Predicting date of entry

Introduction

Here we try to predict a word's date of entry into a language based on cognitive factors.

For age at which the borrowing occurred, we took the estimate of first entry into the recipient language from the WOLD database. For classifications that were given by historical language origin, rather than date, we took the estimations of the age for English from Grant (2009) and for Dutch from van der Sijs (2009), e.g., Proto-Germanic was taken as 2500BCE, and Proto West-Germanic was taken as 0CE. When the language origin indicated a range of years (e.g., Early Old English had date range 449-900CE), the later date was taken (e.g., for Early Old English, 900CE). Log of years before 2000CE was entered into the analyses as the point of borrowing. For word forms with distinct meanings (but similar borrowing status) we distinguished the earliest date of entry and the latest date of entry, e.g., for the English word "palm" the meaning related to the hand was first attested around 1300CE, and the meaning related to the tree was first attested in Early Old English (900CE). There were few words, however, with distinct ages of borrowing, and over the set of words for which there was clear evidence of borrowing, the correlation between earliest date of entry and latest date of entry was $r > .99$, $p < .001$, for both languages, and so for the remaining analyses, we only used the earliest date of entry measure.

Results

The first set of analyses investigated whether a word was borrowed or not. However, a more fine-grained measure of time since a word is borrowed can also be computed from the WOLD, though this provides only an analysis of the time since last change for vocabulary items that have a documented borrowing or language origin, not taking into account the words that remain stable. In order to investigate whether the point of first evidence of borrowing could be predicted by the psycholinguistic variables, we took only those words with clear evidence of borrowing and predicted the log of the years prior to 2000CE at which the borrowing was first attested. The same psycholinguistic variables as in the first analysis were predictors, and we included a random effect for part of speech.

For English, the date of entry was significantly predicted by word length (584 words, edf = 2.5, F = 7.43, $p < 0.0001$) and frequency (edf = 2.2, F = 0.14, p = 0.002, R2 for full model = 0.18). Longer words and less frequent words are more likely to have been borrowed more recently.

For Dutch, the date of entry was significantly predicted by age of acquisition (202 words, edf = 2.7, F = 3.2, p = 0.02). Words learned later are more likely to have been borrowed more recently. Frequency and length were not significant predictors. Indeed, the trend for frequency went in the opposite direction from the trend for English. The differences in results may be because there were only 202 borrowed words to analyse in Dutch, half as many as for English.

Model code

Load libraries

```
library(mgcv)
library(sjPlot)
library(lattice)
library(ggplot2)
library(dplyr)
library(party)
```

```
library(lmtest)
library(gridExtra)
library(scales)
library(itsadug)
library(ggfortify)
library(factoextra)
library(gridExtra)
library(reshape2)
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

English data

```
dataloan <- read.csv("../data/loanword8.csv",stringsAsFactors = F)
dataloan$bor15 <- ifelse(dataloan$borrowing==1,1, ifelse(dataloan$borrowing==5,0,NA))
dataloan$bor15.cat <- factor(dataloan$bor15)
```

Convert to numbers.

```
dataloan$subtlexzipf = as.numeric(dataloan$subtlexzipf)
dataloan$AoA = as.numeric(dataloan$AoA)
dataloan$conc = as.numeric(dataloan$conc)

aoaSD = sd(dataloan$AoA,na.rm = T)
aoaMean = mean(dataloan$AoA/aoaSD,na.rm=T)
dataloan$cat = factor(dataloan$cat)
```

Select only complete cases.

```
dataloan2 = dataloan[complete.cases(dataloan[,
               c("phonlength","AoA",
```

```
                "subtlexzipf", "cat",
                'conc','bor15')]),]
```

Scale and center:

```
dataloan2$AoAscale <- scale(dataloan2$AoA)

dataloan2$subtlexzipfscale <- scale(dataloan2$subtlexzipf)

phonlength.center = median(dataloan2$phonlength)
dataloan2$phonlengthscale <-
  dataloan2$phonlength - phonlength.center
phonlength.scale = sd(dataloan2$phonlengthscale)
dataloan2$phonlengthscale = dataloan2$phonlengthscale/phonlength.scale

attr(dataloan2$phonlengthscale,"scaled:scale") = phonlength.scale
attr(dataloan2$phonlengthscale,"scaled:center") = phonlength.center

dataloan2$concscale <- scale(dataloan2$conc)

dataloan2$cat = relevel(dataloan2$cat,"Noun")

dataloan2$AoA_objscaled = scale(dataloan2$AoA_obj)
```

Dutch data

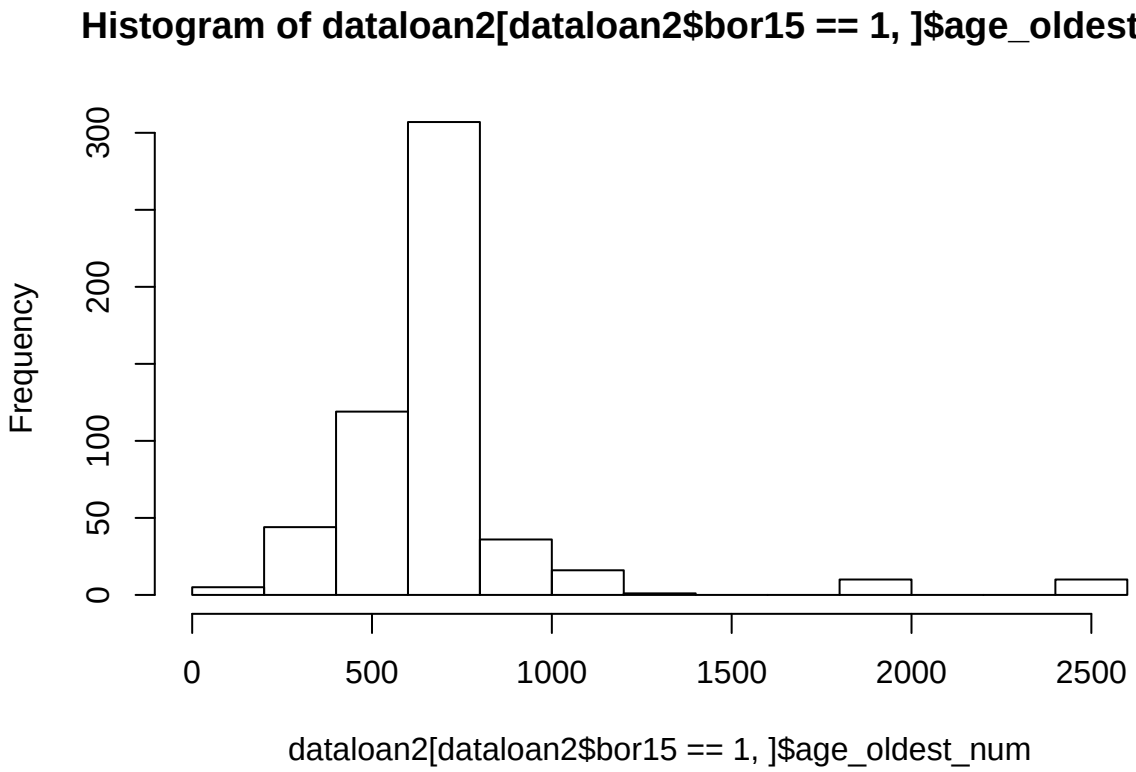
```
load("../data/loanwords_Dutch.Rdat")
```

English analysis

```
dataloan2$age_oldest_num = as.numeric(dataloan2$age_oldest_num)
# remove non-borrowed words
dataloan2[dataloan2$bor15!=1,]$age_oldest_num = NA
# Take log years
dataloan2$age_oldest_num.scaled = log10(dataloan2$age_oldest_num)
# Scale and center
dataloan2$age_oldest_num.scaled = scale(dataloan2$age_oldest_num.scaled)
```

Plot raw data

```
hist(dataloan2[dataloan2$bor15==1,]$age_oldest_num)
```




```
g.ageAoA = ggplot(dataloan2[dataloan2$bor15==1,],
                  aes(x=AoA, y=age_oldest_num))+
  geom_smooth() +
  coord_cartesian(ylim=c(250,1000))
scale_y_reverse()
```

```
## <ScaleContinuousPosition>
##  Range:
##  Limits:    0 --    1
```

```
g.ageLen = ggplot(dataloan2[dataloan2$bor15==1,],
                  aes(x=phonlength, y=age_oldest_num))+
  geom_smooth()+
  coord_cartesian(ylim=c(250,1000))
scale_y_reverse()
```

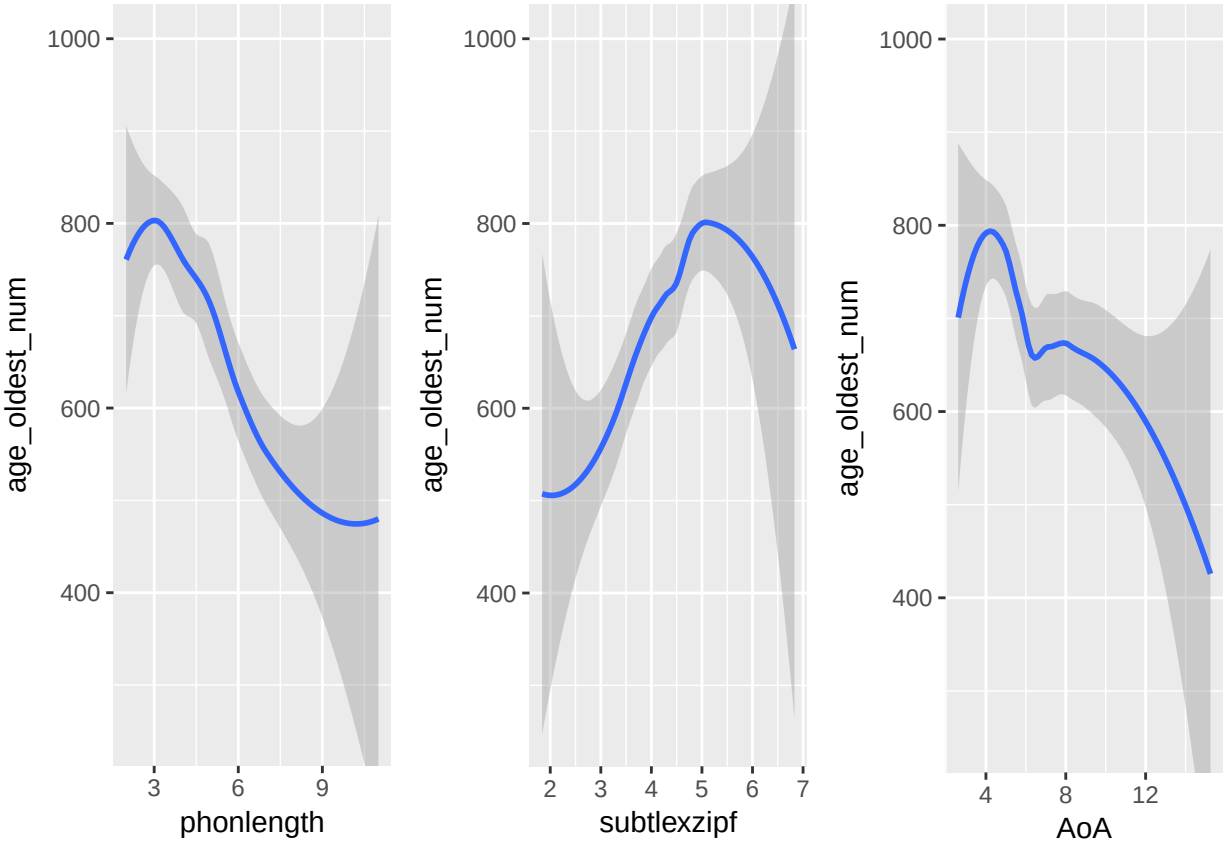
```
## <ScaleContinuousPosition>
##  Range:
##  Limits:    0 --    1
```

```
g.ageFreq = ggplot(dataloan2[dataloan2$bor15==1,],
                   aes(x=subtlexzipf, y=age_oldest_num))+
  geom_smooth()+
  coord_cartesian(ylim=c(250,1000))
scale_y_reverse()
```

```
## <ScaleContinuousPosition>
##  Range:
##  Limits:    0 --    1
```

```
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'
```

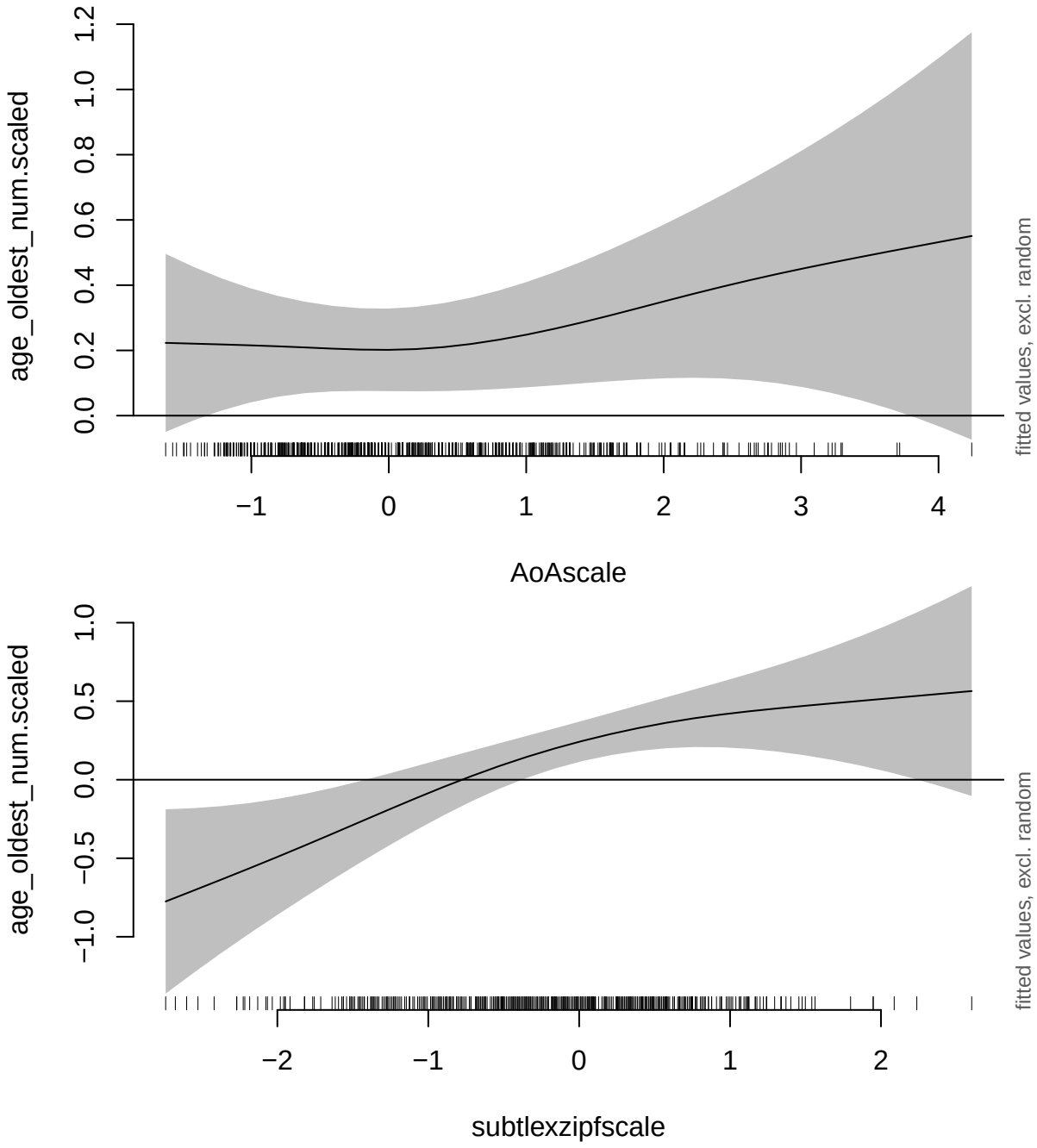


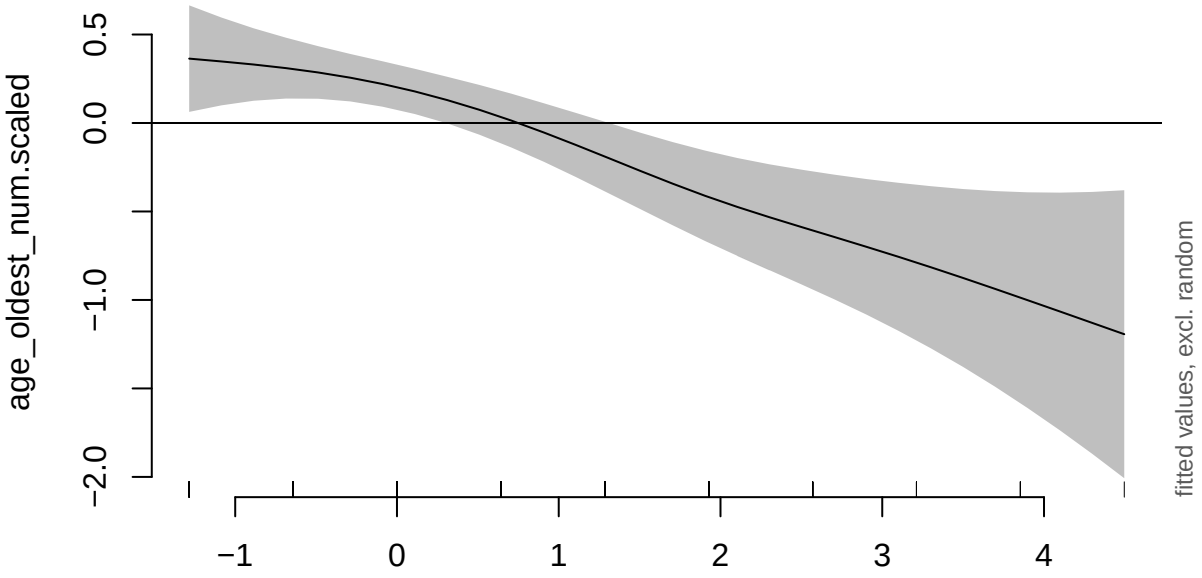
GAM model:

```
r   m0.age = bam(age_oldest_num.scaled~                s(phonlengthscale) +                s(AoAscale)
+                 s(subtlexzipfscale) +                 s(concscale) +                s(cat,bs='re')+
s(cat,phonlengthscale,bs='re')+                s(cat,AoAscale,bs='re')+                s(cat,subtlexzipf
s(cat,concscale,bs='re'),              data = dataloan2[dataloan2$bor15==1,])   summary(m0.age)
```

```
##    ## Family: gaussian    ## Link function: identity    ##    ## Formula:   ## age_oldest_num.scaled
~ s(phonlengthscale) + 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 t value Pr(>|t|)
## (Intercept) -0.004986   0.047084  -0.106    0.916   ##    ## Approximate significance
of smooth terms:   ##                                  edf Ref.df     F  p-value       ##
s(phonlengthscale)      2.450e+00  3.117 7.436 6.06e-05 ***   ## s(AoAscale)             1.804e+00
2.276 0.805  0.40780       ## s(subtlexzipfscale)     2.182e+00  2.797 5.487  0.00164 **
## s(concscale)            1.000e+00  1.000 0.141  0.70762       ## s(cat)                  6.237e-06
8.000 0.000  0.65025       ## s(cat,phonlengthscale)  1.023e+00  7.000 0.297  0.13964
## s(cat,AoAscale)         2.113e-06  8.000 0.000  0.80034       ## s(cat,subtlexzipfscale)
5.838e-01  8.000 0.112  0.22496       ## s(cat,concscale)        7.439e-06  8.000 0.000
0.99367       ## ---   ## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##    ## R-sq.(adj) =  0.175   Deviance explained = 18.8%   ## fREML = 737.37  Scale est.
= 0.82539   n = 548
```

Plot the model estimates. The code is hidden, but you can view it in the Rmd file.


age_oldest_num.scaled
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−1
0
1
2
3
4
AoAscale
fitted values, excl. random
age_oldest_num.scaled
−1.0
−0.5
0.0
0.5
1.0
−2
−1
0
1
2
subtlexzipfscale
fitted values, excl. random


age_oldest_num.scaled
0.5
0.0
−1.0
−2.0
−1
0
1
2
3
4
phonlengthscale
fitted values, excl. random


English
Date of borrowing (years before 2000)
1000
800
600
400
3
6
9
Word Length
2
3
4
5
6
7
Frequency
4
8
12
Age of acquisition

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

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

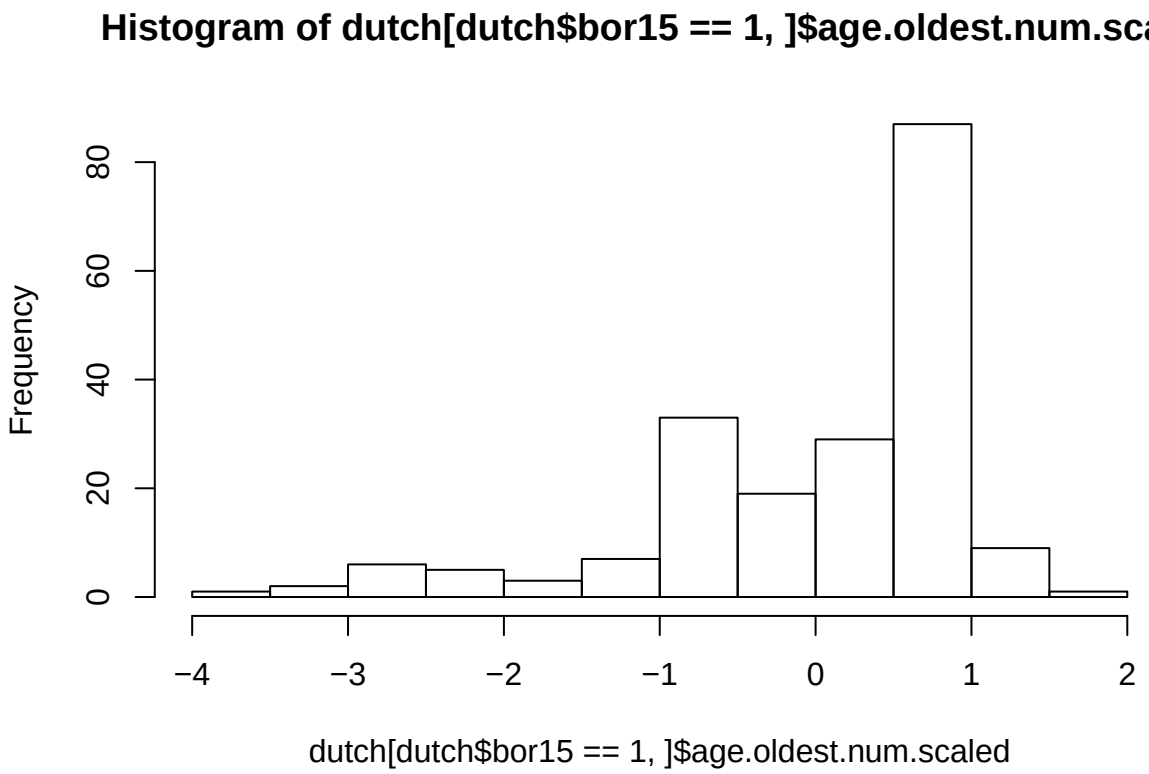
Ductch analysis

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

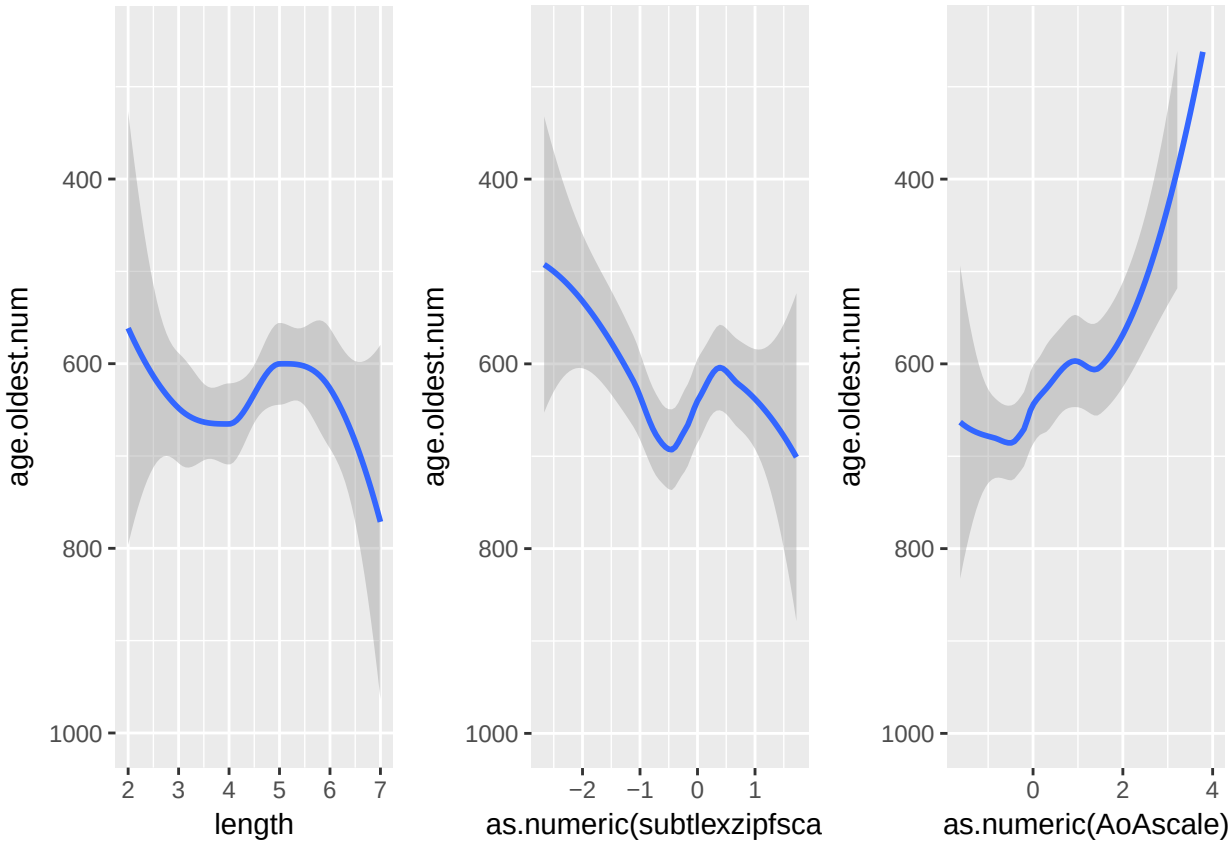
```
hist(dutch[dutch$bor15==1,]$age.oldest.num.scaled)
```




```
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
```

```
# 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
```

```
# 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
```

```
# 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
```

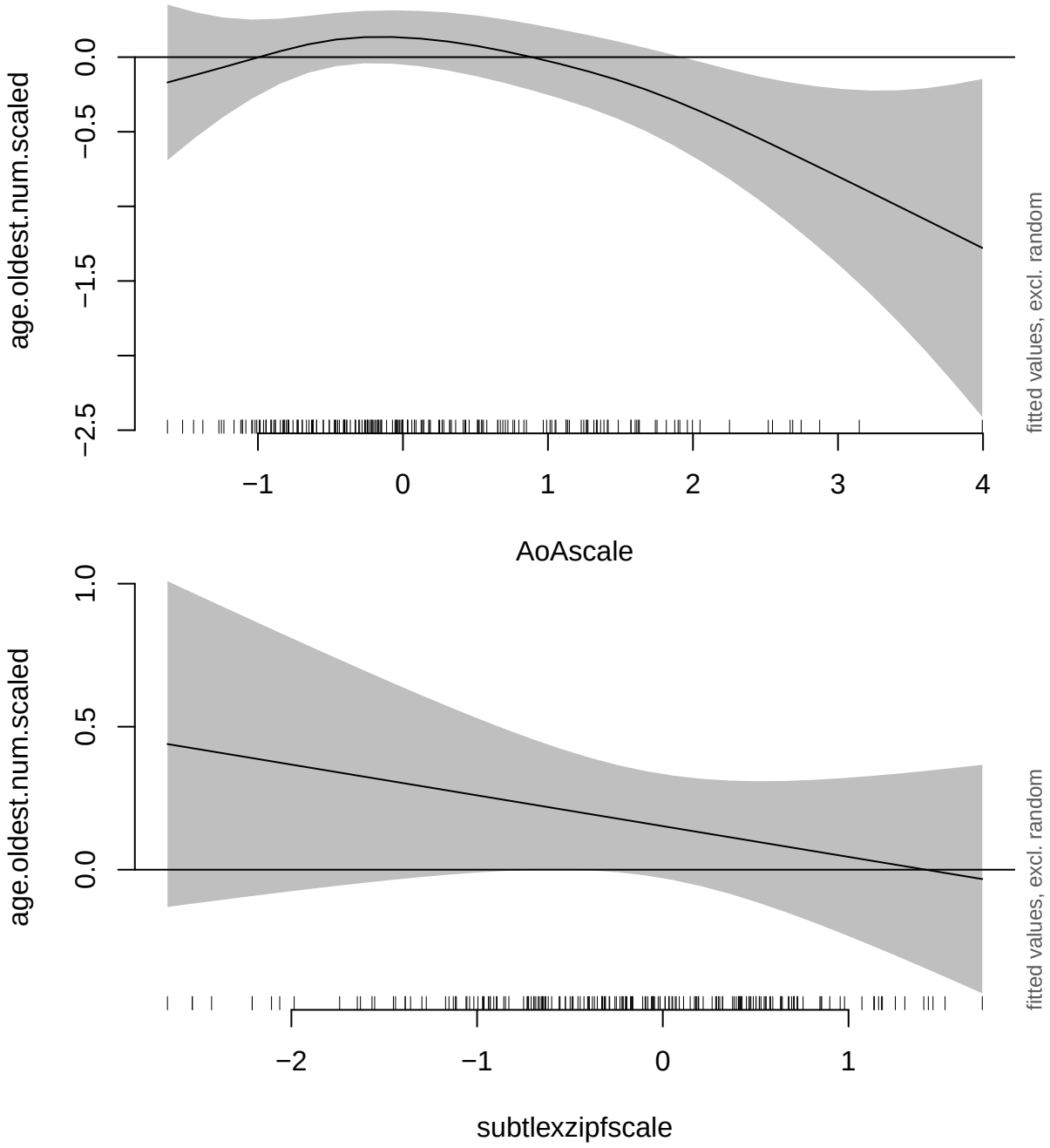
```
# Not sigificant
```

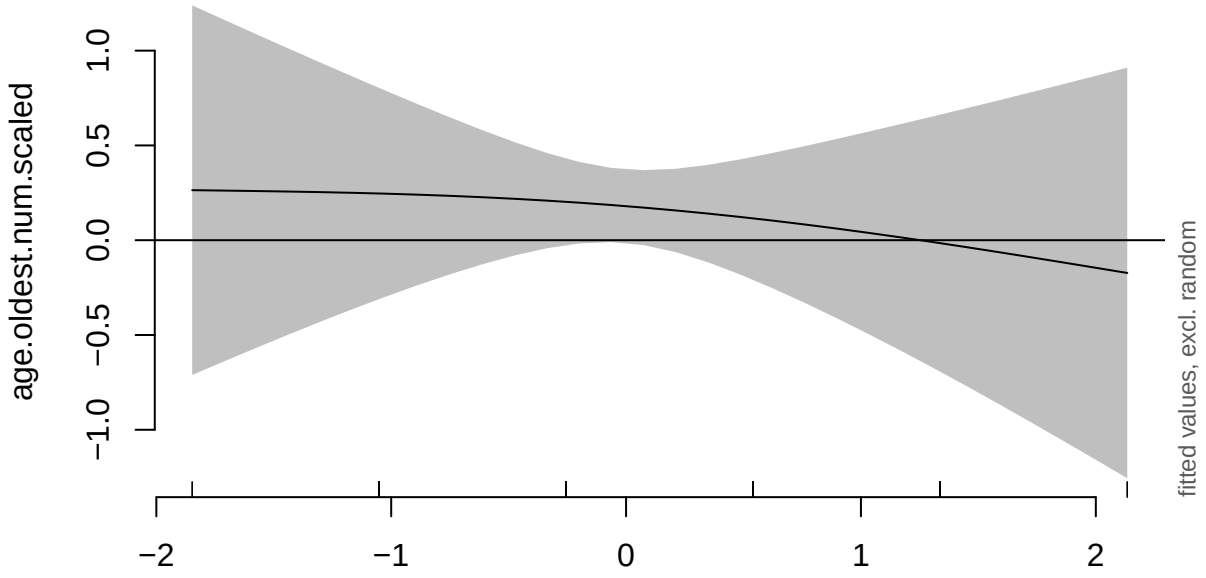
Final model has just age of acquisition as a main effect, so let's take away the random effects for other variables:

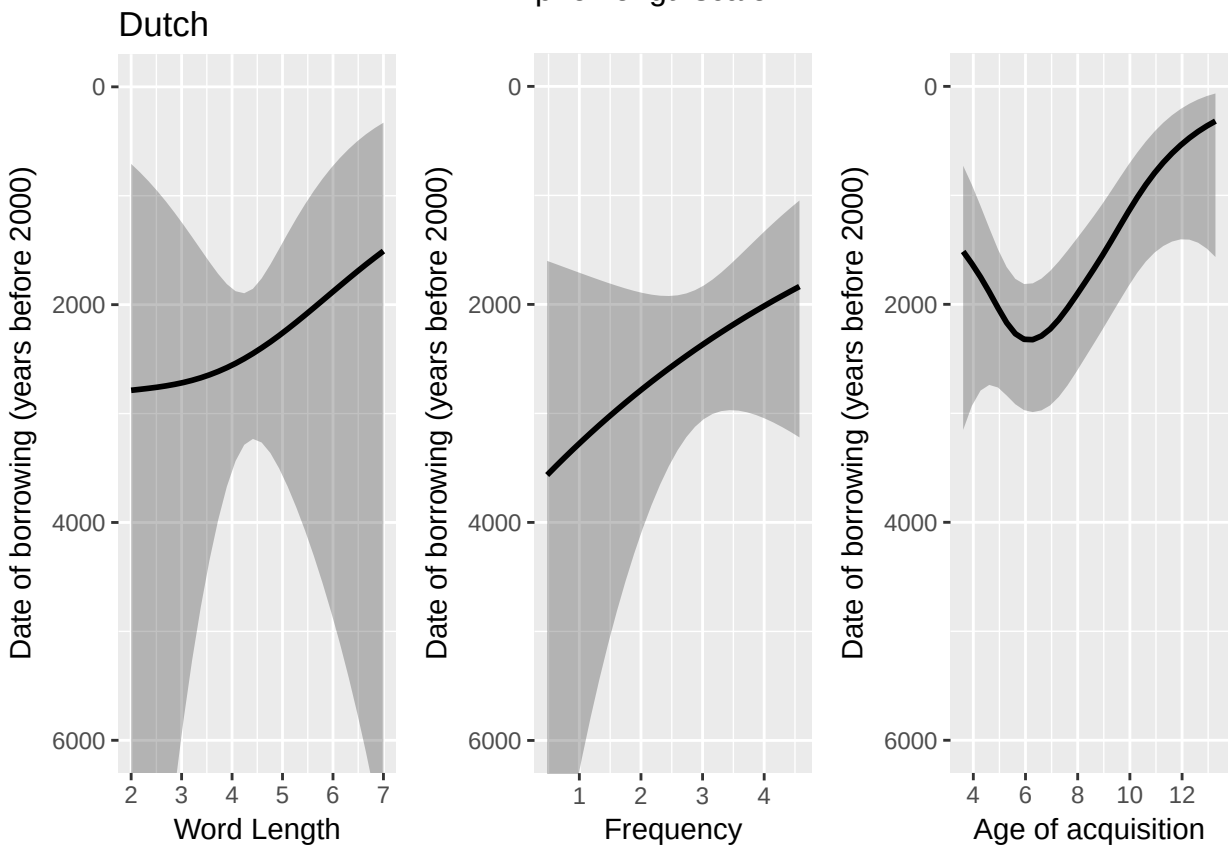
```
m5.age = bam(age.oldest.num.scaled~
s(AoAscale) +
s(cat,bs='re') +
s(cat,AoAscale,bs='re'),
data = dutch[dutch$bor15==1,])
summary(m5.age)
```

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


age.oldest.num.scaled
0.0
−0.5
−1.5
−2.5
−1
0
1
2
3
4
AoAscale
fitted values, excl. random
age.oldest.num.scaled
1.0
0.5
0.0
−2
−1
0
1
subtlexzipfscale
fitted values, excl. random


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


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

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

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