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

Jason Grafmiller, Benedikt Heller, Melanie Röthlisberger & Benedikt Szmrecsanyi

KU Leuven
Quantitative Lexicology and Variational Linguistics

Project summary

- 5-year project (2013-2018), funded by the Odysseus program
  (FWO grant # G.0C59.13N)
- marries the spirit of the Probabilistic Grammar framework (➪ grammatical knowledge is experience-based & probabilistic) to research along the lines of the English World-Wide paradigm
  (➪ sociolinguistics of E-speaking communities)
- innovative potential: synthesizing two hitherto rather disjoint lines of research into one unifying project with a coherent empirical and theoretical focus
- usage-based interest in variation as a "core explanandum" (Adger and Trousdale 2007: 274)

The English World-Wide paradigm

- wide range of postcolonial VoE
  - native mother-tongue (L1) varieties (e.g. New Zealand E)
  - non-native indigenized second-language (L2) varieties (e.g. Hong Kong E)
  - so-called "language-shift" varieties (e.g. Irish E)
- topics: scope, limits, parameters of variation; extent to which structural make-up of VoE can be predicted by communicative needs of colonizers/colonized
  (e.g. Schneider 2007)
- shortcoming: an often primarily descriptive focus on the variable usage frequencies (presence/absence) of linguistic features

The Probabilistic Grammar framework

- explores hidden – though cognitively 'real' – probabilistic constraints on grammatical variation.
- Two crucial assumptions:

1. syntactic variation – and change – is **subtle, gradient & probabilistic** rather than categorical in nature

   (Labov 1982; Bresnan and Hay 2008)

2. linguistic knowledge includes **knowledge of probabilities**, and speakers have powerful predictive capacities

   (Gahl and Garnsey 2006; Gahl and Yu 2006)

Methodological sketch

1. explore a number of well-known alternations in the grammar of English

genitive alternation	dative alternation
particle placement	non-finite/finite complementation

2. create richly annotated datasets
3. use multivariate statistical techniques to model the interplay of probabilistic factors constraining the alternations; check whether there are significant differences between VoE
4. conduct supplementary rating-task experiments

Research questions

- to what extent does speakers' grammatical knowledge vary across VoE?
- six more specific research questions:
  - RQs 1, 2, and 3 adopt a variety-centered (lectal) perspective
  - RQ 4 adopts an alternation-centered perspective
  - RQ 5 adopts a constraint-based perspective
  - RQ 6 concerns the fundamental hypothesis fueling the project

    ➪ are corpus-derived probabilities merely facts about distributions, or are they reflections of the linguistic knowledge possessed by speakers of a community?

RQ 1: core grammar

What is the extent to which VoE share, or do not share, a core grammar that is explanatory across different varieties?

Issue:

- one grammar versus $n > 1$ grammars
  (e.g. Bernaisch et al. to appear)

RQ2: sociohistory

Are lectal differences random, or can they be explained by considering sociohistorical factors?

Issues:

- non-randomness
  ➪ emergence of certain (internally somewhat homogeneous) variety clusters
- some relevant sociohistorical factors:
  1. distinction between L1 varieties, language-shift varieties, and L2 varieties (e.g. Trudgill 2009)
  2. stages in Schneider's (2007) Dynamic Model
  3. attraction to particular varieties (e.g. BrE, AmE, . . .)
  4. substrate effects (e.g. De Cuypere and Verbeke 2013)

RQ3: registers and idiolects

To which extent do register and idiolect differences play a possibly differential role across VoE?

Issue:

- is register and idiolect variation similarly important across VoE?
  ➪ important implications for our understanding of linguistic systematicity (Geeraerts 2010)

RQ4: alternations

To what degree do the grammatical alternations under study exhibit cross-constructional parallelisms?

Issue:

- extent to which grammar is essentially a random collection of independently constrained alternations and/or constructions (Goldberg 2003)

RQ5: constraints

Which of the probabilistic constraints are sociohistorically and/or culturally malleable?

Hypotheses:

- certain constraints, e.g. animacy or verb semantics, are particularly prone to cultural (re-)interpretation across VoE
- other constraints, e.g. end-weight or syntactic priming, are presumably rooted in the architecture of the human speech production system (MacDonald 2013) ➪ stability

RQ6: corpus data vs. experimental data

What is the extent to which corpus-derived probabilistic grammars match up with rating-task experiments?

Hypothesis:

- intuitive judgments about alternation choices should align with corpus-derived probabilities

  ➪ recent work reveals a tight match (Bresnan 2007; Bresnan and Ford 2010)

Corpus material

- The International Corpus of English (ICE)
  - 1 million words of text per ICE sub-corpus
  - 60% of the material spoken, 40% written
  - 12 major text types (e.g. conversation, reportage, student writing, letters)
- ICE9 package:

GB	Canada	NZ	Ireland	
India	Jamaica	HK	Philippines	Singapore

- UZH Dependency Bank 2.0, based on the Pro3Gres Annotation scheme (Schneider 2008)

Work packages 1–3

Three well-studied alternations in English:

1. genitive alternation: *the senator's brother* ∼ *the brother of the senator*
2. dative alternation: *send them a letter* ∼ *send a letter to them*
3. particle placement: *pick the book up* ∼ *pick up the book*

- numerous shared constraints (end-weight, animacy, priming, info status, . . .)
- some evidence for regional differences in all 3 (Hinrichs and Szmrecsanyi 2007; Bresnan and Hay 2008; Haddican and Johnson 2012)

Work package 4: complementation

(1) a. I don't regret$_{CTP}$ [helping her start out]$_{CC}$
(non-finite complementation)
b. I don't regret$_{CTP}$ [that I helped her start out]$_{CC}$
(finite complementation)

- a relatively understudied phenomenon
- Cuyckens, D'hoedt and Sz (2014): first-ever probabilistic analysis, albeit with a focus on historical variation
- regional variation??

Work package 5: A rating experiment

- Bresnan (2007); Bresnan and Ford (2010): regression models match probabilistic intuitions
- converging evidence for psychological reality of the experience-based probabilistic grammars?
  (work packages 1 through 4)
- replicates Experiment 1 in Bresnan (2007: 76-84):
  - recruit $\approx$ 50 native-speaker subjects from different VoE backgrounds
  - subjects rate randomly sampled observations from the corpus database
  - do subjects' ratings match probabilities predicted by the corpus models?

Hundt and Szmrecsanyi (2012)

- study of the genitive alternation in 19th/early 20th century New Zealand English vis-à-vis BrE (building on Bresnan and Hay 2008)

➪ possessor animacy is hugely more important in early NZE than in BrE

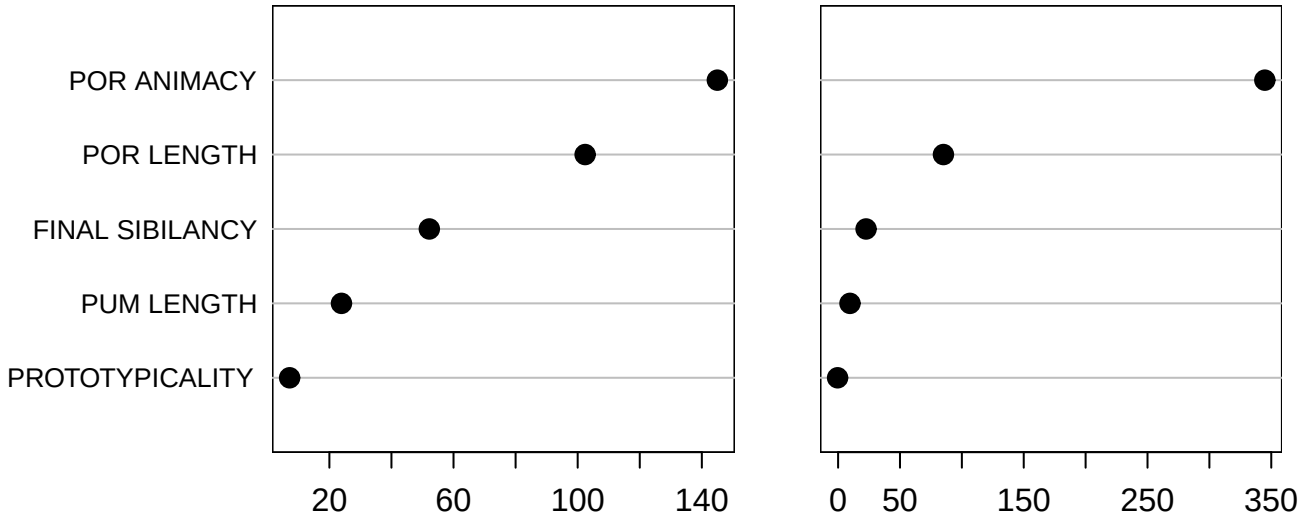

Figure: Importance of factors in model: increase in AIC if factor removed. Left: ARCHER (BrE), right: CENZE (NZE)

Wolk, Bresnan, Rosenbach and Sz. (2013)

- study of the dative alternation in LateModE in USA and UK
- in AmE, increasing theme length disfavors the prepositional dative more robustly than in BrE

➪ end-weight more important in AmE

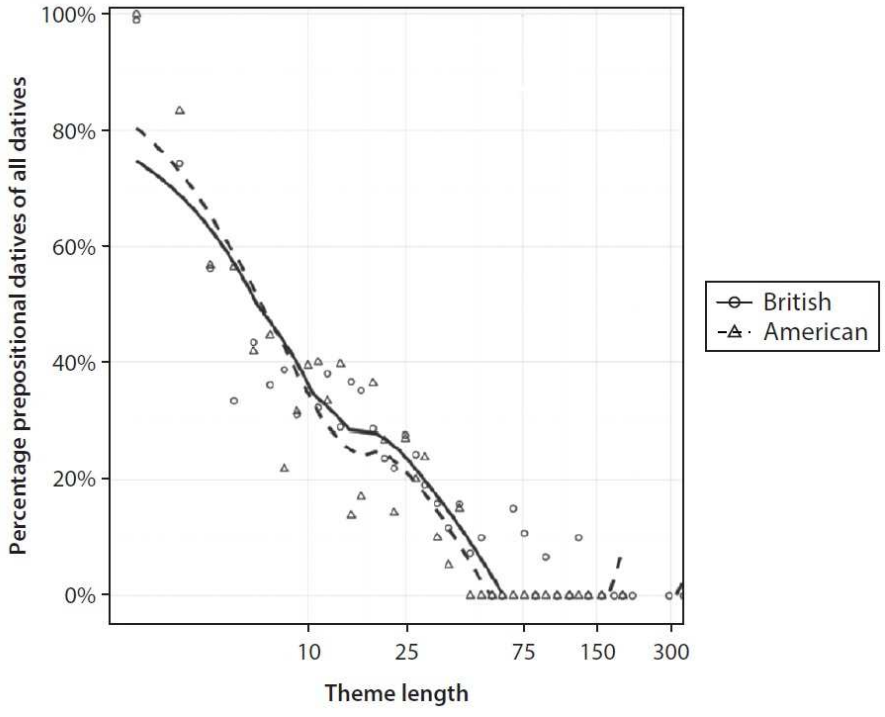

Figure: : Percentage of PP datives (y-axis) as a function of theme length (x-axis) and variety

Innovative potential

- emphasize probabilistic, usage- and experience-based nature of linguistic variation
- assume that language users implicitly learn the probabilistic effects of constraints on variation by constantly (re-)assessing input throughout their lifetimes
- combine a variationist interest in probabilistic modeling with a sociolinguistic/cognitive-linguistic interest in socially contextualized language usage
- bridge gaps between different strands of theoretically oriented usage-based linguistics

Extensions

- more varieties of English
- the catalogue of alternations to be analyzed is open-ended
- not in principle restricted to syntactic variables; morphological or phonological variation may be addressed at later stages

Team members

Jason Grafmiller
PhD, 2013, Stanford University

Daniel-Benedikt Heller
MA, 2013, University of Giessen

Melanie Röthlisberger
MA, 2011, University of Zurich

Thank you!

http://wwwling.arts.kuleuven.be/qlvl/ProbGrammarEnglish.html

References I

Adger, D. and G. Trousdale (2007). Variation in English syntax: Theoretical implications. *English Language and Linguistics 11*, 261–278.

Bernaisch, T., S. T. Gries, and J. Mukherjee (to appear). The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide*.

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston and W. Sternefeld (Eds.), *Roots: Linguistics in search of its evidential base*, pp. 7596. Berlin, New York: Mouton de Gruyter.

Bresnan, J. and M. Ford (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language 86*(1), 186213.

Bresnan, J. and J. Hay (2008). Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua 118*(2), 245259.

Cuyckens, H., F. D'hoedt, and B. Szmrecsanyi (2014). Variability in verb complementation in Late Modern English: finite vs. non-finite patterns. In M. Hundt (Ed.), *Late Modern English Syntax*. Cambridge: Cambridge University Press.

De Cuypere, L. and S. Verbeke (2013). Dative alternation in indian english: A corpus-based analysis. *World Englishes 32*, 169–184.

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.

Gahl, S. and A. C. Yu (2006). *Special theme issue: Exemplar-based models in linguistics*. The linguistic review. Mouton de Gruyter.

References II

Geeraerts, D. (2010). Schmidt redux: How systematic is the linguistic system if variation is rampant? In H. H. Hock and E. Engberg-Pedersen (Eds.), *Language Usage and Language Structure*, Volume 213, pp. 237–262. Berlin, New York: De Gruyter Mouton.

Goldberg, A. E. (2003). *Constructions: a construction grammar approach to argument structure*. Chicago: Univ. of Chicago Press.

Haddican, B. and D. E. Johnson (2012). Effects on the particle verb alternation across English dialects. In *University of Pennsylvania Working Papers in Linguistics 18*, pp. 31–40. University of Pennsylvania.

Hinrichs, L. and B. Szmrecsanyi (2007). Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics 11*(3), 437–474.

Hundt, M. and B. Szmrecsanyi (2012). Animacy in early New Zealand English. *English World-Wide 33*, 241–263.

Labov, W. (1982). Building on empirical foundations. In W. Lehmann and Y. Malkiel (Eds.), *Perspectives on Historical Linguistics*, pp. 17–92. Amsterdam, Philadelphia: Benjamins.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology 4*, 1–16.

Schneider, E. (2007). *Postcolonial English: Varieties Around the World*. Cambridge, New York: Cambridge University Press.

Schneider, G. (2008). *Hybrid Long-Distance Functional Dependency Parsing*. Ph.D. Thesis, University of Zurich, Zurich.

Trudgill, P. (2009). Vernacular universals and the sociolinguistic typology of English dialects. In M. Filppula, J. Klemola, and H. Paulasto (Eds.), *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*, pp. 302–329. London: Routledge.

Wolk, C., J. Bresnan, A. Rosenbach, and B. Szmrecsanyi (2013). Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica 30*(3), 382–419.

Zipp, L. and T. Bernaisch (2012). Particle verbs across first and second language varieties of English. In M. Hundt and U. Gut (Eds.), *Mapping Unity and Diversity World-Wide: Corpus-Based Studies of New Englishes*, pp. 167–196. Amsterdam: John Benjamins.

Regression analysis

- workhorse analysis technique in corpus-based variation studies
- logistic regression probes the probabilistic conditioning of linguistic choice-making
- predicts a binary outcome (i.e. a linguistic choice) given several independent predictor variables (a.k.a. constraints):
  - contextual (language-internal) factors (e.g. animacy of genitive possessors)
  - language-external factors (e.g. genre, variety of English)
- multivariate control

Corpus-derived dative model

Probability of the prepositional dative = $1 / 1 + e^{-(X\beta + u_i)}$
where

$$\hat{X\beta} = \begin{aligned} & 1.1583 \\ & -3.3718\{\text{pronominality of recipient = pronoun}\} \\ & +4.2391\{\text{pronominality of theme = pronoun}\} \\ & +0.5412\{\text{definiteness of recipient = indefinite}\} \\ & -1.5075\{\text{definiteness of theme = indefinite}\} \\ & +1.7397\{\text{animacy of recipient = inanimate}\} \\ & +0.4592\{\text{number of theme = plural}\} \\ & +0.5516\{\text{previous = prepositional}\} \\ & -0.2237\{\text{previous = none}\} \\ & +1.1819 \cdot [\log(\text{length(recipient)}) - \log(\text{length(theme)})] \end{aligned}$$

and $\hat{u}_i \sim N(0, 2.5246)$

Figure 1. The model formula for datives

(Ford and Bresnan 2013)

Bresnan's 100-split task

Using actual corpus examples, . . .

❝ . . . participants rate the naturalness of alternative forms as continuations of a context by distributing 100 points between the alternatives. Thus, for example, participants might give pairs of values to the alternatives like 25–75, 0–100, or 36–64. From such values, one can determine whether the participants give responses in line with the probabilities given by the model and whether people are influenced by the predictors in the same manner as the model. ❞

(Ford and Bresnan 2013)

The 100-split task: an example

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

(1) just give them the wrong medicine

(2) just give the wrong medicine to them

- ➪ the model suggests a 98–2 split in favor of the ditransitive in (1)
  (Ford and Bresnan 2013)