

Age of acquisition and borrowing: Dutch

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)

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

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

Dutch

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(subtlelexzipfscale) +
  s(concscale) +
  s(cat,bs='re')+
  s(cat,phonlengthscale,bs='re')+
  s(cat,AoAscale,bs='re')+
  s(cat,subtlelexzipfscale,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(subtlelexzipfscale) +
##      s(concscale) + s(cat, bs = "re") + s(cat, phonlengthscale,
##      bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlelexzipfscale,
##      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(subtlelexzipfscale)  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,subtlelexzipfscale) 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(subtlelexzipfscale) +
  s(concscale) +
  s(cat,bs='re')+
  s(cat,phonlengthscale,bs='re')+
  s(cat,AoAscale,bs='re')+
  s(cat,subtlelexzipfscale,bs='re')+
  s(cat,concscale,bs='re') +
  te(AoAscale,subtlelexzipfscale),
  data = dutch,
  family='binomial')

lrtest(m0.dutch,m1.dutch)

## Likelihood ratio test
##
## Model 1: bor15.cat ~ s(phonlengthscale, k = 3) + s(AoAscale) + s(subtlelexzipfscale) +
##   s(concscale) + s(cat, bs = "re") + s(cat, phonlengthscale,
##   bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlelexzipfscale,
##   bs = "re") + s(cat, concscale, bs = "re")
## Model 2: bor15.cat ~ s(phonlengthscale, k = 3) + s(AoAscale) + s(subtlelexzipfscale) +
##   s(concscale) + s(cat, bs = "re") + s(cat, phonlengthscale,
##   bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlelexzipfscale,
##   bs = "re") + s(cat, concscale, bs = "re") + te(AoAscale,
##   subtlelexzipfscale)
##   #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(subtlelexzipfscale) +
  s(concscale) +
  s(cat,bs='re')+
  s(cat,phonlengthscale,bs='re')+
  s(cat,AoAscale,bs='re')+
  s(cat,subtlelexzipfscale,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
```

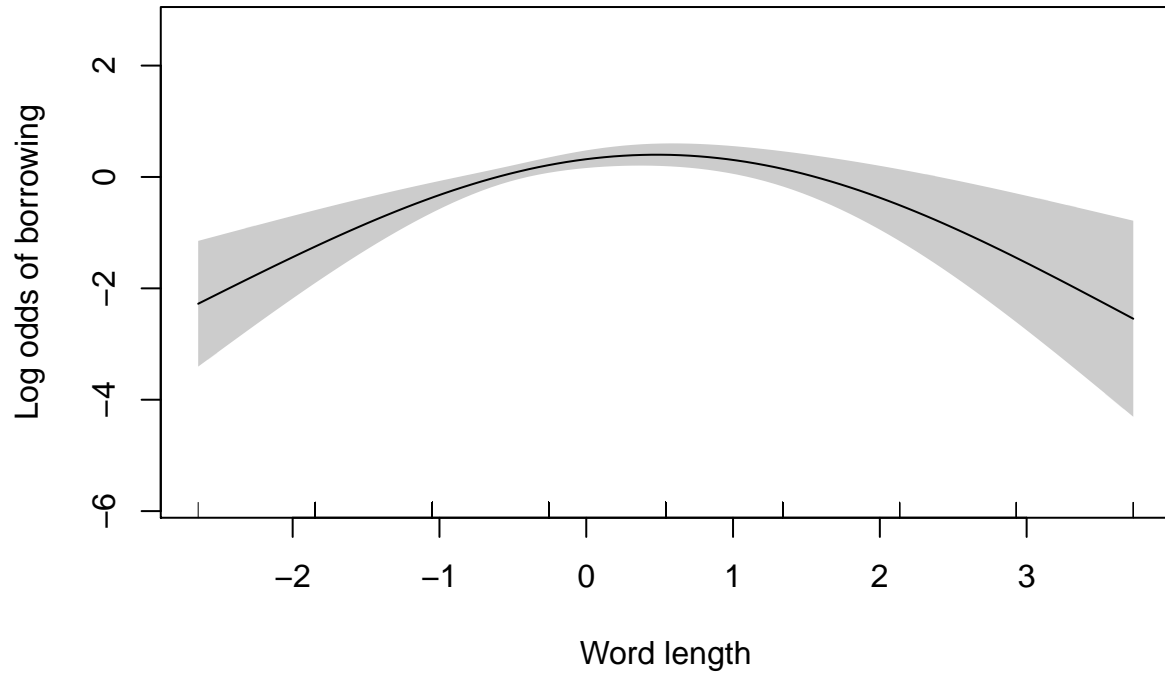
No significant improvement.

So no interactions are necessary.

Model estimates

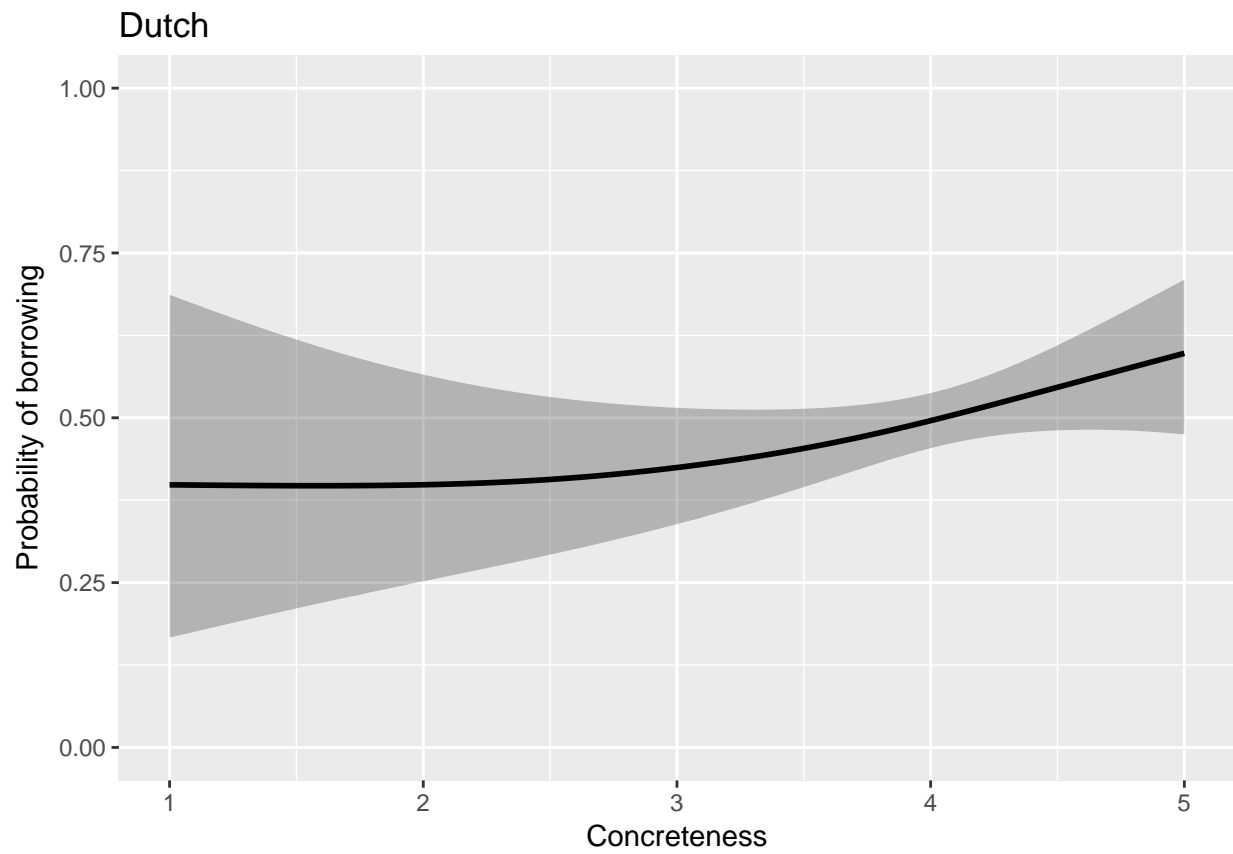
Plot the model estimates, changing the dependent scale to probability and the independent variables to their original scales.

```
px = plot.gam(m0.dutch,select=1, xlab="Word length", ylab="Log odds of borrowing",shade = T)
```

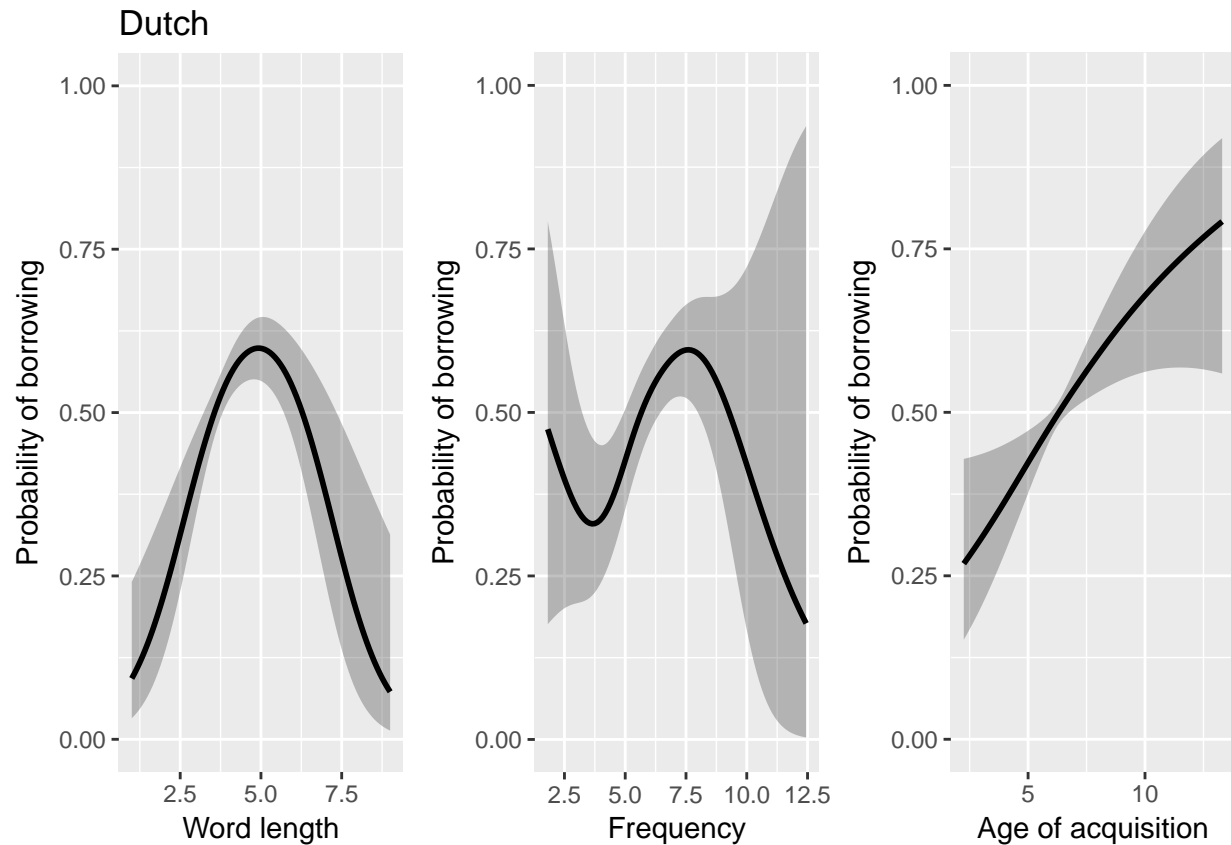


```
plen = rescaleGam(px, 1, dutch$phonlengthscale,"Word length") + ggtitle("Dutch")
paoa = rescaleGam(px, 2, dutch$AoAscale,"Age of acquisition") + ggtitle("")
pfreq = rescaleGam(px, 3, dutch$AoAscale,"Frequency") + ggtitle("")
pconc = rescaleGam(px, 4, dutch$concscale,"Concreteness") + ggtitle("Dutch")

pconc
```



```
grid.arrange(plen,pfreq,paoa, nrow=1)
```



```
pdf(file='../results/graphs/Dutch_ModelResults.pdf',
    height = 3, width = 8)
grid.arrange(plen, pfreq, paoa, nrow = 1)
dev.off()
```

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

Predicting date of entry

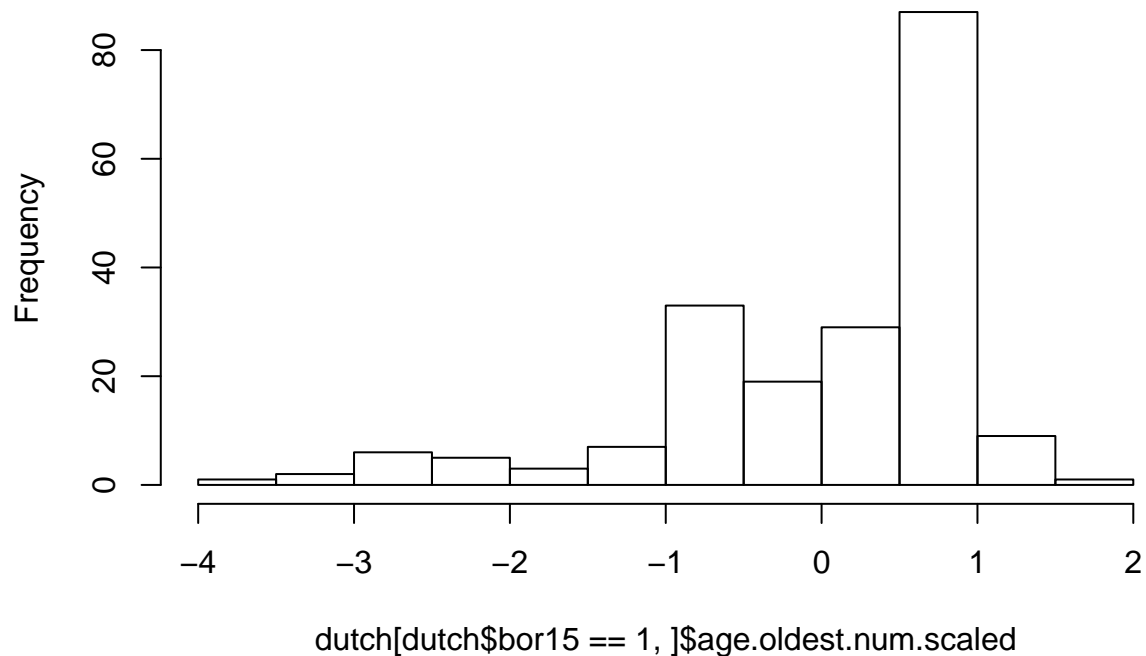
The dates of entry for Dutch are much more dispersed than for English, so we transform them further using the Box-Cox method:

```
# 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)
```

Histogram of dutch[dutch\$bor15 == 1,]\$age.oldest.num.scaled



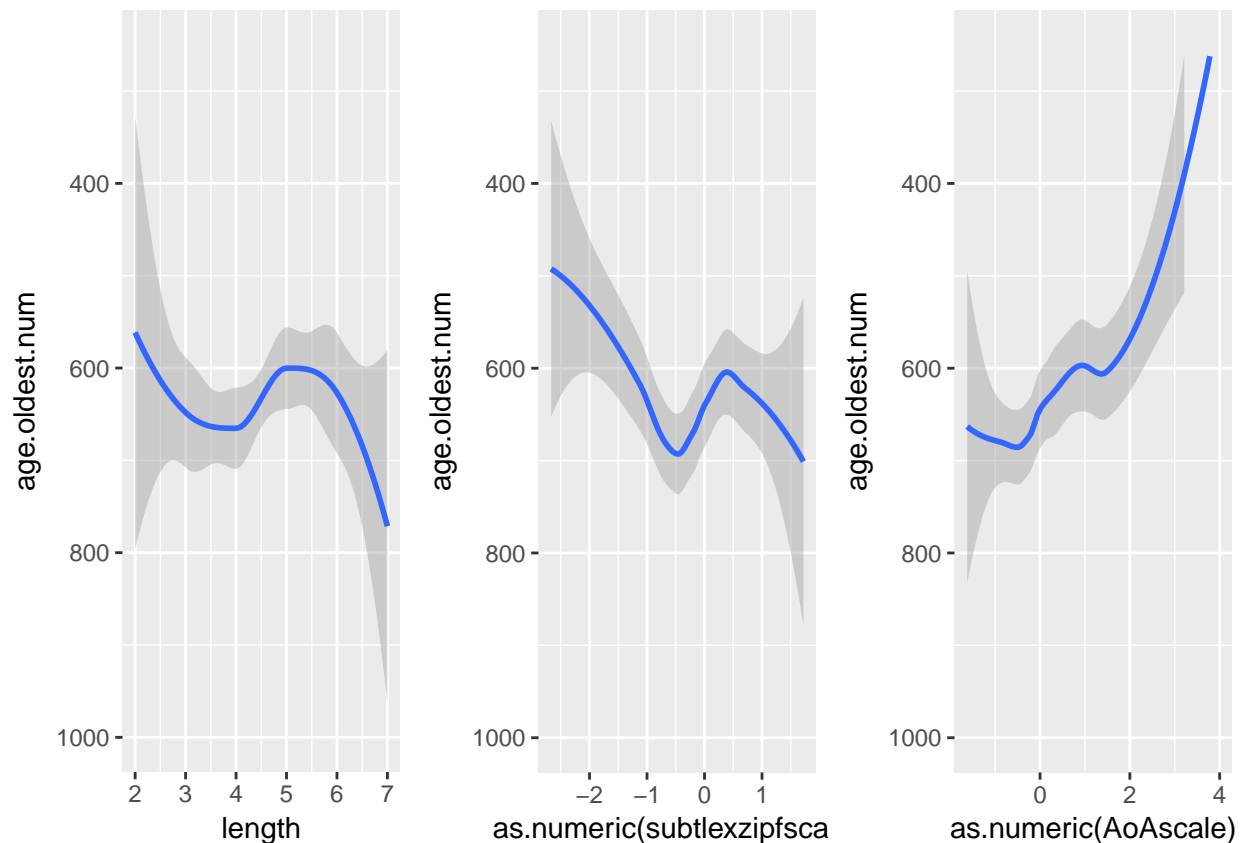
```
g.ageAoA = ggplot(dutch[dutch$bor15==1,],
  aes(x=as.numeric(AoAscale), y=age.oldest.num))+
  geom_smooth() +
  scale_y_reverse(lim=c(1000,250))

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(subtlezipfscale), 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 model 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,subtlelexzipfscale,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, subtlelexzipfscale,
##      bs = "re")
## Model 2: age.oldest.num.scaled ~ s(cat, bs = "re") + s(cat, phonlengthscale,
##      bs = "re") + s(cat, AoAscale, bs = "re") + s(cat, subtlelexzipfscale,
##      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
```

```
# 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, subtexzipfscale,
##      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, subtexzipfscale,
##      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
```

```
# Not significant
m3.age = update(m1.age, ~.+ s(subtexzipfscale))
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, subtexzipfscale,
##      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, subtexzipfscale,
##      bs = "re") + s(AoAscale) + s(subtexzipfscale)
##      #Df  LogLik      Df  Chisq Pr(>Chisq)
## 1 8.2630 -276.41
## 2 8.9716 -275.96 0.70864 0.917      0.3383
```

```
# 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, subtexzipfscale,
##      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, subtexzipfscale,
##      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
```

```
# Not significant
```

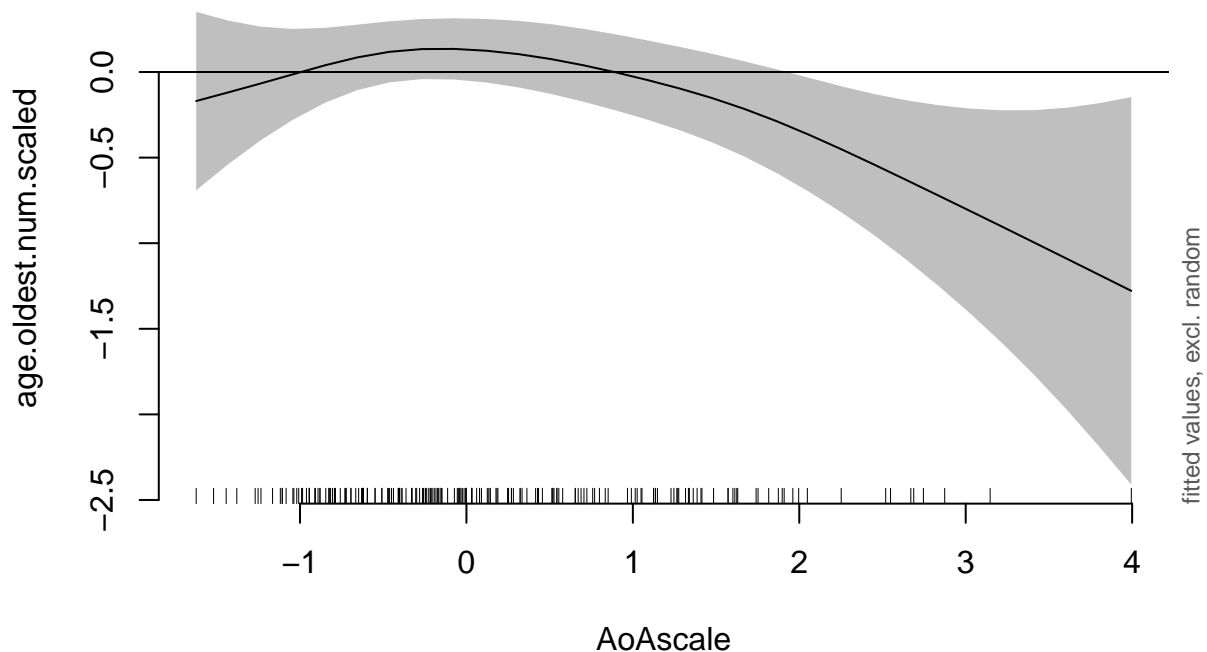
Final model has just age of acquisition as a main effect, so let's take away the random effects for other variables:

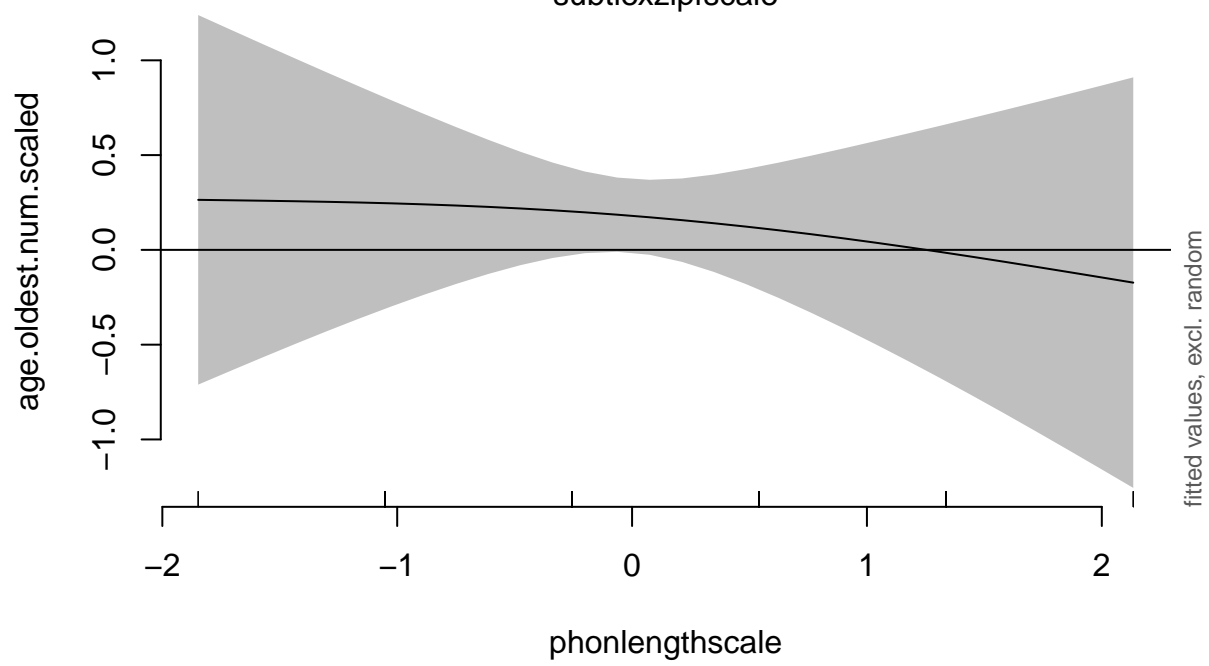
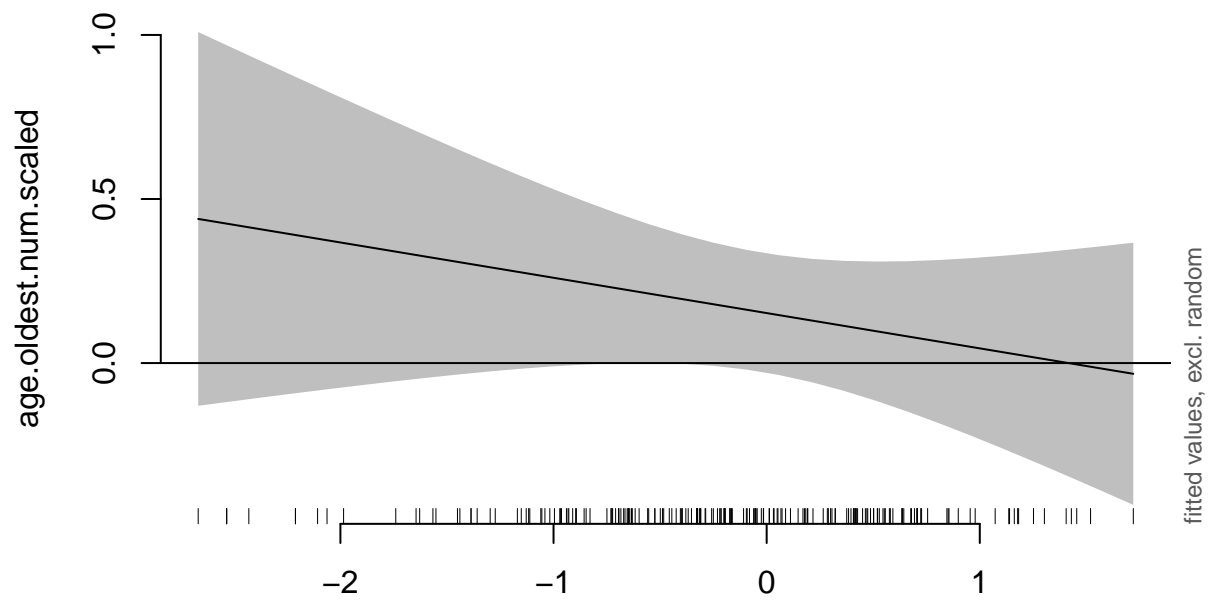
```
m5.age = bam(age.oldest.num.scaled~
  s(AoAscale) +
  s(cat,bs='re') +
  s(cat,AoAscale,bs='re'),
```

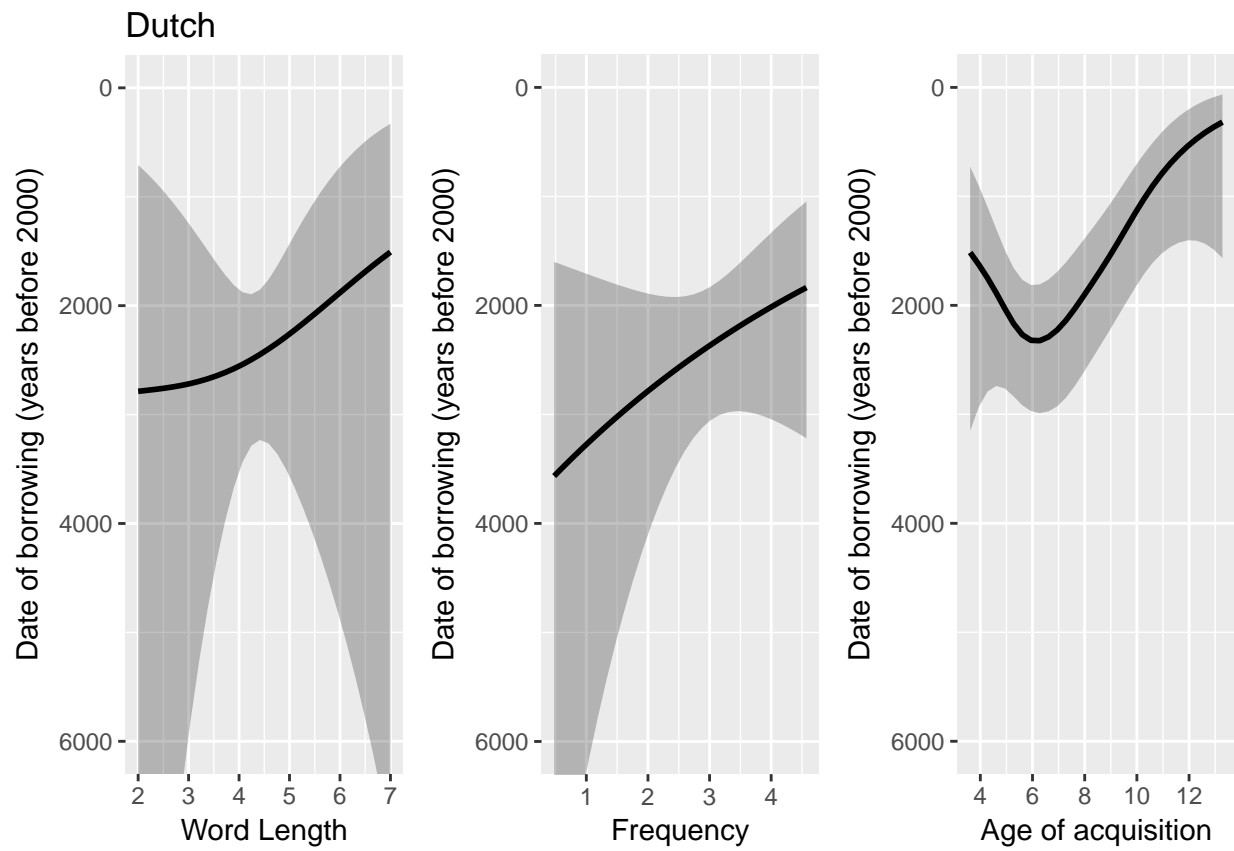
```
data = dutch[dutch$bor15==1,]
summary(m5.age)
```

```
##
## 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 acquisition result is relevant to the main results in the paper, but the others are shown for illustration.







pdf
2