Exploring probabilistic grammar(s) in varieties of English around the world

Jason Grafmiller, Benedikt Heller, Melanie Röthlisberger & Benedikt Szmrecsanyi (KU Leuven)

Background

A 5-year project (2013-2018) founded by the FWO, grant # G.0C59.13N (PI: Szmrecsanyi).

- Offers a usage-based emphasis on variation as a "core explanandum" by synthesizing two hitherto rather disjoint lines of research into one project with a coherent empirical and theoretical focus.

English in global context

Research on postcolonial varieties of English (VoEs) examines the scope and parameters of variation in English around the world.

- Explores the extent to which features of different VoEs are shaped by the communicative needs of their speakers. (e.g. Schneider 2007)

Probabilistic grammar

Probabilistic grammar frameworks explore the hidden—though cognitively 'real'—probabilistic constraints on grammatical variation.

- Syntactic variation and change is subtle, gradient & probabilistic rather than categorical in nature
- Linguistic knowledge includes knowledge of probabilities which provides speakers with powerful—though mostly implicit—predictive capacities (Gahl and Garnsey 2006; Bresnan and Ford 2010)

Research Questions

- To what extent do VoEs share, or not share, a core probabilistic grammar that can explain cross-lectal patterns?
- Are lectal differences random, or can they be explained by considering sociohistorical factors?
- To what extent do corpus-derived probabilities reflect the linguistic knowledge possessed by speakers of a community?

Data

ICE corpora tapped into

In addition to the ICE corpora, we tap into data from ***GloWbE*** (Corpus of global web-based English, cf. Davies 2013).

Syntactic alternations studied

We explore the probabilistic influence of various features on users' choices in three syntactic alternations common to all varieties of English. Recent work finds evidence for regional and register variation in the influence of certain features in each of these constructions.

1. ***Genitive alternation*** (Investigator: B. Heller):
   the senator's brother ~ the brother of the senator
2. ***Dative alternation*** (Investigator: M. Röthlisberger):
   send them a letter ~ send a letter to them
3. ***Particle placement*** (Investigator: J. Grafmiller):
   pick the book up ~ pick up the book

Extracting corpus data

Extraction and selection of tokens for each dataset proceeds in several stages.

1. Possible tokens identified automatically using the CLAWS 7 tagged version of the nine ICE corpora. Accuracy (precision and recall) of scripts is assessed and refined using the manually parsed ICE-GB as baseline.
2. After initial extraction, non-interchangeable tokens are automatically filtered out where possible.
3. Resulting datasets are manually filtered using html-based tools developed for rapid editing. Criteria for inclusion/exclusion of tokens follow methods laid out in previous literature (Rosenbach 2002; Bresnan et al. 2007)


Dismissed of-genitives

Key / Function

X Move highlighted dataset to *Genitive*
C Move highlighted dataset to *Undecided*
Q Move highlighted dataset to *Genitive* and flag it as *questionable*
← Move to the previous instance
→ Move to the following instance

Number of instances: 492

No.	Unit marker	Context left	Node	Context right	Constraint	Violation
			w2e-001.pco			
1	w993:1>	[...] had no need , said Aelius Aristides of 23 -	of	garrisons in the cities	general	general
2	w1100:1>	the upper classes were privileged , enjoying a sort	of	benefit of clergy 31	word left	sort
3	w1100:1>	[...] upper classes were privileged , enjoying a sort of benefit	of	clergy 31	np left	a sort of benefit
			w2e-011.pco			
4	w98:1>	In this chapter we shall develop some	of	the issues dealt with in Chapter	word left	some
5	w923:1>	[...] changes in markets and technologies to turn imply a variety	of	organisational effects , and therefore effects on managers	np left	a variety
6	w925:1>	[...] increased living off and sub-contracting to smaller firms represents one	of	the major organisational changes taking place in the manufacturing sector [...]	word left	one
7	w926:1>	There are thus a number	of	discernible trends towards decentralisation , reintegration of skills , smaller [...]	np left	a number
8	w929:1>	[...] 1988 has pointed out , the actions of management sort	of	market forces are not the same thing	word left	and
9	w931:1>	[...] significant factor even where new technology is introduced as part	of	an integrated strategic approach	word left	part
10	w933:1>	[...] important to emphasise that management attitudes are crucial in terms	of	whether or how such changes take place	np right	missing
11	w939:1>	[...] information is more open , and there is a blurring	of	organisational barriers	np left	a blurring
12	w943:1>	[...] of team work are likely to become an enduring feature	of	organisation where new technology is concerned	np left	an enduring feature
13	w946:1>	A different scenario is	of	course possible	word right	course

Annotation

For each construction, numerous linguistic variables are coded, based on previous literature.

- Coding schema for common predictor variables are kept consistent across alternations.
  - Animacy (human ~ collective ~ temporal ~ locative ~ inanimate)
  - Definiteness (definite ~ indefinite ~ proper noun ~ def. pronoun)
  - Length (orthographic words and letters)
  - Information status (given ~ new)
  - Persistence (type of Cx last used; distance to last usage)
  - Thematicity (text frequency of head)
  - Lexical density of local context (type-token ratio)
  - Rhythmic structure
- Automated coding methods (Perl/Python scripts) are used wherever possible.
- For features requiring manual coding (e.g. animacy), inter-rater reliability tests are conducted (Cohen's/Fleiss' κ).

Statistical Analysis

Mixed-effects regression

Workhorse technique in corpus-based syntactic variation studies (e.g. Bresnan et al. 2007).

- Binary logistic regression probes the probabilistic effects of independent variables (a.k.a. constraints) on linguistic choice-making.
  - contextual (language-internal) factors (animacy, information status, end weight, structural priming, etc.)
  - language-external factors (genre, variety of English)
- Allows for control of multiple variables simultaneously, including effects of individual register/text/speaker variation (i.e. random effects).

Conditional inference trees & Random forests

Model syntactic choices using non-parametric, recursive partitioning methods, e.g. decision trees.

- ***random forests***: sets of trees calculated on random subsets of the data using randomly selected and permuted predictors for each split (Strobl et al. 2009)
- superior to standard methods (e.g. regression)
  - robust to effects of multicollinearity
  - better estimation of the contribution of individual predictors
  - more accurate predictions

Supplementary experiments

Participants are presented excerpts from actual corpora, and asked to rate the naturalness of alternative forms.

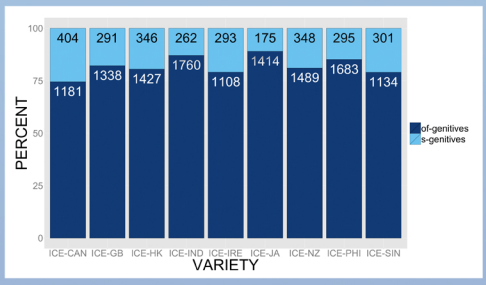
- Participants' responses are compared to the probabilities derived by the corpus model to determine whether participants' ratings are influenced by the predictors in the same manner as the production data from the corpus.
- Example excerpt:

```
I'm in college, and I'm only twenty-one but I had
a speech class last semester, and there was a
girl in my class who did a speech on home care of
the elderly. And I was so surprised to hear how
many people, you know, the older people, are
like, fastened to their beds so they can't get
out just because, you know, they wander the
halls. And they get the wrong medicine, just
because, you know, the aides or whatever just
  (1) give them the wrong medicine [98 pts]
  (2) give the wrong medicine to them [2 pts]
```

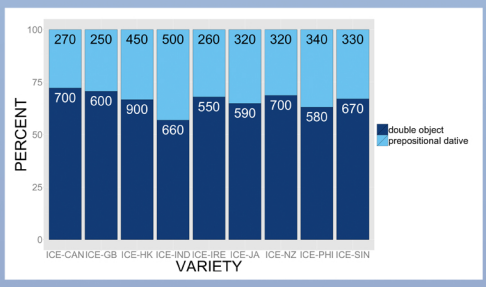
- In prior studies, participants "give ratings of naturalness of the alternative dative forms that turn out to be a function of the probabilities of occurrence and associated predictors found in corpus data" (Ford and Bresnan, 2013).

A frequency overview

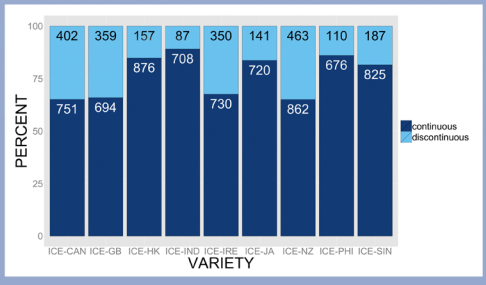
Genitive alternation



Dative alternation



Particle alternation



References

Bresnan, J., A. Cueni, T. Nikitina, and H. Baayen (2007). Predicting the dative alternation. In G. Boume, I. Kraemer, and Z. Zwarts (Eds.), *Cognitive Foundations of Interpretation*, pp. 69–94. Amsterdam: Royal Netherlands Academy of Science.
Bresnan, J. and M. Ford (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language 86*(1), 186–213.
Davies, M. (2013). Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries. Available online at **http://corpus2.byu.edu/glowbe/**.
Ford, M. and J. Bresnan (2013). Studying syntactic variation using convergent evidence from psycholinguistics and usage. In M. Krug and J. Schlüter (Eds.), *Research Methods in Language Variation and Change*. Cambridge: Cambridge University Press.
Gahl, S. and S. M. Garnsey (2006). Knowledge of grammar includes knowledge of syntactic probabilities. *Language 82* (2), 405410.
Rosenbach, A. (2002). *Genitive Variation in English*. Berlin, New York: Mouton de Gruyter.
Schneider, E. (2007). *Postcolonial English: Varieties Around the World*. Cambridge, New York: Cambridge University Press.
Strobl, C., J. Malley, and G. Tutz (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods 14*(4), 323–348.

KU LEUVEN QLVL fwo Research Foundation Flanders Opening new horizons

ISLE 3 Conference, Zurich, 25/08/2014

For further information/contact please visit
http://wwwling.arts.kuleuven.be/qlvl/ProbGrammarEnglish.html