DEVIANT DIACHRONY: EXPLORING NEW METHODS FOR ANALYZING LANGUAGE CHANGE

New Developments in the Quantitative Study of Languages, Helsinki

August 29, 2015

Jason Grafmiller
jason.grafmiller@kuleuven.be

GOALS

adapt/extend recent innovations in multivariate statistical methods—Gries & Deshors'[2] MuPDAR method—to diachronic variationist research

○ take an outcome-centered rather than constraint-centered focus on modeling changes in syntactic variation

➪ examine how speakers' linguistic choices in specific contexts vary over time

○ integrate quantitative hypothesis testing with qualitative exploration and hypothesis generation

[1,2] see handout for references

OUTLINE

1. Methodological background

2. Case studies

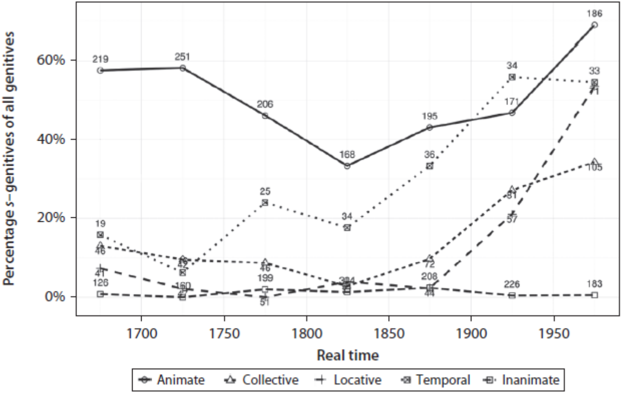
3. Conclusion

METHODOLOGICAL BACKGROUND

VARIATIONIST APPROACH

Traditional variationist studies of diachronic syntactic variation focus on changes in influence of individual factors (‘constraints’) over time.

e.g. Wolk et al.[8] explore variability in the effect of animacy on genitive choice in LME



THE WHY OF HOW

changes in influence of individual constraints tell us about *how* variation has developed, but not so much about *why*

- e.g. why did animacy effects in genitives change like this?
- ‘fixed effects’ categories often very abstract/coarse-grained
- coefficient estimates say little about variability within factor levels

THE WHY OF HOW

changes in influence of individual constraints tell us about *how* variation has developed, but not so much about *why*

- e.g. why did animacy effects in genitives change like this?
- ‘fixed effects’ categories often very abstract/coarse-grained
- coefficient estimates say little about variability within factor levels

can we use regression (or other classification) techniques to find unsuspected patterns in our data?

MUPDAR METHOD

Gries & Deshors[2] devise multi-step method for comparing choices from different groups *A* and *B*

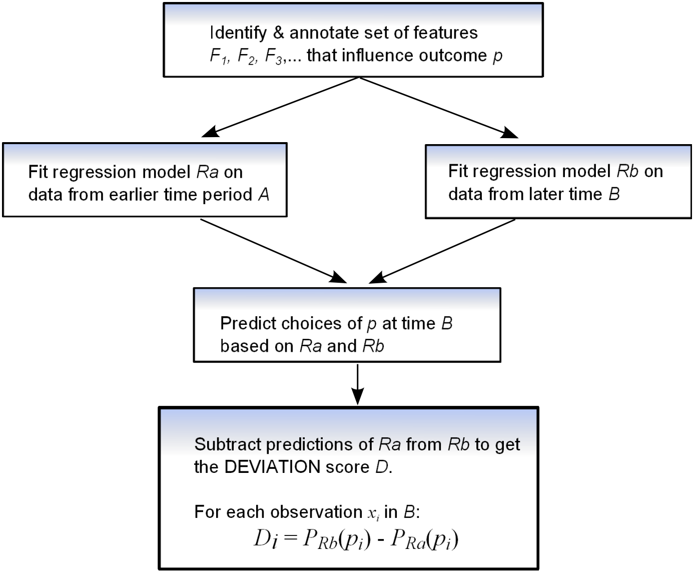
1. fit a model R_a to a reference dataset *A* (e.g. native speaker corpus)
2. use model R_a to predict choices in target dataset *B* (e.g. learner corpus)
3. consider whether speaker from *B* made different choice than speaker from *A* would have
4. fit new model(s) predicting binary and/or finer-grained differences (degree of deviation) in speakers' choices

ITEM-BASED DEVIATION ANALYSIS

an adaptation of MuPDAR for directly comparing predicted probabilities from models fit to separate datasets

- explore how outcome probabilities for specific observations at later times deviate from those of earlier time(s)
- explore deviations for *all* contexts, not just those where groups made different choices
  - usage-based approaches assume gradient change in probabilistic effects
  - large differences in probability of outcome w.r.t. factor *F* may exist even when the actual outcome is the same
  - do speakers make the same choices for the same reasons?

PROCEDURE (SEE ALSO GRIES & DESHORS 2014:127)



Identify & annotate set of features $F_1, F_2, F_3,\ldots$ that influence outcome *p*

Fit regression model *Ra* on data from earlier time period *A*

Fit regression model *Rb* on data from later time *B*

Predict choices of *p* at time *B* based on *Ra* and *Rb*

Subtract predictions of *Ra* from *Rb* to get the DEVIATION score *D*.

For each observation x_i in *B*:

$$D_i = P_{Rb}(p_i) - P_{Ra}(p_i)$$

DEVIATION MODEL

deviation score D represents the difference in outcome probability between A and B

○ $D > 0$: outcome more likely in B than A

○ $D < 0$: outcome more likely in A than B

○ $D = 0$: prob. of outcome exactly the same in A and B

fit linear (mixed) model treating D as the outcome and X = $F_1, \ldots, F_\text{n}$ as predictors

○ `lmer(D` $\sim$ `F`$_1$ `+ F`$_2$ `+ ... + F`$_n$`, data = B)`

DEVIATION MODEL

deviation score D represents the difference in outcome probability between A and B

- $D > 0$: outcome more likely in B than A
- $D < 0$: outcome more likely in A than B
- $D = 0$: prob. of outcome exactly the same in A and B

fit linear (mixed) model treating D as the outcome and $X = F_1, \ldots, F_{\mathrm{n}}$ as predictors

- lmer(D $\sim$ F$_1$ + F$_2$ + ... + F$_n$, data = B)

examine factors yielding the largest changes in deviation scores

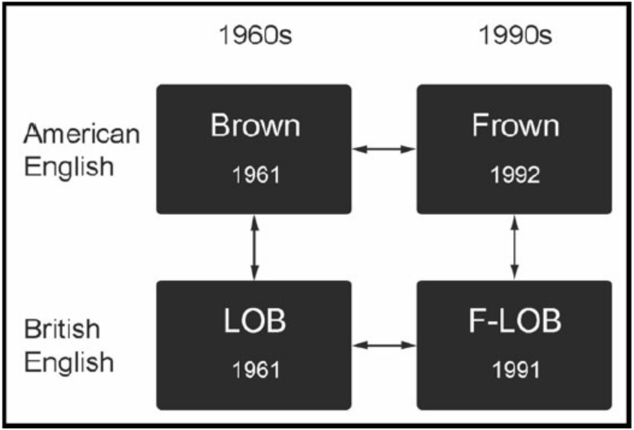
CASE STUDIES

THREE ALTERNATIONS

1. subject relativizer choice (*the cot that caught the tot* vs. *the cot which caught the tot*)
2. genitive choice (*Sally's pet tarantula* vs. *the pet tarantula of Sally*)
3. dative choice (*give the dog a bone* vs. *give a bone to the dog*)

All are known to be changing over time, w.r.t. certain features[4,5,6,7]

BROWN FAMILY



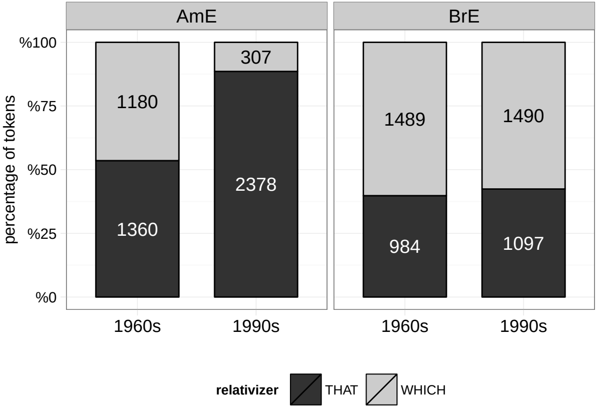
	1960s	1990s
American English	Brown 1961	Frown 1992
British English	LOB 1961	F-LOB 1991

CASE: SUBJECT RELATIVIZERS

- *engineering skills that could be used to construct embankments for a tidal power scheme* [FLOB:J73]

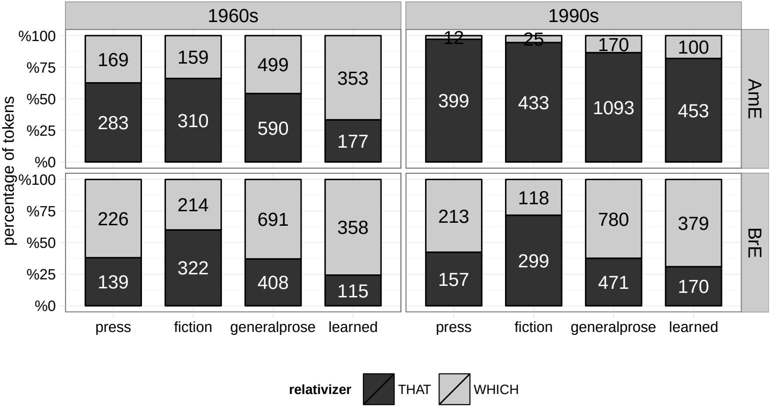
- *routines which continuously check the monitor for various error conditions* [FROWN:J78]

RELATIVIZERS: DISTRIBUTION



○ large reduction in AmE use of *which* from 1960s to 1990s

RELATIVIZERS: BY GENRE



- AmE dropping *which* across the board, but *that* increasing in BrE only in fiction texts

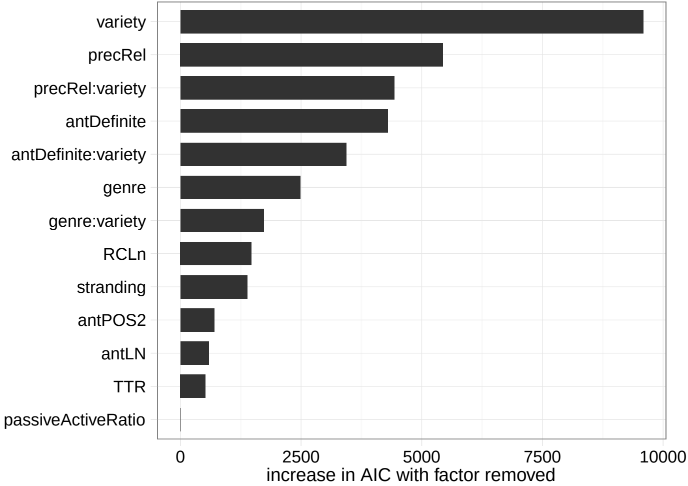
RELATIVIZERS: MODEL PREDICTORS

annotate for various internal and stylistic factors associated with formality[6]

internal	length of RC preceding relativizer antecedent definiteness	length of antecedent antecedent POS
stylistic	lexical density passivization rate	genre P-stranding rate
external	variety	

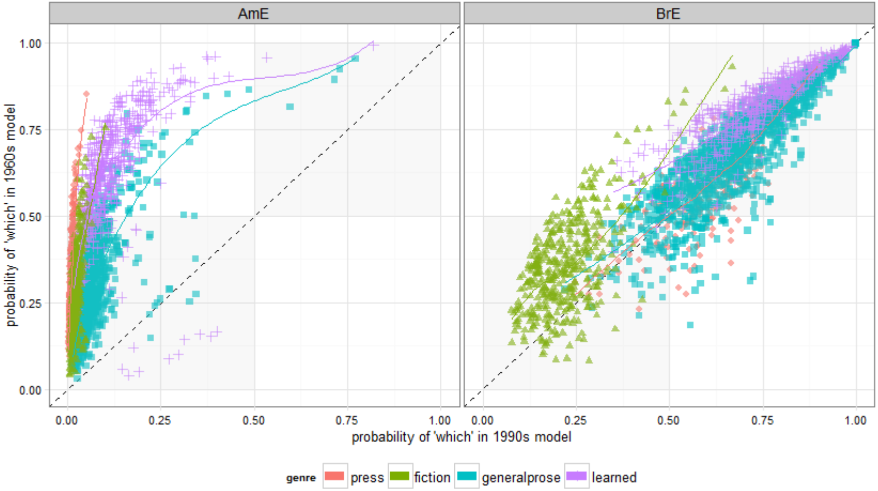
RELATIVIZERS: DEVIATION MODEL

explanatory contribution of predictors influencing deviation score



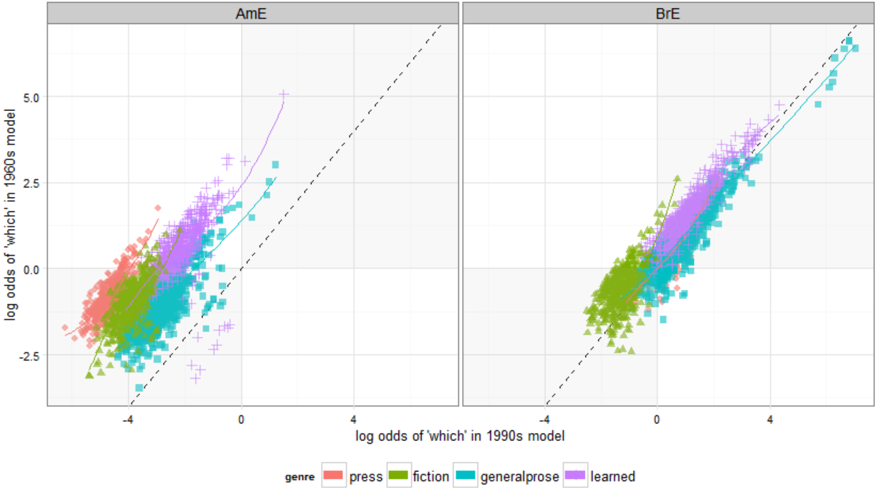
RELATIVIZERS AND GENRE

probability scale



RELATIVIZERS AND GENRE

log odds scale

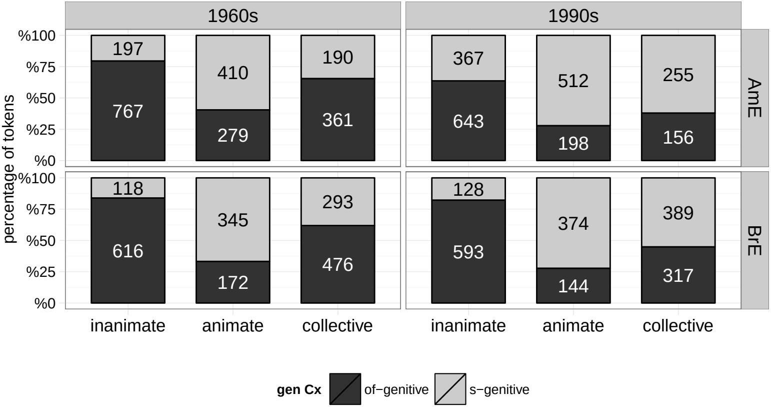


CASE: GENITIVES

- ○ *s*-genitive: *foreign steelmakers'*$_{\text{poss'r}}$ *mouths*$_{\text{poss'm}}$ [BROWN:A43]

- ○ *of*-genitive: *the foreign policies*$_{\text{poss'm}}$ *of her chosen successor*$_{\text{poss'r}}$ [FLOB:B15]

GENITIVES: DISTRIBUTION



- possr animacy by far the single strongest predictor

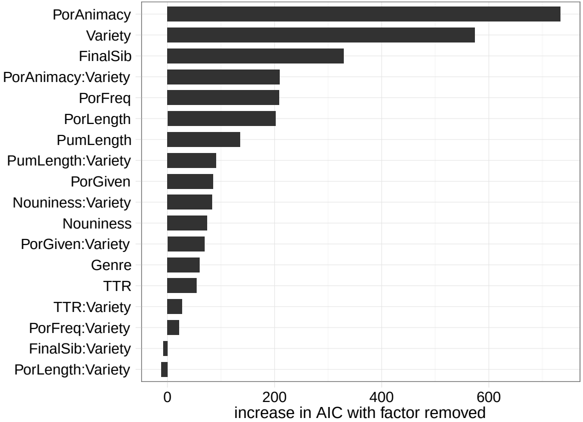
GENITIVES: MODEL PREDICTORS

annotate for internal and context factors associated with formality and ‘economy’[5]

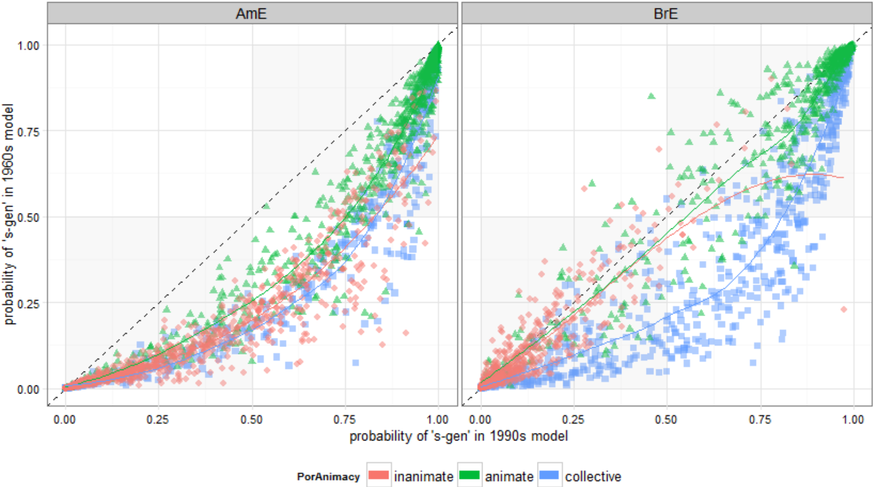
internal	animacy of poss’r	final sibilant
	length of poss’r	length of Poss’m
	frequency of poss’r	givenness of poss’r
stylistic	lexical density	genre
	nouniness	
external	variety	

GENITIVES: DEVIATION MODEL

- predictors influencing deviation score parallels previous research[]
- possr animacy esp. shows significant interactions with variety and time



POSSR ANIMACY ACROSS TIME AND THE POND



GENITIVES: A CLOSER LOOK

inspection of collective poss'rs with large deviation scores shows increased use of locative-as-collective nouns in BrE, e.g. *North Korea's contention*

○ sig. different from AmE ($p_{\text{fisher}} < 0.001$)

	locative	non-locative
AmE	7	87
BrE	25	37

○ suggestive locus for further exploration of stylistic changes across varieties

○ collective poss'rs have been changing for some time[7,8]

CONCLUSION

SUMMING UP

- advantages
    - results compatible with traditional variationist methods
    - offers fine-grained perspective on data driving larger trends
    - quantitative hypothesis confirmation ➪ qualitative hypothesis exploration
- disadvantages
    - (arguably) more complicated than standard methods
    - how to deal with more than 2 (ordered) groups, e.g. multiple centuries?

FUTURE DIRECTIONS

- adapt method to data covering multiple time periods[3,6]
- synchronic applications
  - ESL/EFL contexts[1,2]
  - regional variation
  - other sociolinguistic dimensions
  - . . .
- apply to non-syntactic variables
- . . .
- suggestions?

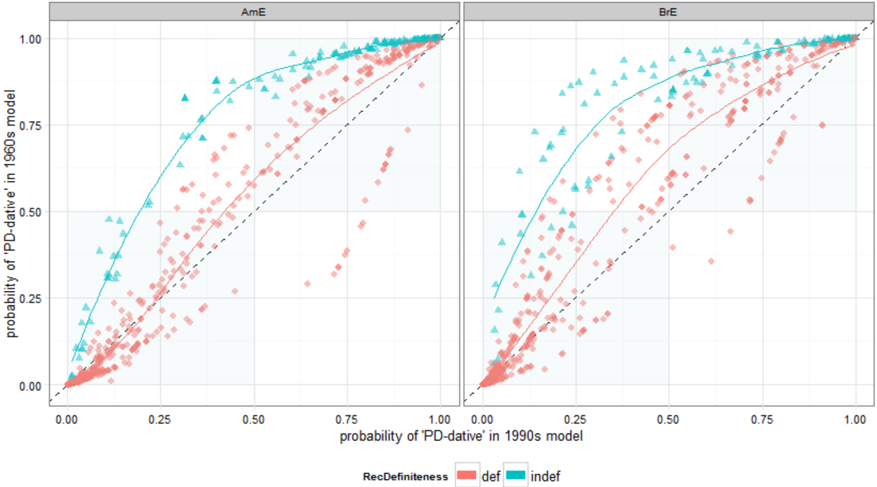
Thank you!

jason.grafmiller@kuleuven.be

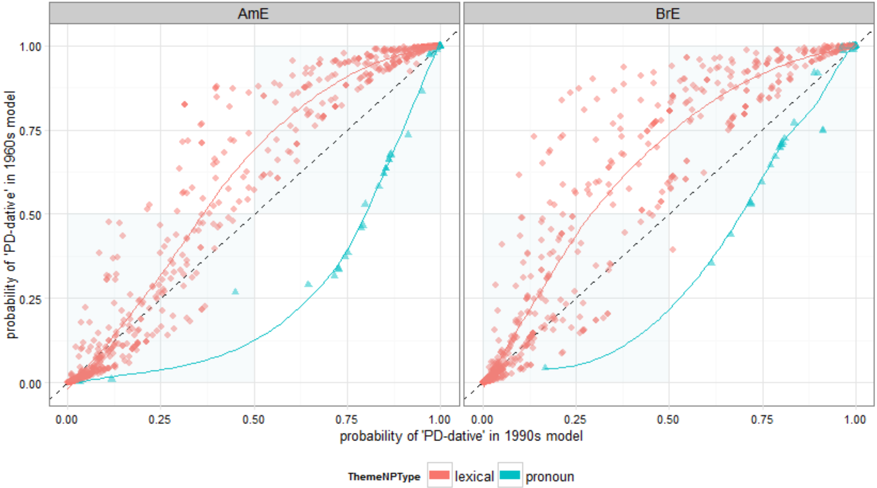
Additional thanks to Lars Hinrichs, Benedikt Szmrecsanyi, Axel Bohmann, Scott Grimm, and Joan Bresnan for sharing their datasets.

KU LEUVEN

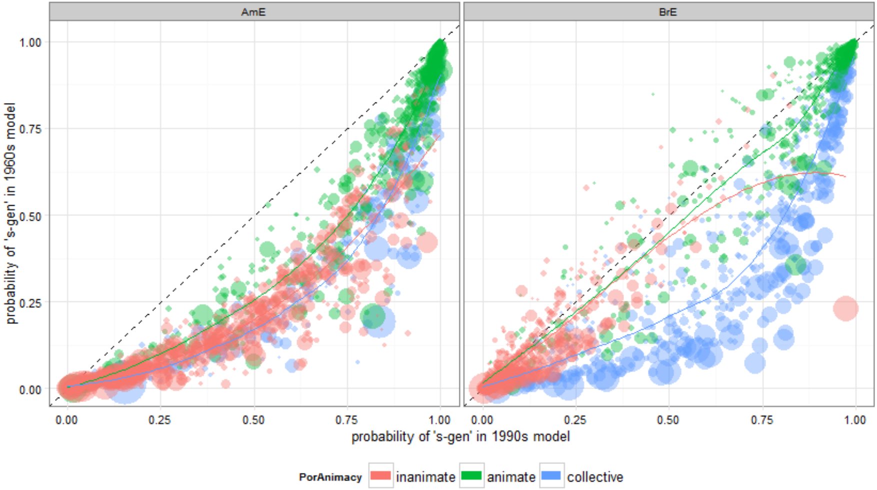
RECIPIENT DEFINITENESS IN DATIVES



THEME PRONOMINALITY IN DATIVES



ANIMACY AND LENGTH IN GENITIVES


AmE
BrE
probability of 's-gen' in 1960s model
probability of 's-gen' in 1990s model
PorAnimacy
inanimate
animate
collective