Comprehending Your Neighbour's English Errors: A Mixed Models Analysis

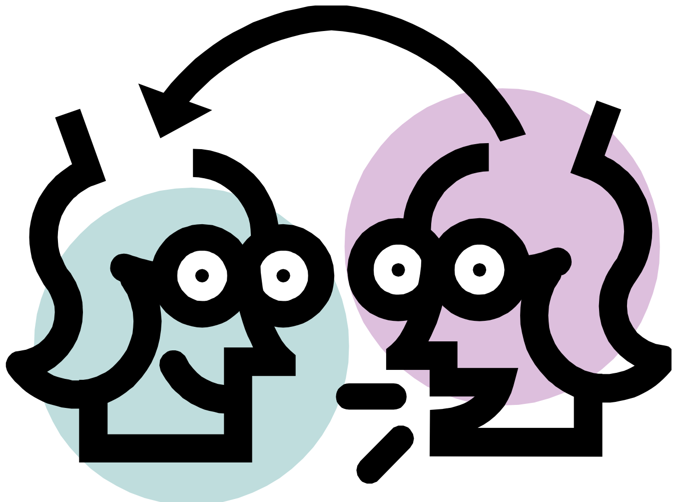
Research Methodology and Statistics 2014

Kristy James

Speaking in a lingua franca

- English as current choice
- Different norms from native speech
- CLI from L1 and limited proficiency can result in errors

Speaker **and** Listener effects:
Present the story accurately in comprehensible English (create stimulus properties), so listener can decipher meaning (dependent on listener factors)



Three language combinations

Family	Listeners	Speakers
Romance	Portuguese	Spanish
Slavic	Slovenian	Croatian
Germanic	Danish	Swedish

vs. English produced by these speakers, eg Spanish-accented English

Contents

› Introduce Experiment and Motivation
› Options for statistical approaches
› Mixed Models – a little theory
› A Generated Data analysis
› Conclusion

Experimental Overview

› Model a storytelling situation – Speaker participants tell one story each in L2 English and L1.

› Listener comprehension tested with multiple-choice questions via an online platform.

› (Conclude whether comprehension is better in English or the related language.)

› **Explain which errors in English most negatively affect comprehension – pronunciation, vocabulary or grammar.**

Motivation: Comprehending a related language

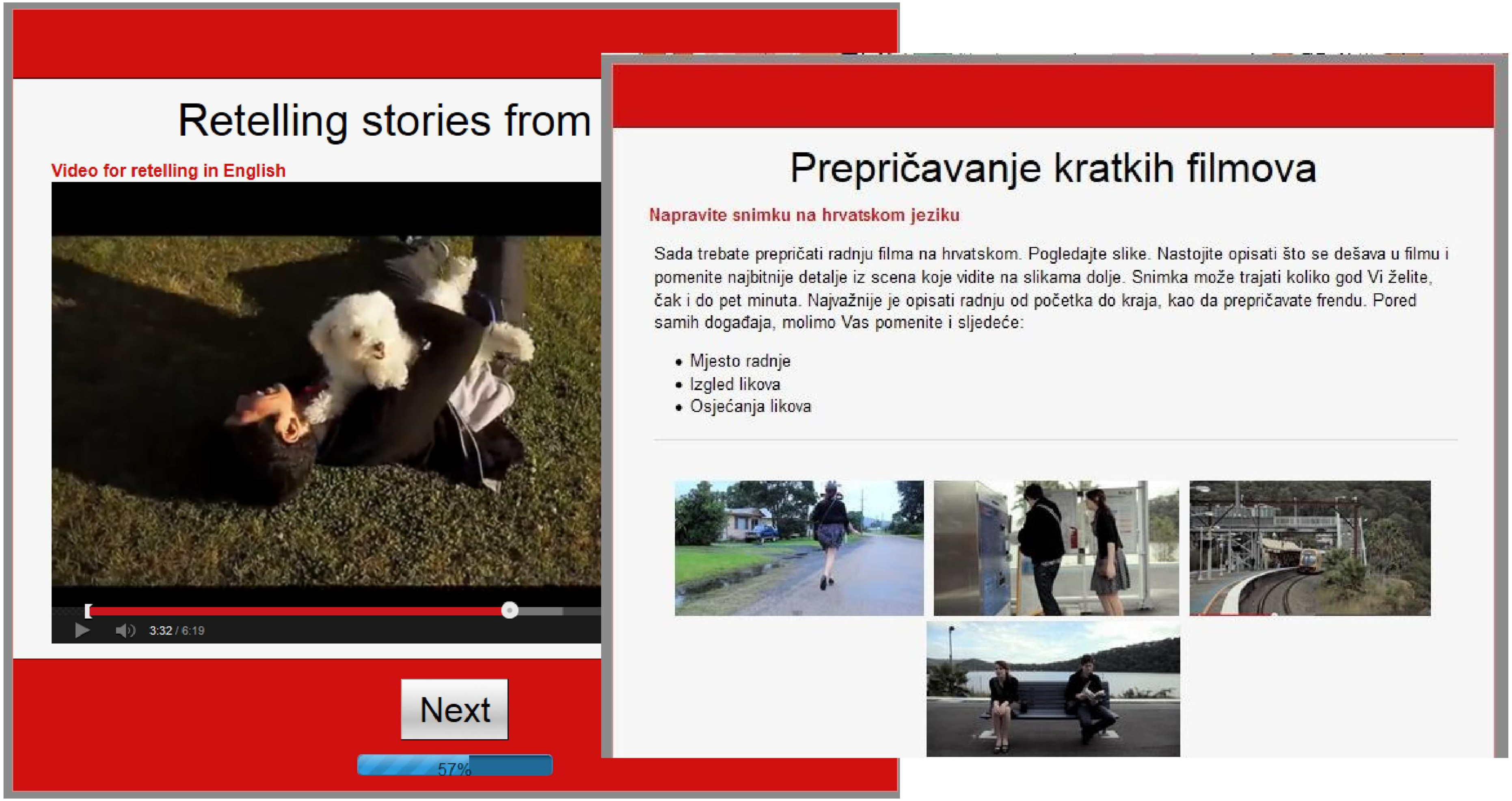
› Trending theme – cross-border communication
› Investigations into lexical/phonological distance
› Effects of prior exposure, schooling etc, as well as individual factors
› Our choice of normal speech (not foreigner-oriented)

Listener factors: Speaker speaks easily in native language (SP), listener factors determine comprehension (LF).

Methodology

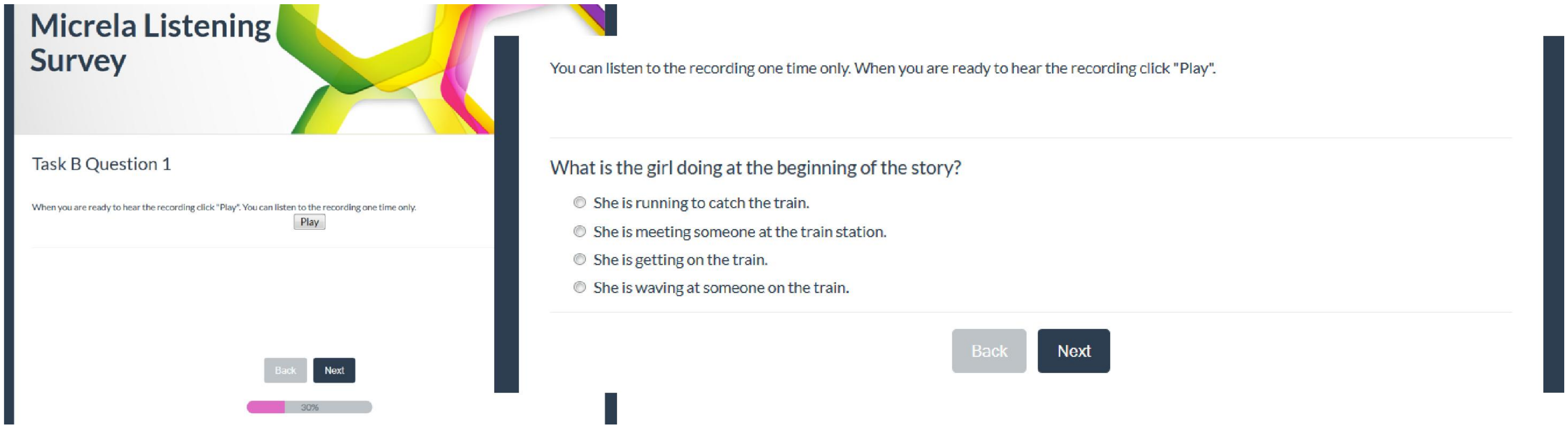
› Elicit retellings based on two silent short films from 20-30 participants per language combination
› Select core speakers that cover canonical topics – approx. 20 stimuli selected
› Segment recordings into audio fragments that are relevant to a particular topic (12 'questions' generated)
› Assign participants (listeners) to a random speaker (approx. 5 per stimuli), expose to audio fragment once and reveal comprehension question

Eliciting Retellings



Measuring Comprehension

- 12 multiple-choice questions, 6 in English, 6 in related language
- Crossed-design – Film A and Film B

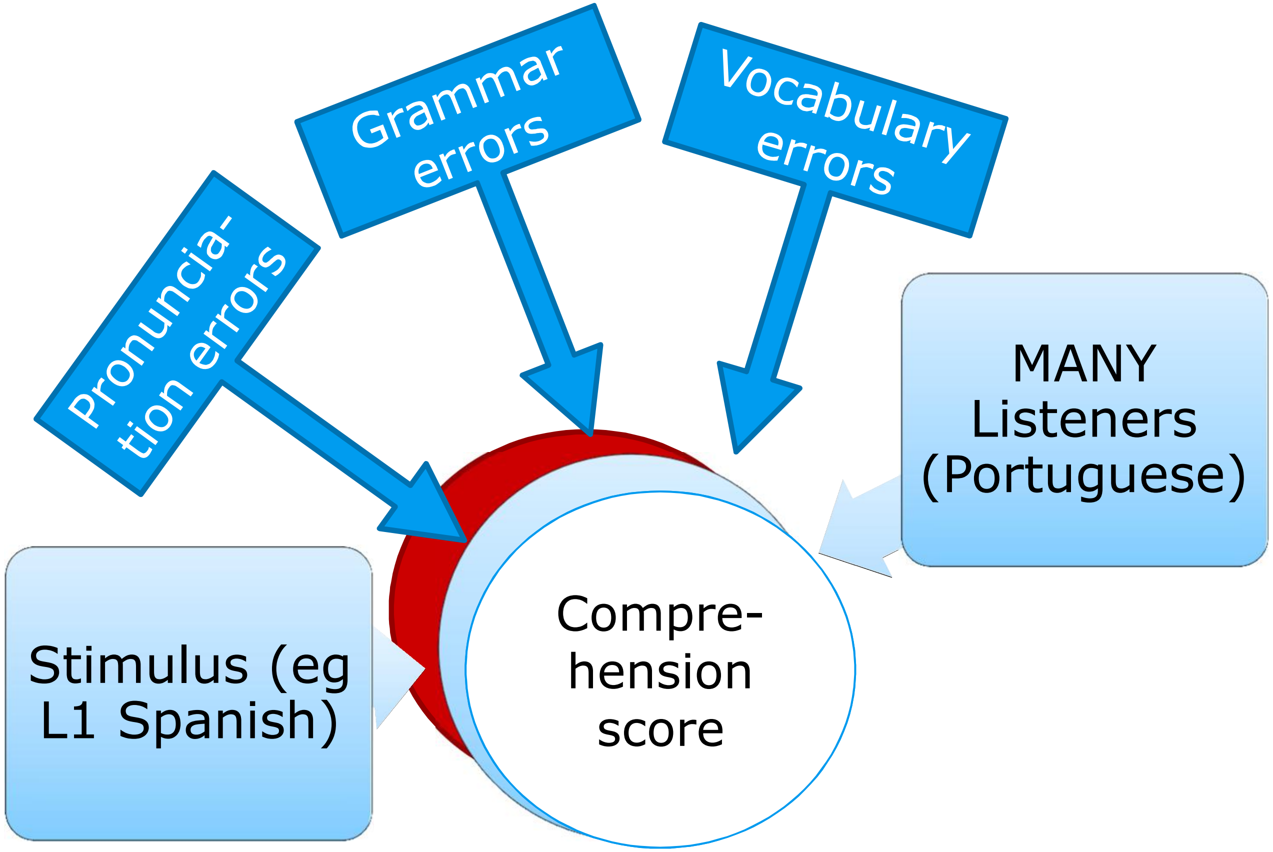


Variation at:

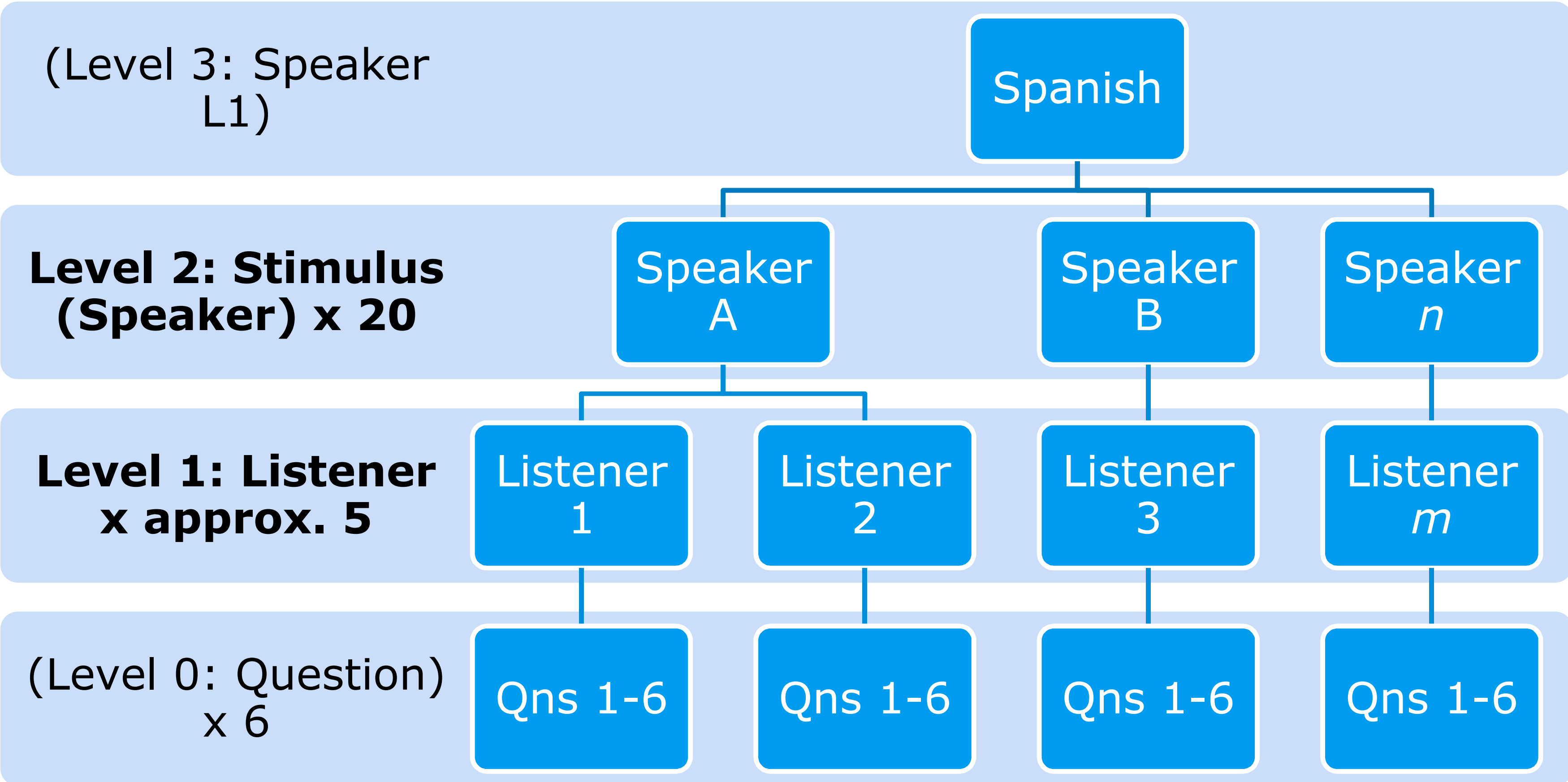
› Speaker Level:
  - S01
  - S06
  - S09
  - Among others no. pronunciation errors, no. grammar errors, no. vocabulary errors

› Listener Level:
  - English proficiency, exposure to English, exposure to a Spanish accent

Explaining the variance



Nested Data


(Level 3: Speaker L1)
Spanish
Level 2: Stimulus (Speaker) x 20
Speaker A
Speaker B
Speaker n
Level 1: Listener x approx. 5
Listener 1
Listener 2
Listener 3
Listener m
(Level 0: Question) x 6
Qns 1-6
Qns 1-6
Qns 1-6
Qns 1-6

Possible statistical methods

	RM ANOVA	Linear/Multiple Regression	Logistic regression	Mixed Models
When used in this experimental context:	For comparing overall success of ELF vs ReLa	When one listener hears each speaker (numerical response)	When one listener hears each speaker (categorical response)	Multiple listeners hear each speaker, can handle either numerical/categorical response
Assumptions:	Assumptions of independence; homogeneity of variance, balanced design	Assumes homogeneity of regression slopes (eg vocabulary errors may aid understanding); Error rating per stimulus	See left; ideally requires annotation of errors per question	No assumption of independence Robust against missing data

Linear Regression

$$y_i = a + bx_i + e_i$$

- a - intercept
- b - slope
- e - error
- Numeric (or categorical) independent variables, numerical response variable
- Numerical response: eg how many correct answers
- Intercept and slope represent average over all data points
- This experiment: with repetition of stimuli (these vary randomly) and potentially unbalanced numbers, therefore require taking averages – throwing away data

Logistic Regression

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 x_1)}}$$

- Numeric (or categorical) independent variable, one categorical response variable
- Eg for each question, P that answer is correct

Advantages of MM

Level 1: $\text{speech.rate}_i = a_{j[i]} + b \times \text{context}_i + e_i$

Level 2: $a_j = \mu_{\text{subject}} + \epsilon_j$

- Can deal with nested variables (stimuli are repeated)
- No need to average data – retain information
- Random slopes could account for account for fatigue/becoming accustomed to accent
- Represents variance as coefficients
- Has equations to handle both categorical and numerical response variables (allows both linear and logistic analyses of data)

What are random and fixed effects?

› Random effects:
  - Levels randomly sampled from a larger population
  - Varies over time/between samples
  - May expect a different slope/intercept for each instance
  - Useful for analysing items tested (avoiding language as a fixed effect fallacy, here difficulty of question), or capturing subject variation

› Fixed effects:
  - Fixed number of levels
  - Expect the variable to contribute equally regardless of context
  - Generally do not vary over time

Random and Fixed Effects

Random Effects	Fixed Effects
Speaker	English exposure
Listener	Errors in pronunciation, grammar, vocabulary
Stimulus (movie vs question)	Speaker Age

```
> summary(lm.mod1)

Call:
lm(formula = tot ~ lst.enexposure + lst.enlevel + spk.age + spk.enyears +
    spk.enfreq + spk.enpron + spk.engrammar + spk.envocab + movie.id,
    data = d5) #d5 is numerical response variable (qns correct)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3488 -1.1858  0.1416  1.2433  3.4159

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.008783   3.573124   0.002  0.99804
lst.enexposure  0.456863   0.189410   2.412  0.01789 *
lst.enlevel     0.293529   0.175791   1.670  0.09844 .
spk.age         0.009743   0.119449   0.082  0.93518
spk.enyears    -0.104554   0.092604  -1.129  0.26188
spk.enfreq      0.641164   0.248287   2.582  0.01143 *
spk.enpron      0.504496   0.222272   2.270  0.02561 *
spk.engrammar  -0.366587   0.527467  -0.695  0.48885
spk.envocab    -0.156958   0.204041  -0.769  0.44376
movie.idm2     -1.279662   0.456806  -2.801  0.00623 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.82 on 90 degrees of freedom
Multiple R-squared:  0.2856,    Adjusted R-squared:  0.2142
F-statistic: 3.998 on 9 and 90 DF,  p-value: 0.0002525

>
```

```
> summary(lmer.mod1)
Linear mixed model fit by REML ['merModLmerTest']
Formula: tot ~ (1 | spk.id) + (1 | movie.id) + lst.enexposure + lst.enlevel +       spk.age +
spk.enyears + spk.enfreq + spk.enpron + spk.engrammar +       spk.envocab
   Data: d5

REML criterion at convergence: 375.7

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.99752 -0.63924  0.03119  0.53328  2.13735

Random effects:
 Groups   Name        Variance Std.Dev.
 spk.id   (Intercept) 2.5916   1.6098
 movie.id (Intercept) 0.2684   0.5181
 Residual             1.6514   1.2850
Number of obs: 100, groups: spk.id, 20; movie.id, 2

Fixed effects:
               Estimate Std. Error       df t value Pr(>|t|)
(Intercept)     0.46277    7.35172 12.85000   0.063 0.950778
lst.enexposure  0.32004    0.13944 80.90000   2.295 0.024311 *
lst.enlevel     0.48521    0.13343 81.03000   3.636 0.000485 ***
spk.age        -0.01497    0.24949 12.12000  -0.060 0.953127
spk.enyears    -0.19302    0.17544 11.85000  -1.100 0.293081
spk.enfreq      0.63005    0.51935 11.95000   1.213 0.248506
spk.enpron      0.55820    0.46437 12.01000   1.202 0.252498
spk.engrammar  -0.42779    1.10097 12.00000  -0.389 0.704408
spk.envocab    -0.19506    0.42619 12.08000  -0.458 0.655304
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) lst.nx lst.nl spk.ag spk.ny spk.nf spk.np spk.ng
lst.enexpsr -0.012
```

```
> lmer.mod1
Linear mixed model fit by REML ['merModLmerTest']
Formula: tot ~ (1 | spk.id) + (1 | movie.id) + lst.enexposure + lst.enlevel +      spk.age +
spk.enyears + spk.enfreq + spk.enpron + spk.engrammar +      spk.envocab
   Data: d5
REML criterion at convergence: 375.6679
Random effects:
 Groups   Name        Std.Dev.
 spk.id   (Intercept) 1.6098
 movie.id (Intercept) 0.5181
 Residual             1.2850
Number of obs: 100, groups: spk.id, 20; movie.id, 2
Fixed Effects:
   (Intercept)  lst.enexposure     lst.enlevel         spk.age     spk.enyears      spk.enfreq
spk.enpron   spk.engrammar     spk.envocab
       0.46277         0.32004         0.48521        -0.01497        -0.19302         0.63005
0.55820        -0.42779        -0.19506
```

```
> lmer.mod3
Linear mixed model fit by REML ['merModLmerTest']
Formula: tot - mean(d5$tot) ~ (1 | spk.id) + (1 | movie.id) + lst.enexposure
+      lst.enlevel + spk.age + spk.enyears + spk.enfreq + spk.enpron +
    spk.engrammar + spk.envocab
   Data: d5
REML criterion at convergence: 375.6679
Random effects:
 Groups   Name        Std.Dev.
 spk.id   (Intercept) 1.6098
 movie.id (Intercept) 0.5181
 Residual             1.2850
Number of obs: 100, groups: spk.id, 20; movie.id, 2
Fixed Effects:
   (Intercept)  lst.enexposure     lst.enlevel         spk.age
spk.enyears       spk.enfreq      spk.enpron   spk.engrammar     spk.envocab
      -2.72724         0.32004         0.48521        -0.01497         -
0.19302         0.63005         0.55820        -0.42779        -0.19506
>
```

```
> anova(lmer.mod1, lmer.mod2)
refitting model(s) with ML (instead of REML)
Data: d5
Models:
object: tot ~ (1 | spk.id) + (1 | movie.id) + lst.enexposure + 
lst.enlevel + 
object:     spk.age + spk.enyears + spk.enfreq + spk.enpron + 
spk.engrammar + 
object:     spk.envocab
..1: tot ~ (1 | spk.id) + (1 | movie.id) + lst.enexposure * 
spk.envocab + 
..1:     lst.enlevel + spk.age + spk.enyears + spk.enfreq + 
spk.enpron * 
..1:     lst.crexposure + spk.engrammar
       Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
object 12 391.87 423.13 -183.94   367.87                        
..1    13 393.80 427.67 -183.90   367.80 0.0726      1     0.7876


> anova(lmer.mod1, lmer.mod3) #equal AIC
```

Explaining the variance

› Random effects – sd

› Fixed effects – visible from model

› Still to consider: centring data

Next steps

- Data collection: release and promote survey to participants in Portugal, Slovenia and Denmark.
- Possible future study: Include L1 English speakers as listeners and see if variance is different
- Model for predicting comprehension? Areas to target in language teaching?

Questions?

- Missing data – ecological validity or to nullify subject?
- Averaging ‘listeners’ to still use linear regression?

References

- Field, Andy. *Discovering statistics using SPSS*. Sage publications, 2009.
- Baayen, R. Harald. "Analyzing linguistic data." *A practical introduction to statistics using R* (2008).
- Bates, Douglas M. "lme4: Mixed-effects modeling with R." *URL http://lme4. r-forge. r-project. org/book* (2010).