The use of statistical tools/models in cognitive/usage-based linguistics

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
http://tinyurl.com/stgries

Why the field has to become more statistical (in two directions) …

- More and more corpus-linguistic studies are based on
  - increasingly larger (samples from) corpora
  - increasingly complex (samples from) corpora
    - complex in terms of both composition and annotation
  - temporally- or otherwise ordered corpora
- these developments often lead to large multi-dimensional data sets whose size and complexity defies
  - mere eye-balling of the data
  - introspective analysis of the data
- therefore, statistical tools are becoming more important and more frequently used
  - sometimes, statistical applications are used in an exploratory / hypothesis-generating way
    - (which was the topic of the previous talk)
  - sometimes, statistical applications are used in a hypothesis-testing way
    - (which is the topic of the present talk)

On quantitative vs. qualitative

- Sometimes, my view of the importance of quantitative methods is opposed (?!)
- there are those who argue that
  - many things in (cognitive) linguistics are not amenable to quantitative study, but to qualitative analysis …
  - since quantitative analysis needs qualitative analysis/interpretation anyway, why bother with the numbers?
- these views are wrong because
  - qualitative analysis needs quantitative analysis just as much as the other way round – if not more (see below)
  - this is because
    - qualitative analysis implies (if only implicitly) labeling (i.e., annotating) data points and interpreting them
    - this annotation of data points leads to frequencies of (co-)occurrence of annotations: $n=0$, $n>0$, $n\gg0$, $n>m$, $n<m$, …
    - it is only quantitative analysis of these frequencies that makes the overall analysis intersubjective, replicable, falsifiable, and predictive

When (only) a quantitative method shows what one is really saying …

- Let's assume one is interested in how media coverage of the word *Muslim* changes over time (any resemblance to real studies, published or on-going, is not coincidental)
- let's assume a discourse-analytic approach finds that *Muslim* is used with a growing number of negative overtones
- let's assume this is backed up by a correlation measure: r=0.97, p<0.001
- however, one needs the interaction WORD:TIME …

When (only) a quantitative method shows what one is really saying …

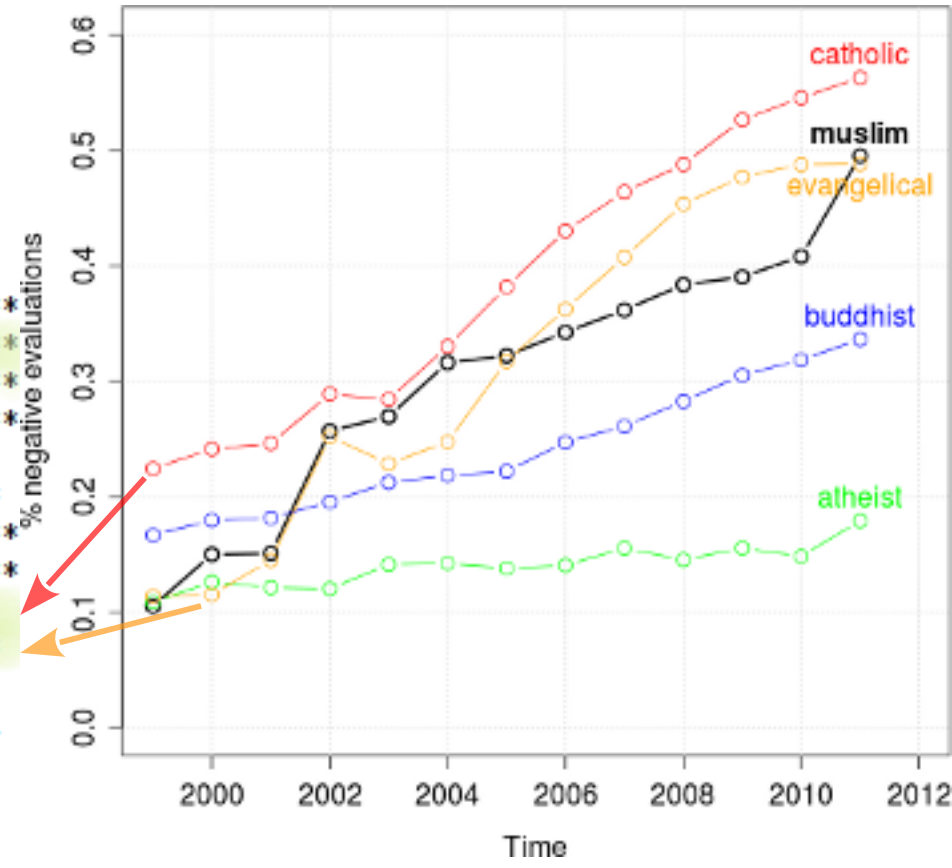
```
> summary(model.01)

Call:
lm(formula = NEGEVAL ~ TIME * WORD)

Residuals:
      Min        1Q    Median        3Q       Max
-0.041519 -0.008056 -0.000604  0.011717  0.044168

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -56.692127   3.043619 -18.627  < 2e-16 ***
TIME                0.028427   0.001518  18.726  < 2e-16 ***
WORDatheist        48.328480   4.304327  11.228 7.49e-16 ***
WORDbuddhist       28.394695   4.304327   6.597 1.72e-08 ***
WORDcatholic       -5.994329   4.304327  -1.393  0.16934
WORDevangelical   -14.686952   4.304327  -3.412  0.00122 **
TIME:WORDatheist   -0.024186   0.002147 -11.266 6.58e-16 ***
TIME:WORDbuddhist  -0.014194   0.002147  -6.612 1.63e-08 ***
TIME:WORDcatholic   0.003030   0.002147   1.412  0.16370
TIME:WORDevangelical 0.007331  0.002147   3.415  0.00121 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02048 on 55 degrees of freedom
Multiple R-squared: 0.9783, Adjusted R-squared: 0.9748
F-statistic:   276 on 9 and 55 DF,  p-value: < 2.2e-16
```



Pitfalls of introspective judgments: a small not quite fair case study

- Topic: the genitive alternation
  - $NP_{Possessor}$ *'s* $NP_{Possessed}$ vs.
    $NP_{Possessed}$ *of* $NP_{Possessor}$
- 'subjects'
  - tenured professors of linguistics
  - native speakers of English
- 'design': the subjects were
  - told I am interested in predicting which construction speakers choose when
  - told I suspect that the following variables influence the choice of construction
    - animacy, length, and givenness of
    - possessor and possessed
  - asked to formulate
    - generalizations as to how strongly the above variables would affect the choice of construction (i.e., an effect size)
    - estimates of the frequencies of variable combinations with a high degree of preference for either *of* or *'s*

Adding observational and experimental data to the mix

- In addition, I collected
  - a sample of corpus data (BNC) which were coded for these variables

	Spoken	Written	Totals
of	75	75	150
s	75	75	150
Totals	150	150	300

  - acceptability judgments from linguistically naïve native speakers of English for which these variables were systematically manipulated, pseudorandomized, interspersed with fillers, etc. (givenness was manipulated with a preceding context sentence)

What the linguists said … and what the other data show

- Linguists' overall response: yes, these variables influence the choice of construction
- linguists' responses re the effect sizes
  - possessors are more important than possesseds
    - in particular with regard to animacy and length
  - possesseds are only important with regard to animacy
  - apart from the above, the answers were diverse
- linguists' responses re frequent combinations
  - estimates for *s*-genitives focused on possessors
    - possessors in *s*-genitives = short, given, animate
  - estimates for *of*-genitives focused on possesseds
    - possesseds in *of*-genitives = long, new, animate, abstract

What the linguists said … and what the other data show

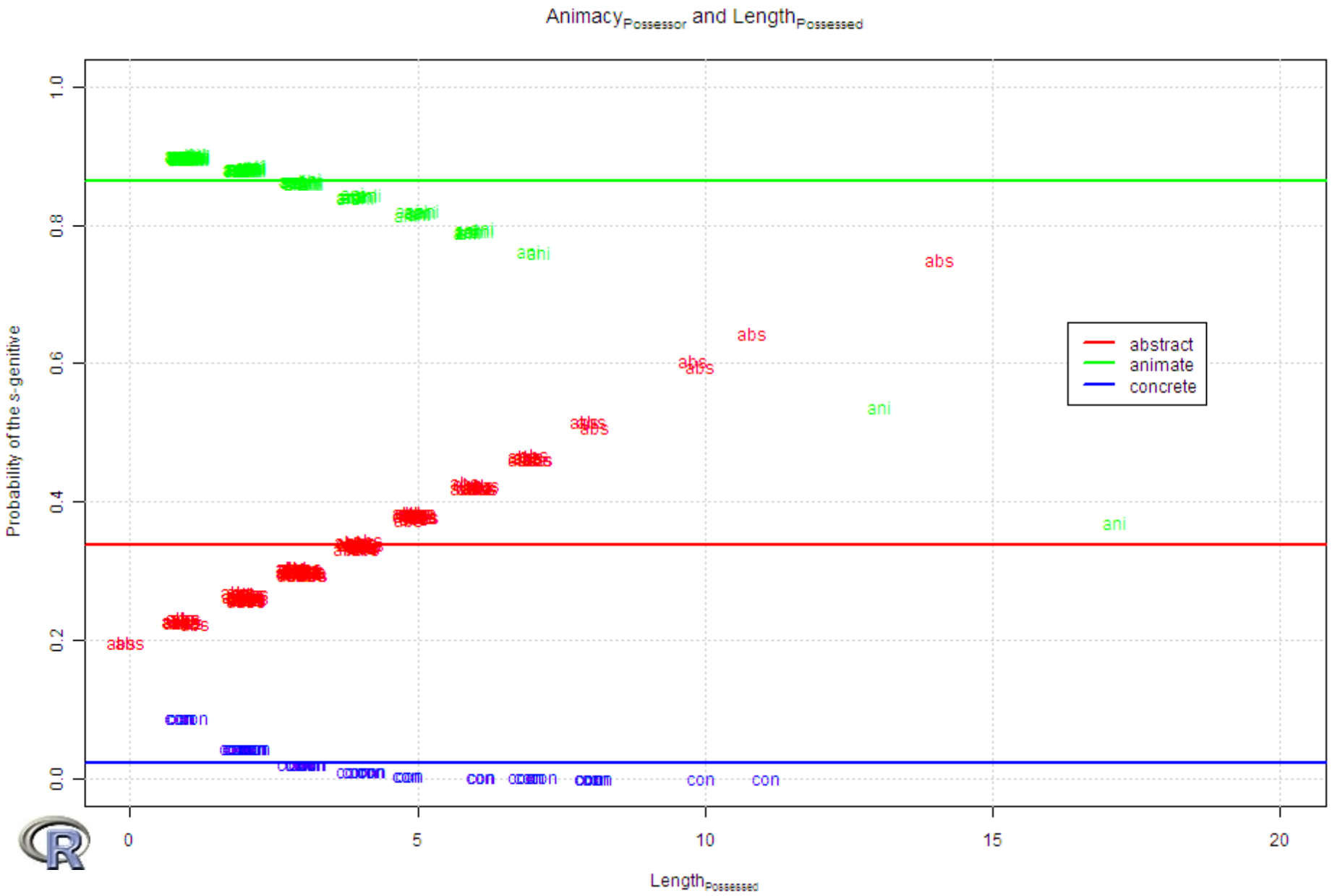
- Linguists' overall responses: yes, these variables influence the choice of construction – correct
  - model L.R. X^2=174.51, df=13; p≈0; C=0.892; R^2=0.588; classification accuracy=81.7%
- linguists' responses re the effect size
  - yes, possessors' properties are more important than possesseds' properties
  - yes, possesseds are important mostly with regard to their animacy
  - yes, animacy is ranked high in importance
  - no, length of the possessor is not important (neither in the corpus data nor in the experimental results)
- note something: these are all just main effects – not a single interaction was mentioned!

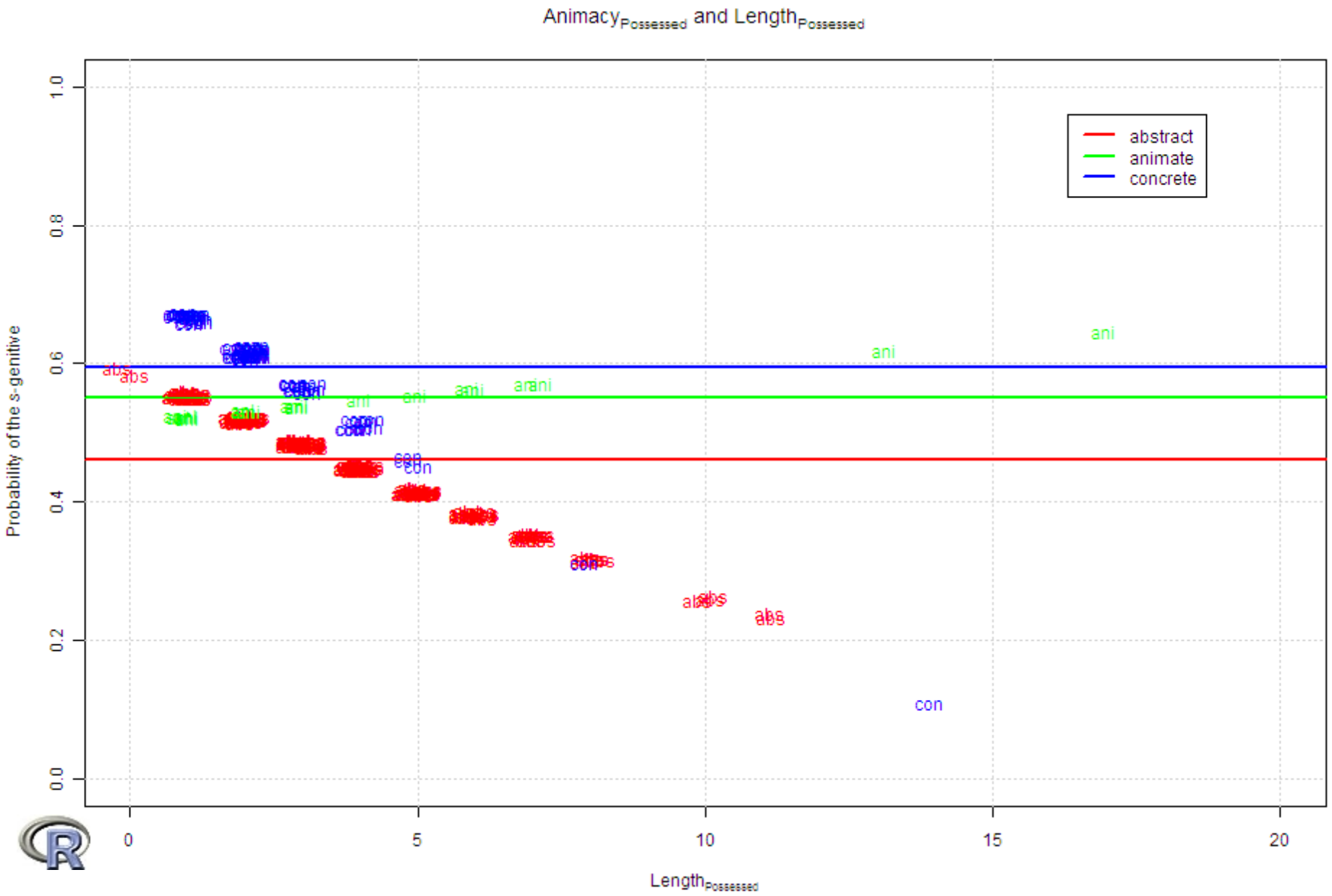
What the linguists said … and what the other data show

- Linguists' responses re frequencies of combinations
  - guesstimates regarding
    - the *s*-genitive focused on possessors (i.e., the first NP)
    - the *of*-genitive focused on possesseds (i.e., the first NP)
  - possessors in *s*-genitives = short, given, animate
    - is this about the two constructions or just a reflection of short>long and given>new?
  - possesseds in *of*-genitives = long, new, animate, abstract
    - violates short>long / given>new
- recall the two constructions' structures

	Element 1		Element 2
	Possessor	*'s*	Possessed
	Possessed	*of*	Possessor
Length	short	>>	long
Givenness	given	>>	new

- note something: these are all just main effects – not a single interactions was mentioned!


Animacy$_{Possessor}$ and Length$_{Possessed}$
Probability of the s-genitive
Length$_{Possessed}$
abstract
animate
concrete
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20


Animacy$_{Possessed}$ and Length$_{Possessed}$
abstract
animate
concrete
Probability of the s-genitive
Length$_{Possessed}$
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20

Conclusion from this slightly unfair experiment

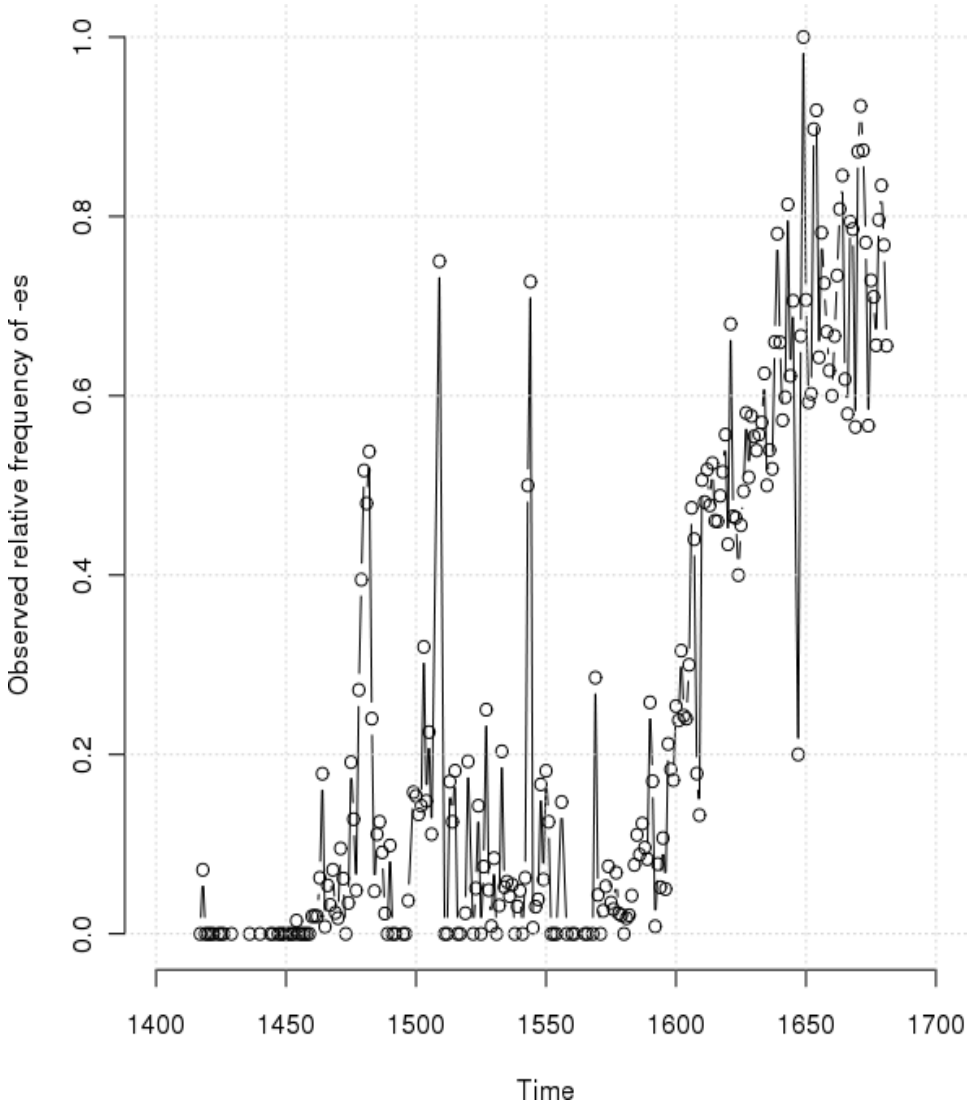
- The introspective judgments regarding the overall effects were ok (at the level of generality at which they were provided, that is)
- the introspective judgments regarding the effect sizes / combinations of variables were
  - partially correct and partially incorrect
  - all (!) monofactorial and, thus, grossly incomplete: most of the main effects assumed by the linguists participate in interactions
    - the role that $\text{Length}_{\text{Possessed}}$ plays is strongly dependent on $\text{Animacy}_{\text{Possessor}}$
    - the role that $\text{Length}_{\text{Possessed}}$ plays is strongly dependent on $\text{Animacy}_{\text{Possessed}}$ (plus other interactions not discussed here)
- these data do not replace a real (cognitive) linguistic account (e.g., Stefanowitsch 2003), but do show the pitfalls of impressionistic linguistics

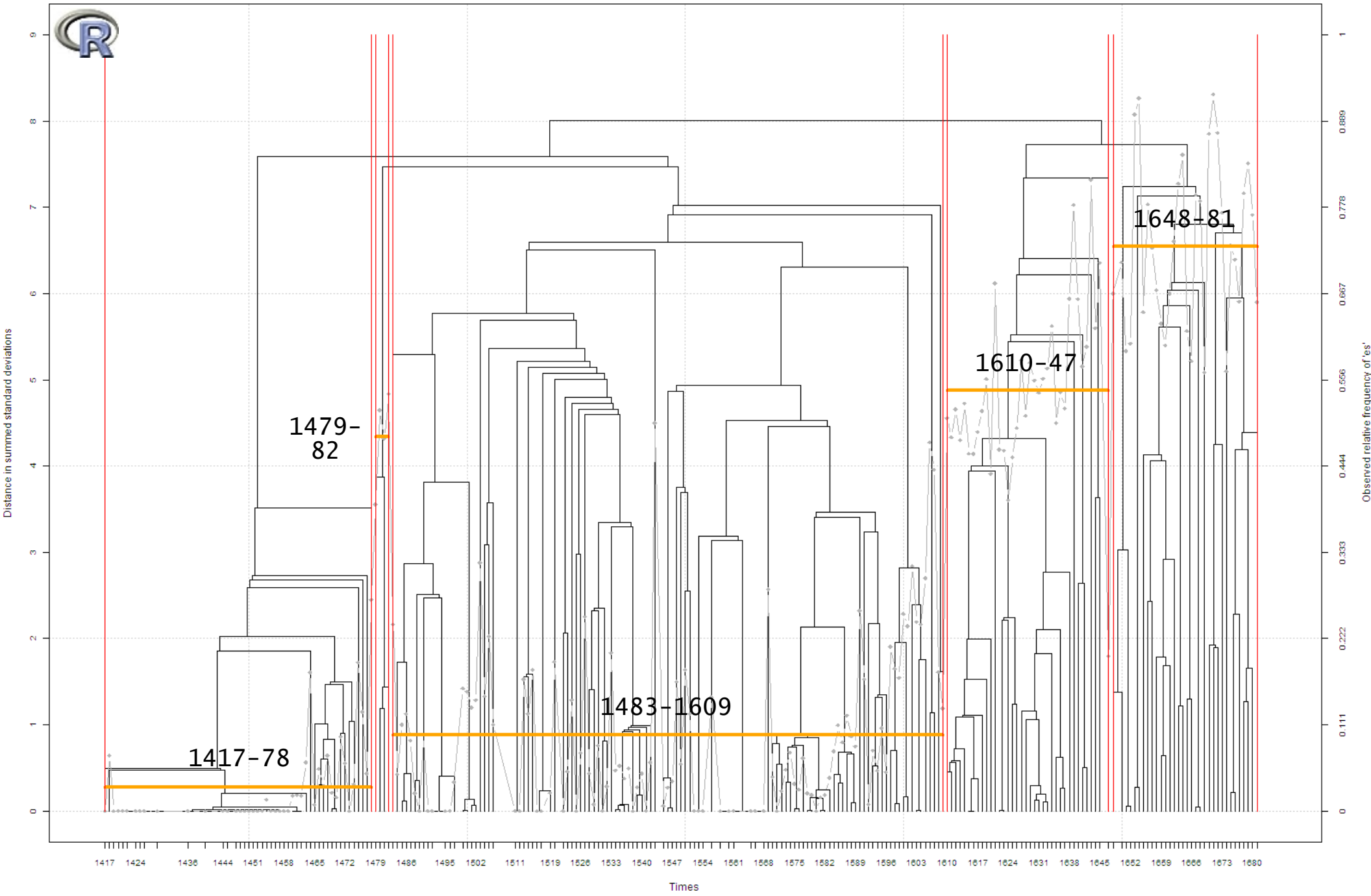
The development from *-(e)th* and *-(e)s* in the CEEC

- Let us now explore the development of the 3SG-PRS Suffix in English between 1400 and 1700 on the basis of VNC stages
  - *doth* → *does*, *giveth* → *gives*, *knoweth* → *knows*, etc. (in PDE, *-(e)th* only survives in archaisms)
- many explanations are conceivable
  - phonological motivations
    - articulatory properties of (the contexts of) verbs
  - syntactic motivations
    - lexical vs. grammatical verbs differed in their preferences
  - semantic motivations
    - verbs of a particular semantic class led the change
  - sociolinguistic motivations
    - women and/or speakers from particular SESs initiated and led the change
    - the change arose in one dialect and spread from there
  - psycholinguistic motivations
    - priming effects

-(e)th and *-(e)s* in the CEEC

- We retrieved from the CEEC
  - ≈13,100 cases of *-(e)th*
  - ≈7,500 cases of *-(e)s*
  - in 233 time periods
- when the proportions of *-(e)s* are plotted against time,
  - there is an overall increasing trend …
  - … which is interrupted by several outliers
- last talk, we applied VNC to these temporal data – now we do the modeling …




1417-78
1479-82
1483-1609
1610-47
1648-81
Distance in summed standard deviations
Observed relative frequency of 'es'
Times

Predicting *-(e)th* and *-(e)s*: a generalized linear mixed effects model

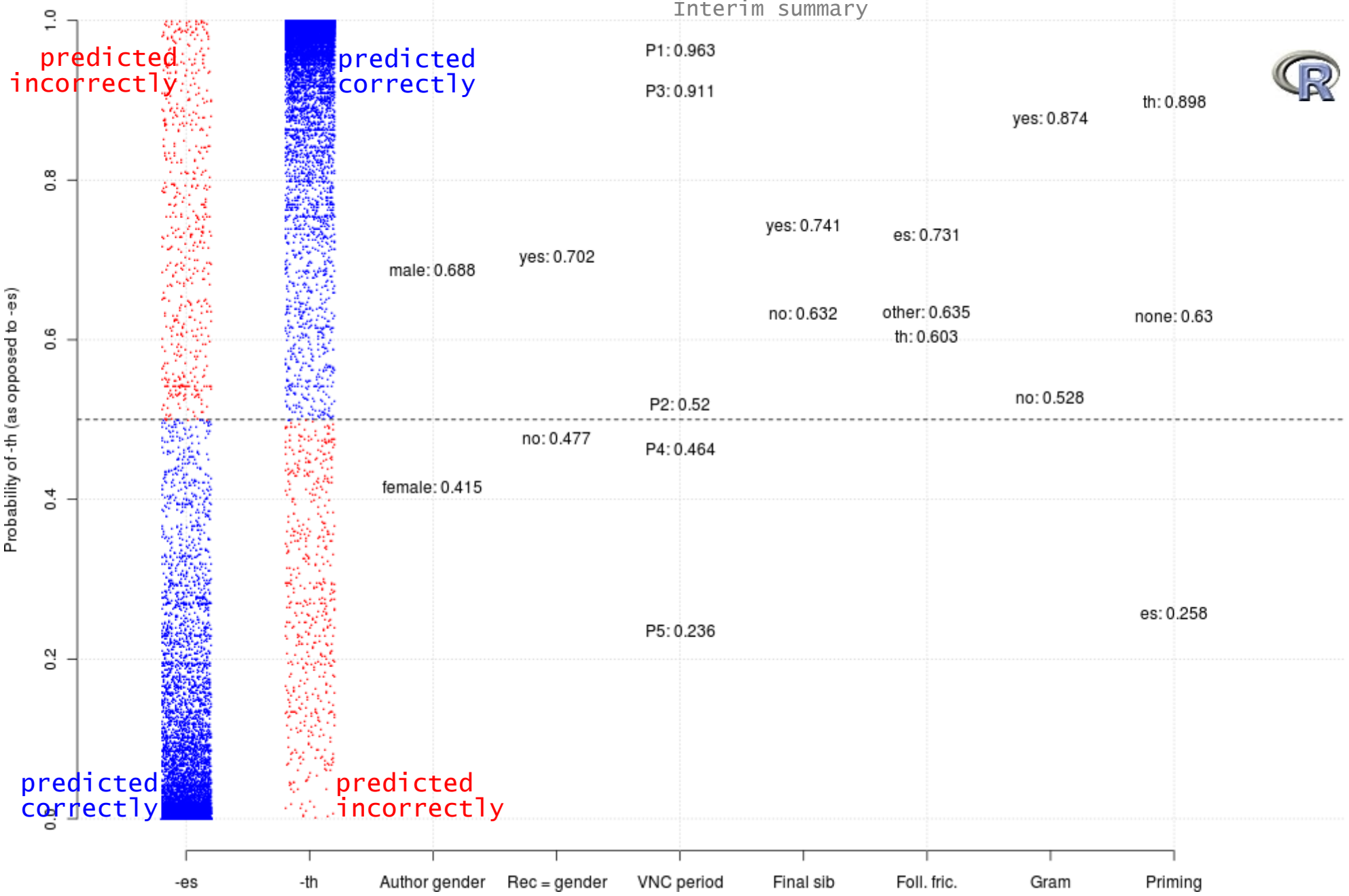
- To determine how *-(e)th* changed to *-(e)s*, we used a generalized linear mixed effects model (lmer in R)
- dependent variable (DV): VARIANT: *-(e)th* vs. *-(e)s*
- independent variables (IVs)
  - 'fixed' effects:
    - VNCPERIOD: 1 vs. 2 vs. 3 vs. 4 vs. 5
    - AUTH_GEND: male vs. female
    - REC_SAME_GENDER: yes vs. no
    - PRIMING: *-(e)th* vs. *-(e)s* vs. none
    - FIN_SIB: yes vs. no
    - CLOSE_FAM: yes vs. no
    - FOL_FRIC: *-s* vs. *-th* vs. other
    - GRAM: yes vs. no
    - interactions of the previous seven IVs with VNCPERIOD
  - 'random' effects:
    - author-specific adjustments to intercept(s) because, e.g.: John Jones (50% *-(e)th*) vs. Winefrid Thimelby (6% *-(e)th*)
    - verb-specific adjustments to intercept(s) because, e.g.: *make* (30% *-(e)th*) vs. *know* (62% *-(e)th*)

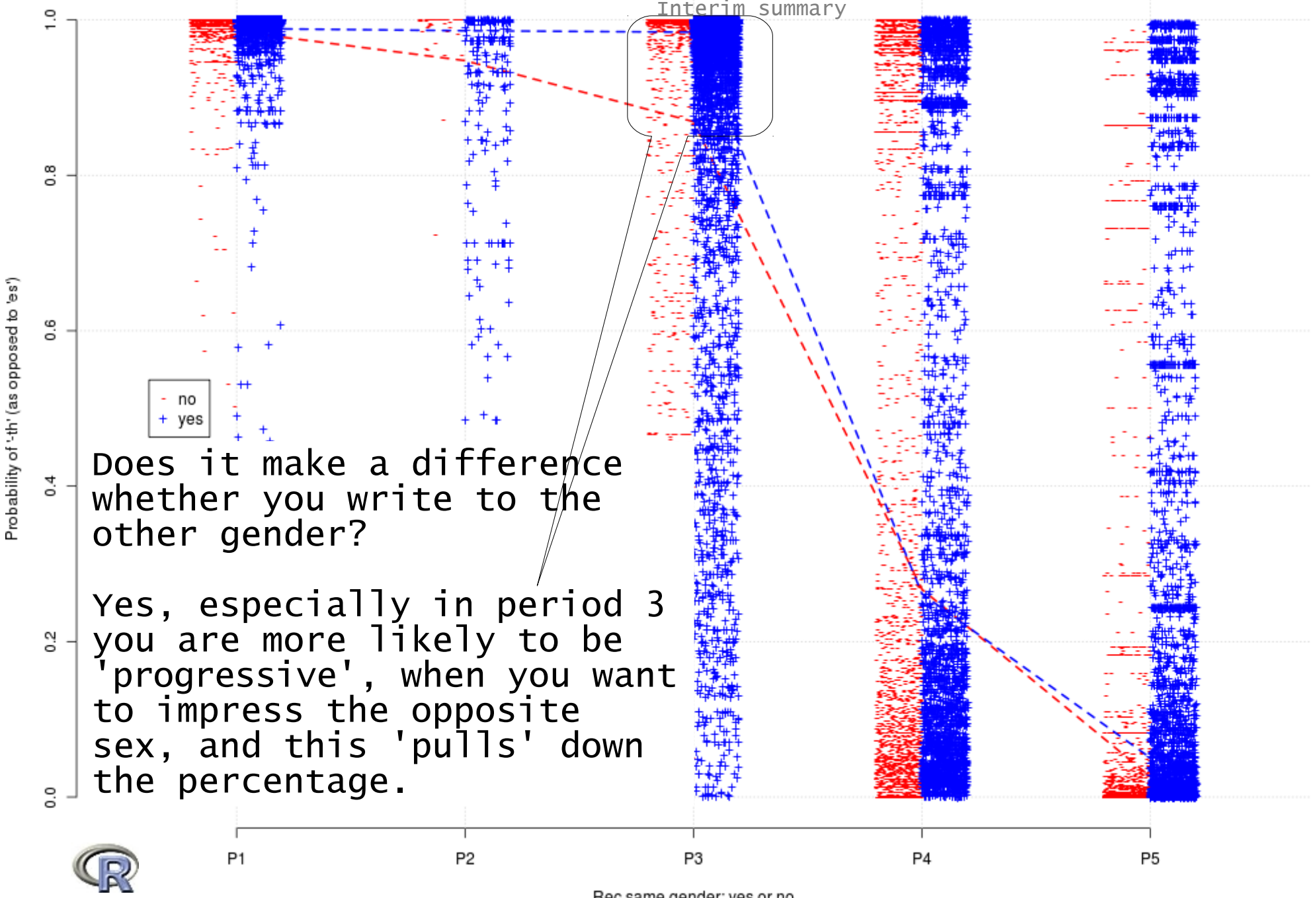
Example coding

- Sentence: *So prayeth he that* promiseth *always to be at your ladiship's command.*
- dependent variable (DV): VARIANT: *-(e)th*
- independent variables (IVs)
  - 'fixed' effects:
    - VNCPERIOD: 4
    - AUTH_GEND: male
    - REC_SAME_GENDER: no
    - PRIMING: *-(e)th*
    - FIN_SIB: yes
    - CLOSE_FAM: no
    - FOL_FRIC: no
    - GRAM: no
  - 'random' effects:
    - author-specific adjustment to intercept(s): James Harrison
    - verb-specific adjustment to intercept(s): *promise*

Results from the GLMEM: overview

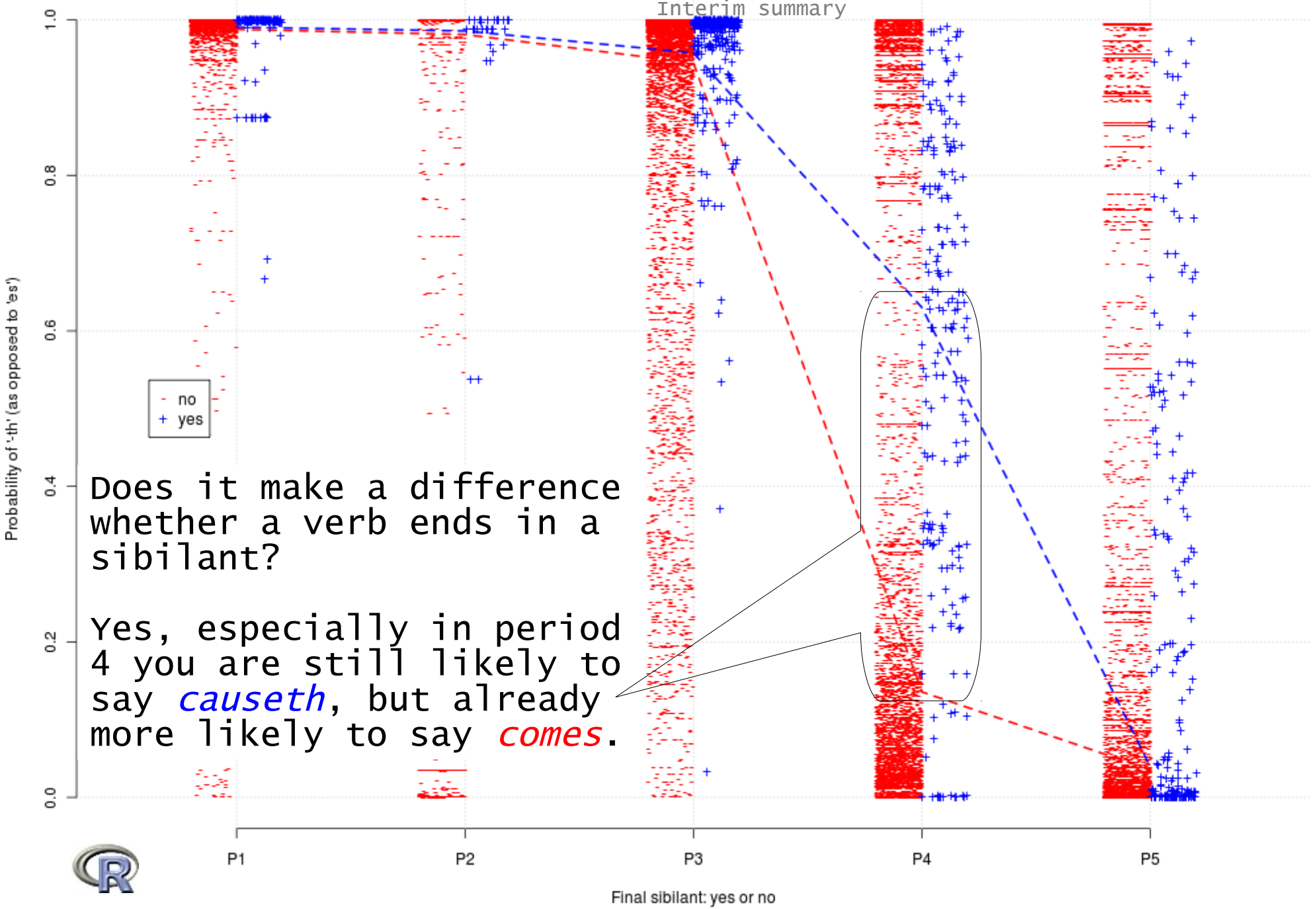
- We used a stepwise model selection procedure to weed out insignificant predictors of the alternation
- predictors that were discarded (in that order)
  - PRIMING:VNCPERIOD ($p \approx 0.45$)
  - AUTH_GEND:VNCPERIOD ($p \approx 0.3$)
  - CLOSE_FAM:VNCPERIOD ($p \approx 0.09$)
  - CLOSE_FAM ($p \approx 0.15$)
- summary statistics of final minimal adequate model
  - *AIC*=7946
  - *BIC*=8223
  - Log-likelihood=-3938; deviance=7876
  - classification accuracy w/ rand effects: 94.59%
  - classification accuracy w/out rand effects: 86.37%


predicted incorrectly
predicted correctly
predicted correctly
predicted incorrectly
Probability of -th (as opposed to -es)
1.0
0.8
0.6
0.4
0.2
0.0
P1: 0.963
P3: 0.911
P2: 0.52
P4: 0.464
P5: 0.236
male: 0.688
female: 0.415
yes: 0.702
no: 0.477
yes: 0.741
no: 0.632
es: 0.731
other: 0.635
th: 0.603
yes: 0.874
no: 0.528
th: 0.898
none: 0.63
es: 0.258
-es
-th
Author gender
Rec = gender
VNC period
Final sib
Foll. fric.
Gram
Priming



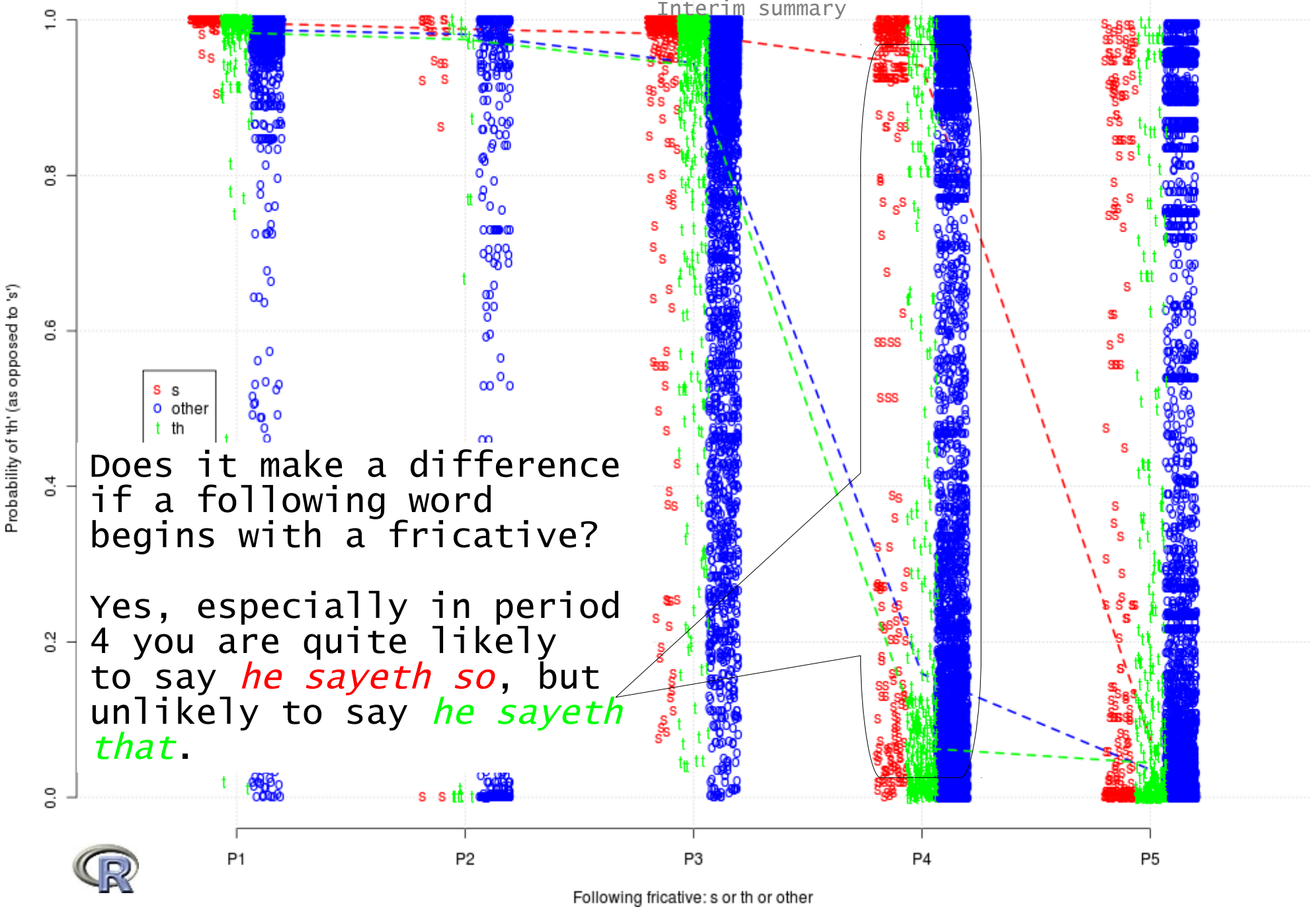
Does it make a difference whether you write to the other gender?

Yes, especially in period 3 you are more likely to be 'progressive', when you want to impress the opposite sex, and this 'pulls' down the percentage.



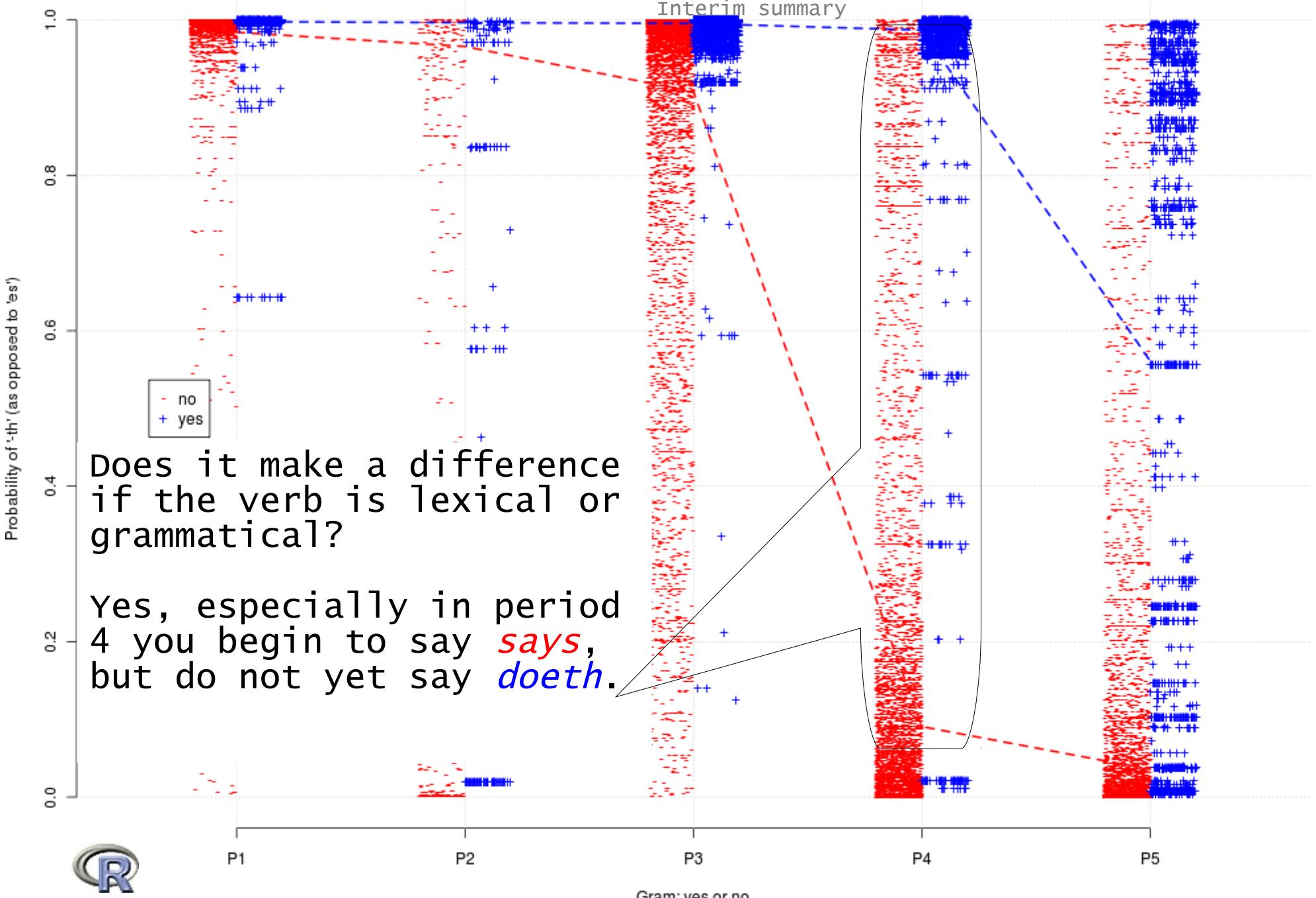
Does it make a difference whether a verb ends in a sibilant?

Yes, especially in period 4 you are still likely to say *causeth*, but already more likely to say *comes*.



Does it make a difference if a following word begins with a fricative?

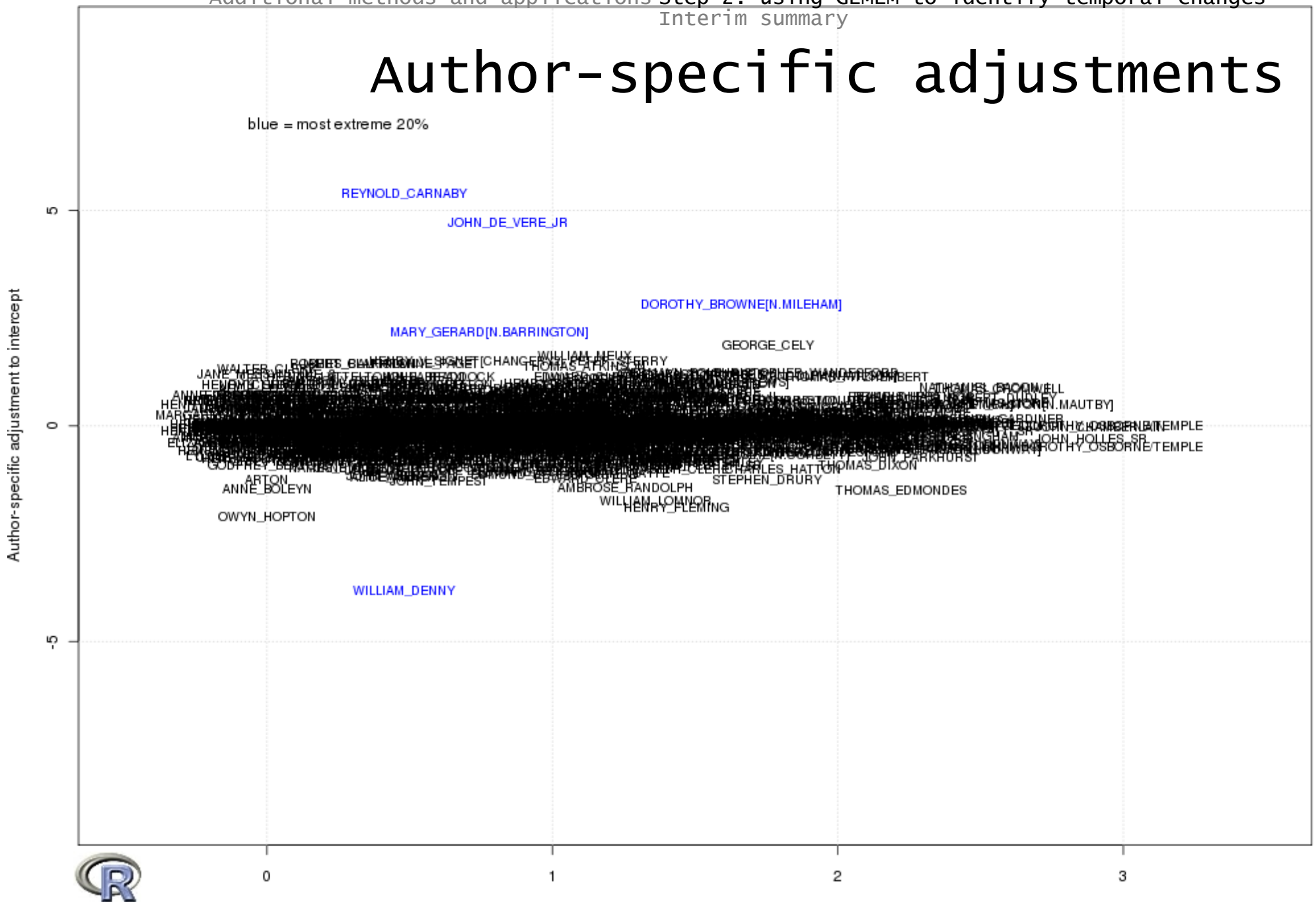
Yes, especially in period 4 you are quite likely to say *he sayeth so*, but unlikely to say *he sayeth that*.



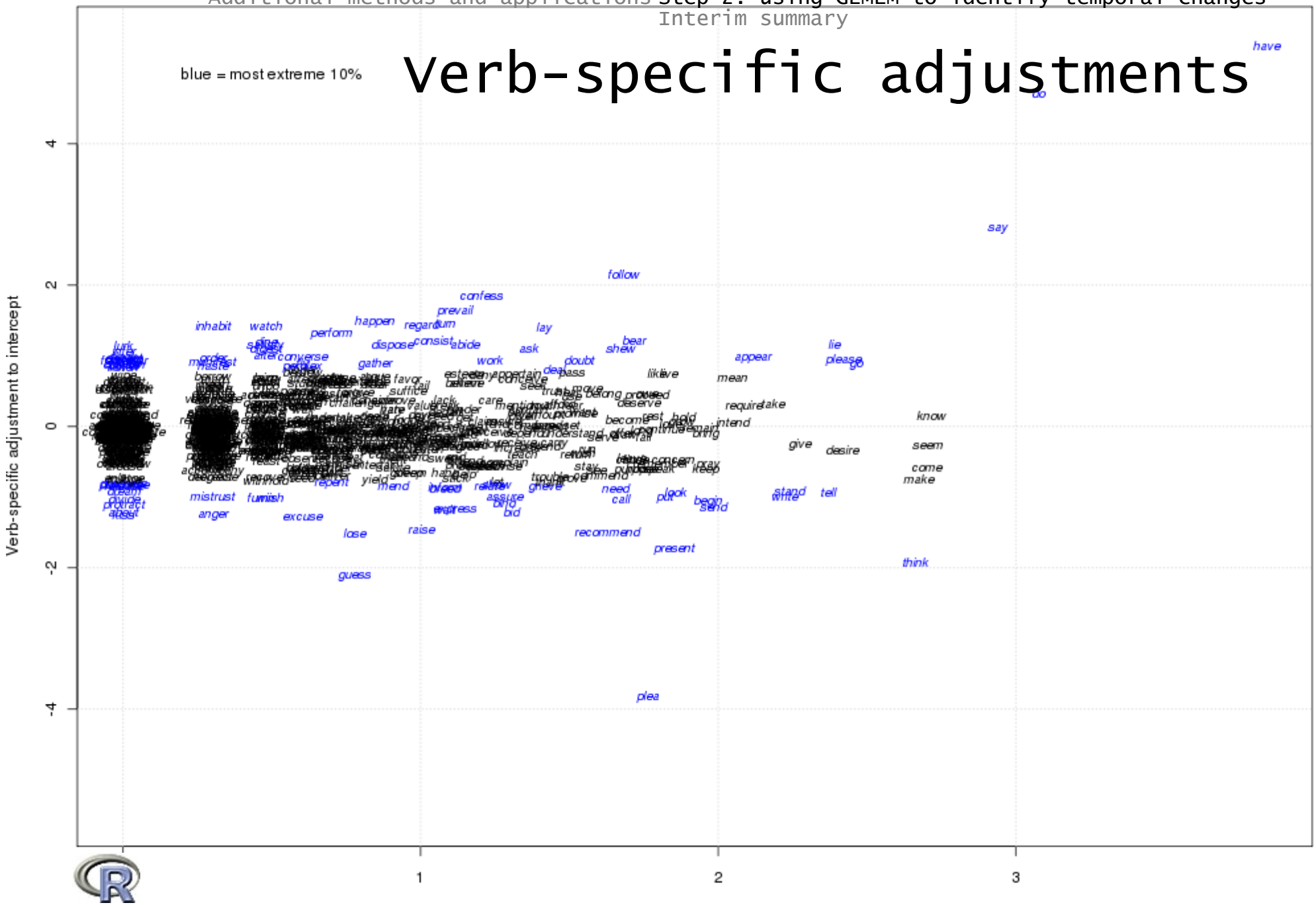
Does it make a difference if the verb is lexical or grammatical?

Yes, especially in period 4 you begin to say *says*, but do not yet say *doeth*.

Author-specific adjustments



Verb-specific adjustments



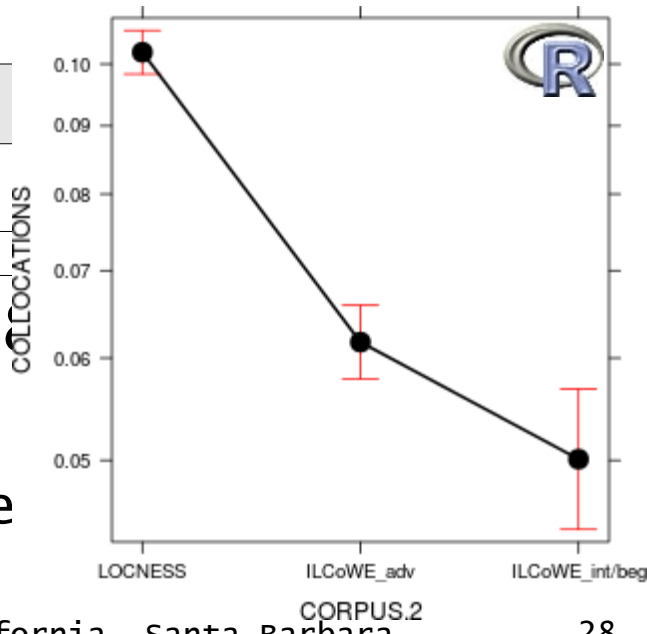
Summary and interpretation

- If we allow for the idiosyncrasies of particular authors and lexical items, we can predict very accurately (94.5%) whether *-(e)s* or *-(e)th* will be chosen in a given context
- upbeat conclusion: we have caught the most important determinants of the alternation (which do not even include variables pertaining to region or SES)
- you are more likely to use *-(e)s* if you
  - are born late
  - are a woman
  - try to impress the opposite sex
  - use verbs without final sibilants (*come*, not *cause*)
  - use a lexical verb, not a grammatical verb
  - are primed with *-(e)s*
  - use a word such as *the* or *that* after the verb
  - as we have shown, these effects do not stay constant across time

Such regression methods are useful in more contexts than you would think

- I hope to have shown that
  - data sets are too large and multi-dimensional to allow for mere eye-balling
  - often, relations between linguistic variables are too complex to allow for mere introspection
- "but if I don't have data that are that complex, isn't this all a massive overkill?" - it depends …
- an example from ~~cognitive~~ corpus linguistics (Laufer & Waldman's 2012 Table 2)

	LOCNESS	ILCoWE: advanced	ILCoWE: intermed.	ILCoWE: basic	Totals
V-N collocations	2527	852	162	68	3609
Non-collocations	22242	12953	2895	1465	39555
Totals	24769	13805	3057	1533	43164



  - they report wrong (!) results from
  - it is a regression that shows that
    - there are only 3 speaker groups: native speakers, adv., and intermed/basic learne
    - the effect size is minuscule: R^2=0.015

Using such methods is not a burden – it's an opportunity

- There are now many fascinating new methods out there
  - generalized linear models are slowly becoming mainstream
    - predicting one variable on the basis of many others
  - mixed-effects models are (more) slowly becoming mainstream
    - predicting one variable on the basis of many others and
    - taking subject/speaker and item-specific variation/dependence into consideration
  - naïve discriminative learning is an interesting alternative similar to the above but cognitively more realistic (cf. Baayen 2011)
  - Bayesian networks are interesting because they force/allow the researcher to test very specific causal hypotheses (cf. Theijssen et al., to app.)
- thus, it is time that the discipline as a whole makes (much) more use of advanced quantitative methods

Thank you!

http://tinyurl.com/stgries