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 overview

- 5-year project (2013-2018), funded by the Research Foundation Flanders (FWO) (PI: Szmrecsanyi)
- marries the spirit of the Probabilistic Grammar framework (➪ grammar is experience-based & probabilistic) to research along the lines of the "English World-Wide" paradigm (➪ sociolinguistics of E-speaking communities)
- usage-based interest in variation as a "core explanandum" (Adger and Trousdale 2007: 274)
- innovative potential: synthesizing two hitherto rather disjoint lines of research into one project with a coherent empirical and theoretical focus

The English World-Wide paradigm

- wide range of postcolonial varieties of English (VoE)
  ➪ focus on English in a global context
- 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)
- often a 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; Bresnan and Ford 2010)

Big research questions

- to what extent do VoE share, or not share, a core grammar that is explanatory across varieties?
- are lectal differences random, or can they be explained by considering sociohistorical factors?
  - distinction between L1, language-shift, and L2 varieties (e.g. Trudgill 2009)
  - stages in Schneider's (2007) Dynamic Model
  - attraction to particular varieties (e.g. BrE, AmE, . . .)
  - substrate effects (e.g. De Cuypere and Verbeke 2013)
- do corpus-derived probabilities truly reflect the linguistic knowledge possessed by speakers of a community?

Methodological outline

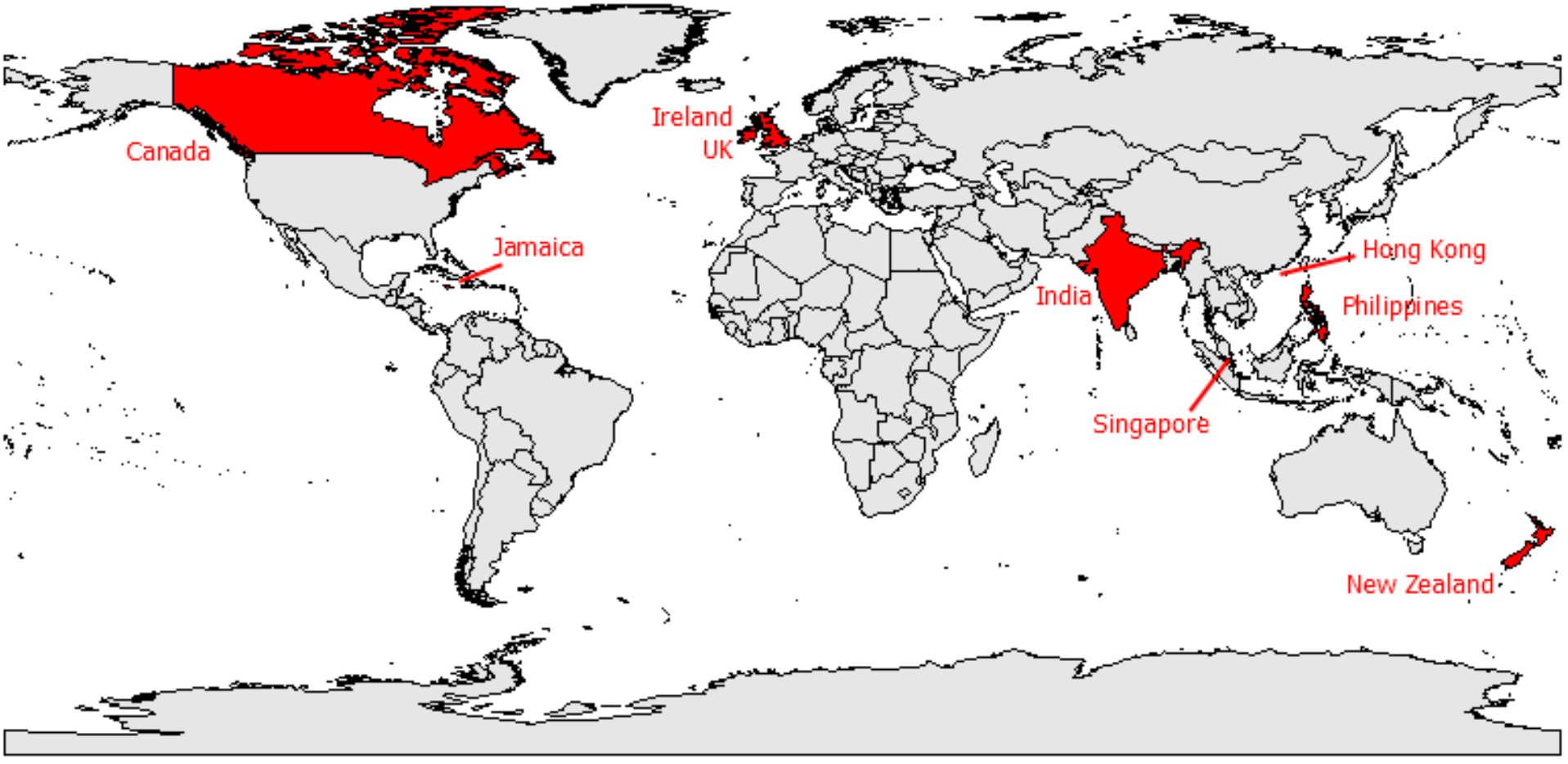
1. create richly annotated datasets using data from The International Corpus of English (ICE)

2. explore 4 common alternations in the grammar of English

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

3. use multivariate statistical and experimental techniques to model the interplay of probabilistic factors constraining the alternations both within and across VoE

9 ICE Varieties



Syntactic alternations

Three well-studied alternations in English:

1. genitive alternation: *the senator's brother* ∼ *the brother of the senator* (B. Heller)
2. dative alternation: *send them a letter* ∼ *send a letter to them* (M. Röthlisberger)
3. particle placement: *pick the book up* ∼ *pick up the book* (J. Grafmiller)

- 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)

Syntactic alternations, cont.

4. finite/non-finite complementation
   (B. Szmrecsanyi)

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

b. I don't regret$_{CTP}$ [helping her start out]$_{CC}$
(non-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??

Supplementary experiments

- Bresnan (2007); Bresnan and Ford (2010): regression models match probabilistic intuitions
- converging evidence for psychological reality of the experience-based probabilistic grammars?
- replicate Experiment 1 in Bresnan (2007: 76-84):
  - recruit 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?

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 ➪ ICE is continually expanding...
- 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

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 (2014). The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide 35*, 7–31.

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. 75–96. 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), 186–213.

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), 245–259.

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.

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.

References II

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.

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

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)