Cognitive influences in language evolution: Dutch data

Load libraries

```
library(mgcv)
library(sjPlot)
library(lattice)
library(ggplot2)
library(gplots)
library(dplyr)
library(party)
library(lmtest)
library(gridExtra)
library(itsadug)
library(car)
library(caret)
library(scales)


logit2per = function(X){
  return(exp(X)/(1+exp(X)))
}

rescaleGam = function(px, n, xvar, xlab=""){
  y = logit2per(px[[n]]$fit)
  x = px[[n]]$x *attr(xvar,"scaled:scale") + attr(xvar,"scaled:center")
  se.upper = logit2per(px[[n]]$fit+px[[n]]$se)
  se.lower = logit2per(px[[n]]$fit-px[[n]]$se)
  dx = data.frame(x=x,y=y,ci.upper=se.upper,ci.lower=se.lower)
  plen = ggplot(dx, aes(x=x,y=y))+
    geom_ribbon(aes(ymin=ci.lower,ymax=ci.upper), alpha=0.3)+
    geom_line(size=1) +
    xlab(xlab)+
    ylab("Probability of borrowing")+
    coord_cartesian(ylim = c(0,1))
  return(plen)
}
```

Load data

The Dutch data is processed very similarly to the English data. The full process can be found in the `processing` folder, but here we just load the final prepared data frame:

```
load("../data/loanwords_Dutch.Rdat")
```

Part of speech

```
catx = data.frame(
  PoS = tapply(dutch$cat, dutch$cat, function(X){as.character(X[1])}),
  mean = tapply(dutch$bor15, dutch$cat, mean),
  n = tapply(dutch$bor15, dutch$cat, length)
)
catx = catx[order(catx$mean, decreasing = T),]
catx$PoS = factor(catx$PoS, levels = catx[order(catx$mean, decreasing = T),]$PoS)

posg = ggplot(catx, aes(x=mean, y=PoS)) +
  geom_point(size=2) +
  ylab("Part of speech") +
  xlab("Proportion of words borrowed")+
  scale_x_continuous(labels=percent_format()) +
  geom_text(aes(label=n), nudge_y=0.4)

pdf("../results/graphs/POS_Borrowing_Dutch.pdf",
    width = 6,
    height = 4)
posg
dev.off()
```

```
## pdf
##   2
```

```
catx$mean= catx$mean*100
write.csv(catx, "../results/Dutch_POS_BorrowingProportions.csv", row.names = F)
```

GAM model

Dutch data has 1028 datapoints.

The range of the length variable limits the number of knots that the gam model can fit:

```
m0.dutch = bam(bor15.cat ~
      s(phonlengthscale, k=3) +
      s(AoAscale) +
      s(subtlexzipfscale) +
      s(concscale) +
      s(cat,bs='re')+
      s(cat,phonlengthscale,bs='re')+
      s(cat,AoAscale,bs='re')+
      s(cat,subtlexzipfscale,bs='re')+
      s(cat,concscale,bs='re'),
    data = dutch,
    family='binomial')
```

```
summary(m0.dutch)
```

```
##
## Family: binomial
## Link function: logit
##
```

```
## Formula:
## bor15.cat ~ s(phonlengthscale, k = 3) + s(AoAscale) + s(subtlexzipfscale) +
##     s(concscale) + s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re") + s(cat, concscale, bs = "re")
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.3389     0.3779  -6.189 6.04e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                              edf Ref.df Chi.sq  p-value
## s(phonlengthscale)      1.942e+00  1.996 18.147 0.000149 ***
## s(AoAscale)             1.310e+00  1.561 11.260 0.005448 **
## s(subtlexzipfscale)     3.630e+00  4.559 11.863 0.022505 *
## s(concscale)            1.663e+00  2.060  2.980 0.241790
## s(cat)                  3.724e+00 10.000 39.054 1.98e-08 ***
## s(cat,phonlengthscale)  1.721e-01 10.000  0.194 0.293757
## s(cat,AoAscale)         9.855e-06 10.000  0.000 0.938779
## s(cat,subtlexzipfscale) 1.883e-05 10.000  0.000 0.669506
## s(cat,concscale)        1.184e+00 10.000  3.069 0.098963 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.112   Deviance explained = 13.4%
## fREML =   1435  Scale est. = 1         n = 1028
```

Interactions

Test whether an interaction between AoA and frequency is warranted:

```
m1.dutch = bam(bor15.cat ~
      s(phonlengthscale, k=3) +
      s(AoAscale) +
      s(subtlexzipfscale) +
      s(concscale) +
      s(cat,bs='re')+
      s(cat,phonlengthscale,bs='re')+
      s(cat,AoAscale,bs='re')+
      s(cat,subtlexzipfscale,bs='re')+
      s(cat,concscale,bs='re') +
      te(AoAscale,subtlexzipfscale),
    data = dutch,
    family='binomial')

lrtest(m0.dutch,m1.dutch)
```

```
## Likelihood ratio test
##
## Model 1: bor15.cat ~ s(phonlengthscale, k = 3) + s(AoAscale) + s(subtlexzipfscale) +
##     s(concscale) + s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re") + s(cat, concscale, bs = "re")
## Model 2: bor15.cat ~ s(phonlengthscale, k = 3) + s(AoAscale) + s(subtlexzipfscale) +
##     s(concscale) + s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re") + s(cat, concscale, bs = "re") + te(AoAscale,
##     subtlexzipfscale)
##      #Df  LogLik     Df  Chisq Pr(>Chisq)
## 1 18.472 -441.31
## 2 20.348 -439.66 1.8758 3.2968     0.1924
```

No significant improvement.

Test whether an interaction between AoA and length is warranted:

```
m2.dutch = bam(bor15.cat ~
      s(phonlengthscale, k=3) +
      s(AoAscale) +
      s(subtlexzipfscale) +
      s(concscale) +
      s(cat,bs='re')+
      s(cat,phonlengthscale,bs='re')+
      s(cat,AoAscale,bs='re')+
      s(cat,subtlexzipfscale,bs='re')+
      s(cat,concscale,bs='re') +
      te(AoAscale,phonlengthscale),
    data = dutch,
    family='binomial')

lrtest(m0.dutch,m2.dutch)
```

```
## Likelihood ratio test
```

```
##
## Model 1: bor15.cat ~ s(phonlengthscale, k = 3) + s(AoAscale) + s(subtlexzipfscale) +
##     s(concscale) + s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re") + s(cat, concscale, bs = "re")
## Model 2: bor15.cat ~ s(phonlengthscale, k = 3) + s(AoAscale) + s(subtlexzipfscale) +
##     s(concscale) + s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re") + s(cat, concscale, bs = "re") + te(AoAscale,
##     phonlengthscale)
##      #Df  LogLik     Df  Chisq Pr(>Chisq)
## 1 18.472 -441.31
## 2 21.220 -439.07 2.7477 4.4682     0.2151
```

There is no improvement in log likelihood.

Test whether an interaction between Frequency and length is warranted:

```
m3.dutch = bam(bor15.cat ~
      s(phonlengthscale, k=3) +
      s(AoAscale) +
      s(subtlexzipfscale) +
      s(concscale) +
      s(cat,bs='re')+
      s(cat,phonlengthscale,bs='re')+
      s(cat,AoAscale,bs='re')+
      s(cat,subtlexzipfscale,bs='re')+
      s(cat,concscale,bs='re') +
      te(subtlexzipfscale,phonlengthscale),
    data = dutch,
    family='binomial')

lrtest(m0.dutch,m3.dutch)
```

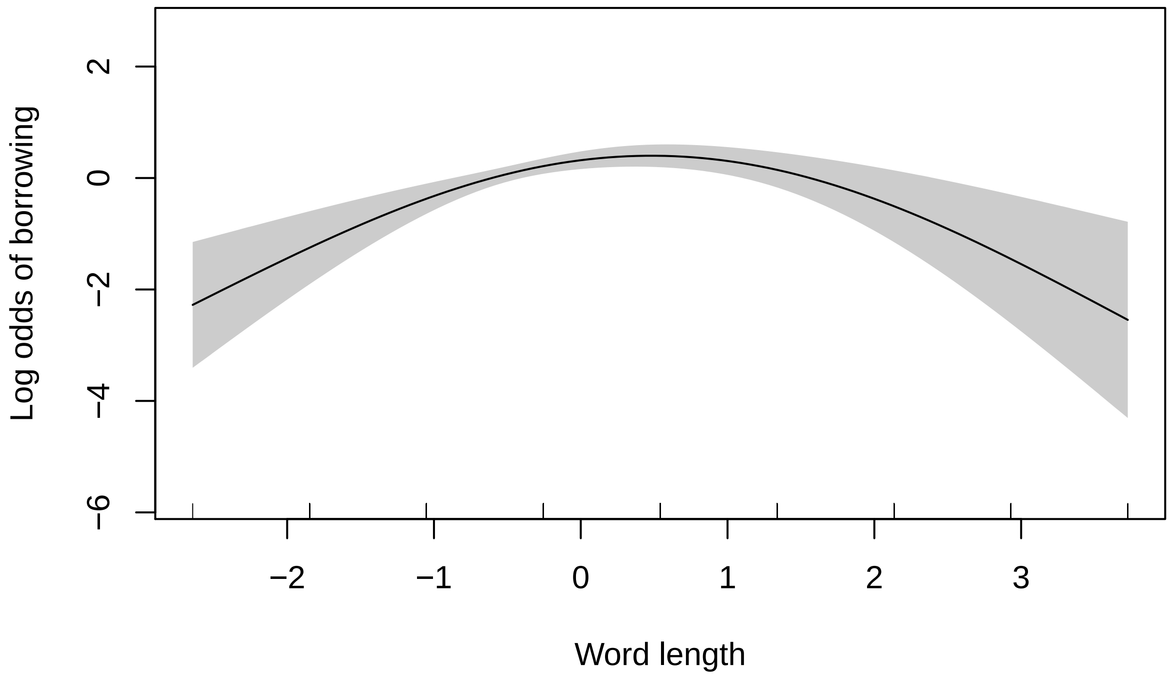
```
## Likelihood ratio test
##
## Model 1: bor15.cat ~ s(phonlengthscale, k = 3) + s(AoAscale) + s(subtlexzipfscale) +
##     s(concscale) + s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re") + s(cat, concscale, bs = "re")
## Model 2: bor15.cat ~ s(phonlengthscale, k = 3) + s(AoAscale) + s(subtlexzipfscale) +
##     s(concscale) + s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re") + s(cat, concscale, bs = "re") + te(subtlexzipfscale,
##     phonlengthscale)
##      #Df  LogLik     Df  Chisq Pr(>Chisq)
## 1 18.472 -441.31
## 2 22.040 -437.60 3.5681 7.4151     0.1155
```

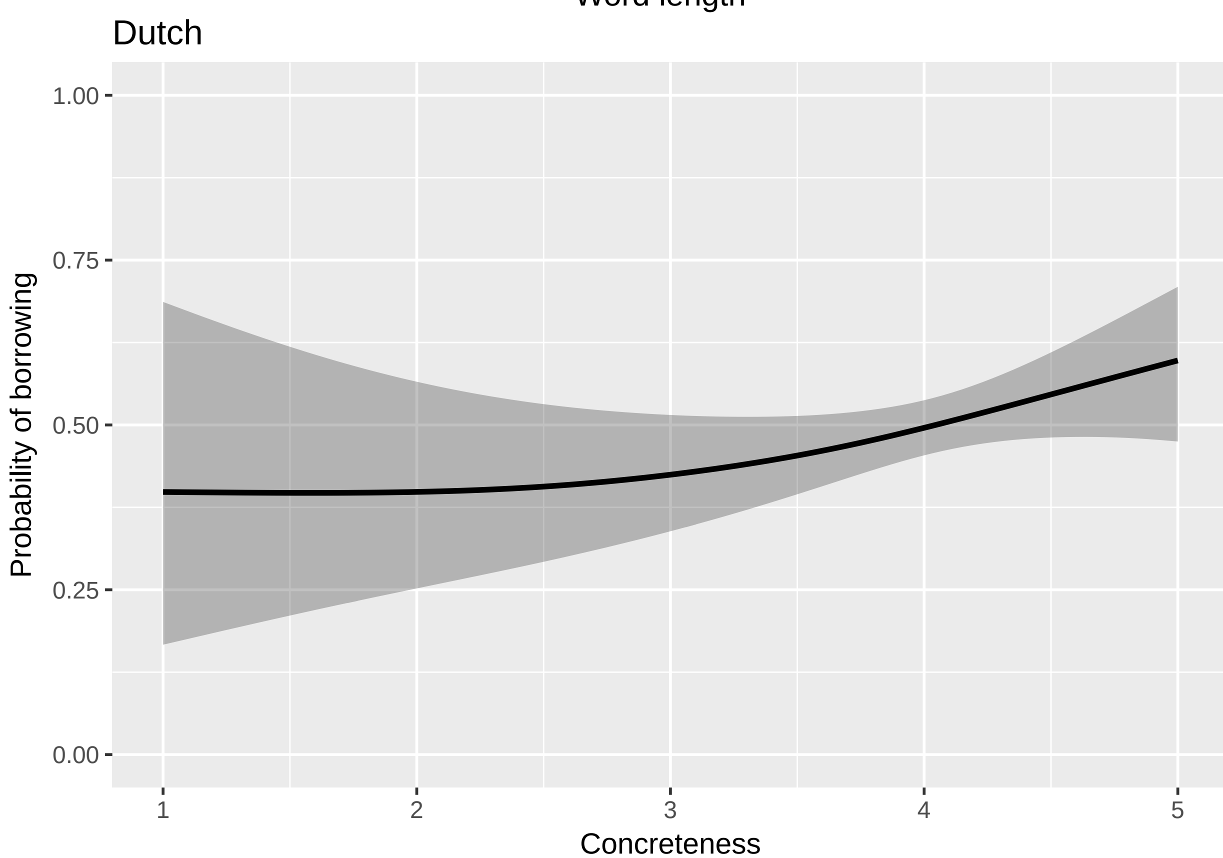
No significant improvement.

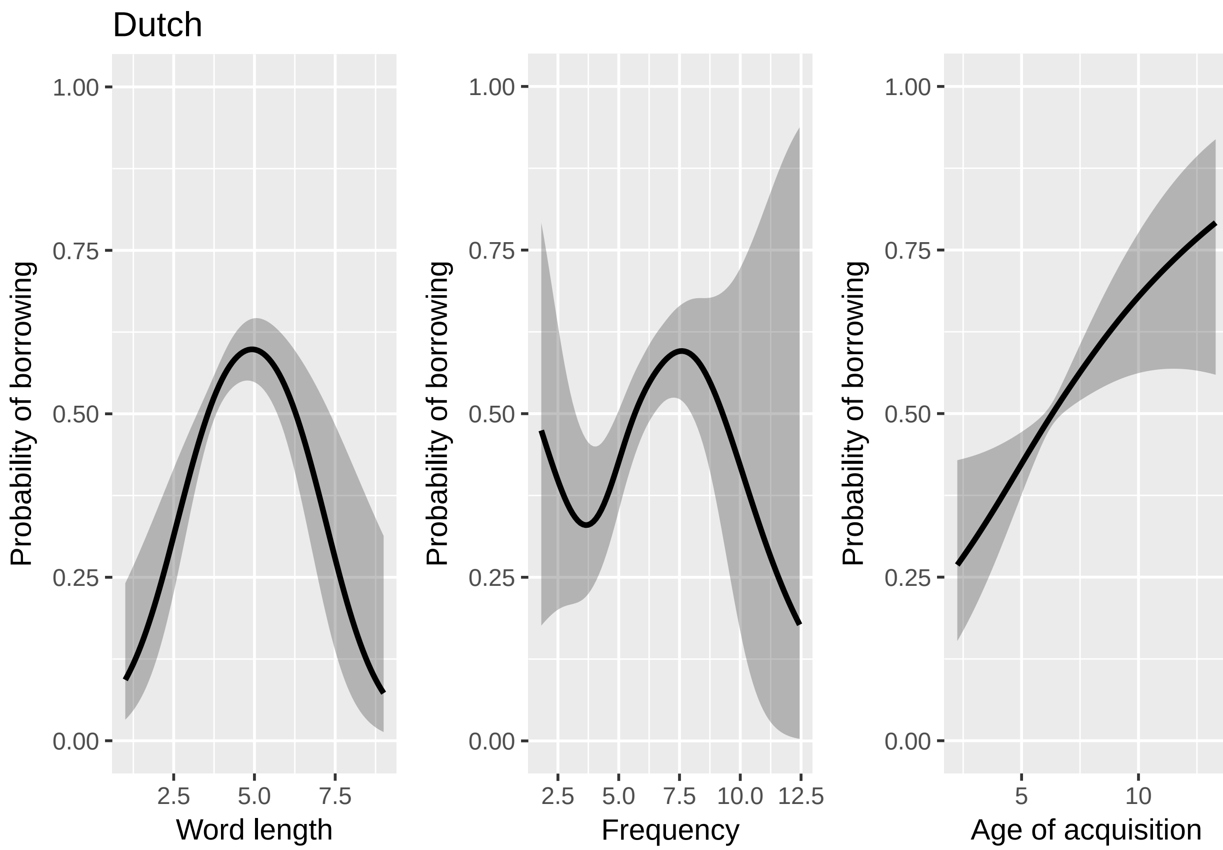
So no interactions are necessary.

Model estiamtes

Plot the model estimates, changing the dependent scale to probability and the independent variables to their original scales (code is hidden, but available in the Rmd file).

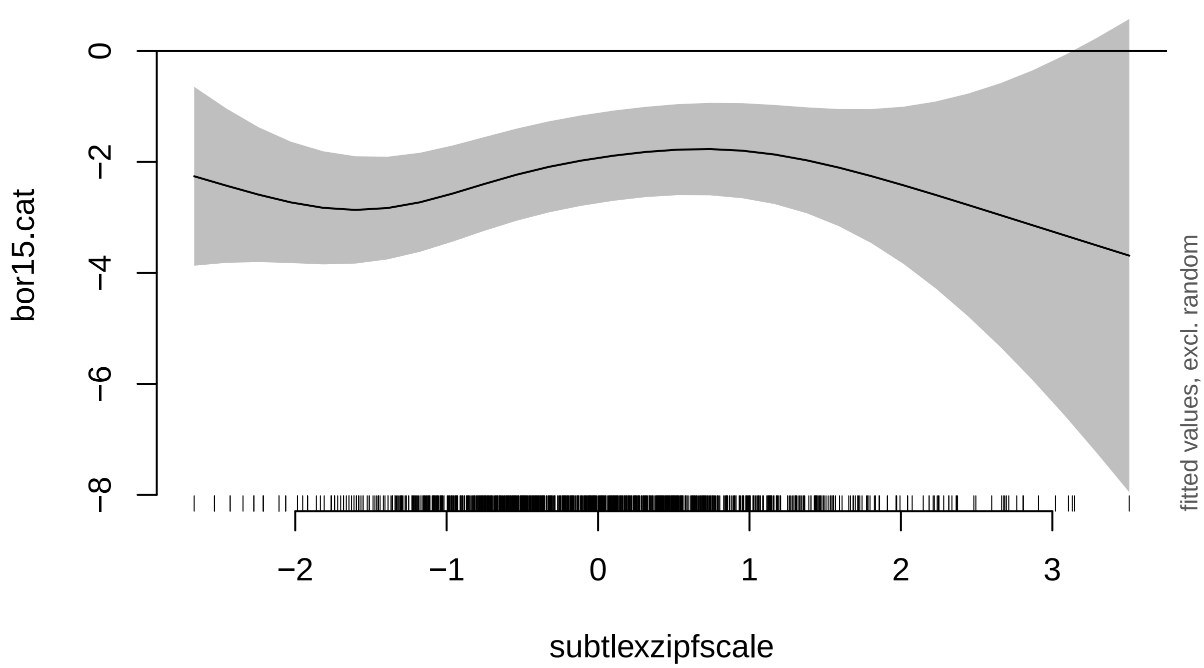


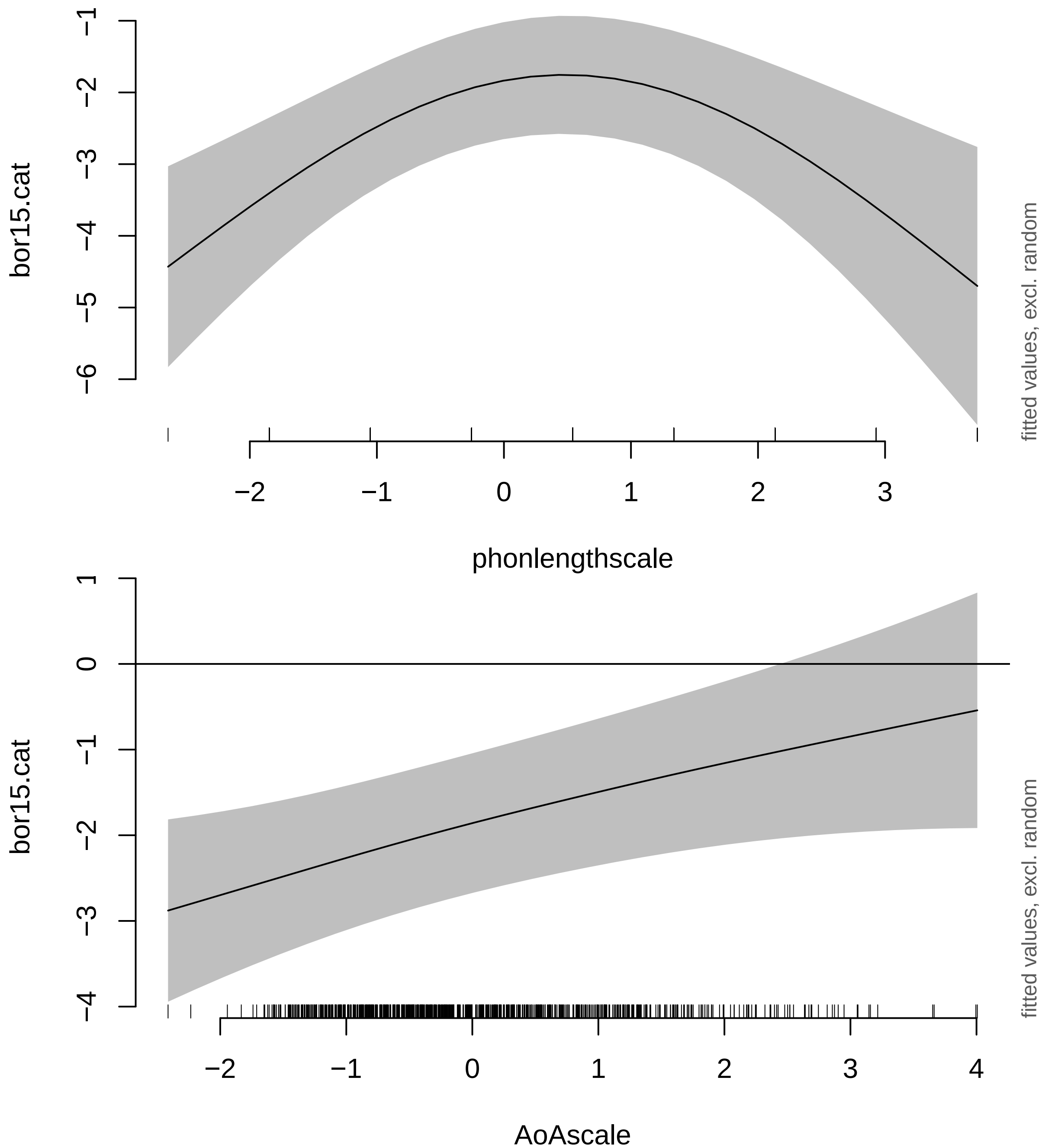


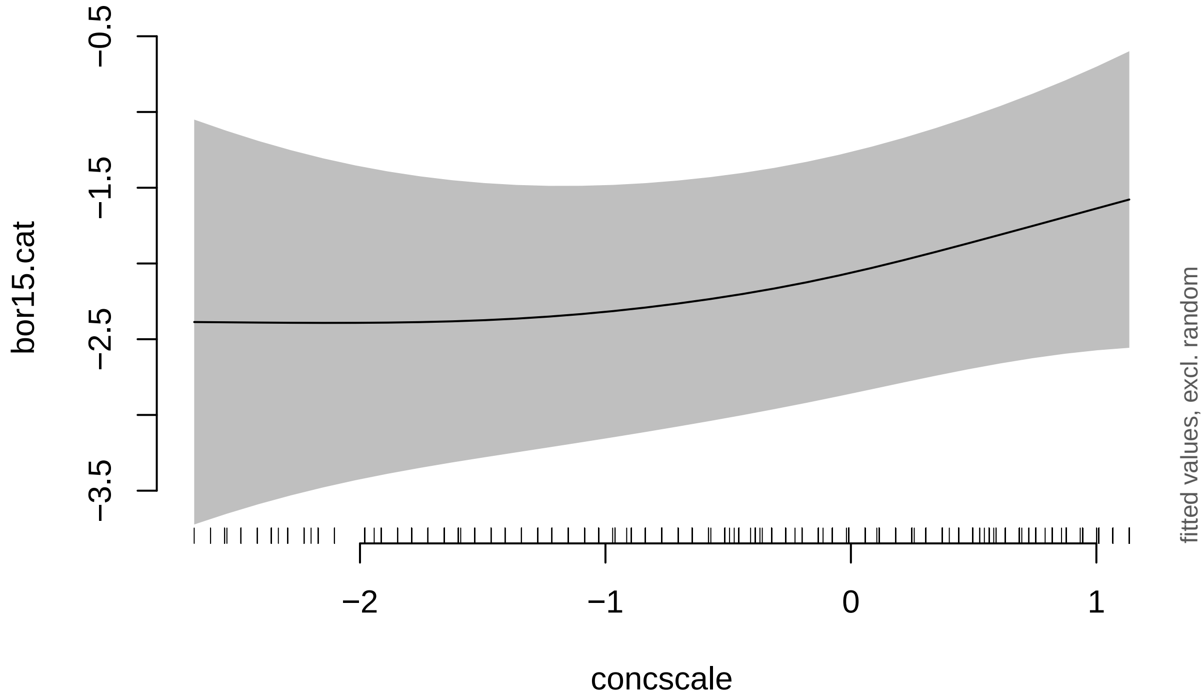


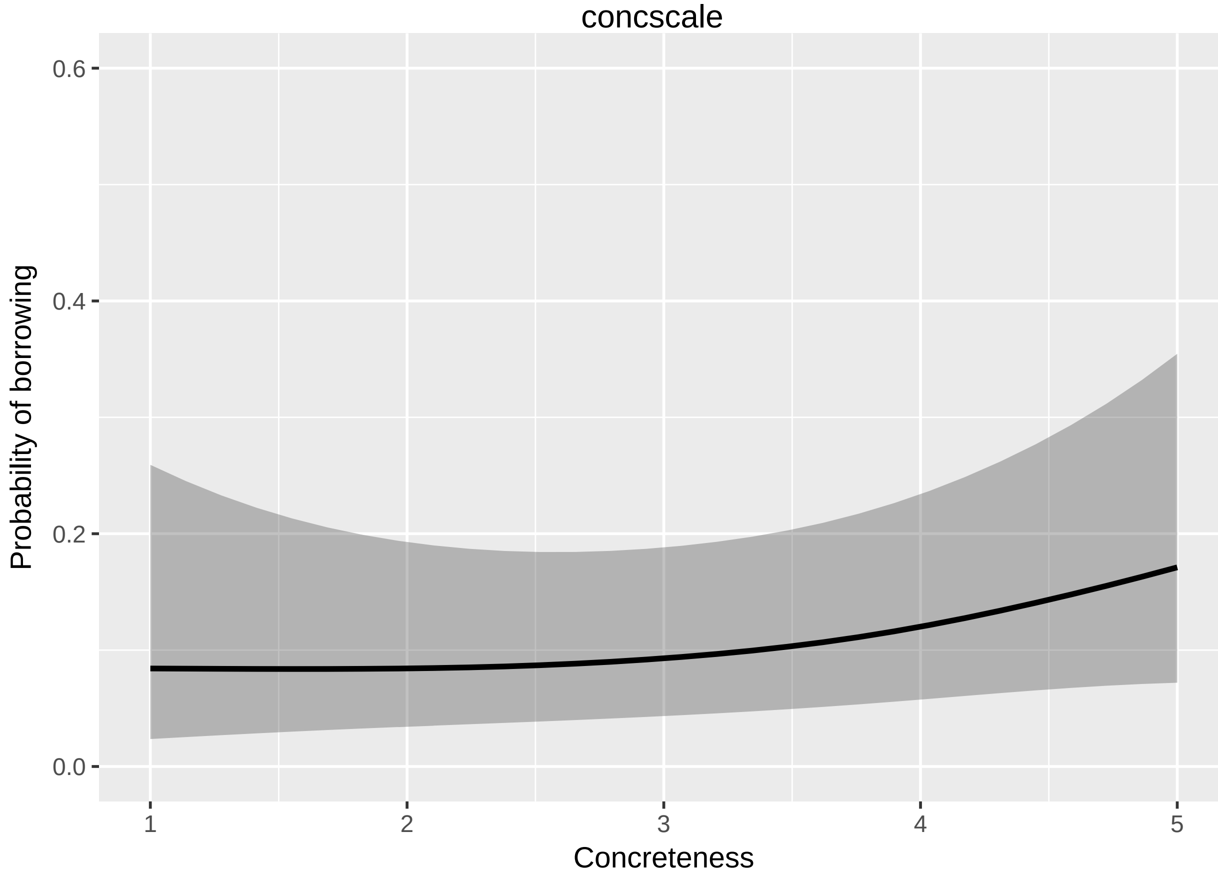

```
## pdf
##   2
```

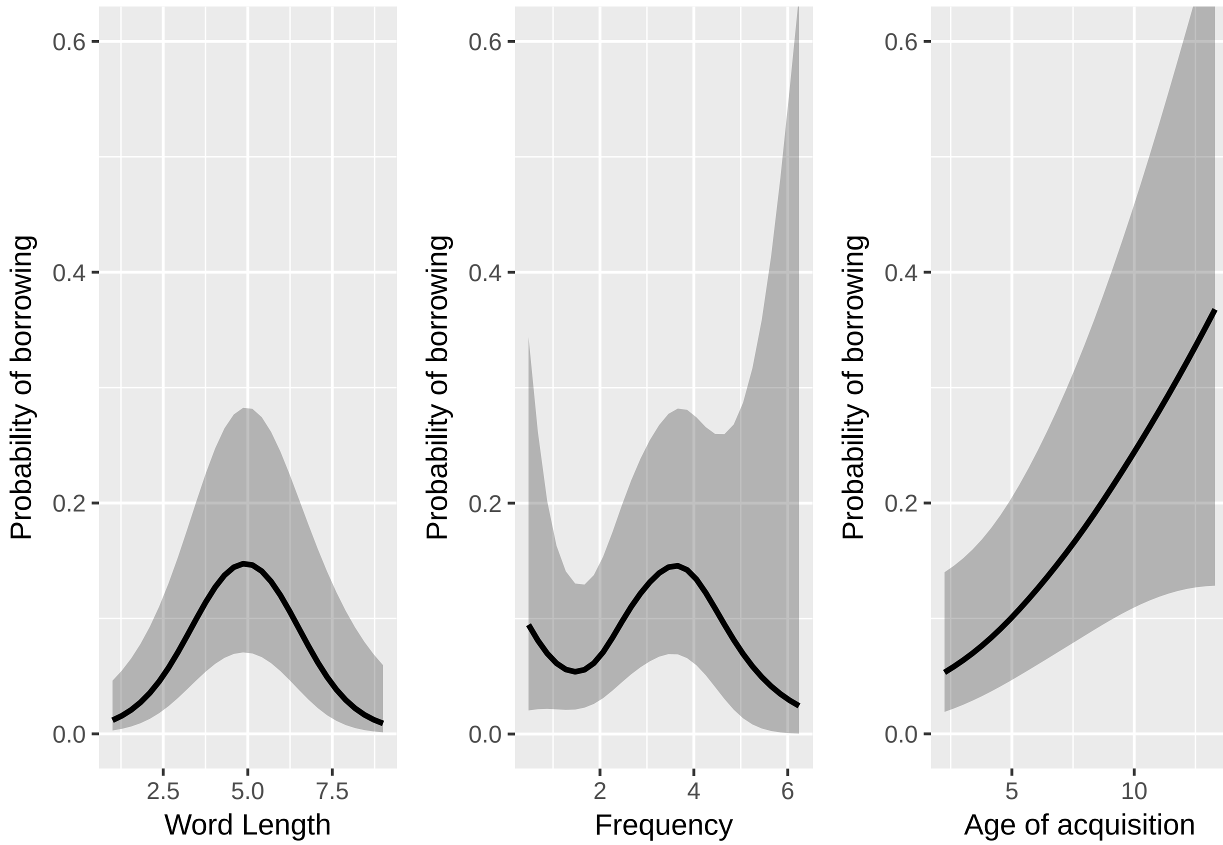
Plot the model estimates, removing the influence of the random effects using the library `itsadug` (code is hidden, but available in the Rmd file).




bor15.cat
−1
−2
−3
−4
−5
−6
fitted values, excl. random
−2
−1
0
1
2
3
phonlengthscale
bor15.cat
1
0
−1
−2
−3
−4
fitted values, excl. random
−2
−1
0
1
2
3
4
AoAscale


bor15.cat
−0.5
−1.5
−2.5
−3.5
−2
−1
0
1
concscale
fitted values, excl. random


Probability of borrowing
0.6
0.4
0.2
0.0
1
2
3
4
5
Concreteness




```
## pdf
##   2
```

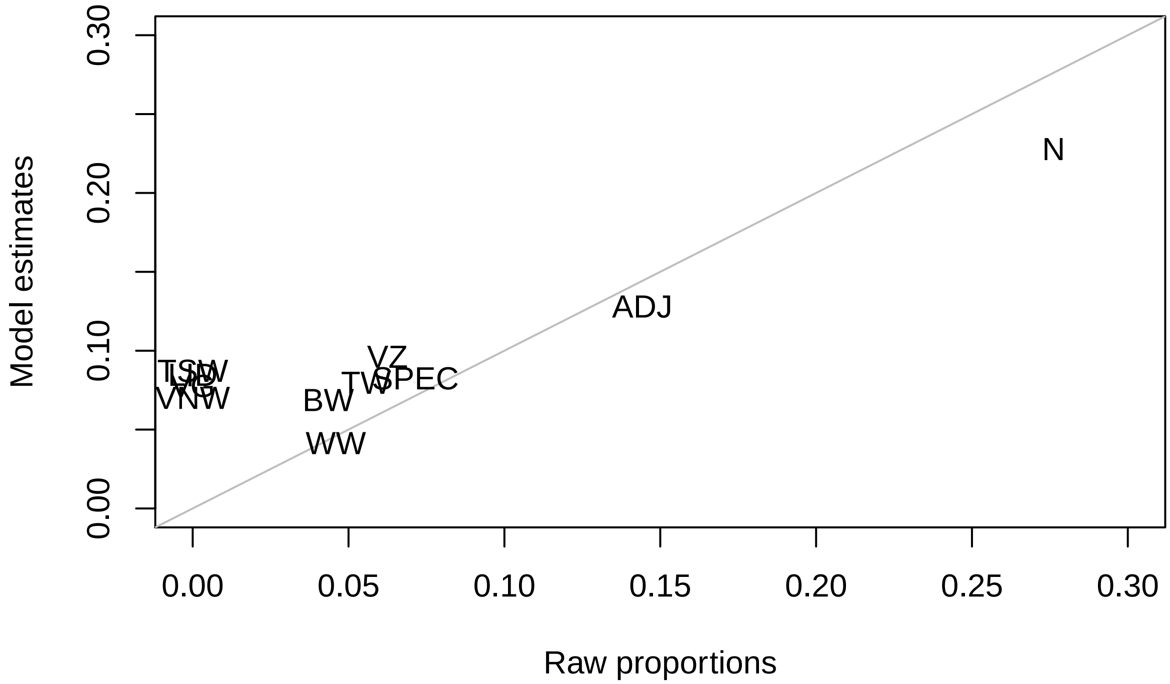
Random effects for Part of speech

```
mc = m0.dutch$coefficients
mc[grepl("s\\(cat\\)",names(mc))]
```

```
##     s(cat).1     s(cat).2     s(cat).3     s(cat).4     s(cat).5
##  1.119986827  0.421313755 -0.262839333 -0.037142605 -0.066124188
##     s(cat).6     s(cat).7     s(cat).8     s(cat).9    s(cat).10
## -0.004943456 -0.103771803 -0.131386158 -0.241569599  0.101695214
##    s(cat).11
## -0.795218653
```

```
raw = tapply(dutch$bor15,dutch$cat,mean)
model.est = logit2per(m0.dutch$coefficients[1] +
              mc[grepl("s\\(cat\\)",names(mc))])

plot(raw, model.est,
  xlab="Raw proportions",
  ylab="Model estimates",
  col="white",
     ylim=c(0,0.3),
     xlim=c(0,0.3))
abline(0,1,col='gray')
text(raw, model.est, names(raw))
```



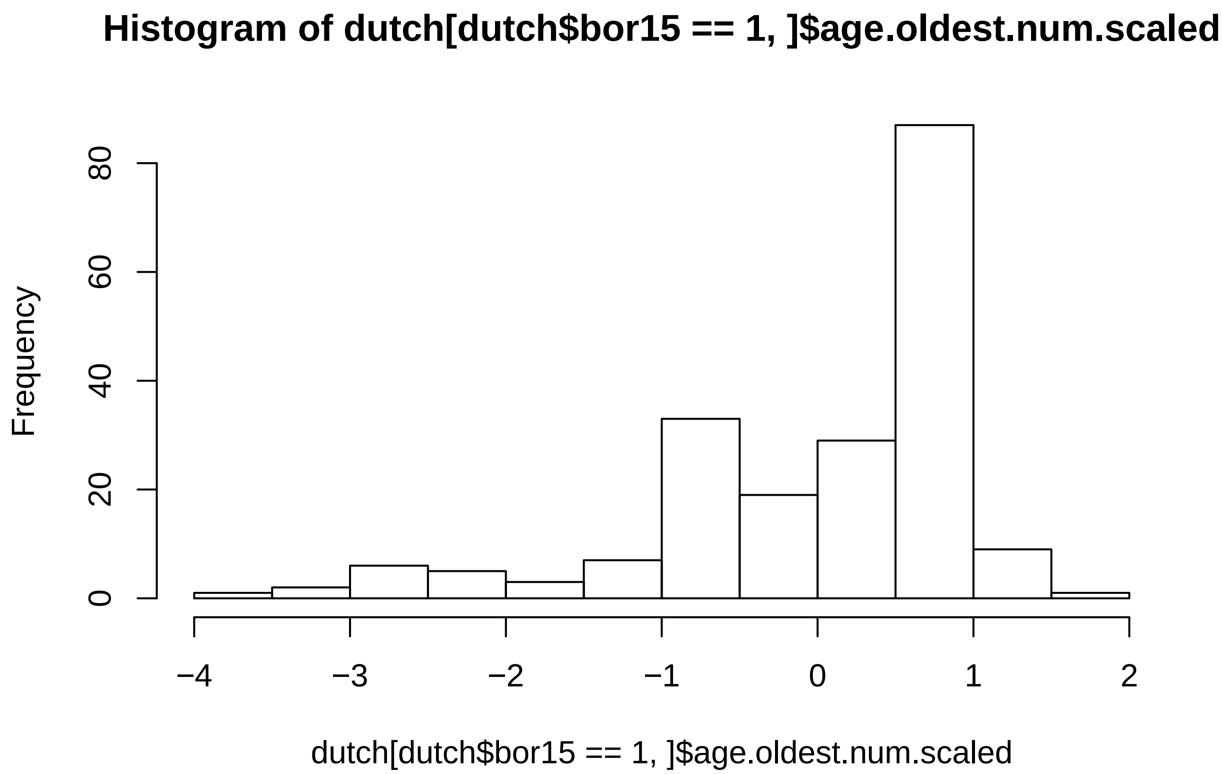
Predicting date of entry

The dates of entry for Dutch are much more dispersed than for English, so we transform them further using the Box-Cox method:

```
# remove non-borrowed words
dutch[dutch$bor15!=1,]$age.oldest.num = NA
# Take log years
dutch$age.oldest.num.scaled = log10(dutch$age.oldest.num)
# scale with boxcox
pp = preProcess(dutch[,c('age.oldest.num.scaled','AoAscale')], method="BoxCox")
dutch$age.oldest.num.scaled = bcPower(dutch$age.oldest.num.scaled, lambda = pp$bc$age.oldest.num.scaled$
dutch$age.oldest.num.scaled = scale(dutch$age.oldest.num.scaled)
```

Plot raw data

```
hist(dutch[dutch$bor15==1,]$age.oldest.num.scaled)
```



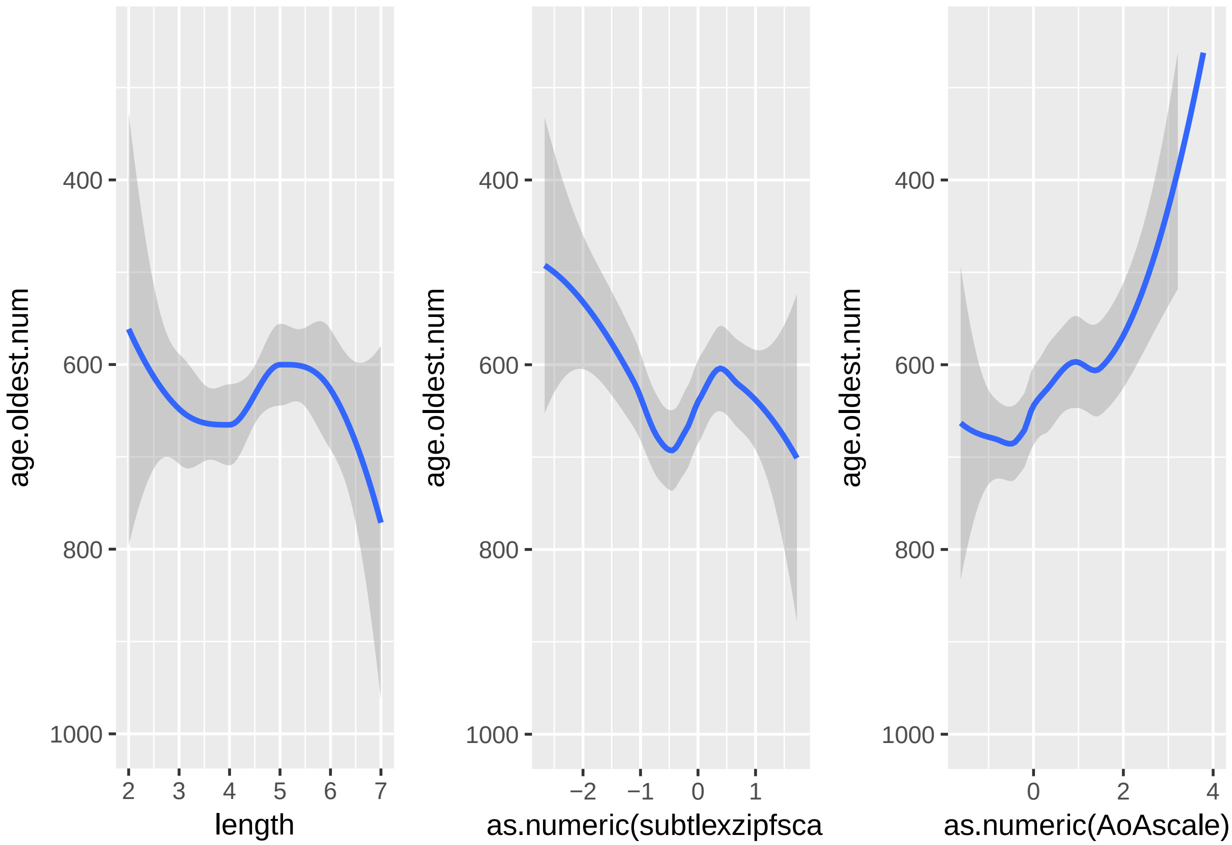

```
g.ageAoA = ggplot(dutch[dutch$bor15==1,],
      aes(x=as.numeric(AoAscale), y=age.oldest.num))+
         geom_smooth() +
  scale_y_reverse(lim=c(1000,250))


g.ageLen = ggplot(dutch[dutch$bor15==1,],
  aes(x=length, y=age.oldest.num))+
         geom_smooth()+
  scale_y_reverse(lim=c(1000,250))

g.ageFreq = ggplot(dutch[dutch$bor15==1,],
  aes(x=as.numeric(subtlexzipfscale), y=age.oldest.num))+
         geom_smooth()+
  scale_y_reverse(lim=c(1000,250))

grid.arrange(g.ageLen,g.ageFreq,g.ageAoA, nrow=1)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



GAM model: Because there are fewer datapoints, we build up the mode one variable at a time, keeping the variable if it significantly improves the fit of the model.

```
m0.age = bam(age.oldest.num.scaled~
    1 +
    s(cat,bs='re')+
    s(cat,phonlengthscale,bs='re')+
    s(cat,AoAscale,bs='re')+
    s(cat,subtlexzipfscale,bs='re'),
    data = dutch[dutch$bor15==1,])
m1.age = update(m0.age, ~.+ s(AoAscale))
lrtest(m0.age,m1.age)
```

```
## Likelihood ratio test
##
## Model 1: age.oldest.num.scaled ~ 1 + s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re")
## Model 2: age.oldest.num.scaled ~ s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re") + s(AoAscale)
##      #Df  LogLik     Df  Chisq Pr(>Chisq)
## 1 6.0707 -280.78
## 2 8.2630 -276.41 2.1923 8.7286    0.01272 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Significant
m2.age = update(m1.age, ~.+ s(phonlengthscale, k=3))
lrtest(m1.age,m2.age)
```

```
## Likelihood ratio test
##
## Model 1: age.oldest.num.scaled ~ s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re") + s(AoAscale)
## Model 2: age.oldest.num.scaled ~ s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re") + s(AoAscale) + s(phonlengthscale, k = 3)
##     #Df  LogLik    Df  Chisq Pr(>Chisq)
## 1 8.263 -276.41
## 2 9.357 -275.72 1.094 1.3926      0.238
```

```r
# Not significant
m3.age = update(m1.age, ~.+ s(subtlexzipfscale))
lrtest(m1.age,m3.age)
```

```
## Likelihood ratio test
##
## Model 1: age.oldest.num.scaled ~ s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re") + s(AoAscale)
## Model 2: age.oldest.num.scaled ~ s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re") + s(AoAscale) + s(subtlexzipfscale)
##      #Df  LogLik      Df Chisq Pr(>Chisq)
## 1 8.2630 -276.41
## 2 8.9716 -275.96 0.70864 0.917     0.3383
```

```r
# Not significant
m4.age = update(m1.age, ~.+ s(concscale))
lrtest(m1.age,m4.age)
```

```
## Likelihood ratio test
##
## Model 1: age.oldest.num.scaled ~ s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re") + s(AoAscale)
## Model 2: age.oldest.num.scaled ~ s(cat, bs = "re") + s(cat, phonlengthscale,
##     bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlexzipfscale,
##     bs = "re") + s(AoAscale) + s(concscale)
##      #Df  LogLik      Df  Chisq Pr(>Chisq)
## 1 8.2630 -276.41
## 2 9.0694 -276.46 0.80643 0.0922     0.7614
```

```r
# Not sigificant
```

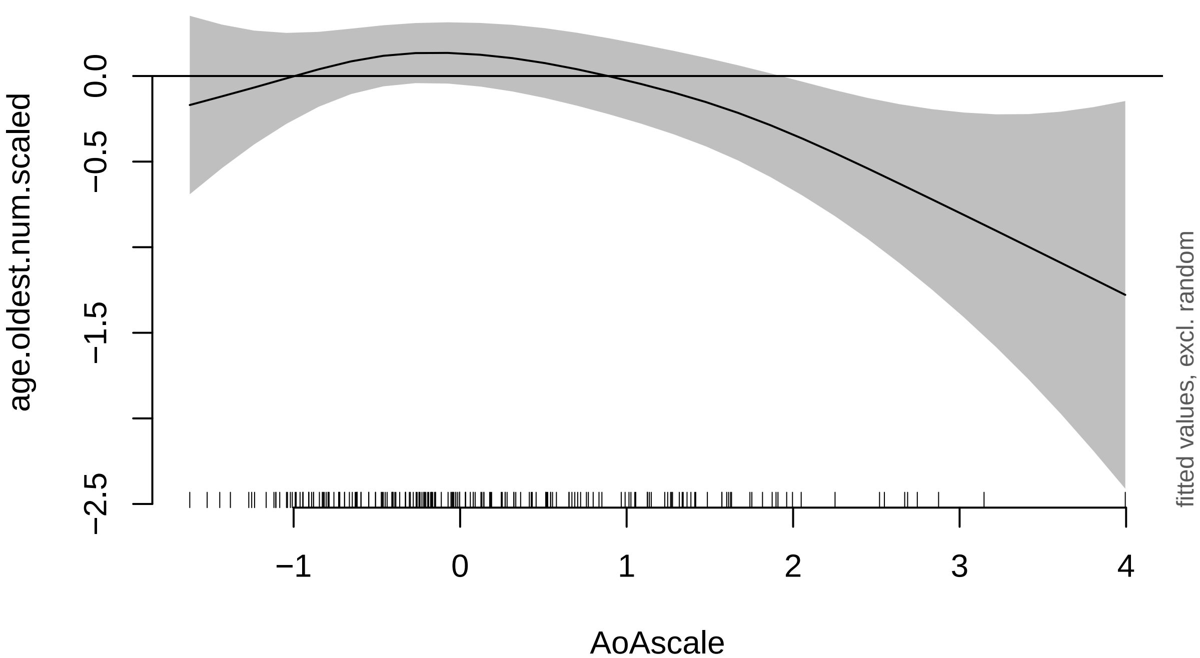
Final model has just age of acquisition as a main effect, so let's take away the random effects for other variables:

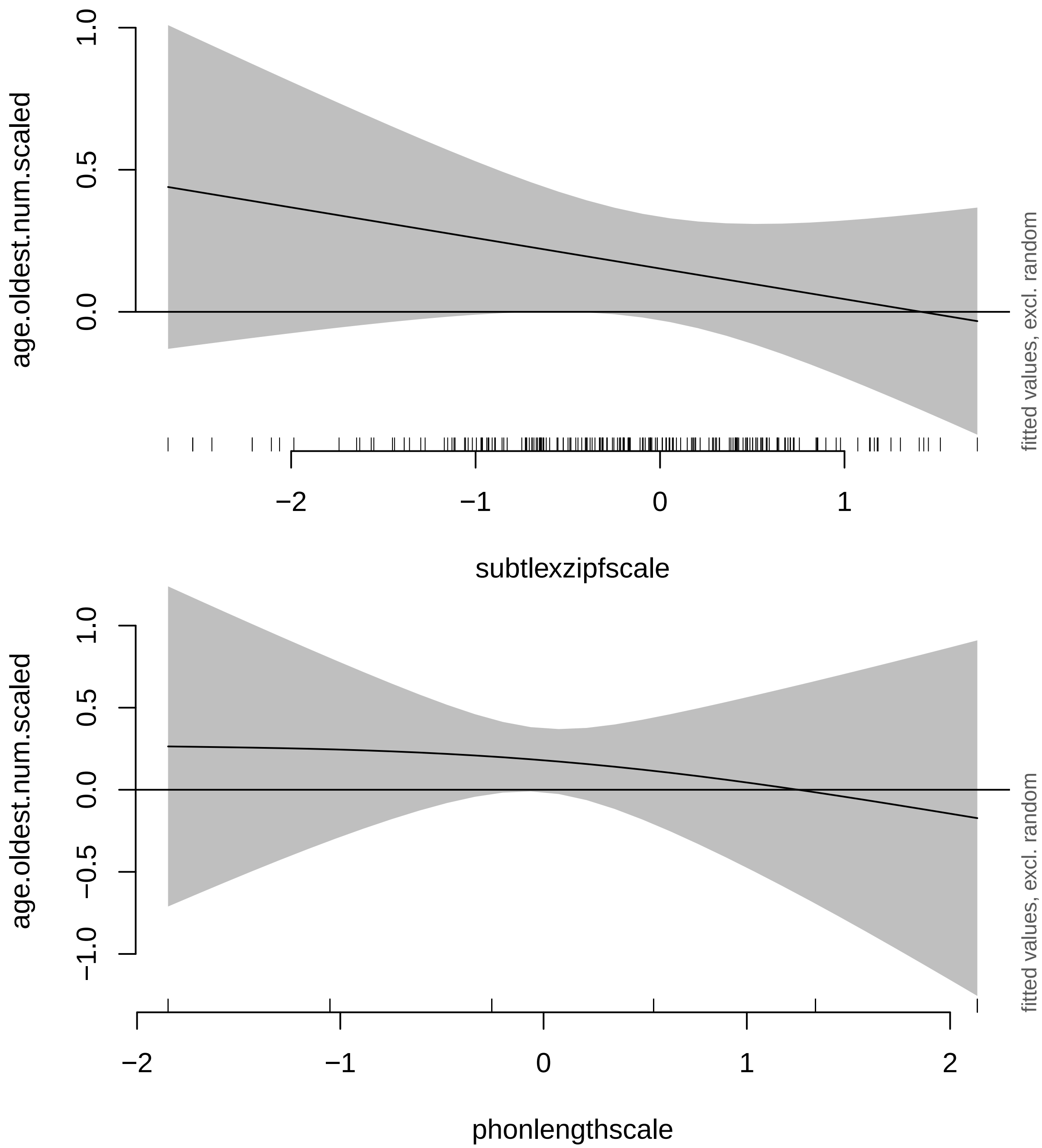
```r
m5.age = bam(age.oldest.num.scaled~
    s(AoAscale) +
    s(cat,bs='re') +
    s(cat,AoAscale,bs='re'),
```

```r
    data = dutch[dutch$bor15==1,])
summary(m5.age)
```

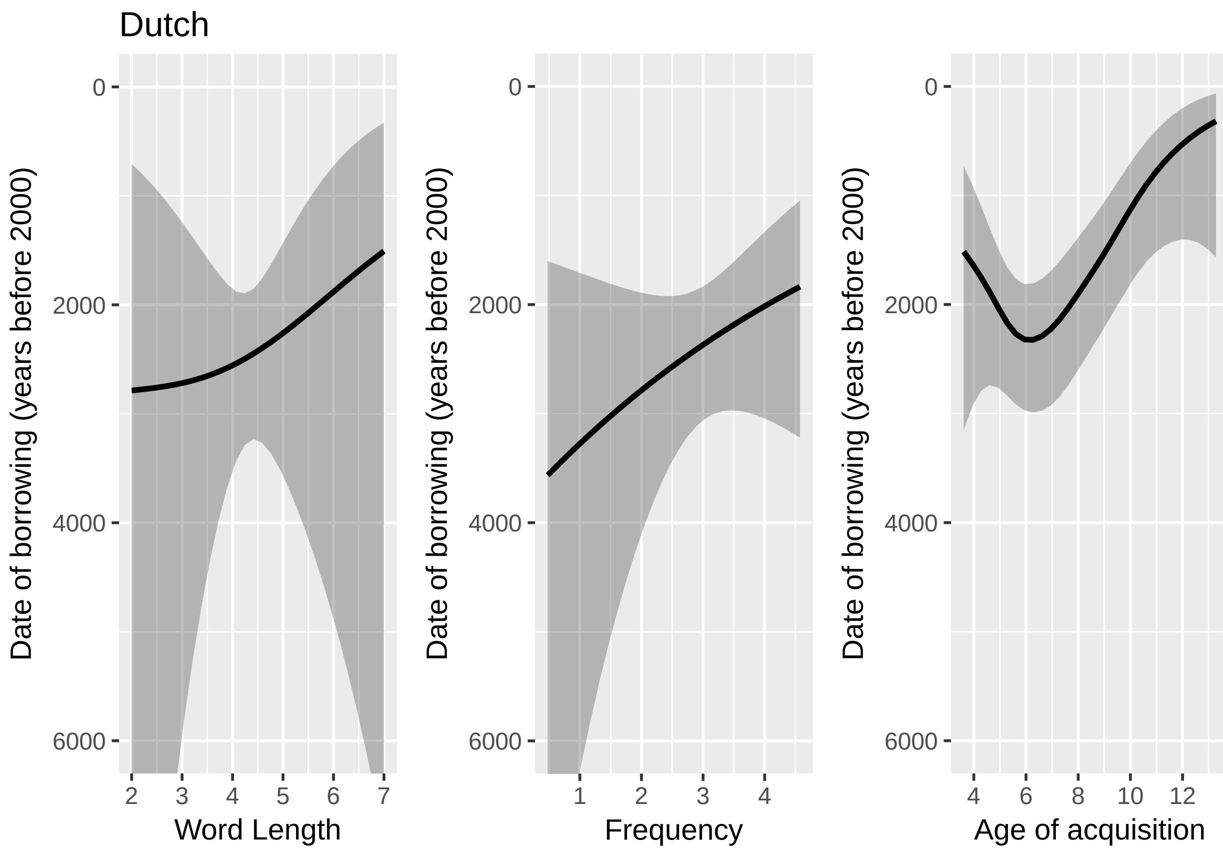
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## age.oldest.num.scaled ~ s(AoAscale) + s(cat, bs = "re") + s(cat,
##     AoAscale, bs = "re")
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.799e-07  6.887e-02       0        1
##
## Approximate significance of smooth terms:
##                        edf Ref.df     F p-value
## s(AoAscale)     2.520e+00  3.205 2.893  0.0317 *
## s(cat)          5.599e-06  6.000 0.000  0.5359
## s(cat,AoAscale) 8.264e-07  6.000 0.000  1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0419   Deviance explained = 5.39%
## fREML = 286.04  Scale est. = 0.95812   n = 202
```

Plot the model estimates. The code is hidden, but you can view it in the Rmd file. Note that the estimates actually come from different models. Only the age of acuqisition result is relevant to the main results in the paper, but the others are shown for illustration.





age.oldest.num.scaled
1.0
0.5
0.0
fitted values, excl. random
−2
−1
0
1
subtlexzipfscale
age.oldest.num.scaled
1.0
0.5
0.0
−0.5
−1.0
fitted values, excl. random
−2
−1
0
1
2
phonlengthscale


Dutch
Date of borrowing (years before 2000)
0
2000
4000
6000
2
3
4
5
6
7
Word Length
Date of borrowing (years before 2000)
0
2000
4000
6000
1
2
3
4
Frequency
Date of borrowing (years before 2000)
0
2000
4000
6000
4
6
8
10
12
Age of acquisition

```
## pdf
##   2
```

```
## pdf
##   2
```