Deviant diachrony: Exploring new methods for analyzing language change

Jason Grafmiller
KU Leuven

Present a novel method for analyzing syntactic change within a probabilistic framework: **I**tem-based **D**eviation **A**nalysis (IDA).

- Variationist studies on diachronic syntactic variation focus on aggregate trends in historical corpora using standard regression-with-interaction models [5, 8].
- This approach takes a more fine-grained, outcome-centered perspective on syntactic variation in diachrony, inspired by recent work on syntactic variation in ESL and EFL, i.e. the MuPDAR method [2, 1]
- Explore how the probability of a construction in a specific context varies across speakers from different time periods. In essence, it asks, *"Given the same syntactic choice(s) in the same context(s), how would the probability of speakers' choice of a Cx at one time have differed from the probability of speakers' choice of that same Cx at a later time?"*

Method

Outline of procedure

1. Fit a model R_a to dataset A from earliest time period.
2. Fit a model R_b with same structure as R_a to the comparison dataset(s) B from later time(s).
3. Generate predicted values from both model R_a AND model R_b on dataset B, giving two sets of predicted probabilities for observations in B. We can now ask:
   "For a given context, how likely is the predicted outcome according to model R_a, and how likely is it according to model R_b?"
4. Compare predictions by subtracting the predicted probabilities (or log odds) for R_a from R_b for each observation. This is the DEVIATION SCORE for that token.

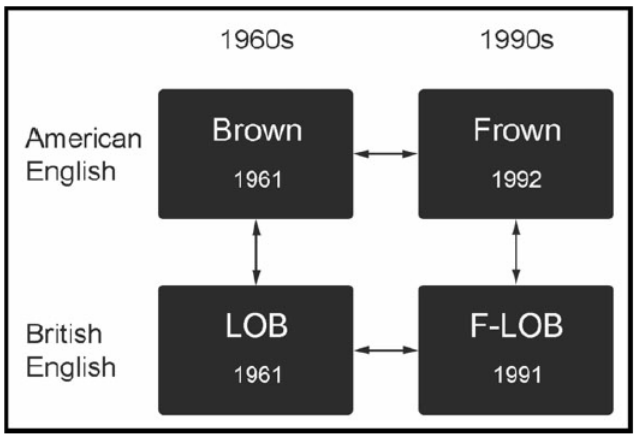
   For observation i in dataset B,

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

   Deviation scores below 0 reflect contexts where the the probability/log odds of the outcome is greater in the earlier time (R_a) than in later time(s) (R_b).
5. Explore deviations between time periods
   (a) univariate: examine correlations of predictors with deviation scores
   (b) multivariate: linearly regress deviation score against predictors
   (c) qualitative: inspect observations with extreme or atypical scores for unknown/hidden features

Case studies

Three cases studies using data from the brown family of corpora



- Subject relativizer choice

(1) a. "that": *engineering skills **that** could be used to construct embankments for a tidal power scheme* [FLOB:J73]

b. "which": *routines **which** continuously check the monitor for various error conditions* [Frown:J78]

- Genitive Cx choice

(2) a. *s*-genitive: *her mother's hospital room* [Frown:A23]

b. *of*-genitive: *the political survival of his two colleagues* [FLOB:A04]

- Dative Cx choice

(3) a. DO-dative: *Douglass gave black male suffrage a much higher priority than white female suffrage* [Frown:G08]

b. PD-dative: *the State Board of Education should be directed to "give priority" to teacher pay raises* [Brown:A01]

Relativizer choice

Data ($N = 10285$) collected and annotated by Hinrichs, Szmrecsanyi & Bohmann [6].

Predictor	Description
`precRel`	rel. of preceding RC ('that', 'which', 'zero', 'none')
`antDefinite`	definiteness of antecedent ('def', 'indef')
`RCLength`	# of words in relative clause
`antLength`	# of words in antecedent NP
`antPOS`	part-of-speech of antecedent ('noun', 'other')
`TTR`	100 word context
`passiveActiveRatio`	ratio of passive to active verbs in text
`stranding`	per 10k words
`Genre`	('press', 'learned', 'fiction', 'generalprose')
`Variety`	AmE, BrE

Table 1: Predictors of relativizer choice (all numeric predictors were z-score standardized)

PROCEDURE

1. Fit GLMM models to 1960s data (R_a;$N = 5013$) and to 1990s data (R_b;$N = 5272$) predicting log odds of *which*
2. Predict use of *which* in the 1990s data using both R_a and R_b
3. Compare differences in predictions for the two models: the DEVIATION SCORES ($R_b - R_a$)
4. Inspect and/or model patterns in deviation scores.

(4) **Relativizer choice model formula:**

```
Response ~ (1|CorpusFile) + (1 + Variety|Category) +
           Variety * (Genre + RCLength + antLength + precRel +
           antDefinite) + antPOS + TTR + passiveActiveRatio +
           stranding
```

Fit of both 1960s and 1990s models is very good (Table 2).

	C	D_{xy}	AIC	BIC	logLik	deviance	df.resid
1960s model	0.90	0.80	5394.96	5564.47	-2671.48	5342.96	4987
1990s model	0.94	0.88	4290.99	4461.82	-2119.50	4238.99	5246

Table 2: Summary statistics for relativizer choice models

Deviation model

The random effects structure of deviation model was simplified due to convergence difficulties. All other predictors were the same. Model diagnostics (multicollinearity, residuals structure, etc.) have yet to be completed.

(5) Relativizers deviation model formula:

```
deviation ~ (1|CorpusFile) +
            Variety * (Genre + RCLength + antLength + precRel +
            antDefinite) + antPOS + TTR + passiveActiveRatio +
            stranding
```

Groups	Name	Std.Dev.	Variance
file	(Intercept)	0.128	0.016
Residual		0.125	0.016

Table 3: Random effects in relativizers deviation model (simplified)

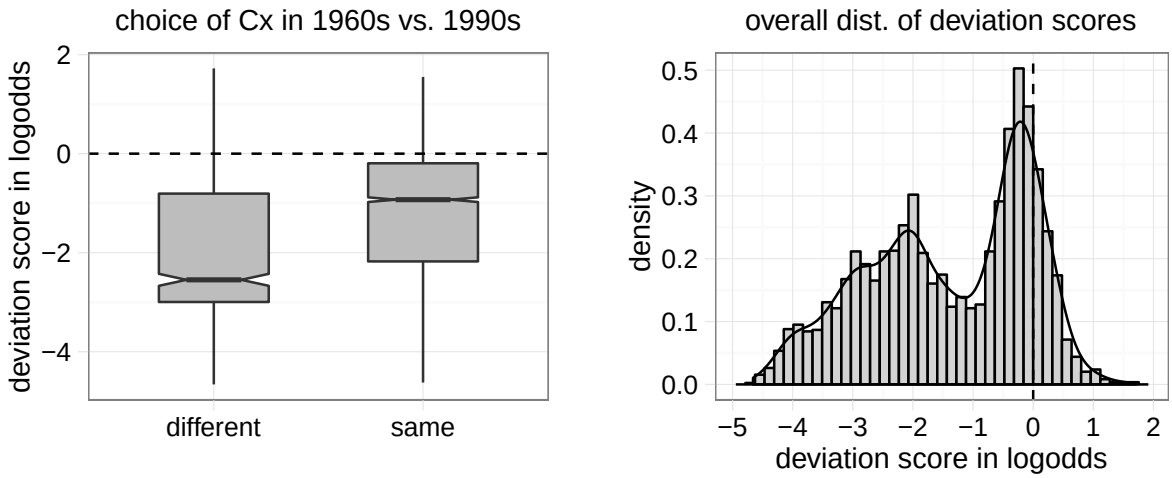


Figure 1: Distribution of deviation scores obtained from the relativizers models.

	Estimate	Std. Error	*t* value
(Intercept)	-3.713	0.019	-195.703
RCLength	-0.080	0.002	-41.606
antLength	-0.048	0.002	-25.027
precRel=THAT	0.526	0.008	62.505
precRel=WHICH	0.126	0.013	9.913
precRel=ZERO	0.272	0.010	26.701
antPOS=other	0.151	0.005	27.497
antDefinite=indef	-0.465	0.006	-81.617
TTR	-0.169	0.006	-26.926
passiveActiveRatio	0.002	0.006	0.316
stranding	0.279	0.005	58.205
genre=fiction	0.537	0.023	23.394
genre=generalprose	1.732	0.020	87.181
genre=learned	0.751	0.025	29.734
variety=BrE	3.877	0.026	152.017
precRelTHAT:varietyBrE	-0.728	0.013	-57.906
precRelWHICH:varietyBrE	-0.211	0.016	-13.607
precRelZERO:varietyBrE	0.124	0.014	8.578
antDefiniteindef:varietyBrE	0.562	0.008	71.280
genrefiction:varietyBrE	-1.719	0.031	-56.210
genregeneralprose:varietyBrE	-1.946	0.027	-71.979
genrelearned:varietyBrE	-1.591	0.033	-48.910

Table 4: Coefficient estimates for relativizers deviation model

Genitive choice

Data ($N = 8300$) collected and annotated by Hinrichs & Szmrecsanyi [5].

(6) a. *s*-genitive: foreign steelmakers'$_{poss'r}$ mouths$_{poss'm}$ [Brown:A43]

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

Predictor	Description
`PorAnimacy`	animacy of poss'r ('human', 'collective', 'inanimate')
`FinalSibilant`	does poss'r end in sibilant? (Y/N)
`PorGiven`	is poss'r given? (Y/N)
`PorLength`	# of words in poss'r
`PumLength`	# of words in poss'm
`PorFrequency`	# obs. of poss'r head per text
`TTR`	type-token ratio of 100 word context
`Nouniness`	# of nouns in 100 word context
`Genre`	text category of obs. ('A', 'B')
`Variety`	variety of obs. ('AmE', 'BrE')

Table 5: Predictors of genitive choice (all numeric predictors were z-score standardized)

PROCEDURE

1. Fit GLMM models to 1960s data (R_a;$N = 4224$) and to 1990s data (R_b;$N = 4076$) predicting log odds of *s*-genitive
2. Predict *s*-genitive in the 1990s data using both R_a and R_b
3. Compare differences in predictions for the two models: the DEVIATION SCORES ($R_b - R_a$)
4. Inspect and/or model patterns in deviation scores.

(7) **Genitive choice model formula:**

```
Response ~ (1|CorpusFile) + (1|PossrHead) + (1|PossmHead) +
           Variety * (PorAnimacy + PorFreq + FinalSibilant +
           PorGiven + PorLength + PumLength + Genre + TTR +
           Nouniness)
```

Fit of both 1960s and 1990s genitive models is also very good (Table **??**).

	C	D_{xy}	AIC	BIC	logLik	deviance	df.resid
1960s model	0.98	0.97	3575.95	3727.41	-1763.95	3527.90	4052
1990s model	0.97	0.95	3444.81	3597.18	-1698.41	3396.81	4200

Table 6: Summary statistics for genitives choice models

Deviation model

Again, full model diagnostics (multicollinearity, residuals structure, etc.) have yet to be completed.

(8) **Genitive deviation model formula:**

```
deviation ~ (1|CorpusFile) + (1|PorHead) + (1|PumHead) +
            Variety * (PorAnimacy + PorFreq + FinalSibilant + PorGiven +
            PorLength + PumLength + Genre + TTR + Nouniness)
```

Groups	Name	Std.Dev.	Variance
PumHead	(Intercept)	0.0199	0.0004
PorHead	(Intercept)	0.0324	0.001
CorpusFile	(Intercept)	0.0100	1e-4
Residual		0.0942	0.008

Table 7: Random effects in genitive deviation model (simplified)

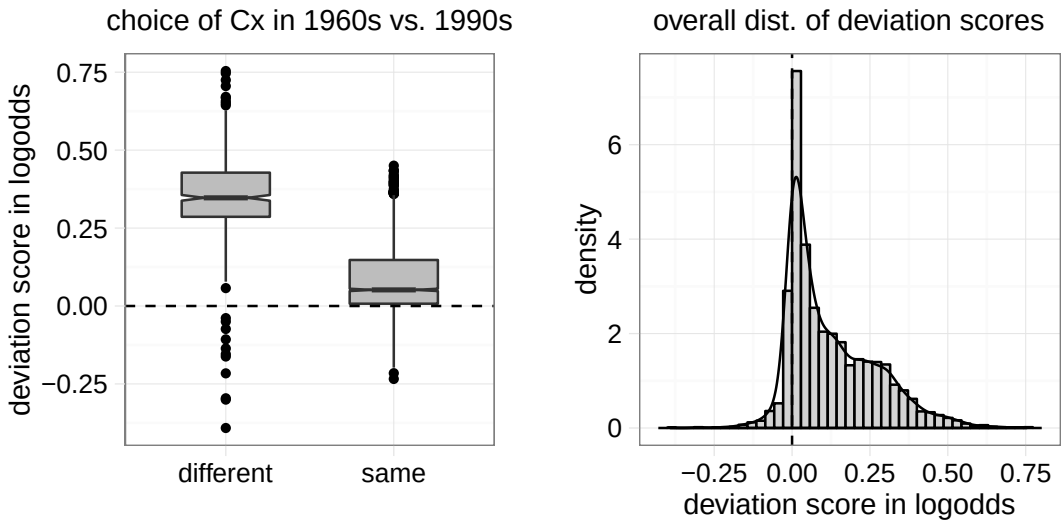


Figure 2: Distribution of deviation scores obtained for the genitive models.

Predictor	Estimate	Std. Error	t value
(Intercept)	0.175	0.005	38.123
PorAnimacy=animate	-0.023	0.006	-3.982
PorAnimacy=collective	0.037	0.006	5.747
PorFreq	0.037	0.002	14.843
FinalSib=yes	-0.089	0.006	-15.139
PorLn	-0.026	0.002	-10.800
PorGiven=yes	-0.056	0.006	-9.921
Genre=B	0.043	0.004	9.764
Nouniness	0.020	0.003	6.683
Variety=BrE	-0.164	0.006	-26.740
PorAnimacyanimate:VarietyBrE	0.041	0.008	4.854
PorAnimacycollective:VarietyBrE	0.129	0.008	15.392
PorFreq:VarietyBrE	-0.022	0.004	-5.791
PumLn:VarietyBrE	0.034	0.003	10.214
PorGivenyes:VarietyBrE	0.075	0.008	8.927
TTR:VarietyBrE	-0.021	0.003	-6.320
Nouniness:VarietyBrE	-0.044	0.004	-11.694

Table 8: Coefficient estimates for genitive deviation model

- Curious effect of poss'r animacy in BrE: differences in models almost entirely localized to collective poss'rs

(9) BrE locative-collective examples:
North Korea's contention, North Korea's defense, China's aging despots, Britain's biggest electronics company, Britain's colonial child, California's pioneering wind turbines, the daunting challenge of Australia, Hong Kong's growing prosperity, India's huge population, the prosecuting authorities of Newcastle, Hong Kong's capital markets

Dative choice

Data ($N = 3100$) collected and annotated by Grimm & Bresnan [4].

(10) a. DO-dative: which gave him$_{recipient}$ inferiority complexes$_{theme}$ [Brown:B13]]

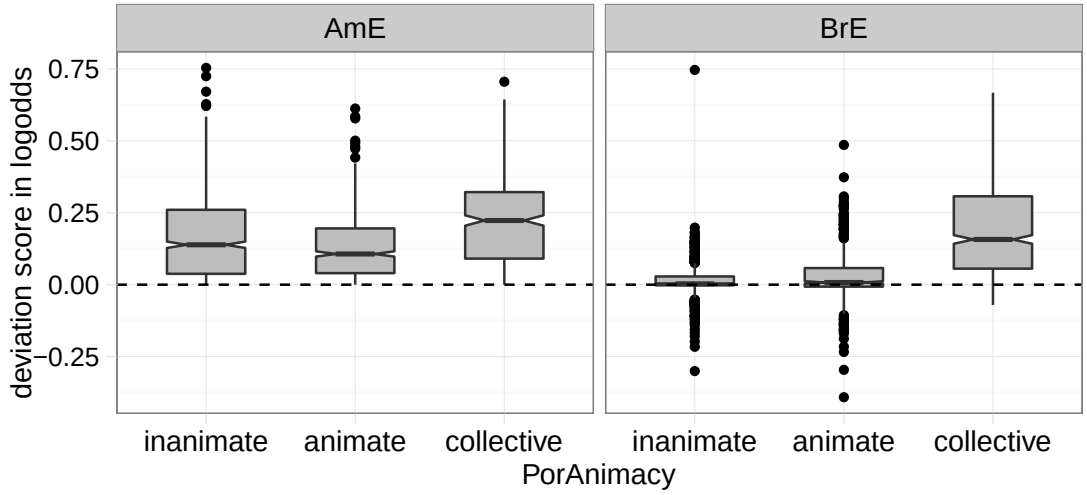


Figure 3: Distribution of deviation scores obtained for the genitive models.

b. PD- dative: teach the law$_{theme}$ to the people$_{recipient}$ [FLOB:D03]

Predictor	Description
`RecNPType`	NP type of recipient (‘lexical’, ‘pronoun’)
`RecDefinite`	definiteness of recipient (‘def’, ‘indef’)
`RecGiven`	is recipient given? (Y/N)
`RecFrequency`	# obs. of recipient head noun in text
`RecAnimacy`	animacy of recipient (‘animate’, ‘collective’, ‘inanimate’)
`ThemeNPType`	NP type of theme (‘lexical’, ‘pronoun’)
`ThemeDefinite`	definiteness of theme (‘def’, ‘indef’)
`ThemeGiven`	is theme given? (Y/N)
`ThemeConcrete`	concreteness of theme (‘concrete’, ‘non-concrete’)
`LengthRatio`	# words in rec. divided by # words in theme
`Variety`	variety of obs. (‘AmE’, ‘BrE’)

Table 9: Predictors of dative choice (all numeric predictors were z-score standardized)

PROCEDURE

1. Fit GLMM models to 1960s data (R_a;$N = 1579$) and to 1990s data (R_b;$N = 1521$) predicting log odds of PD-dative
2. Predict PD-dative in the 1990s data using both R_a and R_b
3. Compare differences in predictions (deviation scores) for the two models

(11) **Dative choice model formula:**

```
Response ~ (1|Verb) + (1|Category) +
           Variety * (RecAnimacy) + RecNPType + RecDefinite +
           RecGiven + RecFrequency + ThemeNPType + ThemeDefinite +
           ThemeGiven + ThemeConcrete + LengthRatio
```

Fit of both 1960s and 1990s genitives models is also very good (Table **??**).

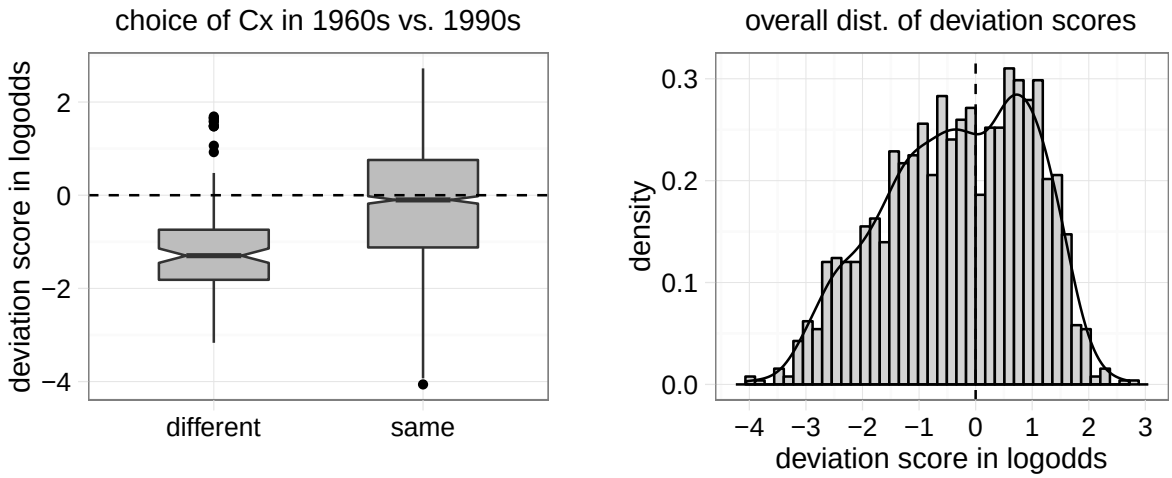


Figure 4: Distribution of deviation scores obtained for the dative models.

	C	D_{xy}	AIC	BIC	logLik	deviance	df.resid
1960s model	0.97	0.95	730.07	821.27	-348.03	696.07	1562
1990s model	0.96	0.92	864.62	955.18	-415.31	830.62	1504

Table 10: Summary statistics for datives choice models

Deviation model

Again, full model diagnostics (multicollinearity, residuals structure, etc.) have yet to be completed.

(12) **Dative deviation model formula:**

```
deviation ~ (1|Verb) +
            Variety * (RecAnimacy) + RecNPType + RecDefinite +
            RecGiven + RecFrequency + ThemeNPType + ThemeDefinite +
            ThemeGiven + ThemeConcrete + LengthRatio
```

Groups	Name	Std.Dev.	Variance
VerbLemma	(Intercept)	0.043	0.002
Residual		0.176959	0.031

Table 11: Random effects in dative deviation model (simplified)

	Estimate	Std. Error	t value
(Intercept)	-0.881	0.020	-42.987
RecNPType=pronoun	1.282	0.012	105.862
RecDefinite=indef	-1.353	0.013	-101.654
RecGiven=N	-0.044	0.011	-3.888
ThemeNPType=pronoun	1.284	0.023	55.624
ThemeDefinite=indef	0.681	0.011	61.412
ThemeGiven=N	0.053	0.010	5.322
RecFrequency	0.072	0.005	14.015
LengthRatio	-0.504	0.006	-90.398
RecAnimacy=collective	0.382	0.018	21.218
RecAnimacy=inanimate	0.481	0.011	43.698
ThemeConcrete=on-concrete	-0.403	0.013	-31.345
Variety=BrE	-0.398	0.009	-42.759

Table 12: Coefficient estimates for dative deviation model

Conclusion

The IDA method is a natural extension of standard and more recent (e.g. MuPDAR) regression modeling techniques:

1. It gives results that parallel closely those found through more traditional regression-with-interaction models
2. It focuses on gradient differences in response probabilities for all observations, not just those where groups make different choices
3. It provides researchers with a more detailed picture of the (types of) data that are driving larger historical trends observable through traditional interaction models, and
4. It establishes a more direct link between the quantitative and qualitative facets of diachronic linguistics by providing a quantitatively robust method for homing in on the important and/or understudied subsets of data
5. Conceptually

Future directions to explore:

- Use of the method for studies spanning multiple time periods. It's not obvious how to calculate, and model, deviation scores from models of several (ordered) time periods, e.g. $R_b - R_a$, $R_c - R_b$, ... How do we model multiple deviation scores as the outcome in the final stage of the analysis? Multivariate/multi-response regression is a possibility, but not commonly used and more complicated to interpret (worth the effort?).
- Sociolinguistically more salient variables, e.g. copula deletion, *-ing*/*-in'*, and many other morphological, phonological, or phonetic variables.
- Other dimensions of synchronic variation, e.g. native vs. non-native, regional variation
- ...

Contact info

Jason Grafmiller
`jason.grafmiller@kuleuven.be`

Quantitative Lexicology and Variational Linguistics (QLVL)
KU Leuven – University of Leuven

References

[1] Stefan Th. Gries and Allison S. Adelman. Subject realization in Japanese conversation by native and non-native speakers: Exemplifying a new paradigm for learner corpus research. In Jesús Romero-Trillo, editor, *Yearbook of Corpus Linguistics and Pragmatics 2014*, volume 2, pages 35–54. Springer International Publishing, Cham, 2014.

[2] Stefan Th. Gries and Sandra C. Deshors. Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora*, 9(1):109–136, 2014.

[3] Stefan Th. Gries and Martin Hilpert. Modeling diachronic change in the third person singular: a multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics*, 14(03):293–320, November 2010.

[4] Scott Grimm and Joan Bresnan. Spatiotemporal variation in the dative alternation: a study of four corpora of British and American English. In *Grammar & Corpora 2009*, Mannheim, Germany, 2009. September.

[5] Lars Hinrichs and Benedikt Szmrecsanyi. Recent changes in the function and frequency of standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics*, 11:437–474, 2007.

[6] Lars Hinrichs, Benedikt Szmrecsanyi, and Axel Bohmann. *Which*-hunting and the Standard English relative clause. *Language*, to appear.

[7] Anette Rosenbach. Emerging variation: determiner genitives and noun modifiers in English. *English Language and Linguistics*, 11:143–189, 2007.

[8] Christoph Wolk, Joan Bresnan, Anette Rosenbach, and Benedikt Szmrecsányi. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica*, 30:382–419, 2013.