
##  2  3  5  6 13 21 27 30
##  1  1  2  2  1  1  1  1

table(breakpointsIE)

## breakpointsIE
##   3   4   5   6   7   8   9  10  11  12  14  15  16  17  18  19  20  21  22  23  24
##  27  31  36  92  28  11   2   6   3   6   1   7 302  82 130  28  18   2  10  15   2
##  25  26  27  28  29  30
##   5  71   6   3   7  69

# number of significant breakpoints
sum(AICvalsIE >= 2)

## [1] 783

sum(pvalsIE < 0.05)

## [1] 797

# distribution of non significant breakpoints
table(breakpointsIE[which(AICvalsIE<2)])

##
##  3  4  5  6  7  8 10 11 12 15 16 17 18 19 20 22 23 25 26 27 28 30
## 19 13 11 23  5  6  6  3  6  3 51  5 20  7  2  3  1  2 10  2  1 18

# visualize significant breakpoints
par(mfrow=c(1,2))
sigbpNIE = breakpointsNIE[which(AICvalsNIE>=2)]
hist(sigbpNIE,main='Non-IE breakpoints',xlab='Age of English onset',
     breaks=seq(0,30)); box()

sigbpIE = breakpointsIE[which(AICvalsIE>=2)]
hist(sigbpIE,main='IE breakpoints',xlab='Age of English onset',
     breaks=seq(0,30)); box()
```



3.5 Model building: model with breakpoints

```
# model building including breakpoints: significant predictors
model0 <- lmer(LD.log ~ (1|Country), data=accents, REML=F)
summary(model0)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: LD.log ~ (1 | Country)
##    Data: accents
##
##      AIC      BIC   logLik deviance df.resid
##   -298.4   -284.3    152.2   -304.4      803
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -2.753 -0.658 -0.023  0.640  4.230
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Country  (Intercept) 0.0107   0.103
##  Residual             0.0347   0.186
## Number of obs: 806, groups:  Country, 139
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  -4.8183     0.0122    -395

breakpointIE = 16
breakpointNIE = 6
accents$ShiftedAEO = NA
accents[accents$IsIESpeaker==T,]$ShiftedAEO =
```

```
accents[accents$IsIESpeaker==T,]$AEO - breakpointIE
accents[accents$IsIESpeaker==F,]$ShiftedAEO =
accents[accents$IsIESpeaker==F,]$AEO - breakpointNIE
accents$PastBreakPoint = as.factor(accents$ShiftedAEO > 0)

model1 <- lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + (1|Country),
               data=accents, REML=F)
summary(model1)$coefficients
```

```
##                                                 Estimate Std. Error  t value
## (Intercept)                                    -4.798860   0.012976 -369.838
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE  0.053394   0.009912    5.387
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE   0.016938   0.002125    7.972
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE   0.007616   0.001273    5.981
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE    0.009686   0.001736    5.581
```

```
AIC(model0) - AIC(model1)
```

```
## [1] 184
```

```
anova(model0,model1)
```

```
## Data: accents
## Models:
## model0: LD.log ~ (1 | Country)
## model1: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + (1 | Country)
##        Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model0  3 -298 -284    152     -304
## model1  7 -482 -450    248     -496   192      4     <2e-16
```

```
model2 <- lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + (1|Country),
               data=accents, REML=F)
summary(model2)$coefficients
```

```
##                                                 Estimate Std. Error  t value
## (Intercept)                                    -4.776144  0.0138235 -345.509
## LR                                             -0.002520  0.0005651   -4.460
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE  0.051528  0.0097993    5.258
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE   0.017450  0.0021030    8.298
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE   0.007322  0.0012597    5.813
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE    0.009878  0.0017148    5.761
```

```
AIC(model1) - AIC(model2)
```

```
## [1] 17.64
```

```
anova(model1,model2)
```

```
## Data: accents
## Models:
## model1: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + (1 | Country)
## model2: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + (1 | Country)
##        Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model1  7 -482 -450    248     -496
## model2  8 -500 -463    258     -516  19.6      1    9.3e-06
```

```
model3 <- lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age + (1|Country),
               data=accents, REML=F)
summary(model3)$coefficients
```

```
##                                                Estimate Std. Error  t value
## (Intercept)                                   -4.833632  0.0227181 -212.766
## LR                                            -0.003923  0.0007139   -5.495
## Age                                            0.002122  0.0006671    3.181
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE  0.046845  0.0098468    4.757
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE   0.016939  0.0020971    8.078
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE   0.006784  0.0012633    5.370
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE    0.008455  0.0017617    4.799
```

```
AIC(model2) - AIC(model3)
```

```
## [1] 8.054
```

```
anova(model2,model3)
```

```
## Data: accents
## Models:
## model2: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + (1 | Country)
## model3: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age + (1 |
## model3:     Country)
##        Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model2  8 -500 -463    258     -516
## model3  9 -508 -466    263     -526  10.1      1     0.0015
```

```
model4 <- lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age +
                        CountryEduYrs + (1|Country), data=accents, REML=F)
summary(model4)$coefficients
```

```
##                                                Estimate Std. Error  t value
## (Intercept)                                   -4.681213  0.0309006 -151.493
## LR                                            -0.004015  0.0006876   -5.839
## Age                                            0.002223  0.0006510    3.416
## CountryEduYrs                                 -0.019187  0.0027877   -6.883
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE  0.050267  0.0096827    5.191
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE   0.015233  0.0020055    7.596
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE   0.005666  0.0012344    4.590
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE    0.009331  0.0017344    5.380
```

```
AIC(model3) - AIC(model4)
```

```
## [1] 35.23
```

```
anova(model3,model4)
```

```
## Data: accents
## Models:
## model3: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age + (1 |
## model3:     Country)
```

```
## model4: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age + CountryEduYrs +
## model4:     (1 | Country)
##        Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model3  9 -508 -466    263     -526
## model4 10 -543 -496    282     -563  37.2      1      1e-09

model5 <- lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age +
                        CountryEduYrs + NrLang + (1|Country), data=accents, REML=F)
summary(model5)$coefficients

##                                                 Estimate Std. Error  t value
## (Intercept)                                   -4.668294   0.030653 -152.297
## LR                                            -0.004179   0.000685   -6.101
## Age                                            0.002495   0.000654    3.814
## CountryEduYrs                                 -0.019087   0.002707   -7.050
## NrLang                                        -0.015917   0.005573   -2.856
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE  0.049352   0.009643    5.118
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE   0.014851   0.001989    7.466
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE   0.005591   0.001226    4.561
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE    0.009056   0.001729    5.238

AIC(model4) - AIC(model5)

## [1] 6.012

anova(model4,model5)

## Data: accents
## Models:
## model4: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age + CountryEduYrs +
## model4:     (1 | Country)
## model5: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age + CountryEduYrs +
## model5:     NrLang + (1 | Country)
##        Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model4 10 -543 -496    282     -563
## model5 11 -549 -498    286     -571  8.01      1     0.0046

# PopSize.log is significant here, but after excluding a single speaker
# in the model criticism phase (below), it is not significant
# anymore (t = 1.86; not shown), so we exclude it here as well
model5b <- lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age +
                        CountryEduYrs + NrLang + PopSize.log + (1|Country),
                        data=accents, REML=F)
summary(model5b)$coefficients

##                                                 Estimate Std. Error  t value
## (Intercept)                                   -4.709077  0.0353496 -133.215
## LR                                            -0.004053  0.0006827   -5.937
## Age                                            0.002422  0.0006526    3.711
## CountryEduYrs                                 -0.017734  0.0026833   -6.609
## NrLang                                        -0.014950  0.0055729   -2.683
```

```
## PopSize.log                                          0.009948  0.0045364    2.193
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE  0.050276  0.0096246    5.224
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE   0.015347  0.0019863    7.726
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE   0.005585  0.0012181    4.585
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE    0.009225  0.0017245    5.349
```

```
AIC(model5) - AIC(model5b)
```

```
## [1] 2.612
```

```
anova(model5,model5b)
```

```
## Data: accents
## Models:
## model5: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age + CountryEduYrs +
## model5:     NrLang + (1 | Country)
## model5b: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age + CountryEduYr...
## model5b:     NrLang + PopSize.log + (1 | Country)
##         Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model5  11 -549 -498    286     -571
## model5b 12 -552 -496    288     -576  4.61      1      0.032
```

```
# IsNaturalLearner is not significant
model5b <- lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age +
                        CountryEduYrs + NrLang + IsNaturalLearner + (1|Country),
                data=accents, REML=F)
summary(model5b)$coefficients
```

```
##                                                  Estimate Std. Error  t value
## (Intercept)                                     -4.668204  0.0305022 -153.045
## LR                                              -0.003971  0.0007028   -5.651
## Age                                              0.002401  0.0006576    3.651
## CountryEduYrs                                   -0.018830  0.0026934   -6.991
## NrLang                                          -0.015745  0.0055680   -2.828
## IsNaturalLearnerTRUE                            -0.026019  0.0203617   -1.278
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE  0.045853  0.0100161    4.578
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE   0.014597  0.0019929    7.325
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE   0.005850  0.0012404    4.716
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE    0.009535  0.0017662    5.399
```

```
AIC(model5) - AIC(model5b)
```

```
## [1] -0.3769
```

```
anova(model5,model5b)
```

```
## Data: accents
## Models:
## model5: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age + CountryEduYrs +
## model5:     NrLang + (1 | Country)
## model5b: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age + CountryEduYr...
## model5b:     NrLang + IsNaturalLearner + (1 | Country)
##         Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model5  11 -549 -498    286     -571
## model5b 12 -549 -493    286     -573  1.62      1        0.2
```

```
# IsMale, NrLangIE, NrLangNIE and GNI.log are also not significant (not shown)
```

```
# model building: significant interactions

# center LR and Age to facilitate interpretation of interaction
accents$LR.c = accents$LR - mean(accents$LR)
accents$Age.c = accents$Age - mean(accents$Age)

model6 <- lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR.c*Age.c +
                        CountryEduYrs + NrLang + (1|Country),
                        data=accents, REML=F)
summary(model6)$coefficients
```

```
##                                                 Estimate Std. Error  t value
## (Intercept)                                   -4.6212863  2.556e-02 -180.782
## LR.c                                          -0.0062342  8.746e-04   -7.128
## Age.c                                          0.0024400  6.488e-04    3.761
## CountryEduYrs                                 -0.0203029  2.735e-03   -7.424
## NrLang                                        -0.0159796  5.530e-03   -2.890
## LR.c:Age.c                                     0.0001278  3.421e-05    3.734
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE  0.0441856  9.659e-03    4.575
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE   0.0146145  1.978e-03    7.389
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE   0.0056072  1.217e-03    4.609
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE    0.0093360  1.716e-03    5.440
```

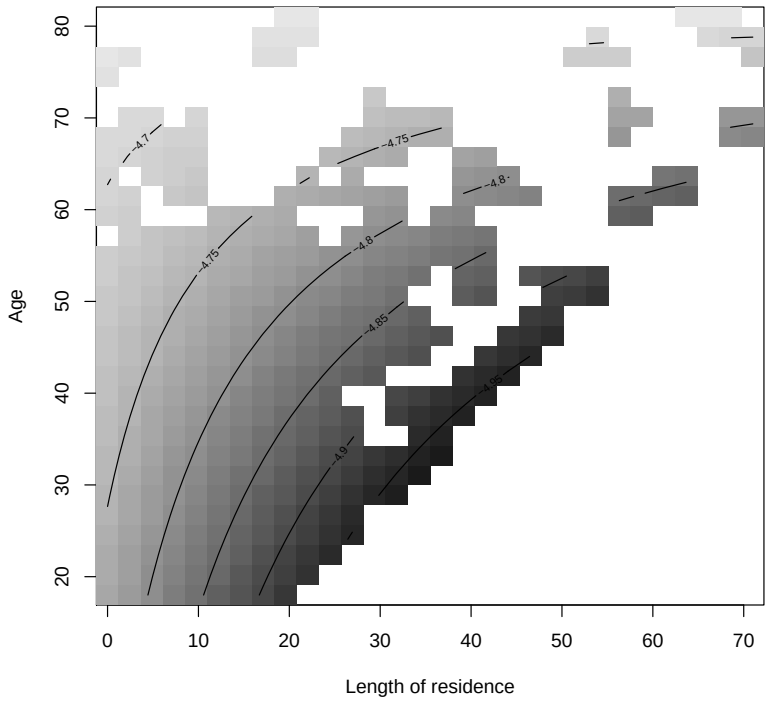
```
AIC(model5) - AIC(model6)
```

```
## [1] 11.81
```

```
anova(model5,model6)
```

```
## Data: accents
## Models:
## model5: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR + Age + CountryEduYrs +
## model5:     NrLang + (1 | Country)
## model6: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR.c * Age.c +
## model6:     CountryEduYrs + NrLang + (1 | Country)
##        Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model5 11 -549 -498    286     -571
## model6 12 -561 -505    293     -585  13.8      1      2e-04
```

```
# The interaction is visualized below.
# No other interactions improve the model (not shown).
```




```
# model building: testing for the presence of random slopes (with REML=T), e.g.:
model6 <- lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR.c*Age.c +
               CountryEduYrs + NrLang  + (1|Country),
               data=accents, REML=T)

model6b <- lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR.c*Age.c  +
                         CountryEduYrs + NrLang + (1|Country) +
                         (0+LR.c|Country),
                data=accents, REML=T)

AIC(model6) - AIC(model6b)

## [1] -1.992

anova(model6, model6b, refit=F)

## Data: accents
## Models:
## model6: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR.c * Age.c +
## model6:     CountryEduYrs + NrLang + (1 | Country)
## model6b: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR.c * Age.c +
## model6b:     CountryEduYrs + NrLang + (1 | Country) + (0 + LR.c | Country)
##         Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model6  12 -449 -393    236     -473
## model6b 13 -447 -386    236     -473  0.01      1       0.93

# no random slopes / correlation parameters are significant (not shown)
```

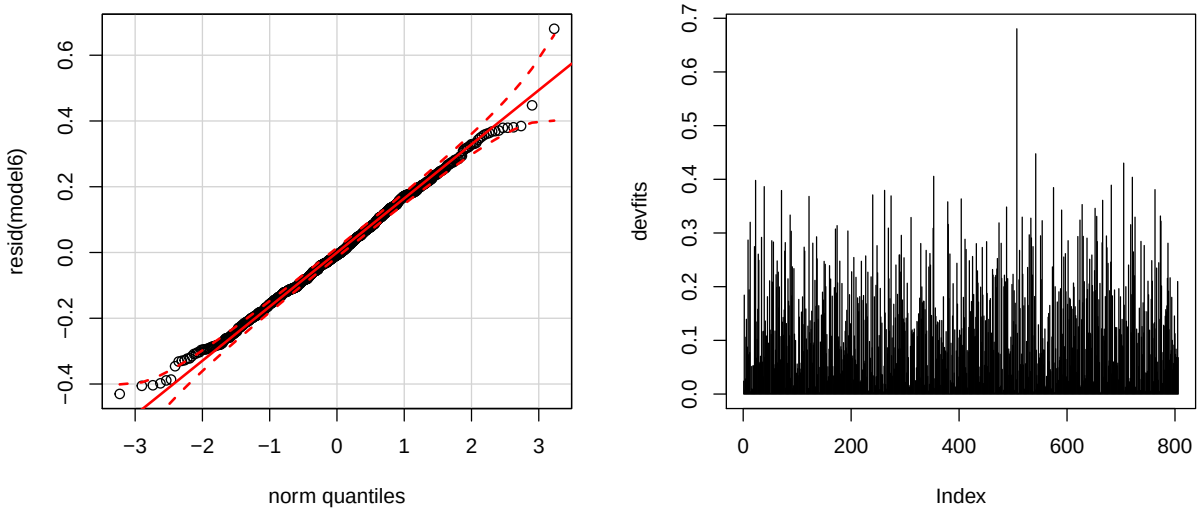
```r
# summary of the current best model
summary(model6)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR.c * Age.c +
##     CountryEduYrs + NrLang + (1 | Country)
##    Data: accents
##
## REML criterion at convergence: -473.1
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -2.622 -0.677 -0.049  0.677  4.148
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Country  (Intercept) 0.00246  0.0496
##  Residual             0.02691  0.1641
## Number of obs: 806, groups:  Country, 139
##
## Fixed effects:
##                                                Estimate Std. Error t value
## (Intercept)                                   -4.62e+00   2.61e-02  -177.4
## LR.c                                          -6.24e-03   8.81e-04    -7.1
## Age.c                                          2.44e-03   6.53e-04     3.7
## CountryEduYrs                                 -2.03e-02   2.80e-03    -7.3
## NrLang                                        -1.59e-02   5.57e-03    -2.9
## LR.c:Age.c                                     1.28e-04   3.44e-05     3.7
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE  4.42e-02   9.72e-03     4.5
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE   1.46e-02   2.00e-03     7.3
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE   5.61e-03   1.23e-03     4.6
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE    9.32e-03   1.73e-03     5.4
##
## Correlation of Fixed Effects:
##                     (Intr) LR.c   Age.c  CntrEY NrLang LR.:A. SAEO:IIESFALSE:PBPF
## LR.c                -0.054
## Age.c                0.074 -0.468
## ContryEdYrs         -0.848  0.106 -0.024
## NrLang              -0.272  0.071 -0.142 -0.014
## LR.c:Age.c          -0.020 -0.629 -0.021 -0.118 -0.008
## SAEO:IIESFALSE:PBPF  0.158  0.197 -0.154 -0.033  0.036 -0.142
## SAEO:IIESTRUE:PBPF   0.144  0.025 -0.076  0.115  0.057 -0.037  0.201
## SAEO:IIESFALSE:PBPT -0.356  0.099 -0.127  0.159  0.021  0.004 -0.195
## SAEO:IIESTRUE:PBPT  -0.069  0.084 -0.258 -0.084  0.061  0.046 -0.043
##                     SAEO:IIESTRUE:PBPF SAEO:IIESFALSE:PBPT
## LR.c
## Age.c
## ContryEdYrs
## NrLang
## LR.c:Age.c
## SAEO:IIESFALSE:PBPF
```

```
## SAEO:IIESTRUE:PBPF
## SAEO:IIESFALSE:PBPT -0.280
## SAEO:IIESTRUE:PBPT  -0.209              0.158
```

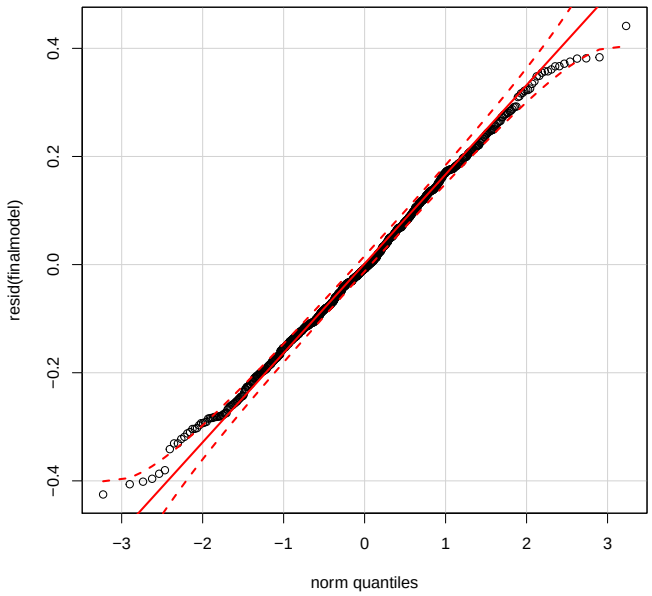
3.6 Model criticism

```
# model criticism: 1 clear outlier
par(mfrow=c(1,2))
qqp(resid(model6))
devfits = abs(resid(model6,"deviance"))
plot(devfits,type="h")
```

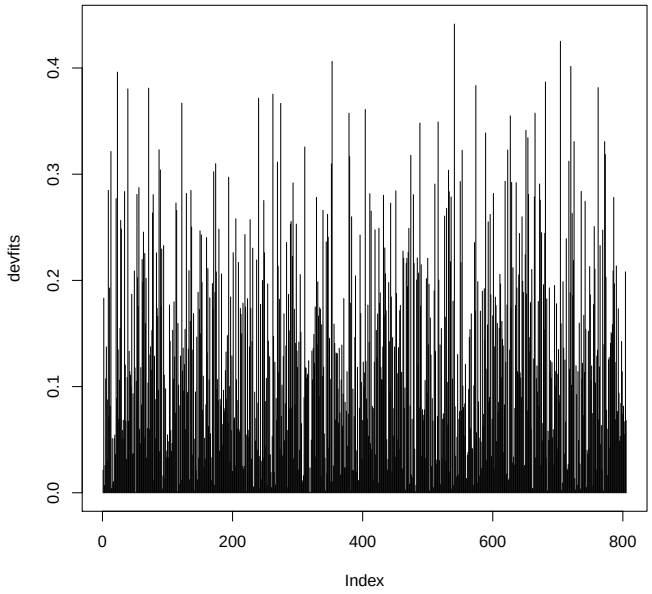



```
# exclude outlier and refit
accents2 = accents[-c(which(devfits==max(devfits))),]
nrow(accents) - nrow(accents2) # number of outliers

## [1] 1

finalmodel <- lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR.c*Age.c  +
                           CountryEduYrs + NrLang + (1|Country),
                   data=accents2, REML=T)

# performance increase
cor(fitted(model6),accents$LD.log)^2

## [1] 0.4246

cor(fitted(finalmodel),accents2$LD.log)^2
```

```
## [1] 0.4363

# residuals normally distributed
qqp(resid(finalmodel))
```




```
devfits = abs(resid(finalmodel,"deviance"))
plot(devfits,type="h")
```



3.7 Best model with breakpoints

```
# summary of the final model
summary(finalmodel)$coefficients

##                                                    Estimate Std. Error  t value
## (Intercept)                                      -4.6250052  2.634e-02 -175.612
## LR.c                                             -0.0061675  8.744e-04   -7.053
## Age.c                                             0.0023163  6.470e-04    3.580
## CountryEduYrs                                    -0.0202782  2.838e-03   -7.145
## NrLang                                           -0.0153671  5.517e-03   -2.785
## LR.c:Age.c                                        0.0001296  3.406e-05    3.804
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE  0.0438424  9.608e-03    4.563
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE   0.0142658  1.988e-03    7.177
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE   0.0057672  1.216e-03    4.742
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE    0.0094959  1.709e-03    5.557
```

```
# approximate test (on linear model) for heteroscedasticity
m <- lm(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR.c*Age.c  +
        CountryEduYrs + NrLang + Country, data=accents2)
ncvTest(m) # no heteroscedasticity

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.992    Df = 1     p = 0.1582
```

```
# determine effect size of interaction by calculating the range of the total effect
mx = -Inf
mn = Inf
accents2$LR.c_Age.c = accents2$LR.c * accents2$Age.c
fx = fixef(finalmodel)
for (i in 1:nrow(accents2)) {
    val = fx["LR.c"]*accents2[i,"LR.c"] +
          fx["Age.c"]*accents2[i,"Age.c"] +
          fx["LR.c:Age.c"]*accents2[i,"LR.c_Age.c"]
    names(val) = "LR.c*Age.c"
    if (val < mn) {
        mn = val
    }
    if (val > mx) {
        mx = val
    }
}
round(mx - mn,3)

## LR.c*Age.c
##      0.272

# effect size of AEO (separate for IE vs. non-IE speakers and before/after breakpoint)
```

```
subdat = accents2[accents2$IsIESpeaker==F & accents2$PastBreakPoint==F,]
round((max(subdat$ShiftedAEO) - min(subdat$ShiftedAEO))*
       fixef(finalmodel)["ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE"],10)

## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE
##                                          0.2192

subdat = accents2[accents2$IsIESpeaker==T & accents2$PastBreakPoint==F,]
round((max(subdat$ShiftedAEO) - min(subdat$ShiftedAEO))*
       fixef(finalmodel)["ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE"],10)

## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE
##                                         0.1997

subdat = accents2[accents2$IsIESpeaker==F & accents2$PastBreakPoint==T,]
round((max(subdat$ShiftedAEO) - min(subdat$ShiftedAEO))*
       fixef(finalmodel)["ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE"],10)

## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE
##                                          0.2192

subdat = accents2[accents2$IsIESpeaker==T & accents2$PastBreakPoint==T,]
round((max(subdat$ShiftedAEO) - min(subdat$ShiftedAEO))*
       fixef(finalmodel)["ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE"],10)

## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE
##                                         0.3419

# effect size of the other variables
round((max(accents2$CountryEduYrs) - min(accents2$CountryEduYrs))*
       fixef(finalmodel)["CountryEduYrs"],3)

## CountryEduYrs
##        -0.231

round((max(accents2$NrLang) - min(accents2$NrLang))*
       fixef(finalmodel)["NrLang"],3)

## NrLang
## -0.077
```

3.8 Model with breakpoints better than model without

```
# model building process not shoum
modelNoBP <- lmer(LD.log ~ IsIESpeaker + IsIESpeaker:AEO + LR.c*Age.c +
                           IsNaturalLearner + CountryEduYrs + NrLang +
                           (1+IsNaturalLearner|Country),
                  data=accents2, REML=T)
summary(modelNoBP)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: LD.log ~ IsIESpeaker + IsIESpeaker:AEO + LR.c * Age.c + IsNaturalLearner...
##     CountryEduYrs + NrLang + (1 + IsNaturalLearner | Country)
##    Data: accents2
##
## REML criterion at convergence: -493.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4829 -0.6703 -0.0458  0.6480  2.5727
##
## Random effects:
##  Groups   Name                 Variance Std.Dev. Corr
##  Country  (Intercept)          0.00237  0.0487
##           IsNaturalLearnerTRUE 0.01294  0.1137   -0.16
##  Residual                      0.02546  0.1596
## Number of obs: 805, groups:  Country, 139
##
## Fixed effects:
##                       Estimate Std. Error t value
## (Intercept)          -4.69e+00   2.91e-02  -161.4
## IsIESpeakerTRUE      -1.23e-01   2.62e-02    -4.7
## LR.c                 -5.73e-03   9.22e-04    -6.2
## Age.c                 2.14e-03   6.57e-04     3.3
## IsNaturalLearnerTRUE -5.30e-02   2.45e-02    -2.2
## CountryEduYrs        -2.04e-02   2.86e-03    -7.1
## NrLang               -1.65e-02   5.49e-03    -3.0
## IsIESpeakerFALSE:AEO  7.88e-03   1.27e-03     6.2
## IsIESpeakerTRUE:AEO   1.21e-02   1.23e-03     9.9
## LR.c:Age.c            1.24e-04   3.49e-05     3.6
##
## Correlation of Fixed Effects:
##             (Intr) IsIESTRUE LR.c   Age.c  INLTRU CntrEY NrLang IIESFA IIESTRUE:
## IsIESpkTRUE -0.328
## LR.c        -0.148  0.116
## Age.c        0.163 -0.011    -0.499
## IsNtrlLTRUE -0.010  0.048    -0.192  0.098
## ContryEdYrs -0.697 -0.160     0.071  0.014 -0.061
## NrLang      -0.253 -0.065     0.069 -0.138 -0.030 -0.002
## IIESFALSE:A -0.571  0.546     0.226 -0.231 -0.043  0.047  0.029
## IIESTRUE:AE -0.040 -0.521     0.136 -0.300 -0.107 -0.041  0.087  0.085
## LR.c:Age.c   0.024 -0.067    -0.635  0.015  0.082 -0.100 -0.007 -0.069 -0.001

# full model with breakpoints is better than the full model with breakpoints
# (random-effects structure has been made identical for a valid comparison)
modelNoBP.comp <- lmer(LD.log ~ IsIESpeaker + IsIESpeaker:AEO + LR.c*Age.c +
                               IsNaturalLearner + CountryEduYrs + NrLang +
                               (1+IsNaturalLearner|Country),
                       data=accents2, REML=F)

finalmodel.comp <- lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint +
```

```
                                LR.c*Age.c + CountryEduYrs + NrLang +
                                (1+IsNaturalLearner|Country),
                        data=accents2, REML=F)

AIC(modelNoBP.comp) - AIC(finalmodel.comp)

## [1] 6.92

anova(modelNoBP.comp,finalmodel.comp)

## Data: accents2
## Models:
## modelNoBP.comp: LD.log ~ IsIESpeaker + IsIESpeaker:AEO + LR.c * Age.c + IsNatural...
## modelNoBP.comp:     CountryEduYrs + NrLang + (1 + IsNaturalLearner | Country)
## finalmodel.comp: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR.c * Age.c +
## finalmodel.comp:     CountryEduYrs + NrLang + (1 + IsNaturalLearner | Country)
##                 Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## modelNoBP.comp  14 -572 -506    300     -600
## finalmodel.comp 14 -579 -513    303     -607  6.92      0     <2e-16
```

4 Validation of results using native-likeness ratings

4.1 Data preparation

```
# subset data
nat = accents[!is.na(accents$Nativelikeness),]
nat = droplevels(nat)
natIE = nat[nat$IsIESpeaker==T,]
natIE = droplevels(natIE)
natNIE = nat[nat$IsIESpeaker==F,]
natNIE = droplevels(natNIE)
dim(nat)

## [1] 272  21
```

4.2 Testing if a log-transformation is necessary

```
# approximate test (on linear model) for heteroscedasticity using native-likeness
m <- lm(Nativelikeness ~ AEO, data=nat)
ncvTest(m) # severe heteroscedasticity

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 15.07    Df = 1     p = 0.0001037

# results when log-transforming nativelikeness
m <- lm(log(Nativelikeness) ~ AEO, data=nat)
ncvTest(m) # no heteroscedasticity

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.985    Df = 1     p = 0.1589
```

4.3 Assessing which random-effect factors are needed

```
# testing random intercepts
m0 = lmer(log(Nativelikeness) ~ AEO + (1|Country), data=nat)
m = lmer(log(Nativelikeness) ~ AEO + (1|Country) + (1|Language), data=nat)

# language is not needed as a random intercept
AIC(m0) - AIC(m)

## [1] -2

anova(m0, m, refit=F)
```

```
## Data: nat
## Models:
## m0: log(Nativelikeness) ~ AEO + (1 | Country)
## m: log(Nativelikeness) ~ AEO + (1 | Country) + (1 | Language)
##    Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m0  4 224 238   -108      216
## m   5 226 244   -108      216     0      1          1
```

4.4 Breakpoint analysis: single breakpoint

```
# determine breakpoint by iterating over possible breakpoints between an AEO of 1 and 30
deviances = rep(Inf, 30)

for (i in (min(nat$AEO)+1):min(30,max(nat$AEO)-1)) {
    breakpoint = i
    nat$ShiftedAEO = nat$AEO - breakpoint;
    nat$PastBreakPoint = as.factor(nat$ShiftedAEO > 0)
    m = lmer(log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1|Country),
             data=nat, REML=F)
    deviances[i] = deviance(m)
}

breakpoint = which(deviances == min(deviances))
breakpoint

## [1] 6
```

```
# model parameters of model with breakpoint
nat$ShiftedAEO = nat$AEO - breakpoint
nat$PastBreakPoint = as.factor(nat$ShiftedAEO > 0)
m1 = lmer(log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1|Country), data=nat,
          REML=F)
summary(m1)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1 | Country)
##    Data: nat
##
##      AIC      BIC   logLik deviance df.resid
##    196.3    214.4    -93.2    186.3      267
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2489 -0.5989 -0.0081  0.6641  2.4312
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Country  (Intercept) 0.0297   0.172
```

```
##  Residual             0.0953   0.309
## Number of obs: 272, groups:  Country, 94
##
## Fixed effects:
##                                Estimate Std. Error t value
## (Intercept)                     1.09414    0.03844   28.47
## ShiftedAEO:PastBreakPointFALSE -0.11412    0.02292   -4.98
## ShiftedAEO:PastBreakPointTRUE  -0.01924    0.00377   -5.11
##
## Correlation of Fixed Effects:
##             (Intr) SAEO:PBPF
## SAEO:PBPFAL  0.403
## SAEO:PBPTRU -0.659 -0.329
```

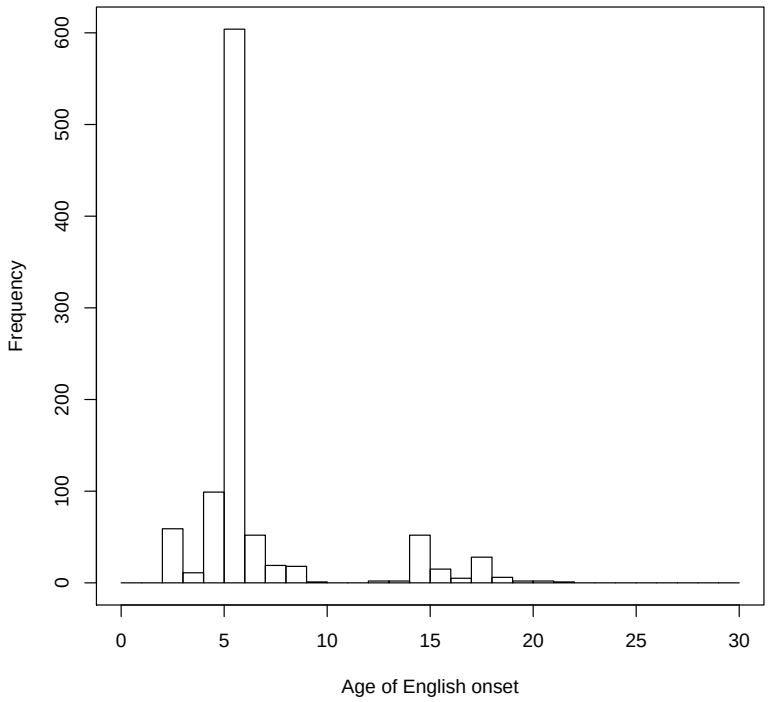
```
# compare models with and without breakpoint (the AIC reduction should be >= 2)
m0 = lmer(log(Nativelikeness) ~ AEO + (1|Country), data=nat, REML=F)
AIC(m0) - AIC(m1)

## [1] 12.65

anova(m0,m1)

## Data: nat
## Models:
## m0: log(Nativelikeness) ~ AEO + (1 | Country)
## m1: log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1 | Country)
##    Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m0  4 209 223 -100.5      201
## m1  5 196 214  -93.2      186  14.6      1    0.00013
```

4.4.1 Breakpoint validation: bootstrapping

```
# validating the breakpoint by bootstrapping (1000 iterations)
breakpointsNat = rep(0, 1000)
AICvalsNat = rep(NA,1000)
pvalsNat = rep(NA,1000)
for (j in 1:1000) { # 1000 iterations
    dat = nat[c(sample(nrow(nat), nrow(nat), replace=T)), ]

    deviances = rep(Inf, 30)
    for (i in (min(dat$AEO)+1):min(30,max(dat$AEO)-1)) {
        breakpoint = i
        dat$ShiftedAEO = dat$AEO - breakpoint
        dat$PastBreakPoint = as.factor(dat$ShiftedAEO > 0)
        m = lmer(log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1|Country),
                 data=dat, REML=F)
        deviances[i] = deviance(m)
    }
```

```
    breakpointsNat[j] = which(deviances == min(deviances))

    # evaluate significance of the best model for this iteration
    breakpoint = breakpointsNat[j]
    dat$ShiftedAEO = dat$AEO - breakpoint
    dat$PastBreakPoint = as.factor(dat$ShiftedAEO > 0)
    m = lmer(log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1|Country),
             data=dat, REML=F)
    m0 = lmer(log(Nativelikeness) ~ AEO + (1|Country), data=dat, REML=F)
    AICvalsNat[j] = AIC(m0) - AIC(m) # AIC comparison
    pvalsNat[j] = anova(m0,m)$"Pr(>Chisq)"[2]
}

# saved as computation takes 30 mins.
save(breakpointsNat,file='results/breakpointsNat.rda')
save(AICvalsNat,file='results/AICvalsNat.rda')
save(pvalsNat,file='results/pvalsNat.rda')
```

```
# breakpoints are not stable
load('results/breakpointsNat.rda')
load('results/AICvalsNat.rda')
load('results/pvalsNat.rda')
table(breakpointsNat)

## breakpointsNat
##   2   3   4   5   6   7   8   9  10  13  14  15  16  17  18  19  20  21  22
##   1  62  13 100 609  53  19  18   1   2   2  55  16   6  29   8   2   3   1

# number of significant breakpoints
sum(AICvalsNat >= 2)

## [1] 978

sum(pvalsNat < 0.05)

## [1] 979

# visualization of significant breakpoints
sigbpNat = breakpointsNat[which(AICvalsNat>=2)]
hist(sigbpNat,main='',xlab='Age of English onset',breaks=seq(0,30)); box()
```



4.5 Breakpoint analysis: two breakpoints

```
# determine separate breakpoints for IE as opposed to non-IE speakers
# by iterating over possible breakpoints between an AEO of 1 and 30

# IE speakers
deviances = rep(Inf, 30)

for (i in (min(natIE$AEO)+1):min(30,max(natIE$AEO)-1)) {
    breakpoint = i
    natIE$ShiftedAEO = natIE$AEO - breakpoint;
    natIE$PastBreakPoint = as.factor(natIE$ShiftedAEO > 0)
    m = lmer(log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1|Country),
             data=natIE, REML=F)
    deviances[i] = deviance(m)
}

breakpointIE = which(deviances == min(deviances))
breakpointIE

## [1] 18

# non-IE speakers
deviances = rep(Inf, 30)
```

```
for (i in (min(natNIE$AEO)+1):min(30,max(natNIE$AEO)-1)) {
    breakpoint = i
    natNIE$ShiftedAEO = natNIE$AEO - breakpoint;
    natNIE$PastBreakPoint = as.factor(natNIE$ShiftedAEO > 0)
    m = lmer(log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1|Country),
             data=natNIE, REML=F)
    deviances[i] = deviance(m)
}

breakpointNIE = which(deviances == min(deviances))
breakpointNIE

## [1] 6
```

```
# model parameters of two separate models with breakpoint
natIE$ShiftedAEO = natIE$AEO - breakpointIE
natIE$PastBreakPoint = as.factor(natIE$ShiftedAEO > 0)
mIE = lmer(log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1|Country), data=natIE,
           REML=F)
summary(mIE)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1 | Country)
##    Data: natIE
##
##      AIC      BIC   logLik deviance df.resid
##     70.1     85.0    -30.0     60.1      142
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1514 -0.6925 -0.0179  0.6604  2.5285
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Country  (Intercept) 0.0181   0.134
##  Residual             0.0746   0.273
## Number of obs: 147, groups:  Country, 48
##
## Fixed effects:
##                                Estimate Std. Error t value
## (Intercept)                     0.88340    0.05516   16.02
## ShiftedAEO:PastBreakPointFALSE -0.03268    0.00617   -5.30
## ShiftedAEO:PastBreakPointTRUE   0.00195    0.01068    0.18
##
## Correlation of Fixed Effects:
##             (Intr) SAEO:PBPF
## SAEO:PBPFAL  0.812
## SAEO:PBPTRU -0.347 -0.314

# model parameters of model with breakpoint
```

```
natNIE$ShiftedAEO = natNIE$AEO - breakpointNIE
natNIE$PastBreakPoint = as.factor(natNIE$ShiftedAEO > 0)
mNIE = lmer(log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1|Country), data=natNIE,
            REML=F)
summary(mNIE)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1 | Country)
##    Data: natNIE
##
##      AIC      BIC   logLik deviance df.resid
##    101.1    115.2    -45.5     91.1      120
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -2.164 -0.494 -0.009  0.695  2.352
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Country  (Intercept) 0.0137   0.117
##  Residual             0.1097   0.331
## Number of obs: 125, groups:  Country, 52
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                      0.93740    0.05092   18.41
## ShiftedAEO:PastBreakPointFALSE -0.17429    0.03024   -5.76
## ShiftedAEO:PastBreakPointTRUE  -0.01801    0.00498   -3.62
##
## Correlation of Fixed Effects:
##             (Intr) SAEO:PBPF
## SAEO:PBPFAL  0.444
## SAEO:PBPTRU -0.661 -0.330
```

```
# comparable results when creating a single model with two separate breakpoints
deviances = rep(Inf, 30^2)
dim(deviances) = c(30,30)
for (i in (min(natIE$AEO)+1):min(30,max(natIE$AEO)-1)) {
    for (j in (min(natNIE$AEO)+1):min(30,max(natNIE$AEO)-1)) {
        breakpointIE = i
        breakpointNIE = j
        nat$ShiftedAEO = NA
        nat[nat$IsIESpeaker==T,]$ShiftedAEO =
            nat[nat$IsIESpeaker==T,]$AEO - breakpointIE
        nat[nat$IsIESpeaker==F,]$ShiftedAEO =
            nat[nat$IsIESpeaker==F,]$AEO - breakpointNIE
        nat$PastBreakPoint = as.factor(nat$ShiftedAEO > 0)

        m = lmer(log(Nativelikeness) ~ ShiftedAEO:IsIESpeaker:PastBreakPoint +
                 (1|Country), data=nat, REML=F)
```

```
        deviances[i,j] = deviance(m)
    }
}

breakpointIE = which(deviances == min(deviances),arr.ind=T)[1]
breakpointNIE = which(deviances == min(deviances),arr.ind=T)[2]

# breakpoint for Indo-European speakers
breakpointIE

## [1] 17

# breakpoint for non-Indo-European speakers
breakpointNIE

## [1] 6
```

```
# model parameters of model with two breakpoints
nat$ShiftedAEO = NA
nat[nat$IsIESpeaker==T,]$ShiftedAEO =
    nat[nat$IsIESpeaker==T,]$AEO - breakpointIE
nat[nat$IsIESpeaker==F,]$ShiftedAEO =
    nat[nat$IsIESpeaker==F,]$AEO - breakpointNIE
nat$PastBreakPoint = as.factor(nat$ShiftedAEO > 0)
m2 = lmer(log(Nativelikeness) ~ ShiftedAEO:IsIESpeaker:PastBreakPoint +
          (1|Country), data=nat, REML=F)
summary(m2)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: log(Nativelikeness) ~ ShiftedAEO:IsIESpeaker:PastBreakPoint +
##     (1 | Country)
##    Data: nat
##
##      AIC      BIC   logLik deviance df.resid
##    168.7    193.9    -77.4    154.7      265
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2798 -0.6228 -0.0095  0.6318  2.6339
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Country  (Intercept) 0.0171   0.131
##  Residual             0.0902   0.300
## Number of obs: 272, groups:  Country, 94
##
## Fixed effects:
##                                                Estimate Std. Error t value
## (Intercept)                                     0.93000    0.03673   25.32
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE -0.17888    0.02672   -6.70
```

```
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE  -0.03081    0.00546   -5.64
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE  -0.01708    0.00417   -4.10
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE   -0.00283    0.01037   -0.27
##
## Correlation of Fixed Effects:
##                     (Intr) SAEO:IIESFALSE:PBPF SAEO:IIESTRUE:PBPF SAEO:IIESFALSE:...
## SAEO:IIESFALSE:PBPF  0.349                                                      ...
## SAEO:IIESTRUE:PBPF   0.649  0.246                                               ...
## SAEO:IIESFALSE:PBPT -0.542 -0.243              -0.358                           ...
## SAEO:IIESTRUE:PBPT  -0.249 -0.089              -0.200              0.143
```

```
# compare model with two breakpoints to the model with one breakpoint (m1)
AIC(m1) - AIC(m2)

## [1] 27.64

anova(m1,m2)

## Data: nat
## Models:
## m1: log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1 | Country)
## m2: log(Nativelikeness) ~ ShiftedAEO:IsIESpeaker:PastBreakPoint +
## m2:     (1 | Country)
##    Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m1  5 196 214  -93.2      186
## m2  7 169 194  -77.4      155  31.6      2    1.3e-07

# compare model with two breakpoints to a model with only a different effect of
# AEO for Indo-European as compared to non-Indo-European speakers
m0alt = lmer(log(Nativelikeness) ~ AEO*IsIESpeaker + (1|Country),
             data=nat, REML=F)
AIC(m0alt) - AIC(m2)

## [1] 28.55

anova(m0alt,m2)

## Data: nat
## Models:
## m0alt: log(Nativelikeness) ~ AEO * IsIESpeaker + (1 | Country)
## m2: log(Nativelikeness) ~ ShiftedAEO:IsIESpeaker:PastBreakPoint +
## m2:     (1 | Country)
##       Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m0alt  6 197 219  -92.6      185
## m2     7 169 194  -77.4      155  30.6      1    3.2e-08

# test if the individual breakpoint is needed for the non-Indo-European speakers
m0nonIE = lmer(log(Nativelikeness) ~ AEO + (1|Country),
               data=natNIE, REML=F)
AIC(m0nonIE) - AIC(mNIE)
```

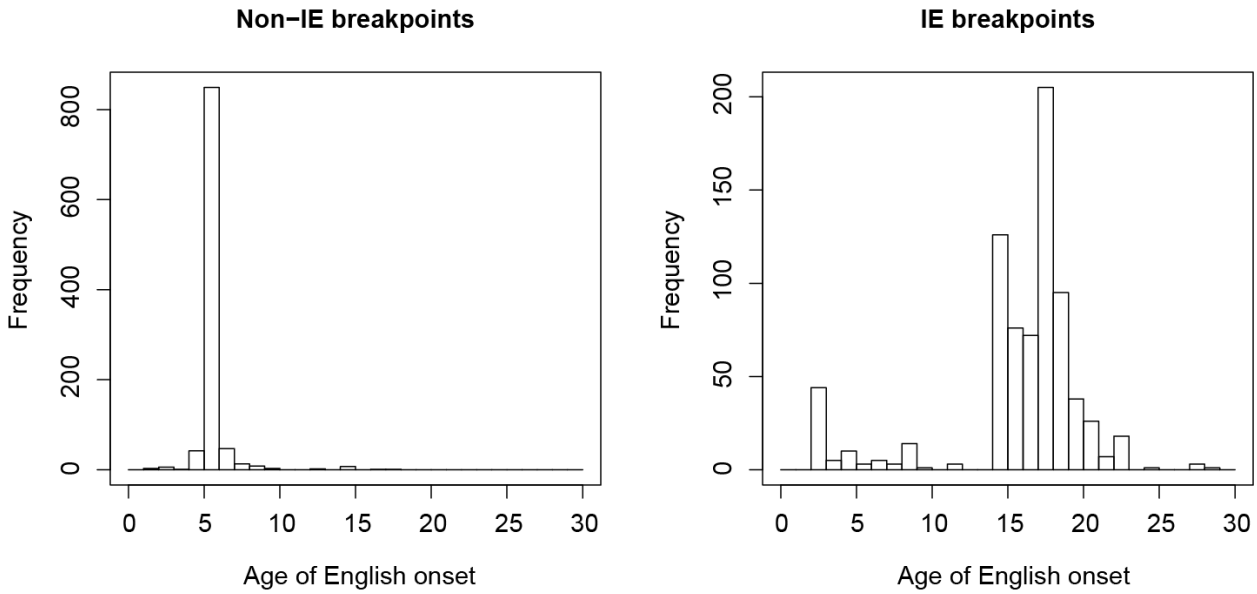
```
## [1] 19.45

anova(m0nonIE,mNIE)

## Data: natNIE
## Models:
## m0nonIE: log(Nativelikeness) ~ AEO + (1 | Country)
## mNIE: log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1 | Country)
##         Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m0nonIE  4 120 132  -56.3    112.5
## mNIE     5 101 115  -45.5     91.1  21.4      1    3.6e-06

# test if the individual breakpoint is also needed for the Indo-European speakers
m0IE = lmer(log(Nativelikeness) ~ AEO + (1|Country),
            data=natIE, REML=F)
AIC(m0IE) - AIC(mIE)

## [1] 4.004

anova(m0IE,mIE)

## Data: natIE
## Models:
## m0IE: log(Nativelikeness) ~ AEO + (1 | Country)
## mIE: log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1 | Country)
##      Df  AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m0IE  4 74.1  86    -33     66.1
## mIE   5 70.1  85    -30     60.1     6      1      0.014
```

4.5.1 Breakpoint validation: bootstrapping

```
# validating the breakpoint for IE speakers by bootstrapping (1000 iterations)
breakpointsNatIE = rep(0, 1000)
AICvalsNatIE = rep(NA,1000)
pvalsNatIE = rep(NA,1000)
for (j in 1:1000) { # 1000 iterations
    dat = natIE[c(sample(nrow(natIE), nrow(natIE), replace=T)), ]

    deviances = rep(Inf, 30)
    for (i in (min(dat$AEO)+1):min(30,max(dat$AEO)-1)) {
        breakpointIE = i
        dat$ShiftedAEO = dat$AEO - breakpointIE
        dat$PastBreakPoint = as.factor(dat$ShiftedAEO > 0)
        m = lmer(log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1|Country),
                 data=dat, REML=F)
        deviances[i] = deviance(m)
    }
    breakpointsNatIE[j] = which(deviances == min(deviances))
```

```
    # evaluate significance of the best model for this iteration
    breakpointIE = breakpointsNatIE[j]
    dat$ShiftedAEO = dat$AEO - breakpointIE
    dat$PastBreakPoint = as.factor(dat$ShiftedAEO > 0)
    m = lmer(log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1|Country),
             data=dat, REML=F)
    m0 = lmer(log(Nativelikeness) ~ AEO + (1|Country), data=dat, REML=F)
    AICvalsNatIE[j] = AIC(m0) - AIC(m) # AIC comparison
    pvalsNatIE[j] = anova(m0,m)$"Pr(>Chisq)"[2]
}

# saved as computation takes 30 mins.
save(breakpointsNatIE,file='results/breakpointsNatIE.rda')
save(AICvalsNatIE,file='results/AICvalsNatIE.rda')
save(pvalsNatIE,file='results/pvalsNatIE.rda')


# validating the breakpoint for NIE speakers by bootstrapping (1000 iterations)
breakpointsNatNIE = rep(0, 1000)
AICvalsNatNIE = rep(NA,1000)
pvalsNatNIE = rep(NA,1000)
for (j in 1:1000) { # 1000 iterations
    dat = natNIE[c(sample(nrow(natNIE), nrow(natNIE), replace=T)), ]

    deviances = rep(Inf, 30)
    for (i in (min(dat$AEO)+1):min(30,max(dat$AEO)-1)) {
        breakpointNIE = i
        dat$ShiftedAEO = dat$AEO - breakpointNIE
        dat$PastBreakPoint = as.factor(dat$ShiftedAEO > 0)
        m = lmer(log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1|Country),
                 data=dat, REML=F)
        deviances[i] = deviance(m)
    }
    breakpointsNatNIE[j] = which(deviances == min(deviances))

    # evaluate significance of the best model for this iteration
    breakpointNIE = breakpointsNatNIE[j]
    dat$ShiftedAEO = dat$AEO - breakpointNIE
    dat$PastBreakPoint = as.factor(dat$ShiftedAEO > 0)
    m = lmer(log(Nativelikeness) ~ ShiftedAEO:PastBreakPoint + (1|Country),
             data=dat, REML=F)
    m0 = lmer(log(Nativelikeness) ~ AEO + (1|Country), data=dat, REML=F)
    AICvalsNatNIE[j] = AIC(m0) - AIC(m) # AIC comparison
    pvalsNatNIE[j] = anova(m0,m)$"Pr(>Chisq)"[2]
}

# saved as computation takes 30 mins.
save(breakpointsNatNIE,file='results/breakpointsNatNIE.rda')
save(AICvalsNatNIE,file='results/AICvalsNatNIE.rda')
save(pvalsNatNIE,file='results/pvalsNatNIE.rda')
```

```r
# breakpoints are (again) not stable
load('results/breakpointsNatNIE.rda')
load('results/AICvalsNatNIE.rda')
load('results/pvalsNatNIE.rda')
load('results/breakpointsNatIE.rda')
load('results/AICvalsNatIE.rda')
load('results/pvalsNatIE.rda')

table(breakpointsNatNIE)

## breakpointsNatNIE
##   2   3   4   5   6   7   8   9  10  11  13  15  17  18
##   8   7   1  43 855  47  13   8   3   1   2   8   1   3

# number of significant breakpoints
sum(AICvalsNatNIE >= 2)

## [1] 983

sum(pvalsNatNIE < 0.05)

## [1] 986

# number of significant breakpoints
sum(AICvalsNatIE >= 2)

## [1] 756

sum(pvalsNatIE < 0.05)

## [1] 775

# distribution of non significant breakpoints
table(breakpointsNatIE[which(AICvalsNatIE<2)])

##
##  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 27 30
## 39  6  5  5  2  3 20  3  2 10  1  3 29 17 19 41 17  9  2  1  8  1  1

# visualize significant breakpoints
par(mfrow=c(1,2))
sigbpNatNIE = breakpointsNatNIE[which(AICvalsNatNIE>=2)]
hist(sigbpNatNIE,main='Non-IE breakpoints',xlab='Age of English onset',
     breaks=seq(0,30)); box()

sigbpNatIE = breakpointsNatIE[which(AICvalsNatIE>=2)]
hist(sigbpNatIE,main='IE breakpoints',xlab='Age of English onset',
     breaks=seq(0,30)); box()
```



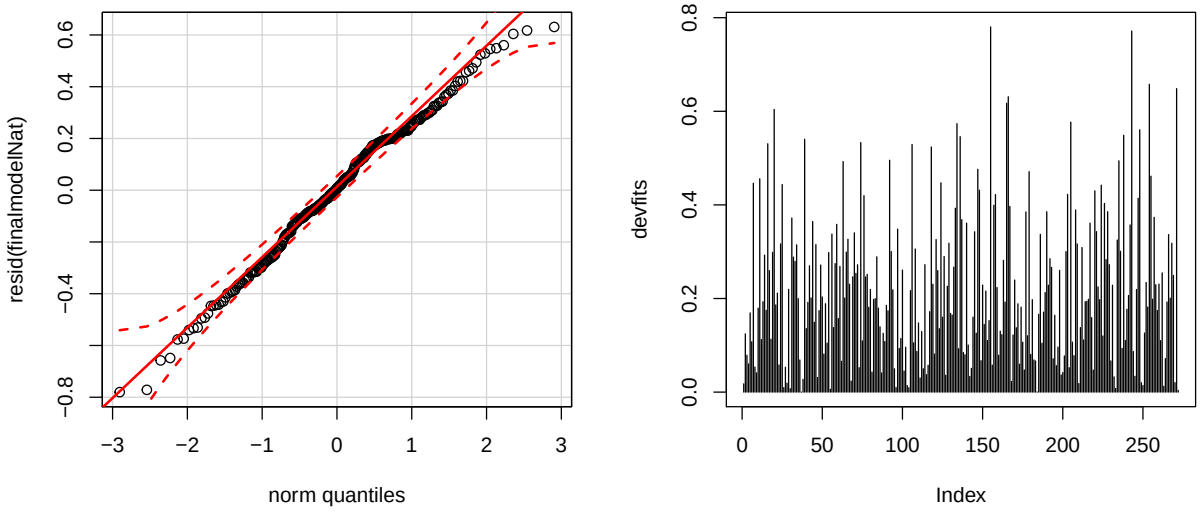
4.6 Best model with breakpoints

```
breakpointIE = 17
breakpointNIE = 6
nat$ShiftedAEO = NA
nat[nat$IsIESpeaker==T,]$ShiftedAEO = nat[nat$IsIESpeaker==T,]$AEO - breakpointIE
nat[nat$IsIESpeaker==F,]$ShiftedAEO = nat[nat$IsIESpeaker==F,]$AEO - breakpointNIE
nat$PastBreakPoint = as.factor(nat$ShiftedAEO > 0)
finalmodelNat <- lmer(log(Nativelikeness) ~ ShiftedAEO:IsIESpeaker:PastBreakPoint +
                                            LR.c*Age.c + IsNaturalLearner +
                                            CountryEduYrs + NrLang + (1|Country),
                      data=nat, REML=F)

summary(finalmodelNat)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: log(Nativelikeness) ~ ShiftedAEO:IsIESpeaker:PastBreakPoint +
##     LR.c * Age.c + IsNaturalLearner + CountryEduYrs + NrLang +      (1 | Country)
##    Data: nat
##
##      AIC      BIC   logLik deviance df.resid
##    111.9    158.8    -43.0     85.9      259
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8326 -0.6167  0.0286  0.7171  2.2908
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Country  (Intercept) 0.00497  0.0705
##  Residual             0.07587  0.2754
```

```
## Number of obs: 272, groups:  Country, 94
##
## Fixed effects:
##                                                 Estimate Std. Error t value
## (Intercept)                                     0.639655   0.064907    9.85
## LR.c                                            0.015542   0.002685    5.79
## Age.c                                          -0.004518   0.001914   -2.36
## IsNaturalLearnerTRUE                            0.130827   0.059229    2.21
## CountryEduYrs                                   0.030286   0.006803    4.45
## NrLang                                          0.077838   0.015789    4.93
## LR.c:Age.c                                     -0.000533   0.000150   -3.55
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointFALSE -0.115203   0.026011   -4.43
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointFALSE  -0.019787   0.004961   -3.99
## ShiftedAEO:IsIESpeakerFALSE:PastBreakPointTRUE  -0.015496   0.003743   -4.14
## ShiftedAEO:IsIESpeakerTRUE:PastBreakPointTRUE   -0.017170   0.010016   -1.71
##
## Correlation of Fixed Effects:
##                     (Intr) LR.c   Age.c  INLTRU CntrEY NrLang LR.:A.
## LR.c                -0.053
## Age.c                0.091 -0.489
## IsNtrlLTRUE          0.023 -0.285  0.187
## ContryEdYrs         -0.810  0.162 -0.036 -0.084
## NrLang              -0.272  0.122 -0.117  0.002  0.002
## LR.c:Age.c          -0.140 -0.568  0.001  0.171 -0.090 -0.016
## SAEO:IIESFALSE:PBPF  0.216  0.219 -0.112  0.244 -0.049  0.079 -0.220
## SAEO:IIESTRUE:PBPF   0.130 -0.003 -0.021  0.187  0.170  0.164 -0.062
## SAEO:IIESFALSE:PBPT -0.363  0.124 -0.135 -0.155  0.117 -0.006  0.019
## SAEO:IIESTRUE:PBPT   0.012  0.115 -0.166 -0.311 -0.075 -0.095 -0.160
##                     SAEO:IIESFALSE:PBPF SAEO:IIESTRUE:PBPF SAEO:IIESFALSE:PBPT
## LR.c
## Age.c
## IsNtrlLTRUE
## ContryEdYrs
## NrLang
## LR.c:Age.c
## SAEO:IIESFALSE:PBPF
## SAEO:IIESTRUE:PBPF   0.294
## SAEO:IIESFALSE:PBPT -0.240              -0.379
## SAEO:IIESTRUE:PBPT  -0.116              -0.235              0.170
```

4.7 Model criticism

```
# model criticism: no outliers
par(mfrow=c(1,2))
qqp(resid(finalmodelNat))
devfits = abs(resid(finalmodelNat,"deviance"))
plot(devfits,type="h")
```



4.8 Model with breakpoints better than model without

```
modelNoBPNat <- lmer(log(Nativelikeness) ~ IsIESpeaker + AEO + LR.c*Age.c +
                                           IsNaturalLearner + CountryEduYrs + NrLang +
                                           (1|Country), data=nat, REML=F)

summary(modelNoBPNat)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: log(Nativelikeness) ~ IsIESpeaker + AEO + LR.c * Age.c + IsNaturalLearne...
##     CountryEduYrs + NrLang + (1 | Country)
##    Data: nat
##
##      AIC      BIC   logLik deviance df.resid
##    118.5    158.1    -48.2     96.5      261
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -3.245 -0.611  0.090  0.717  2.419
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Country  (Intercept) 0.00398  0.0631
##  Residual             0.07985  0.2826
## Number of obs: 272, groups:  Country, 94
##
## Fixed effects:
##                       Estimate Std. Error t value
## (Intercept)           0.831003   0.067612   12.29
## IsIESpeakerTRUE       0.132504   0.042072    3.15
```

```
## AEO                  -0.021677   0.002997   -7.23
## LR.c                  0.016512   0.002737    6.03
## Age.c                -0.004666   0.001962   -2.38
## IsNaturalLearnerTRUE  0.184315   0.056011    3.29
## CountryEduYrs         0.032337   0.007074    4.57
## NrLang                0.082874   0.015925    5.20
## LR.c:Age.c           -0.000616   0.000149   -4.13
##
## Correlation of Fixed Effects:
##             (Intr) IIESTR AEO    LR.c   Age.c  INLTRU CntrEY NrLang
## IsIESpkTRUE  0.022
## AEO         -0.467  0.054
## LR.c        -0.182  0.122  0.295
## Age.c        0.202 -0.053 -0.282 -0.500
## IsNtrlLTRUE  0.001 -0.029 -0.085 -0.340  0.195
## ContryEdYrs -0.688 -0.382 -0.046  0.104  0.007 -0.090
## NrLang      -0.301 -0.130  0.046  0.111 -0.119 -0.054  0.014
## LR.c:Age.c  -0.010 -0.071 -0.180 -0.563 -0.009  0.209 -0.074 -0.010

# full model with breakpoints is better than the full model with breakpoints
modelNoBPNat.comp <- lmer(LD.log ~ IsIESpeaker + IsIESpeaker:AEO + LR.c*Age.c +
                                   IsNaturalLearner + CountryEduYrs + NrLang +
                                   (1+IsNaturalLearner|Country),
                          data=accents2, REML=F)

finalmodelNat.comp <- lmer(LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint +
                                    LR.c*Age.c + CountryEduYrs + NrLang +
                                    (1+IsNaturalLearner|Country),
                           data=accents2, REML=F)

AIC(modelNoBPNat.comp) - AIC(finalmodelNat.comp)

## [1] 6.92

anova(modelNoBPNat.comp,finalmodelNat.comp)

## Data: accents2
## Models:
## modelNoBPNat.comp: LD.log ~ IsIESpeaker + IsIESpeaker:AEO + LR.c * Age.c + IsNatu...
## modelNoBPNat.comp:     CountryEduYrs + NrLang + (1 + IsNaturalLearner | Country)
## finalmodelNat.comp: LD.log ~ ShiftedAEO:IsIESpeaker:PastBreakPoint + LR.c * Age.c +
## finalmodelNat.comp:     CountryEduYrs + NrLang + (1 + IsNaturalLearner | Country)
##                    Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## modelNoBPNat.comp  14 -572 -506    300     -600
## finalmodelNat.comp 14 -579 -513    303     -607  6.92      0     <2e-16
```