Quantitative Methods in Sociolinguistics

Sociolinguistics Summer School 5
University College Dublin
30 July 2014

Daniel Ezra Johnson
Lancaster University
d.e.johnson@lancaster.ac.uk
www.danielezrajohnson.com/sss5.pdf

“the old testament”

There were (quantitative) (sociolinguistic) studies of language variation before Labov:

Terracher 1914.

Duraffour 1927.

Martinet 1945.

Putnam & O’Hern 1955.

Malkiel, Yakov. 1984. Revisionist dialectology and mainstream linguistics. *Language in Society* 13: 29-66.

Joseph, John E. 2002.

Koerner, E. F. K. 2003.

"John The Baptist"

Table I. Number of children favoring *-ing* and *-in* variant suffixes in TAT protocols according to sex.

	-ing>*-in*	*-ing*<*-in*
Boys	5	7
Girls	10	2

Chi square: 2.84 ; 05 <P<.1 (by two-tailed test)

Fischer, John L. 1958. Social influences on the choice of a linguistic variant. *Word* 14(1): 47-56.

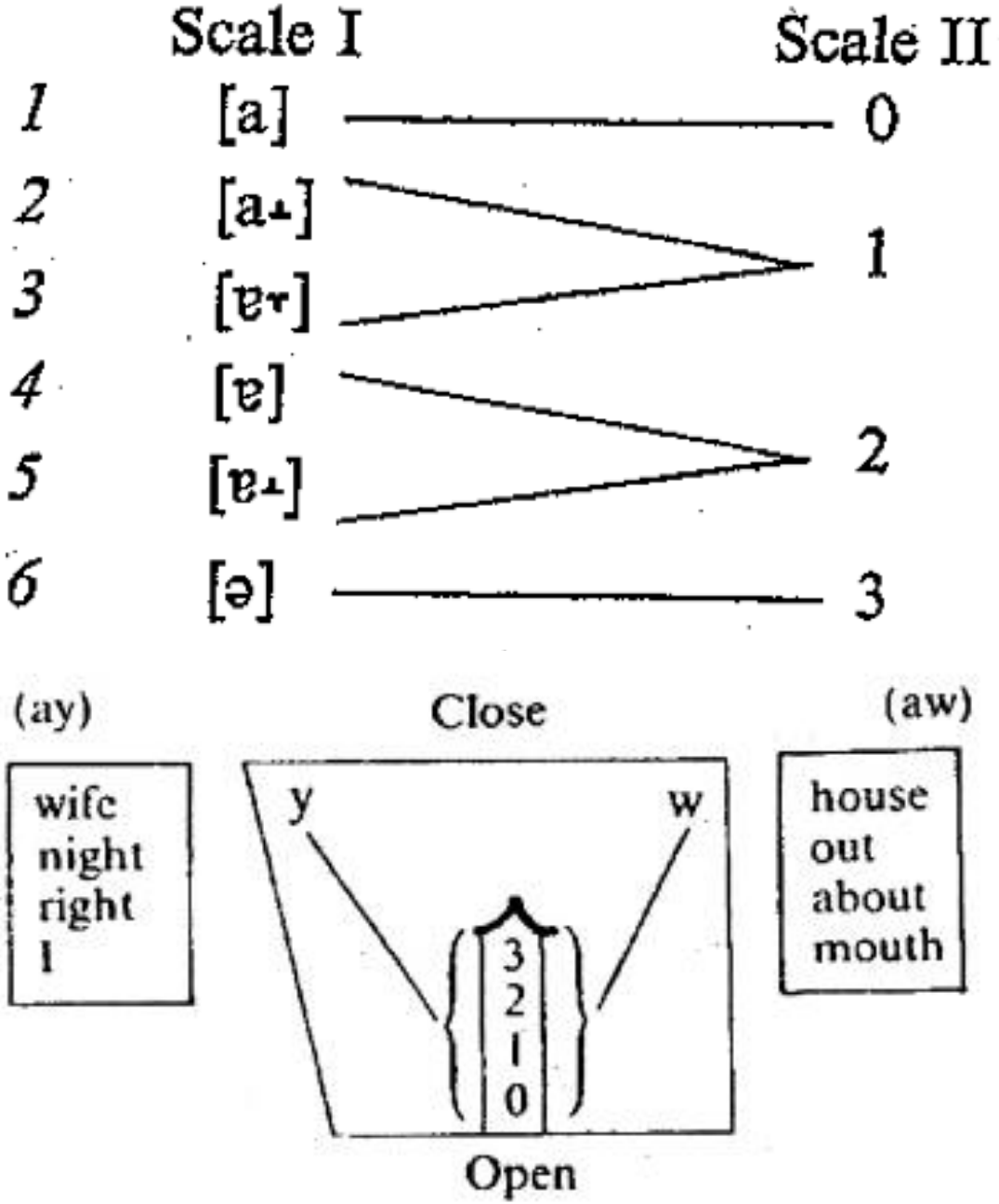

Scale I
Scale II
1
[a]
0
2
[a˧]
3
[ɐ˞]
1
4
[ɐ]
5
[ɐ˧]
2
6
[ə]
3
(ay)
Close
(aw)
wife
night
right
I
y
w
3
2
1
0
house
out
about
mouth
Open

TABLE 3. GEOGRAPHICAL DISTRIBUTION OF CENTRALIZATION

	CI /ai/	CI /au	CI /ai/	CI /au/
Down-island			0.35	0.33
Edgartown	0.48	0.55		
Oak Bluffs	0.33	0.10		
Vineyard Haven	0.24	0.33		
Up-island			0.61	0.66
Oak Bluffs	0.71	0.99		
No. Tisbury	0.35	0.13		
West Tisbury	0.51	0.51		
Chilmark	1.00	0.81		
Gay Head	0.51	0.81		

TABLE 4. CENTRALIZATION BY OCCUPATIONAL GROUPS

	CI /ai/	CI /au/
fishermen	1.00	0.79
farmers	0.32	0.22
others	0.41	0.57

TABLE 5. CENTRALIZATION BY ETHNIC GROUPS

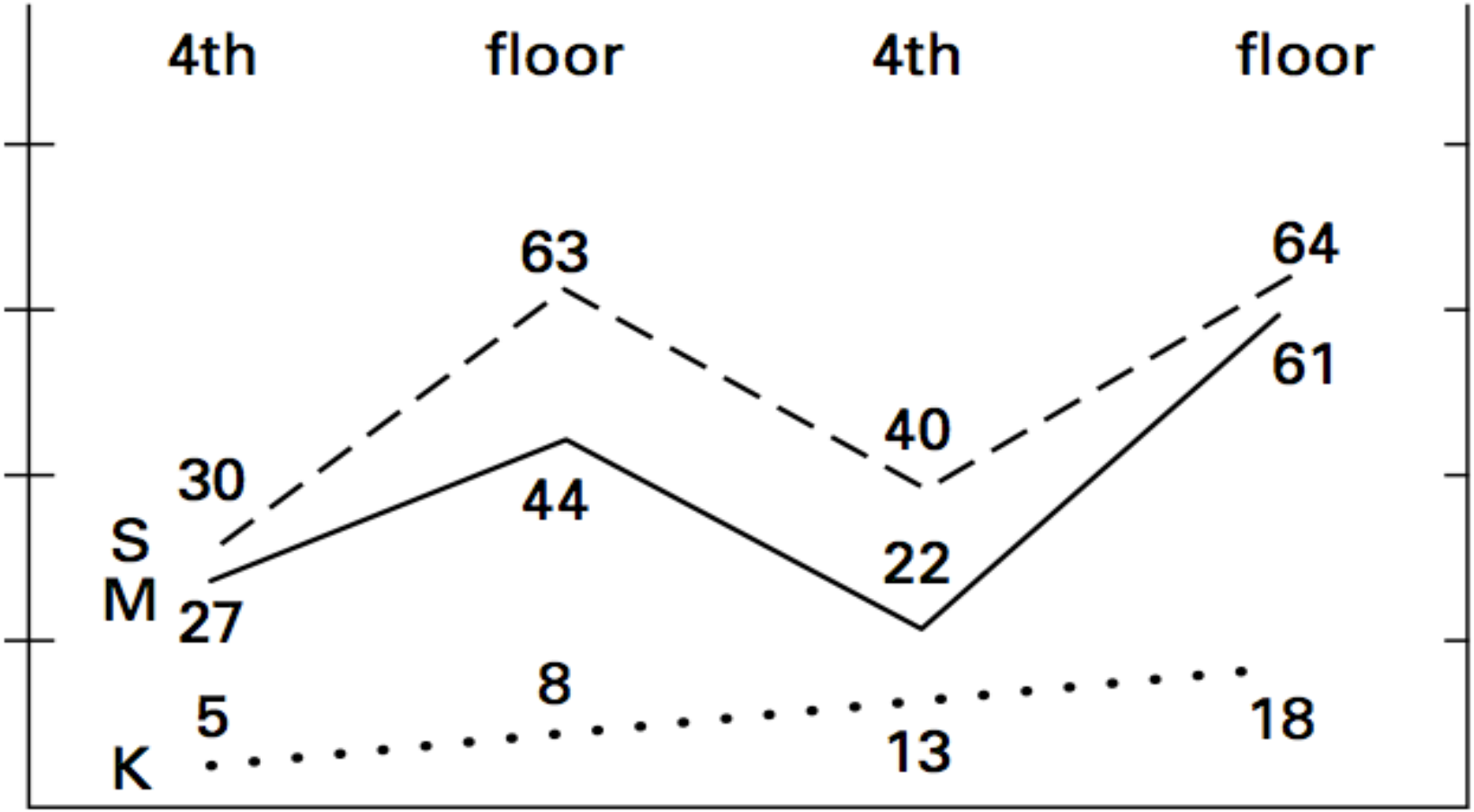
	English		Portuguese		Indian	
Age Level	CI /ai/	CI /au/	CI /ai/	CI /au/	CI /ai/	CI /au/
over 60	0.36–0.34		0.26–0.26		0.32–0.40	
46 to 60	0.85–0.63		0.37–0.59		0.71–1.00	
31 to 45	1.08–1.09		0.73–0.83		0.80–1.33	
under 30	0.35–0.31		0.34–0.52		0.47–0.88	
all ages	0.67–0.60		0.42–0.54		0.56–0.90	

TABLE 6. CENTRALIZATION AND ORIENTATION TOWARDS MARTHA'S VINEYARD

Persons		CI /ai/	CI /au/
40	Positive	0.63	0.62
19	Neutral	0.32	0.42
6	Negative	0.09	0.08

The fact that this table shows us the sharpest example of stratification we have yet seen, indicates that we have come reasonably close to a valid explanation of the social distribution of centralized diphthongs.

Labov, William. 1963. The social motivation of a sound change. *Word* 19: 273-309.



(S=Saks, M=Macy's, K=S. Klein)

Figure 3.2 Percentage of *all (r-1)* by store for four positions

Labov, William. 1966. *The social stratification of English in New York City*. Ph.D. dissertation, Columbia.

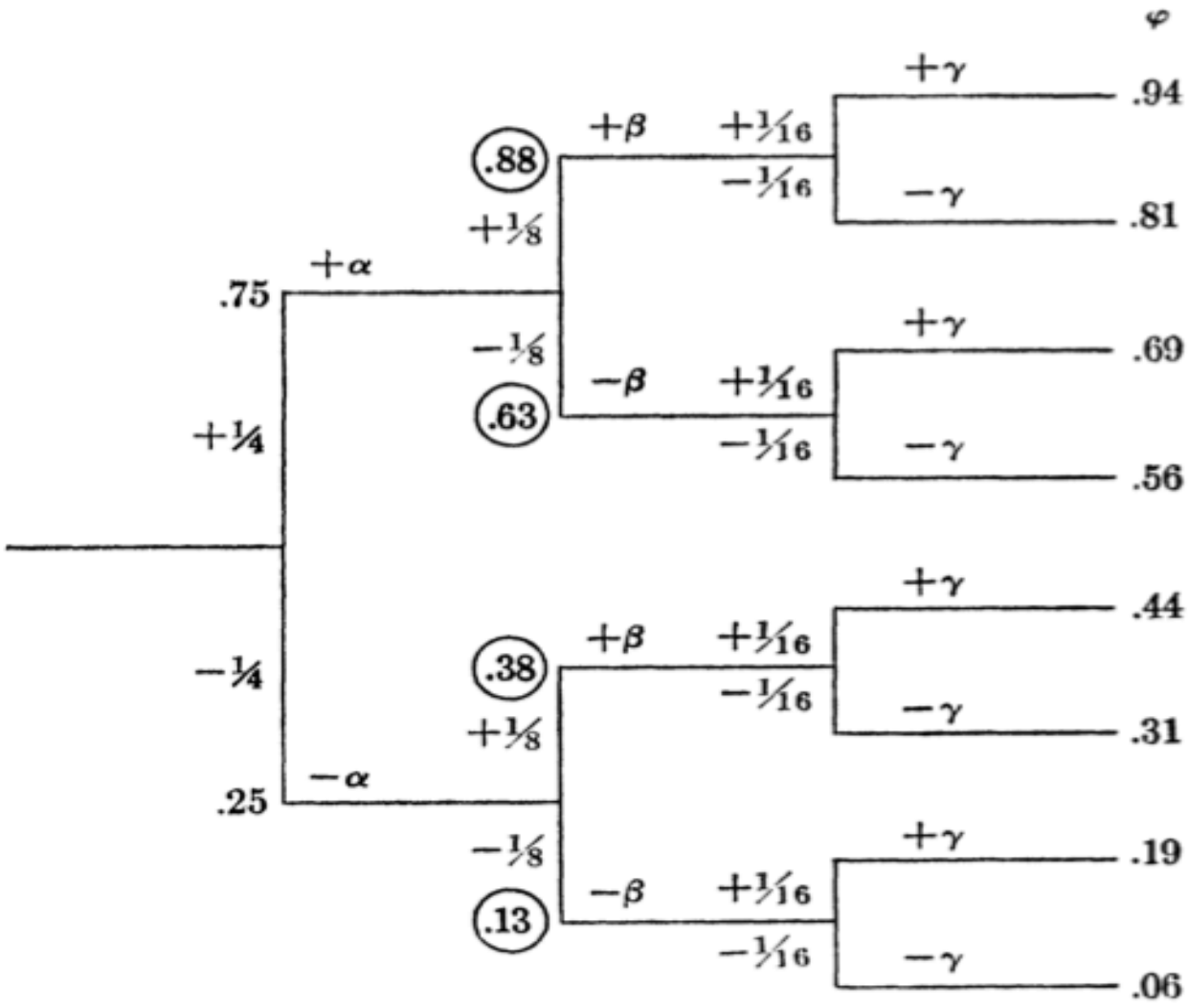


FIGURE 10
Geometric ordering of variable constraints α, β, γ.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language*.

Input probability $p_0 = 1$.

Preceding environment:	[+sib]_	$\begin{bmatrix}+\text{cns}\\-\text{sib}\end{bmatrix}$_	[−cns]_
Effect:	1	0.85	0.37
Following environment:	_[+sib]	_$\begin{bmatrix}+\text{cns}\\-\text{sib}\end{bmatrix}$	_[−cns]
Effect:	1	0.50	0.10
Occupational class:	workers	professional	
Effect:	1	0.35	
Sex:	women	men	
Effect:	1	1	

TABLE 6. Effect of each factor affecting *que* deletion in MF, according to a multiplicative application probabilities model.

Cedergren, Henrietta & Sankoff, David. 1974. Variable rules: performance as a statistical reflection of competence. *Language*.

$$p = \beta_0 + \beta_a + \ldots + \beta_n$$

$$p = p_0 + p_i + p_j + \cdots + p_k$$

$$\log p = \beta_0 + \beta_a + \ldots + \beta_n$$

$$p = p_0 \times p_i \times p_j \times \cdots \times p_k$$

$$\log \frac{p}{1-p} = \beta_0 + \beta_a + \ldots + \beta_n$$

$$\left(\frac{p}{1-p}\right) = \left(\frac{p_0}{1-p_0}\right) \times \left(\frac{p_i}{1-p_i}\right) \times \left(\frac{p_j}{1-p_j}\right) \times \cdots \times \left(\frac{p_k}{1-p_k}\right)$$

$$\left(\frac{p}{1-p}\right)=\left(\frac{p_0}{1-p_0}\right)\times\left(\frac{p_i}{1-p_i}\right)\times\left(\frac{p_j}{1-p_j}\right)\times\cdots\times\left(\frac{p_k}{1-p_k}\right)$$

TABLE 3.2

Contribution of Grammatical Category, Following Phonological Segment, and Following Stress to the Deletion of Plural (s)[a]

Grammatical category		Following phonological segment		Following stress	
Adjective	.69	Pause	.65	Weak	.56
Noun	.57	Consonant	.47	Heavy	.44
Determiner	.26	Vowel	.37		

[a] Input probability = .65.

Poplack, Shana. 1980. The notion of the plural in Puerto Rican Spanish: competing constraints on (s) deletion. In Labov (ed.)

$$\left(\frac{p}{1-p}\right)=\left(\frac{p_0}{1-p_0}\right)\times\left(\frac{p_i}{1-p_i}\right)\times\left(\frac{p_j}{1-p_j}\right)\times\cdots\times\left(\frac{p_k}{1-p_k}\right)$$

TABLE 6. *Social factors contributing to the Canadian Vowel Shift in younger Toronto English speakers, by ethnic group and EO status (excluding tokens preceding a nasal consonant)*

			(ɛ)		(æ)	
		Total N:	2,270		1,404	
		Input:	.205		.201	
Ethnicity and EO Status						
British/Irish			**.68**		.51	
Italian	Low EO		.54		**.60**	
	High EO		**.63**		**.82**	
Chinese	Low EO		.32		.29	
	High EO		.30		.17	
		Range:		*38*		*65*
Speaker Sex						
Women			**.56**		**.60**	
Men			.44		.42	
		Range:		*8*		*18*

Hoffman, Michol F. & Walker, James A. 2010. Ethnolects and the city: ethnic orientation and linguistic variation in Toronto English. *Language Variation and Change* 22: 37-67.

$$\left(\frac{p}{1-p}\right)=\left(\frac{p_0}{1-p_0}\right)\times\left(\frac{p_i}{1-p_i}\right)\times\left(\frac{p_j}{1-p_j}\right)\times\cdots\times\left(\frac{p_k}{1-p_k}\right)$$

TABLE 6. *Social factors contributing to the Canadian Vowel Shift in younger Toronto English speakers, by ethnic group and EO status (excluding tokens preceding a nasal consonant)*

			(ɛ)		(æ)	
		Total N:	2,270		1,404	
		Input:	.205		.201	
Ethnicity and EO Status						
British/Irish			**.68**		.51	
Italian	Low EO		.54		**.60**	
	High EO		**.63**		**.82**	
Chinese	Low EO		.32		.29	
	High EO		.30		.17	
		Range:		*38*		*65*
Speaker Sex						
Women			**.56**		**.60**	
Men			.44		.42	
		Range:		*8*		*18*

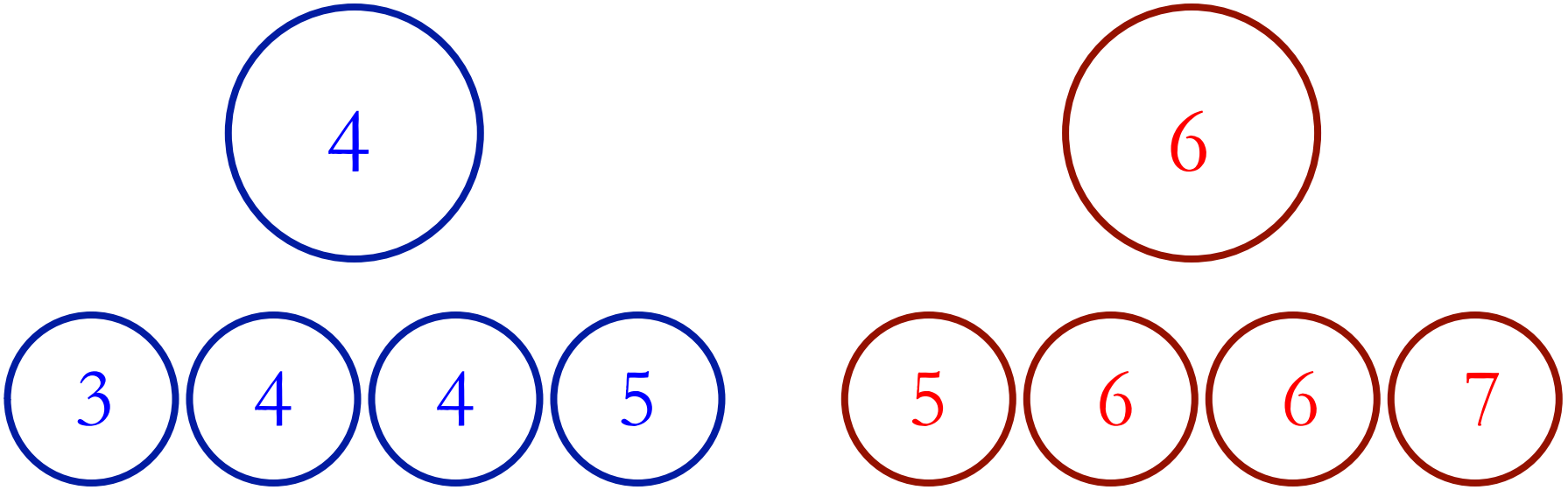
Each token was coded impressionistically as shifted or nonshifted, excluding any tokens where agreement between the authors could not be reached.

EO (ethnic orientation): For example, informants who identified themselves as “Canadian” received a score of 1; those who responded “Italian” received a score of 3; a response of “Italian-Canadian” or “both” received a score of 2.

response and predictors: categorical or continuous

	categorical (binary) response logistic regression	continuous response linear regression
categorical predictors	GoldVarb factor weights or log-odds coefficients	no GoldVarb no factor weights coefficients in same units as response
continuous predictors	no GoldVarb log-odds coefficients	no GoldVarb no factor weights coefficients in same units as response

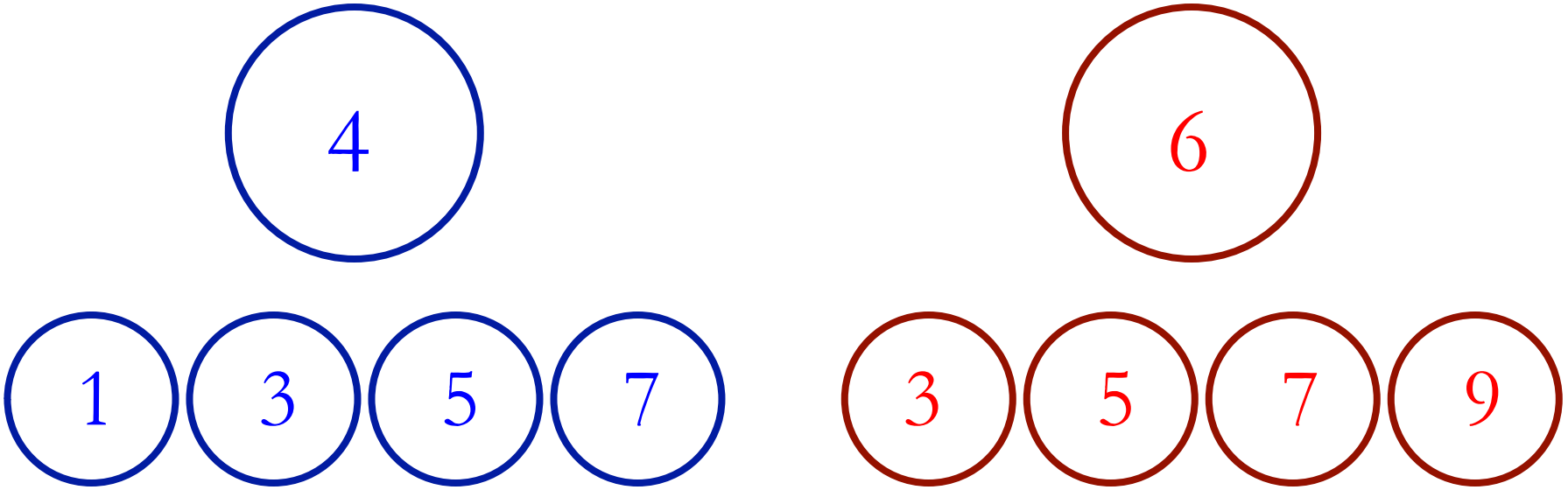
within-group differences affect the significance of between-group differences



small within-group variation

significant between-group difference

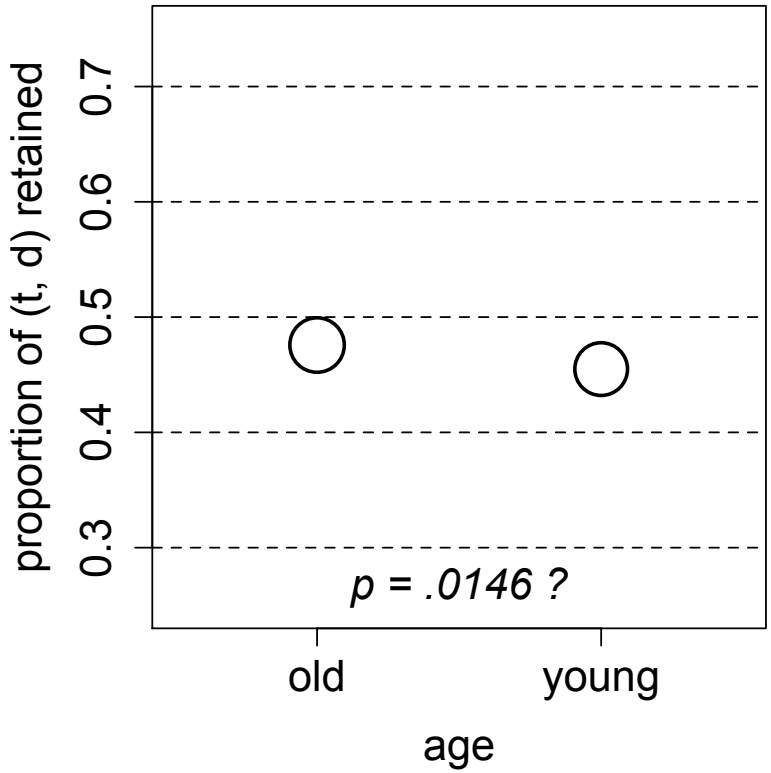
within-group differences affect the significance of between-group differences



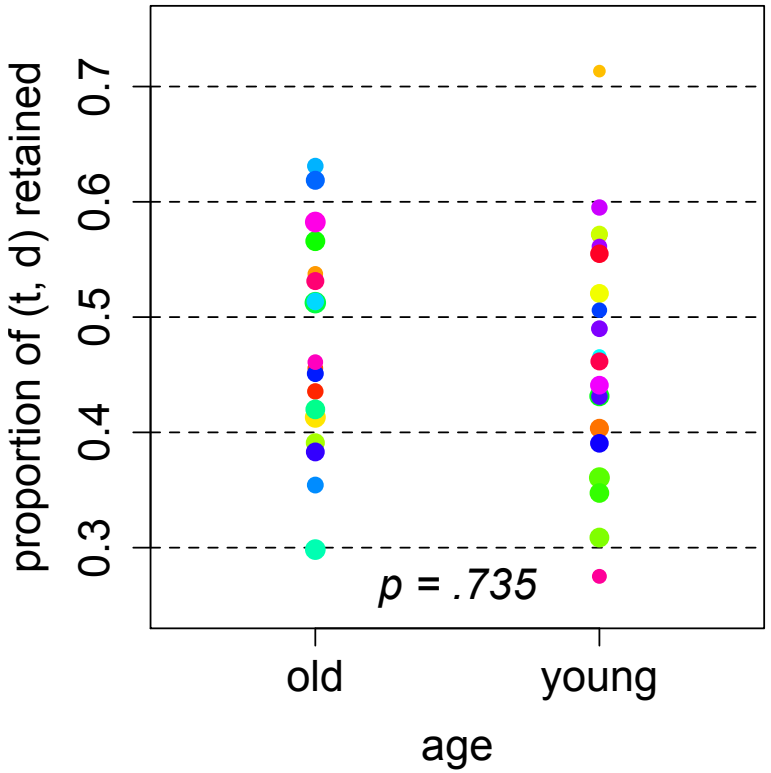
large within-group variation

non-significant between-group difference

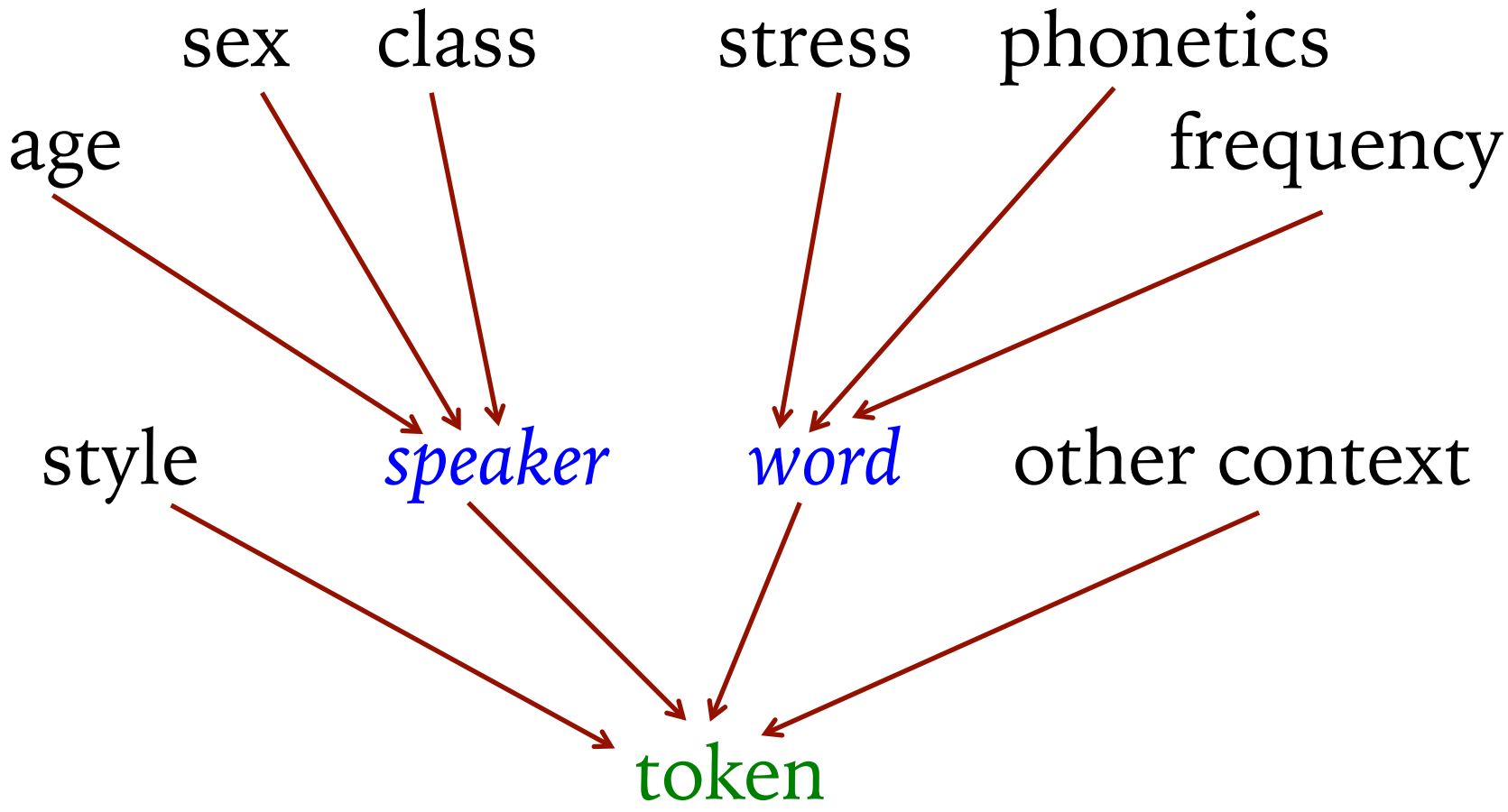
```
> age.fixed <- glm(tq1 ~ Age, buck, family = binomial)
> summary(age.fixed)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.09701    0.02384  -4.070  4.7e-05 ***
Ageyoung    -0.08386    0.03433  -2.443   0.0146 *

ONE-LEVEL ANALYSIS OF RESPONSE tq1 WITH PREDICTOR(S):
 Age (0.0146)

$Age
 factor logodds tokens 1/1+0 centered factor weight
    old   0.042   7056 0.476                   0.51
  young  -0.042   6608 0.455                   0.49
```



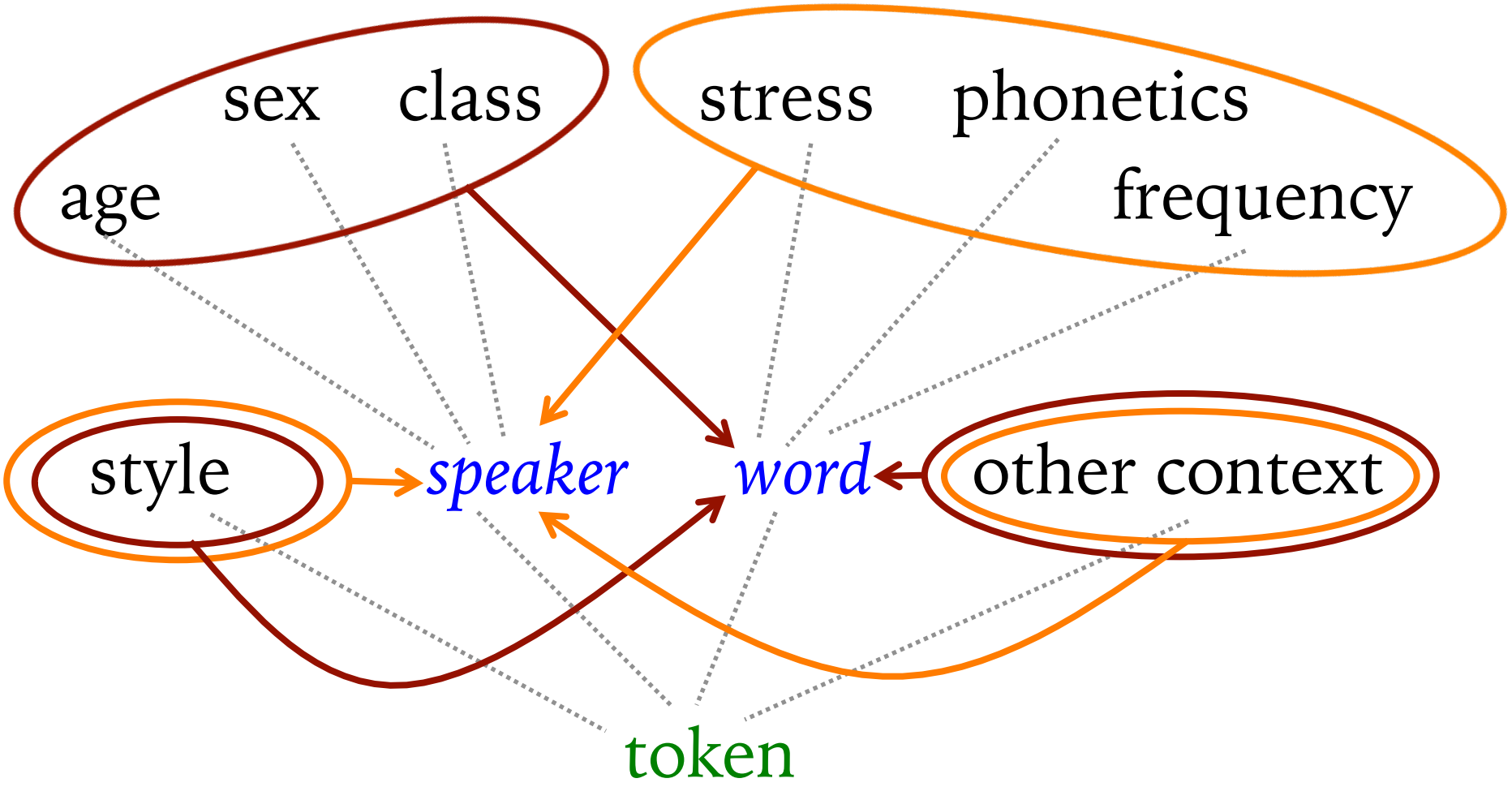

```
> age.mixed <- glmer(tq1 ~ Age + (1 | Speaker), buck,
   family = binomial)
> summary(age.mixed)

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.08842    0.08805  -1.004    0.315
Ageyoung    -0.04229    0.12471  -0.339    0.735

ONE-LEVEL ANALYSIS OF RESPONSE tq1 WITH PREDICTOR(S):
 Speaker [random, not tested] and Age (0.735)

$Age
 factor logodds tokens 1/1+0 centered factor weight
    old   0.021   7056 0.476                   0.505
  young  -0.021   6608 0.455                   0.495
```



nesting: the relationship of variables in the mixed-effects model



random intercepts and random slopes



random intercept *random slope (by speaker)* *random slope (by word)*

individual-speaker variation: are mixed models necessary?

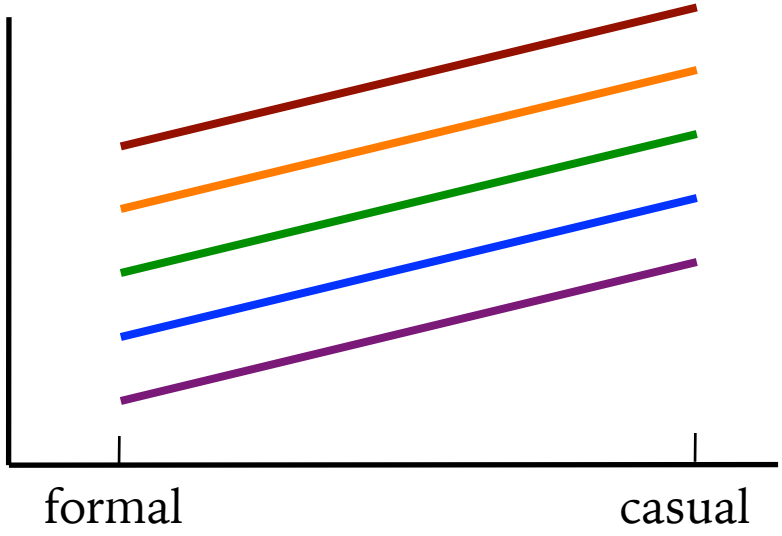
if only one speaker: *no!*

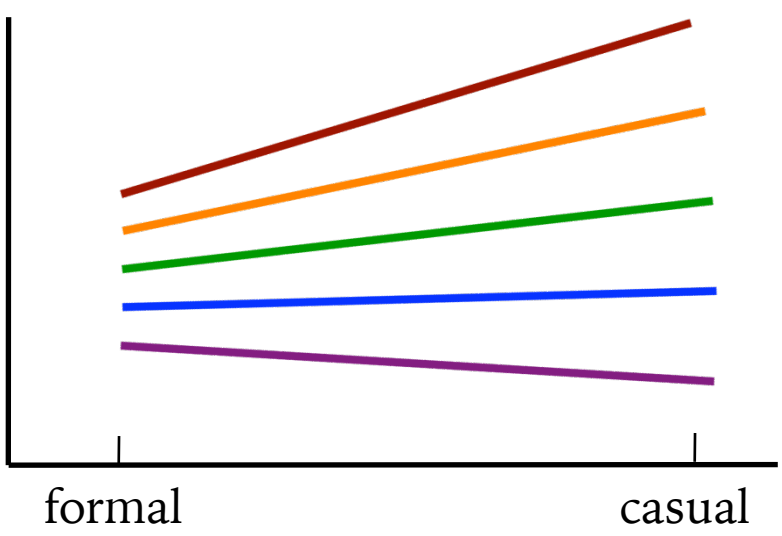
if no repeated measures per speaker: *no! (dep't stores)*

if every by-speaker difference is accounted for: *no!*

if they differ in overall rate/use of variable? *intercept!*

if they differ in the effect of other predictors? *slope(s)!*





individual-word variation: are mixed models necessary?

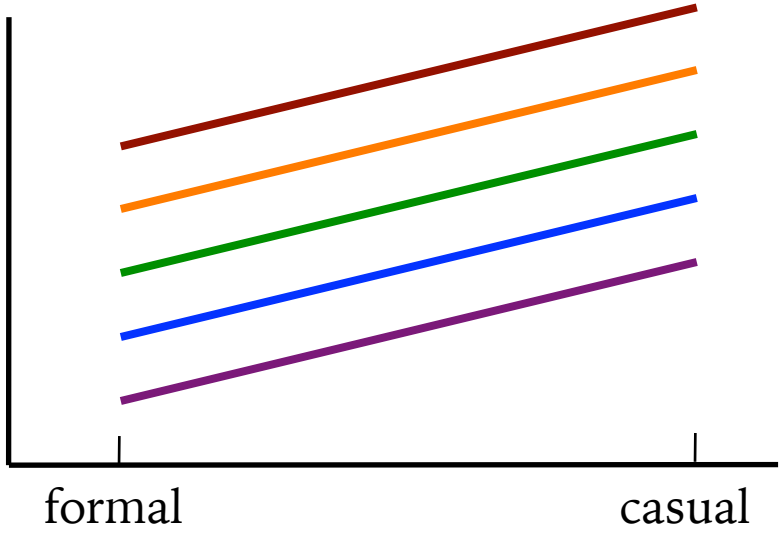
if only one word: *no!*

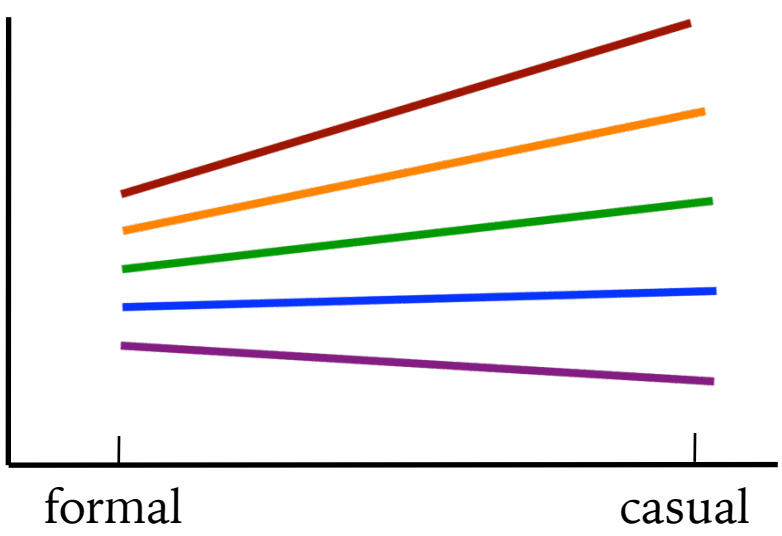
if no repeated measures per word: *no! (dep't stores)*

if every by-word difference is accounted for: *no!*

if words differ in overall rate/use of variable? *intercept!*

if words differ in effect of other predictor(s)? *slope(s)!*





individual-speaker and -word variation: the received wisdom

do speakers differ in overall rate/use of variable?

– acknowledged to be true

– ignored in statistical practice

do speakers differ in the effect of other predictor(s)?

– claimed to be false (within a speech community)

– possibly true, ignored in statistical practice

do words differ overall *or* in the effect of predictor(s)?

– may depend on phon. theory, ignored in practice

is regression enough?

how am I going to do this?

what is R?

what is Rbrul?

what about RStudio?

are there any good books?

who else can I ask for help with statistics?

starting with R and Rbrul

1) To download and install R, go to: http://cran.r-project.org
2) To install "packages" that will be needed, start R and execute these commands by typing at the > prompt:

> install.packages("ggplot2") (this is one graphics package)

> install.packages("lme4") (this is for mixed models)

> install.packages("lmerTest") (this helps lme4 provide p-values)

If these install cleanly, you will not need to install them again.

3) To load packages (needs to be done each time you start R):

> load(ggplot2) > load(lme4) > load(lmerTest)

4) To install Rbrul (needs to be done each time you start R):

> source("http://www.danielezrajohnson.com/Rbrul.R")

5) To start Rbrul:

> rbrul()