Cognitive influences in language evolution: English data

Introduction

We test whether rates of lexical replacement can be predicted by age of acuisition. The key variables in the dataset `pagel/loanword9.csv` are:

- word: orthographic form
- borrowing: whether the word is borrowed into English, according to WOLD
- phonology: phonological form, according to CELEX
- phonlength: number of phonological segments
- AoA: age of acuqisition, according to Kuperman et al. (2012)
- subtlexzipf: frequency from SUBTLEX
- pagel_rate: rate of lexical replacement according to Pagel, Atkinson & Meade (2007)
- cat: part of speech of the word

Pagel, Atkinson & Meade find that lexical replacement rates are higher for less frequent words, though the baseline rate differs by part of speech. We use mixed effects modelling to investigate the effects of length, age of acquisition and concreteness, with random intercepts for part of speech.

Load libraries

```
library(lme4)
library(sjPlot)
library(ggplot2)
library(gplots)
library(gridExtra)
library(mgcv)
```

Load data

Scale and center all variables:

```
d = read.csv("../pagel/loanword9.csv", stringsAsFactors = F)
d$pagel_rate = as.numeric(d$pagel_rate)
```

```
## Warning: NAs introduced by coercion
```

```
d$AoA = as.numeric(d$AoA)
```

```
## Warning: NAs introduced by coercion
```

```
d$subtlexzipf = as.numeric(d$subtlexzipf)
```

```
## Warning: NAs introduced by coercion
```

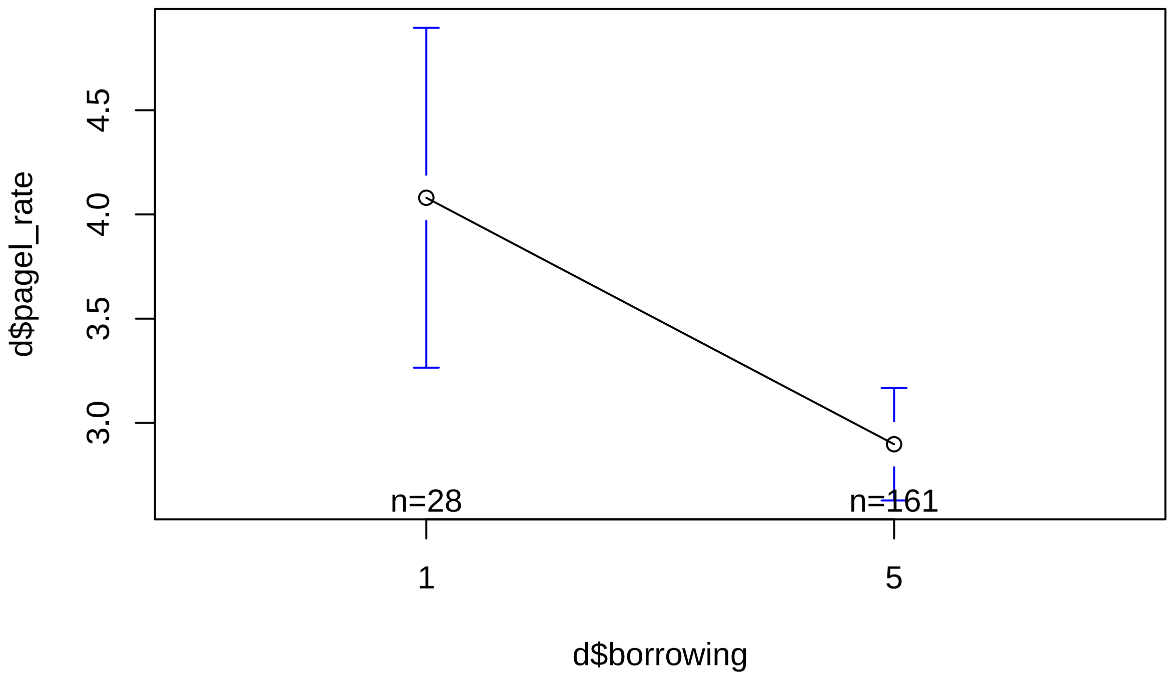
```
d$phonlength = as.numeric(d$phonlength)
d$conc = as.numeric(d$conc)
```

```
## Warning: NAs introduced by coercion
```

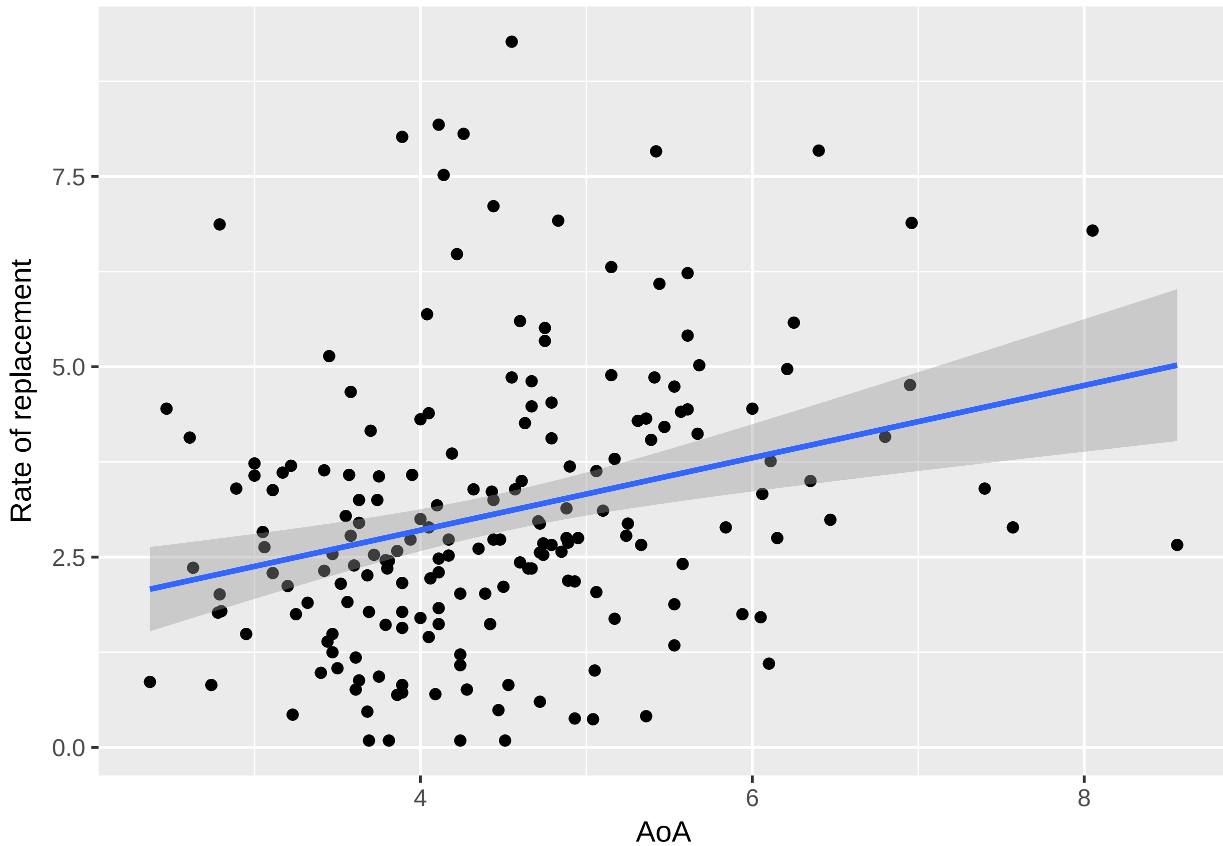
```r
# group borrowing confidence
d$borrowing[d$borrowing==2] = 1
d$borrowing[d$borrowing==3] = NA
d$borrowing[d$borrowing==4] = 5
d$borrowing[d$borrowing==7] = NA

# complete cases only
d = d[complete.cases(d[,c("borrowing",'AoA','phonlength','conc','pagel_rate')]),]
```

Plot raw data:

```r
plotmeans(d$pagel_rate~d$borrowing)
```




```r
ggplot(d, aes(AoA, as.numeric(pagel_rate))) +
  geom_point() + stat_smooth(method="lm") +
  ylab("Rate of replacement")
```



Scale variables for analysis:

```
d$pagel_rate = scale(d$pagel_rate)
d$AoA.scale = scale(d$AoA)
d$subtlexzipf.scale = scale(d$subtlexzipf)
d$phonlength.scale = scale(d$phonlength)
d$conc.scale = scale(d$conc)

# most frequent category as intercept
d$borrowing = factor(d$borrowing,levels=c(5,1),labels = c("no","yes"))
d$borrowing.num = as.numeric(d$borrowing)-1
```

Modelling

Null model:

```
m0 = lmer(pagel_rate ~ 1 + (1|cat), data=d)
```

As expected, borrowed words have a higer rate of change:

```
m1 = lmer(pagel_rate ~
          borrowing +
          (1|cat),
         data = d)
anova(m0,m1)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: d
## Models:
## m0: pagel_rate ~ 1 + (1 | cat)
## m1: pagel_rate ~ borrowing + (1 | cat)
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m0  3 528.17 537.90 -261.09   522.17
## m1  4 518.80 531.77 -255.40   510.80 11.371      1   0.000746 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The original model in Pagel et al. found relationships between rate of change, frequency and length:

```r
m2 = lmer(pagel_rate ~
          borrowing +
          subtlexzipf.scale +
          (1 | cat),
        data = d)
m3 = lmer(pagel_rate ~
          borrowing +
          subtlexzipf.scale +
          phonlength.scale +
          (1 | cat),
        data = d)
anova(m1,m2,m3)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: d
## Models:
## m1: pagel_rate ~ borrowing + (1 | cat)
## m2: pagel_rate ~ borrowing + subtlexzipf.scale + (1 | cat)
## m3: pagel_rate ~ borrowing + subtlexzipf.scale + phonlength.scale +
## m3:     (1 | cat)
##    Df    AIC    BIC  logLik deviance   Chisq Chi Df Pr(>Chisq)
## m1  4 518.80 531.77 -255.40   510.80
## m2  5 509.09 525.30 -249.55   499.09 11.7106      1  0.0006214 ***
## m3  6 505.04 524.49 -246.52   493.04  6.0557      1  0.0138616 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(m3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: pagel_rate ~ borrowing + subtlexzipf.scale + phonlength.scale +
##     (1 | cat)
##    Data: d
##
## REML criterion at convergence: 502.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0268 -0.6162 -0.2251  0.5129  2.7396
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  cat      (Intercept) 0.2484   0.4984
```

```
##  Residual             0.7513   0.8668
## Number of obs: 189, groups:  cat, 10
##
## Fixed effects:
##                    Estimate Std. Error t value
## (Intercept)         0.01825    0.20291   0.090
## borrowingyes        0.49831    0.18476   2.697
## subtlexzipf.scale -0.24474    0.08498  -2.880
## phonlength.scale   0.16917    0.06898   2.452
##
## Correlation of Fixed Effects:
##             (Intr) brrwng sbtlx.
## borrowingys -0.127
## sbtlxzpf.sc -0.308  0.039
## phnlngth.sc  0.050 -0.197  0.215
```

Indeed, both length and freuqnecy are significant predictors of rate of change.

We can test whether there are non-linear effects for frequency and length:

```
m3.freq2 = lmer(pagel_rate ~
          borrowing +
          subtlexzipf.scale +
          I(subtlexzipf.scale^2) +
          phonlength.scale +
          (1 | cat),
        data = d)
m3.len2 = lmer(pagel_rate ~
          borrowing +
          subtlexzipf.scale +
          I(phonlength.scale^2) +
          phonlength.scale +
          (1 | cat),
        data = d)
anova(m3, m3.freq2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: d
## Models:
## m3: pagel_rate ~ borrowing + subtlexzipf.scale + phonlength.scale +
## m3:     (1 | cat)
## m3.freq2: pagel_rate ~ borrowing + subtlexzipf.scale + I(subtlexzipf.scale^2) +
## m3.freq2:     phonlength.scale + (1 | cat)
##          Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m3        6 505.04 524.49 -246.52   493.04
## m3.freq2  7 507.02 529.72 -246.51   493.02 0.0115      1     0.9147
```

```
anova(m3, m3.len2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: d
## Models:
## m3: pagel_rate ~ borrowing + subtlexzipf.scale + phonlength.scale +
## m3:     (1 | cat)
## m3.len2: pagel_rate ~ borrowing + subtlexzipf.scale + I(phonlength.scale^2) +
```

```
## m3.len2:     phonlength.scale + (1 | cat)
##          Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## m3        6 505.04 524.49 -246.52   493.04
## m3.len2   7 505.16 527.86 -245.58   491.16 1.871      1     0.1714
```

There is no significant non-linear (quadratic) effect of frequency or length.

We can now add age of acquisition:

```r
m4 = lmer(pagel_rate ~
           borrowing +
           phonlength.scale +
           subtlexzipf.scale +
           AoA.scale +
          (1 |cat),
         data = d)
anova(m3,m4)
```

```
## refitting model(s) with ML (instead of REML)

## Data: d
## Models:
## m3: pagel_rate ~ borrowing + subtlexzipf.scale + phonlength.scale +
## m3:     (1 | cat)
## m4: pagel_rate ~ borrowing + phonlength.scale + subtlexzipf.scale +
## m4:     AoA.scale + (1 | cat)
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m3  6 505.04 524.49 -246.52   493.04
## m4  7 500.55 523.24 -243.27   486.55 6.4894      1    0.01085 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(m4)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: pagel_rate ~ borrowing + phonlength.scale + subtlexzipf.scale +
##     AoA.scale + (1 | cat)
##    Data: d
##
## REML criterion at convergence: 500.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0850 -0.6595 -0.1920  0.4955  2.8777
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  cat      (Intercept) 0.2093   0.4575
##  Residual             0.7338   0.8566
## Number of obs: 189, groups:  cat, 10
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)       -0.06110    0.19282  -0.317
## borrowingyes       0.50080    0.18250   2.744
## phonlength.scale   0.15176    0.06849   2.216
## subtlexzipf.scale -0.15846    0.08969  -1.767
```

```
## AoA.scale          0.17481    0.06942   2.518
##
## Correlation of Fixed Effects:
##             (Intr) brrwng phnln. sbtlx.
## borrowingys -0.132
## phnlngth.sc  0.067 -0.196
## sbtlxzpf.sc -0.345  0.038  0.163
## AoA.scale   -0.160  0.005 -0.103  0.372
```

Age of acquisition significantly improves the model and has an effect size similar to length (also, frequency has a much weaker effect).

Add concreteness:

```
m5 = lmer(pagel_rate ~
           phonlength.scale +
           subtlexzipf.scale +
           AoA.scale +
           borrowing +
           conc.scale +
            (1|cat),
         data = d)
anova(m4,m5)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: d
## Models:
## m4: pagel_rate ~ borrowing + phonlength.scale + subtlexzipf.scale +
## m4:     AoA.scale + (1 | cat)
## m5: pagel_rate ~ phonlength.scale + subtlexzipf.scale + AoA.scale +
## m5:     borrowing + conc.scale + (1 | cat)
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m4  7 500.55 523.24 -243.27   486.55
## m5  8 501.03 526.96 -242.51   485.03 1.5177      1      0.218
```

Concreteness does not significantly improve the fit of the model.

Final model:

```
finalModel = m4
summary(finalModel)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: pagel_rate ~ borrowing + phonlength.scale + subtlexzipf.scale +
##     AoA.scale + (1 | cat)
##    Data: d
##
## REML criterion at convergence: 500.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0850 -0.6595 -0.1920  0.4955  2.8777
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
```

```
##  cat      (Intercept) 0.2093   0.4575
##  Residual             0.7338   0.8566
## Number of obs: 189, groups:  cat, 10
##
## Fixed effects:
##                    Estimate Std. Error t value
## (Intercept)        -0.06110    0.19282  -0.317
## borrowingyes        0.50080    0.18250   2.744
## phonlength.scale    0.15176    0.06849   2.216
## subtlexzipf.scale -0.15846    0.08969  -1.767
## AoA.scale           0.17481    0.06942   2.518
##
## Correlation of Fixed Effects:
##             (Intr) brrwng phnln. sbtlx.
## borrowingys -0.132
## phnlngth.sc  0.067 -0.196
## sbtlxzpf.sc -0.345  0.038  0.163
## AoA.scale   -0.160  0.005 -0.103  0.372
```

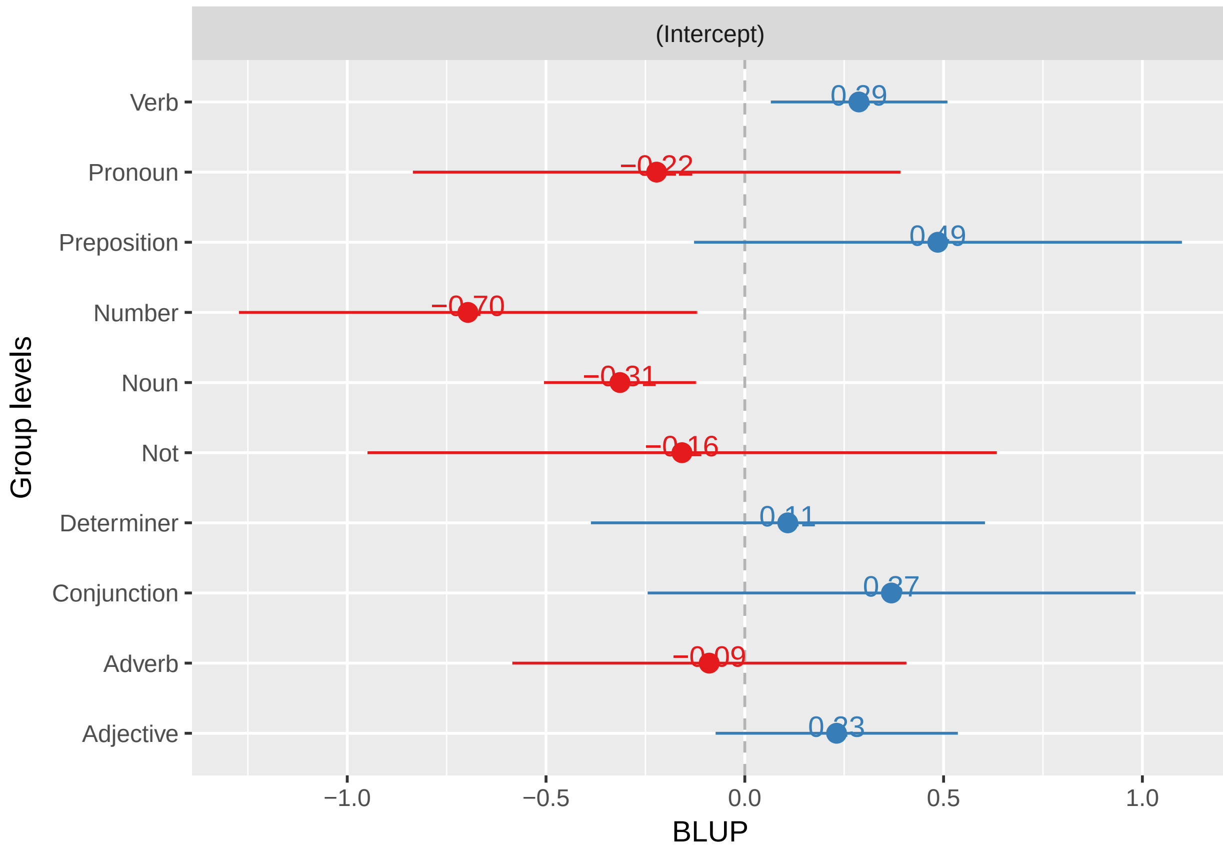
```
# Rough measure of model fit (R2):
cor(predict(finalModel),d$pagel_rate)^2
```

```
##           [,1]
## [1,] 0.3058029
```

Random intercepts for part of speech:

```
sjp.lmer(finalModel)
```

```
## Plotting random effects...
```



Model estimates

```
px =sjp.lmer(finalModel, 'eff', c("AoA.scale","phonlength.scale","subtlexzipf.scale"),
             show.ci = T, prnt.plot = F, facet.grid = F)
```

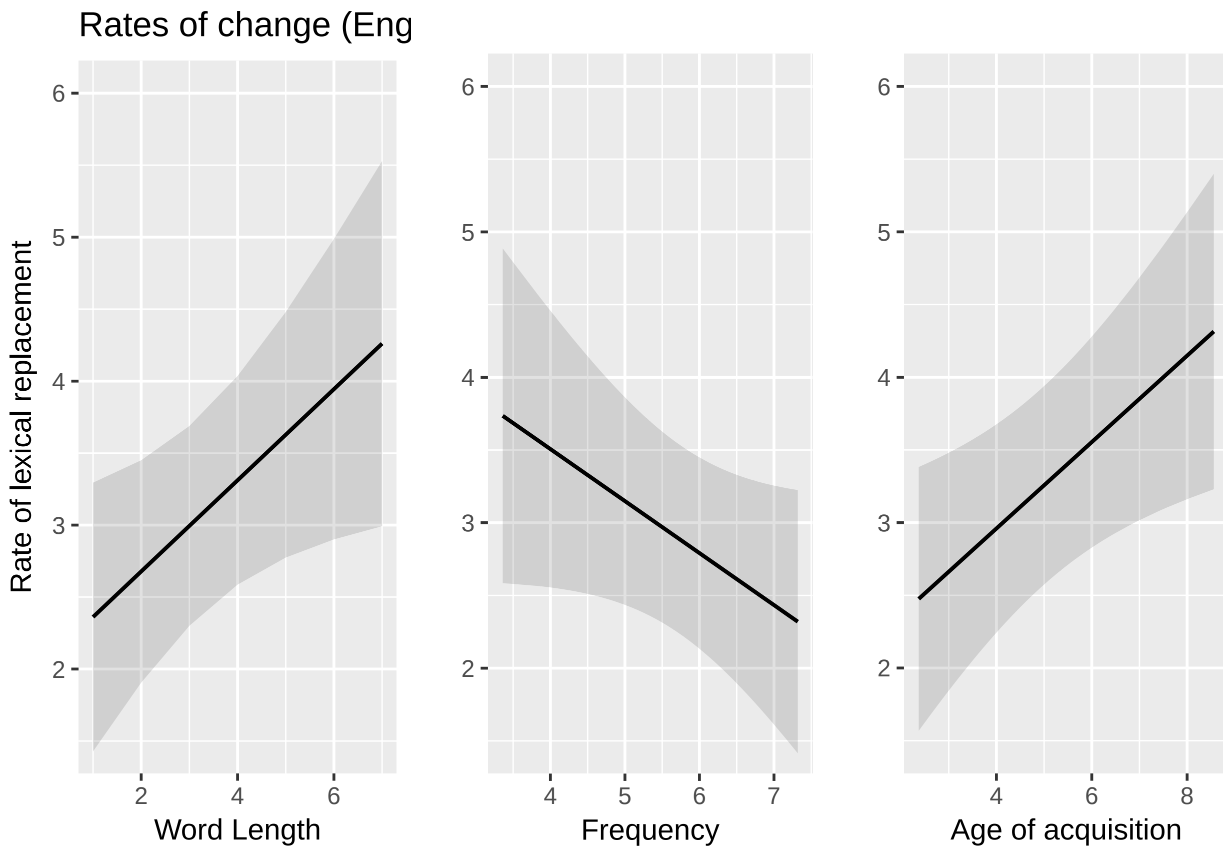
```
## Warning: package 'bindrcpp' was built under R version 3.3.2
```

```
pr.sc = attr(d$pagel_rate,"scaled:scale")
pr.cn = attr(d$pagel_rate,"scaled:center")
# length
p1 = px$plot.list[[1]]
p1$data$x = p1$data$x *attr(d$phonlength.scale,"scaled:scale") +
  attr(d$phonlength.scale,"scaled:center")
p1$data$y = p1$data$y * pr.sc + pr.cn
p1$data$lower = p1$data$lower * pr.sc + pr.cn
p1$data$upper = p1$data$upper * pr.sc + pr.cn
p1 = p1 + ggtitle("Rates of change (English)") +
  xlab("Word Length") +
  ylab("")+
  ylab("Rate of lexical replacement")+
  coord_cartesian(ylim=c(1.5,6))

# frequency
p2 = px$plot.list[[2]]
p2$data$x = p2$data$x *attr(d$subtlexzipf.scale,"scaled:scale") +
  attr(d$subtlexzipf.scale,"scaled:center")
```

```r
p2$data$y = p2$data$y * pr.sc + pr.cn
p2$data$lower = p2$data$lower * pr.sc + pr.cn
p2$data$upper = p2$data$upper * pr.sc + pr.cn
p2 = p2 + ggtitle("") +
  xlab("Frequency") +
  ylab("")+
  coord_cartesian(ylim=c(1.5,6))

# Age of acquisition
p3 = px$plot.list[[3]]
p3$data$x = p3$data$x *attr(d$AoA.scale,"scaled:scale") +
  attr(d$AoA.scale,"scaled:center")
p3$data$y = p3$data$y * pr.sc + pr.cn
p3$data$lower = p3$data$lower * pr.sc + pr.cn
p3$data$upper = p3$data$upper * pr.sc + pr.cn
p3 = p3 + ggtitle("") +
  xlab("Age of acquisition") +
  ylab("") +
  coord_cartesian(ylim=c(1.5,6))
pdf(file='../results/graphs/Pagel_RatesOfChange.pdf',
    height =3,width = 8)
grid.arrange(p1,p2,p3, nrow=1)
dev.off()
```

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

```r
grid.arrange(p1,p2,p3, nrow=1)
```



Non-linear model

We can test a non-linear model

```
d$cat = factor(d$cat)
d$cat = relevel(d$cat,"Noun")
m0.GAM  = bam(pagel_rate~
    s(phonlength.scale, k=5) +
    s(subtlexzipf.scale) +
    s(AoA.scale) +
    s(cat,bs='re'),
    data = d)
summary(m0.GAM)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## pagel_rate ~ s(phonlength.scale, k = 5) + s(subtlexzipf.scale) +
##     s(AoA.scale) + s(cat, bs = "re")
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.008816   0.192210   0.046    0.963
##
```

```
## Approximate significance of smooth terms:
##                         edf Ref.df     F  p-value
## s(phonlength.scale)  1.236  1.434 6.022   0.0134 *
## s(subtlexzipf.scale) 1.000  1.000 3.256   0.0728 .
## s(AoA.scale)         1.000  1.000 6.177   0.0138 *
## s(cat)               5.609  9.000 3.046 1.95e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.241   Deviance explained = 27.6%
## fREML = 252.99  Scale est. = 0.75926   n = 189
```

The results are very similar, and all edf scores are very close to 1 (linear). Given the small amount of data, we prefer the linear mixed effects model above.

Summary

There was a significant main effect of whether the word is borrowed (log likelihood difference = 5.7 , df = 1 , Chi Squared = 11.37 , p = 0.00075).

There was a significant main effect of frequency (log likelihood difference = 5.9 , df = 1 , Chi Squared = 11.71 , p = 0.00062).

There was a significant main effect of word length (log likelihood difference = 3 , df = 1 , Chi Squared = 6.06 , p = 0.014).

There was a significant main effect of age of acquisition (log likelihood difference = 3.2 , df = 1 , Chi Squared = 6.49 , p = 0.011).

There was no significant main effect of concreteness (log likelihood difference = 0.76 , df = 1 , Chi Squared = 1.52 , p = 0.22).

Rates of change are higher for:

- Borrowed words
- Less frequent words
- Longer words
- Words acquired later in childhood

References

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. Behavior Research Methods, 44(4), 978-990.

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Nature, 449(7163), 717-720.