Contrastive Probabilistic Grammar

Benedikt Szmrecsanyi

KU Leuven
Quantitative Lexicology and Variational Linguistics

slides @ http://www.benszm.net/cpg.pdf

Introduction

Today

- contrastive: comparison of variation patterns in 9 geographic/synchronic varieties of English
- probabilistic: focus on the gradient effects that language-internal, contextual predictors have on variation patterns
- grammar: investigate 3 well-known syntactic alternations in the grammar of English

Research context

- project "Exploring probabilistic grammar(s) in varieties of English around the world" @ KU Leuven
  (see Szmrecsanyi, Grafmiller, Heller, and Röthlisberger to appear)
- crossroads of research on dialectology, English as a World Language, sociolinguistics, variationist linguistics & usage-based theoretical linguistics
- synthesize disjoint lines of scholarship into one unifying project with a coherent empirical and theoretical focus

Today

1. Introduction
2. Theoretical and methodological framework
3. Methods & data
4. Variation in contrast
5. Conclusion & outlook

Theoretical and methodological framework

The "English World-Wide Paradigm"

- wide range of postcolonial varieties (➪ "gold mine")
- topics: scope, limits, parameters of variation; extent to which structural make-up of varieties of E can be predicted by communicative needs of colonizers/colonized (e.g. Kachru 1992; Schneider 2007; Mesthrie and Bhatt 2008)

The Probabilistic Grammar framework

rely on variation-centered, usage- and experience-based probabilistic grammar framework developed by Joan Bresnan and collaborators

(e.g. Bresnan 2007; Bresnan and Ford 2010; Wolk et al. 2013)

The Probabilistic Grammar framework

rely on variation-centered, usage- and experience-based probabilistic grammar framework developed by Joan Bresnan and collaborators

(e.g. Bresnan 2007; Bresnan and Ford 2010; Wolk et al. 2013)

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 2004; Gahl and Yu 2006)

Focus on linguistic variation

- adopt the variationist methodology and restrict attention to "alternate ways of saying 'the same' thing" (Labov 1972: 188)

Focus on linguistic variation

- adopt the variationist methodology and restrict attention to "alternate ways of saying 'the same' thing" (Labov 1972: 188)

(1)
a. We sent [the president]$_{recipient}$ [a letter]$_{theme}$
(the ditransitive dative)
b. We sent [a letter]$_{theme}$ to [the president]$_{recipient}$
(the prepositional dative)

- Bresnan, Cueni, Nikitina, and Baayen (2007) ("Predicting the dative alternation"): how do speakers and writers choose between variants?

Regression analysis

- probes the probabilistic conditioning of linguistic choice-making
- on the basis of annotated linguistic observations, investigates the role which constraints play
- checks whether predictors have significant effect; quantifies effect

Regression analysis

- probes stic choice
- on the ons, investi
- checks ect; quantifies effect

Bresnan et al. (2007)

the dative alternation in spoken AmE is constrained by no fewer that 10 probabilistic constraints

A dative model, based on usage data

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

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

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

Figure 1. The model formula for datives

(Ford and Bresnan 2013)

The 100-split task

"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

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) give the wrong medicine to them

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 wand
they get the wrong medicine, just bec
aides or whatever

(1) just give them the wrong medicine

(2) give the wrong medicine to them

Predictions

the model suggests a 98–2 split in favor of the ditransitive dative in (1) – speakers tend to agree!

Some interesting Probabilistic Grammar work

- Bresnan and Hay (2008):
  US-NZ differences
- de Marneffe, Grimm, Arnon, Kirby, and Bresnan (2012):
  development of probabilistic grammars in children
- Wolk, Bresnan, Rosenbach, and Szmrecsanyi (2013):
  real-time dynamics of probabilistic change
- Grafmiller (2014):
  register-induced probabilistic variation

Methods & Data

A methodological sketch of the Leuven project

1. Tap into the International Corpus of English (ICE) and the Corpus of Global Web-Based English (GloWbE) and identify syntactic variants.

A methodological sketch of the Leuven project

1. Tap into the International Corpus of English (ICE) and the Corpus of Global Web-Based English (GloWbE) and identify syntactic variants.
2. Create richly annotated datasets to model the way language users make syntactic choices.

A methodological sketch of the Leuven project

1. Tap into the International Corpus of English (ICE) and the Corpus of Global Web-Based English (GloWbE) and identify syntactic variants.
2. Create richly annotated datasets to model the way language users make syntactic choices.
3. Check if choice making differs as a function of variety, according to corpus data.

A methodological sketch of the Leuven project

1. Tap into the International Corpus of English (ICE) and the Corpus of Global Web-Based English (GloWbE) and identify syntactic variants.
2. Create richly annotated datasets to model the way language users make syntactic choices.
3. Check if choice making differs as a function of variety, according to corpus data.
4. Conduct supplementary rating-task experiments.

Some research questions

- Do the varieties of English under study share a core probabilistic grammar?
- Do the alternations under study differ in terms of their probabilistic sensitivity to variety effects?
- Which of the probabilistic constraints are malleable, which are stable?

Corpora

- The International Corpus of English (ICE)
  (Greenbaum 1991)
- The Corpus of Global Web-based English (GloWbE)
  (Davies and Fuchs 2015)

Nine varieties of English

British E
Canadian E
Irish E
New Zealand E
Hong Kong E
Indian E
Jamaican E
Philippine E
Singapore E

The genitive alternation

(2) a. [The Senator]$_{possessor}$'s [brother]$_{possessum}$
(the *s*-genitive)
b. [The brother]$_{possessum}$ of [the Senator]$_{possessor}$
(the *of*-genitive)

The dative alternation

(3) a. We sent [the president]$_{recipient}$ [a letter]$_{theme}$
(the ditransitive dative)

b. We sent [a letter]$_{theme}$ to [the president]$_{recipient}$
(the prepositional dative)

The particle placement alternation

(4)
a. The president looked$_{verb}$ [the word]$_{NP}$ up$_{particle}$ (V-DO-P)
b. The president looked$_{verb}$ up$_{particle}$ [the word]$_{NP}$ (V-P-DO)

Variation in contrast

Key findings

1. Core probabilistic grammars are stable.
2. Constraint strength is variable.
3. All alternations are not equal.

Core probabilistic grammars are stable

Effect directions are stable

- there clearly are **qualitative** generalizations – predictors tend to consistently favor/disfavor particular linguistic outcomes

  possible exception: theme concreteness in CanE

- e.g. wherever we look, longer constituents follow shorter constituents – a pattern that is known as the principle of end weight, and/or "Easy First" (MacDonald 2013)

Effect directions are stable

- there clearly are **qualitative** generalizations – predictors tend to consistently favor/disfavor particular linguistic outcomes

  possible exception: theme concreteness in CanE

- e.g. wherever we look, longer constituents follow shorter constituents – a pattern that is known as the principle of end weight, and/or "Easy First" (MacDonald 2013)

- note that in our view, weight effects are a part of grammar as well as symptomatic of processing demands – grammar and processing are not mutually exclusive

Constraint strength is variable

Variable constraint strengths

- quantitative differences between varieties with regard to the effect size of the constraints on variation:
  - **genitive alternation**:
    animacy, possessum weight, final sibilancy
  - **dative alternation**:
    recipient pronominality, theme concreteness, constituent weight
  - **particle placement alternation**:
    DO weight, presence directional PP

Particle placement: length effects are variable

(*look up* [*the difficult word*] vs *look* [*the difficult word*] *up*)

Particle placement: length effects are variable

(*look up* [*the difficult word*] vs *look* [*the difficult word*] *up*)

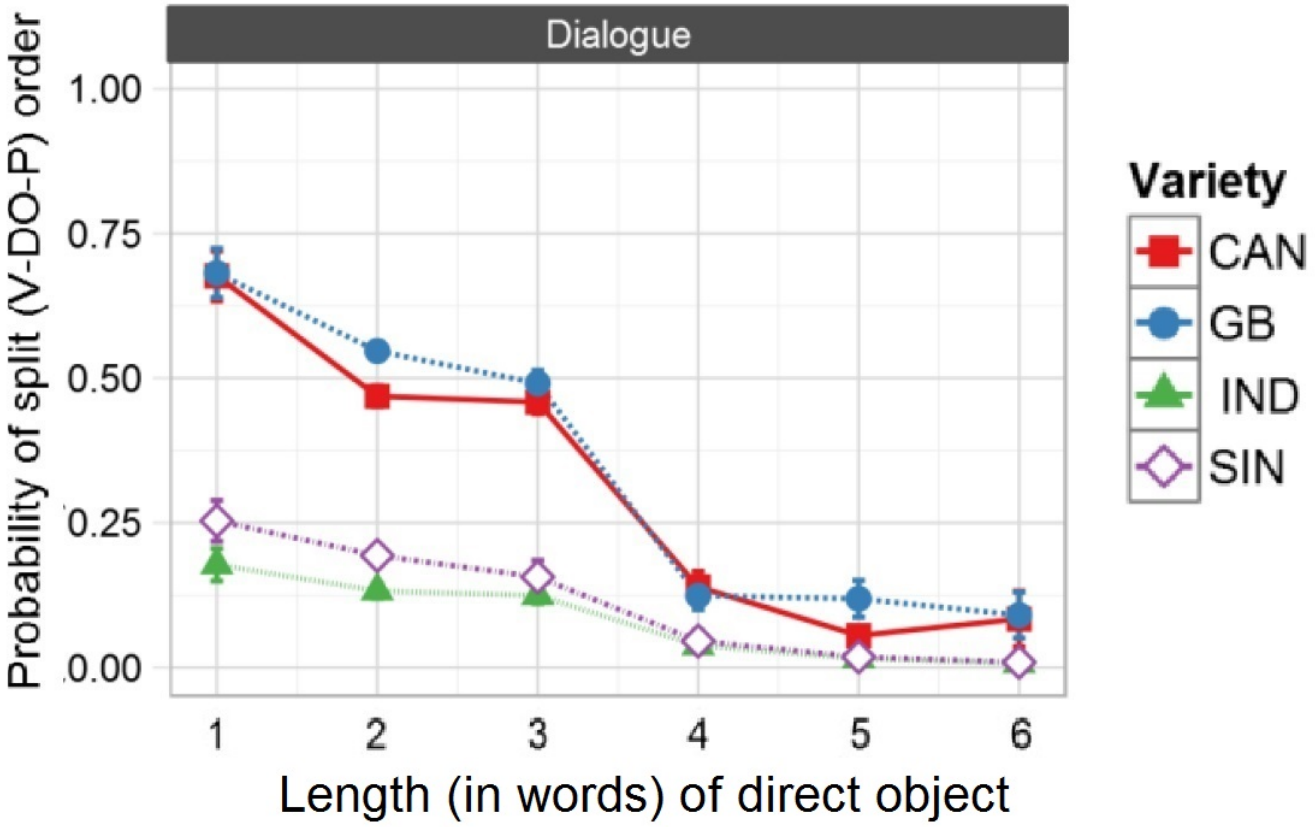


Figure: Predicted probabilities obtained from Conditional Random Forest model on corpus data (with 95% confidence intervals)

Particle e variable

(*look* rd] *up*)

A generalization

cross-variety differences only in contexts where neither alternate is more or less difficult to process

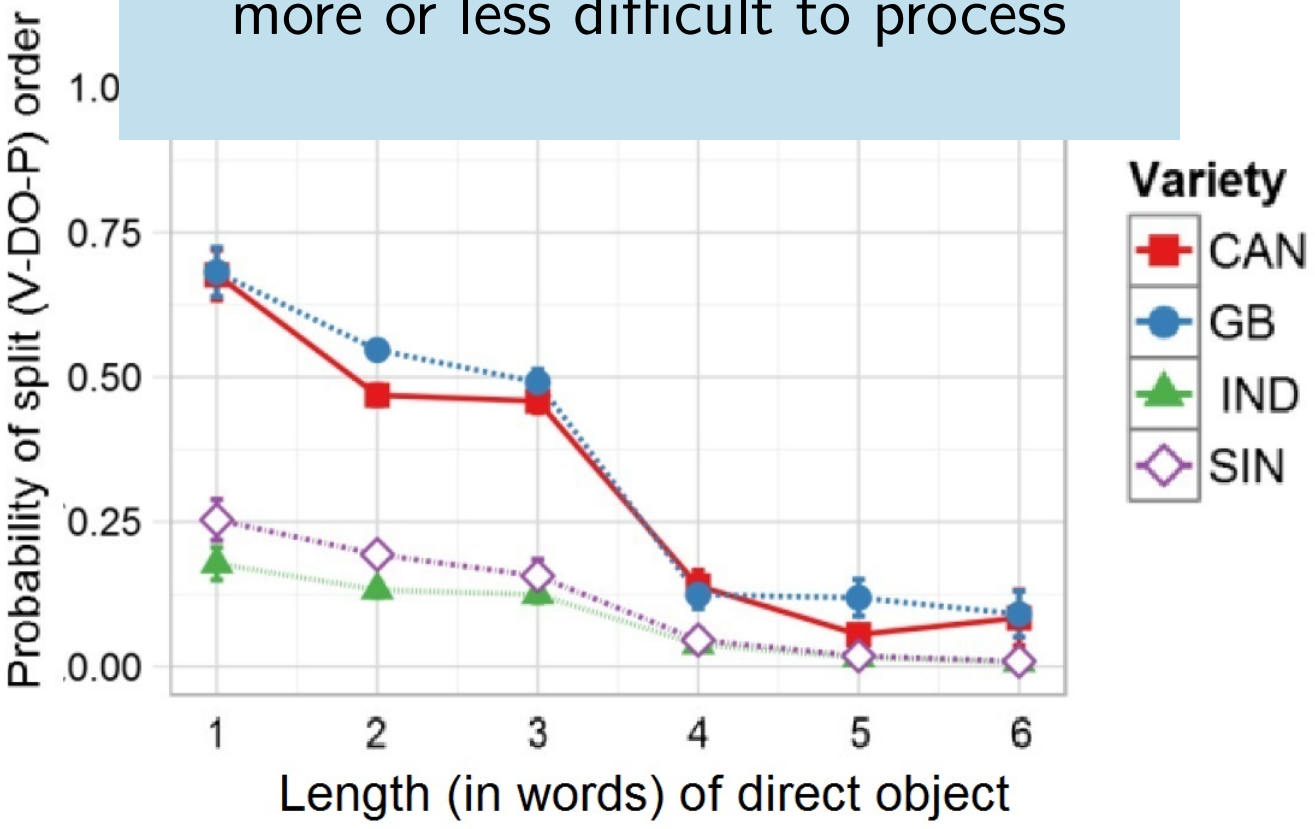


Figure: Predicted probabilities obtained from Conditional Random Forest model on corpus data (with 95% confidence intervals)

Particle placement: length effects are variable

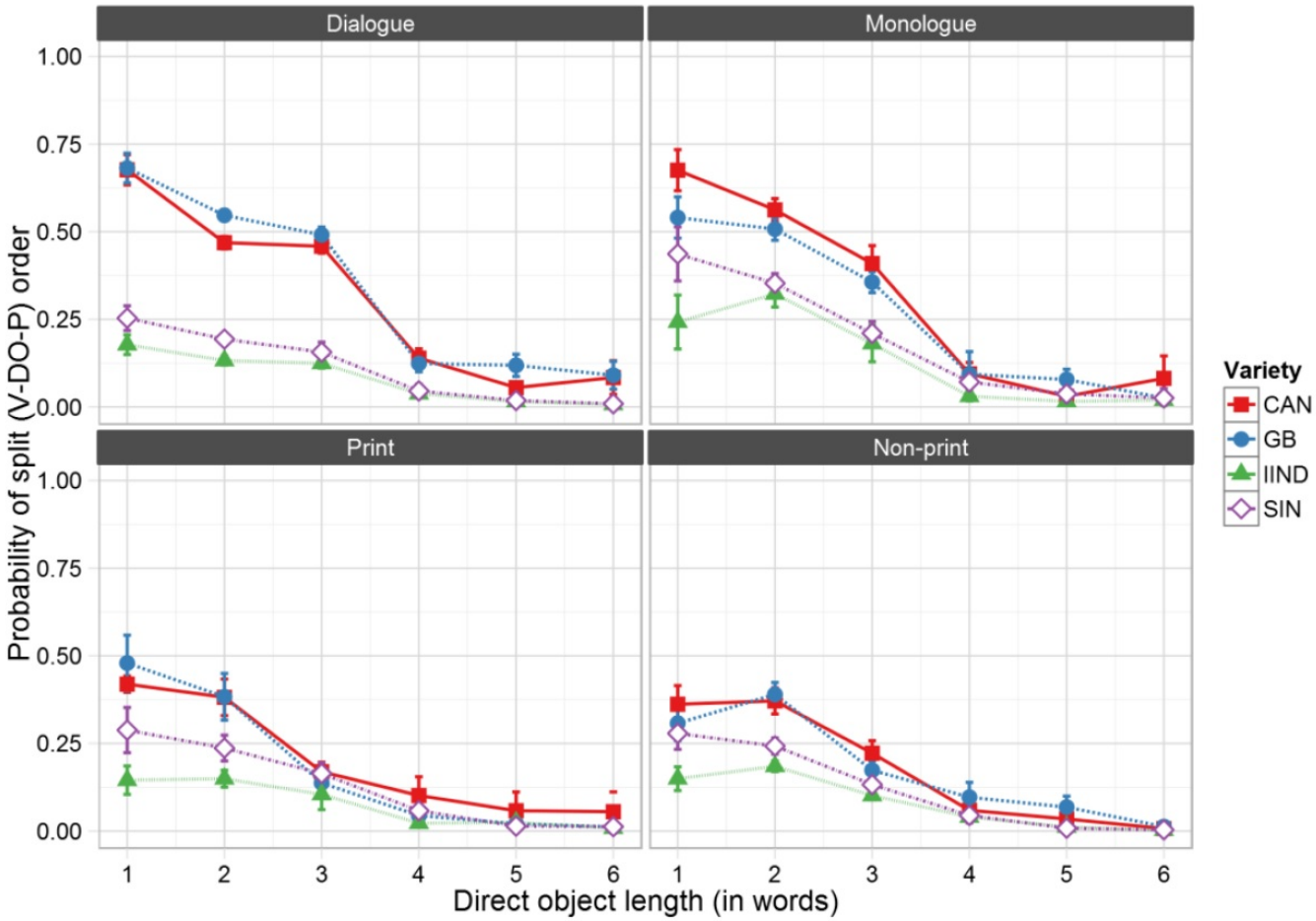


Figure: Predicted probabilities obtained from the Conditional Random Forest model (with 95% confidence intervals)

All alternations are not equal

"Probabilistic indigenization"

- the process whereby stochastic patterns of internal linguistic variation are reshaped by shifting usage frequencies in speakers of post-colonial varieties

"Probabilistic indigenization"

- the process whereby stochastic patterns of internal linguistic variation are reshaped by shifting usage frequencies in speakers of post-colonial varieties
- these patterns need not be consistent or stable (especially in the early stages of nativization), but they nonetheless reflect the emergence of a unique, region-specific grammar.

Probabilistic sensitivity to variety effects

The existence of a core grammar notwithstanding, the three alternations under study differ as to how amenable they are to probabilistic indigenization.

Genitives: predictor importance

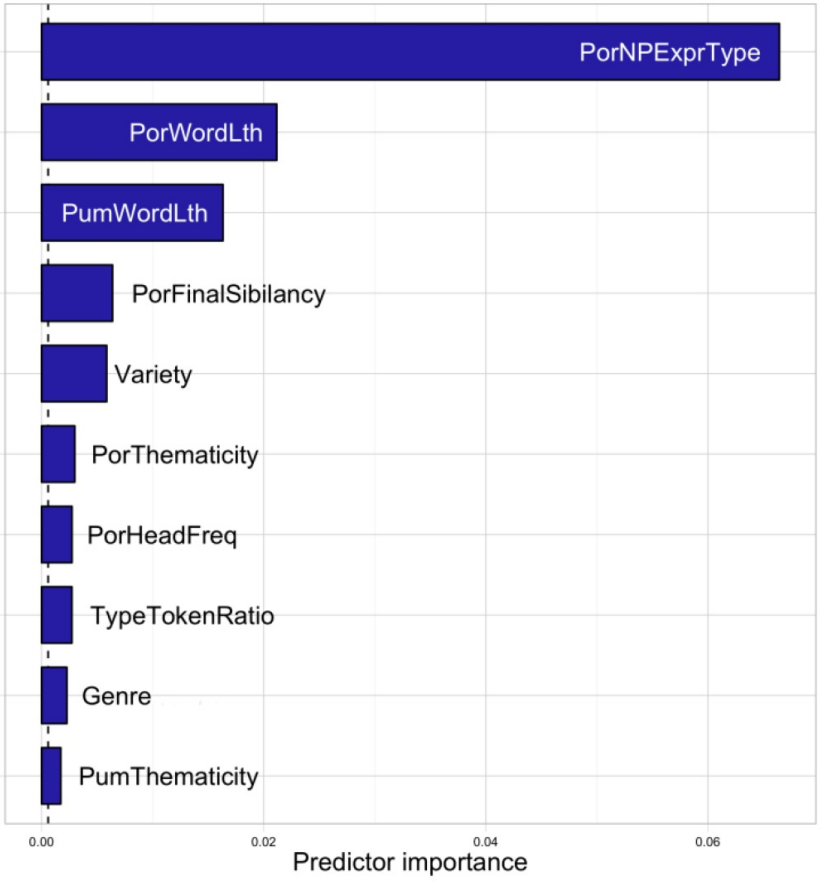


Figure: Predictor importance ranking for CRF analysis of genitive choice (displayed: 10 most important predictors). $C = 0.85$.

Datives: predictor importance

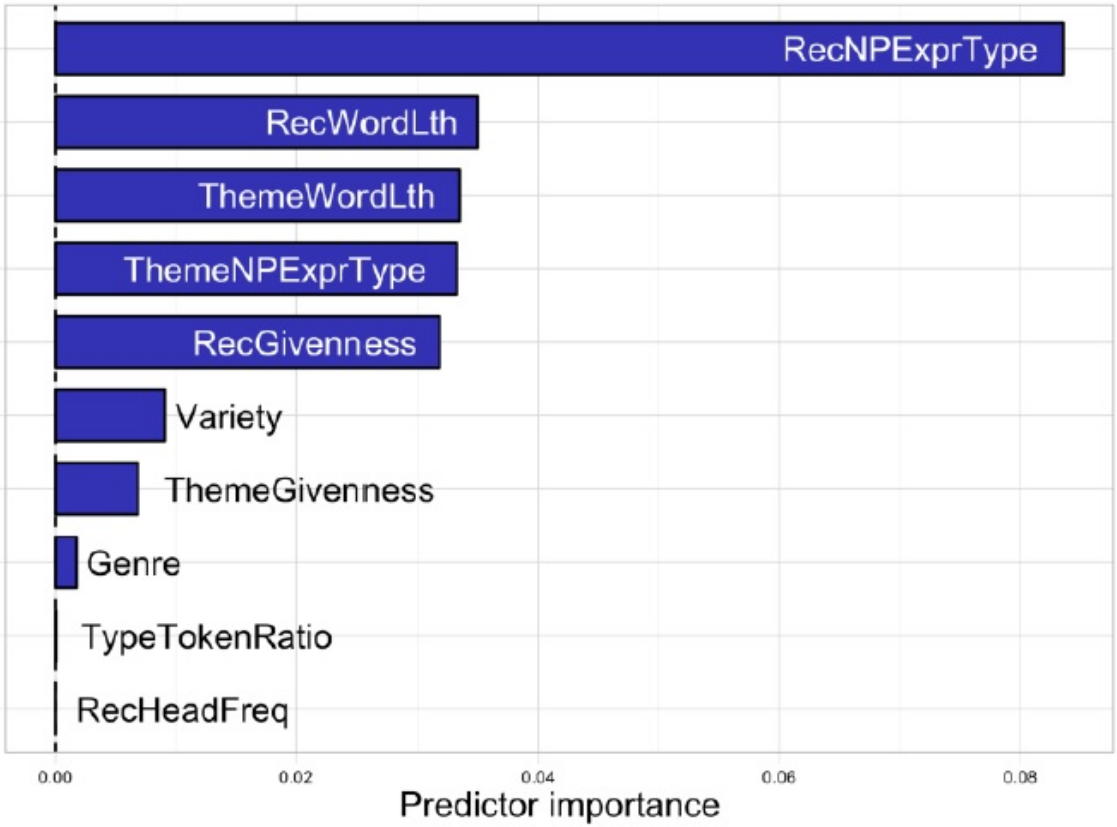


Figure: Predictor importance ranking for CRF analysis of dative choice (displayed: 10 most important predictors). $C = 0.93$.

Particle placement: predictor importance

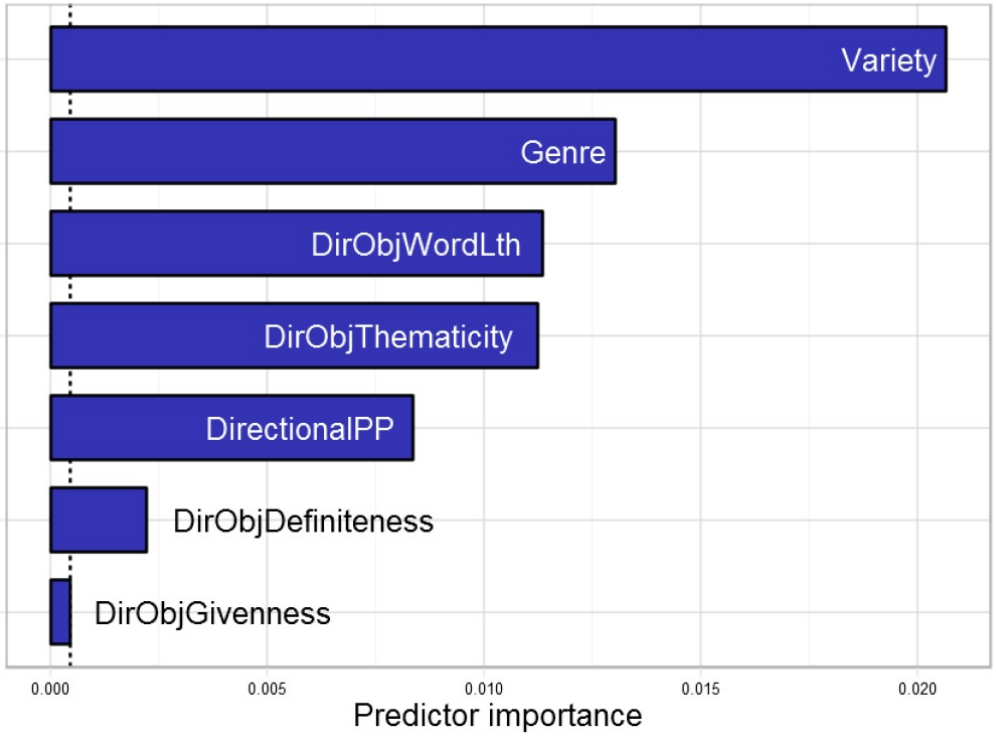


Figure: Predictor importance ranking for CRF analysis of particle placement. $C = 0.87$.

Lexis/grammar

- **ranking amenability to probabilistic indigenization**: particle placement $>$ datives $>$ genitives

Lexis/grammar

- **ranking amenability to probabilistic indigenization**: particle placement > datives > genitives
- Schneider (2003: 249): lexico-grammar is a prime target of early-stage indigenization

Lexis/grammar

- **ranking amenability to probabilistic indigenization**: particle placement > datives > genitives
- Schneider (2003: 249): lexico-grammar is a prime target of early-stage indigenization

  generalization: the more tightly associated a given syntactic alternation is with concrete instantiations involving specific lexical items – consider verb slots in the particle placement and dative alternation – the more likely it is to exhibit cross-varietal indigenization effects.

And by the way . . .

Variety differences are generally more important than register differences.

Conclusion & outlook

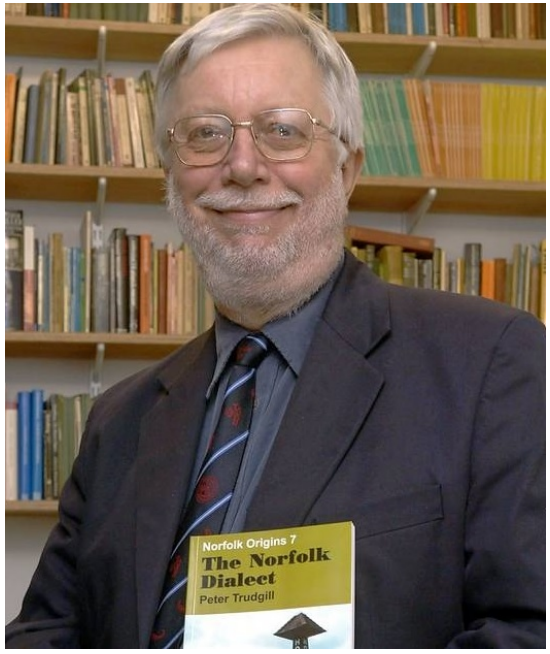
Philosophy

- variation is, or should be, a "core explanandum" (Adger and Trousdale 2007: 274) in linguistic theorizing
- combine a variationist interest in probabilistic modeling with a sociolinguistic interest in socially contextualized language usage
- language users implicitly learn the probabilistic effects of constraints on variation by constantly (re-)assessing input of spoken and written discourses throughout their lifetimes

Some work in progress . . .

- contrastive complexity analysis
- rule-based versus memory-based learning

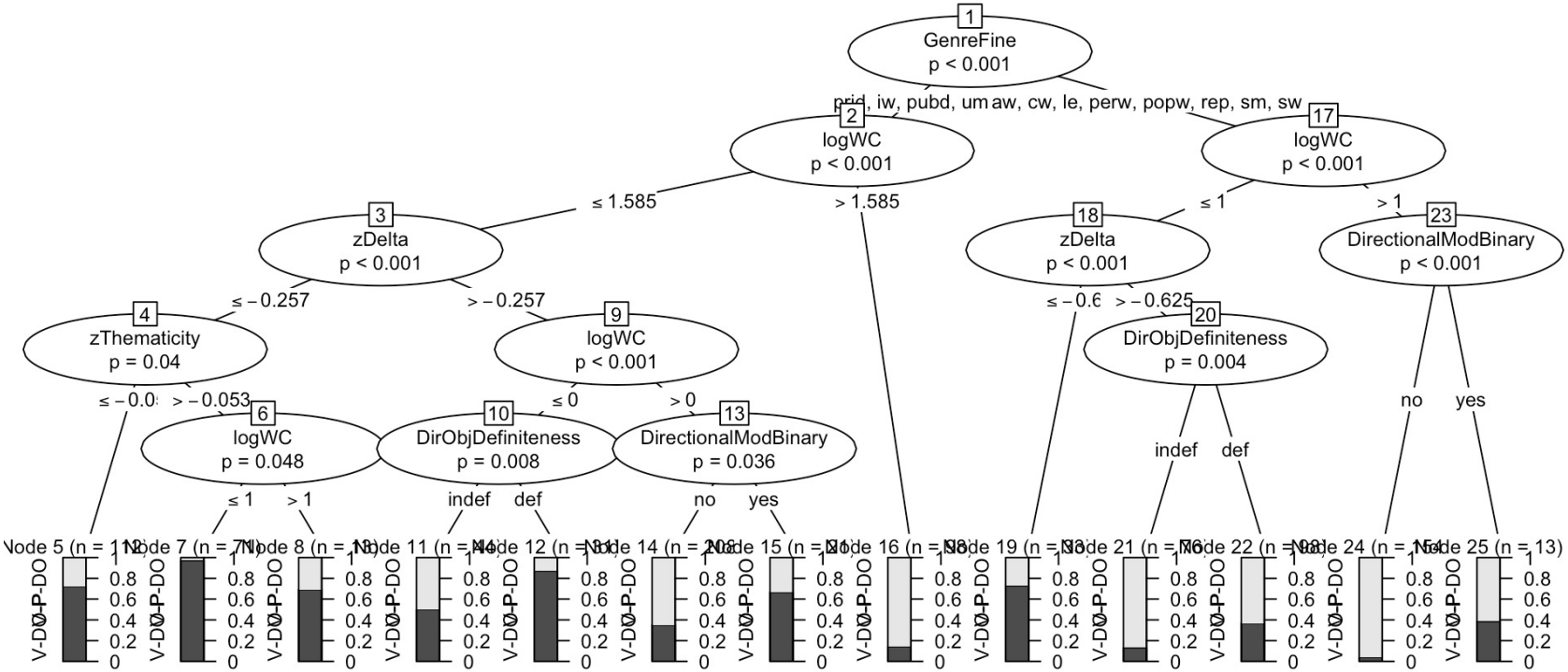
Defining variational complexity



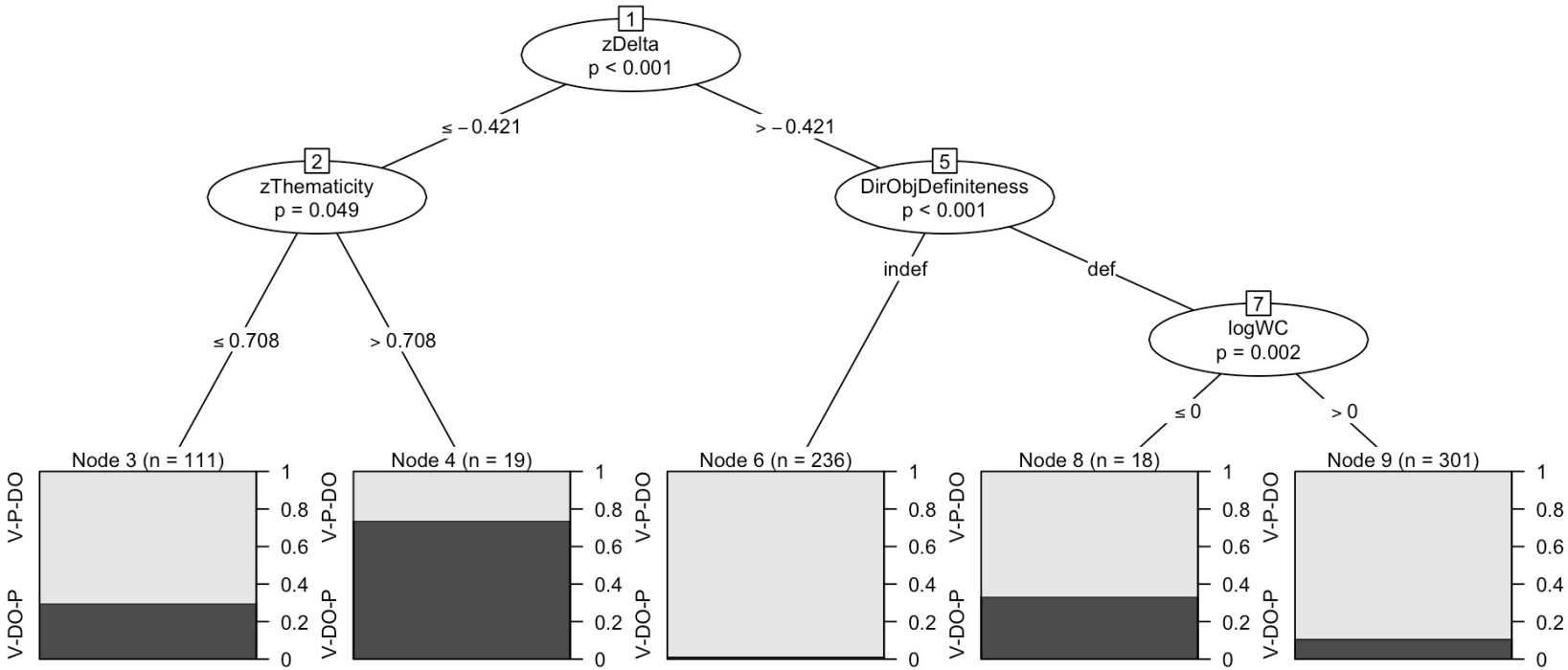
Peter Trudgill

- Trudgill (2011): contact, social instability, adult SLA
  ➪ simplification
- assumption: language or language variety A is more complex than language (variety) B to the extent that linguistic variation in A is more constrained than variation in B
  (see also Shin 2014: 3)

Particle placement in British English: ctree



Particle placement in Indian English: ctree



Regression modeling versus exemplar theory

(O'Reilly et al. 2013: Fig 2)

- regression analysis: rule-based technique drawing on researcher-defined "higher-level" abstract predictors

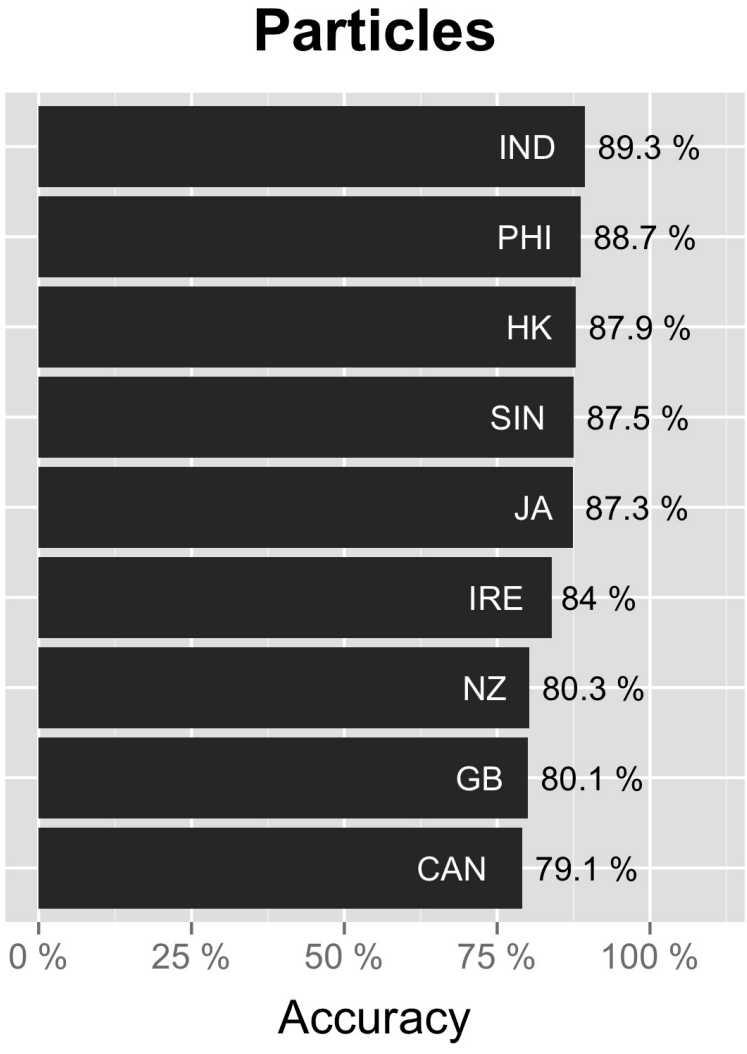
  (e.g. Bresnan et al. 2007)

- memory-based learning (MBL/TiMBL): predicts new instances of an alternation by "surfacy" extrapolation from the most similar cases in a training set

  (Daelemans and Bosch 2005; Theijssen et al. 2013)

TiMBL classification accuracy particle placement

predictors: verb, particle, DO length, DO head, 1st word of DO, 1st word after VP



Team members

Jason Grafmiller
Ph.D., 2013, Stanford University
particle placement

Benedikt Heller
MA, 2013, University of Giessen
the genitive alternation

Melanie Röthlisberger
MA, 2011, University of Zurich
the dative alternation

Thank you!

benszm@kuleuven.be

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

This presentation is based upon work supported by an Odysseus grant of the Research Foundation Flanders (FWO) (grant no. G.0C59.13N).

References I

Adger, D. and G. Trousdale (2007, July). Variation in English syntax: theoretical implications. *English Language and Linguistics 11*(02), 261.

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: Mouton de Gruyter.

Bresnan, J., A. Cueni, T. Nikitina, and H. Baayen (2007). Predicting the Dative Alternation. In G. Boume, I. Kraemer, and J. 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), 168–213.

References II

Bresnan, J. and J. Hay (2008, February). Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua 118*(2), 245–259.

Daelemans, W. and A. v. d. Bosch (2005). *Memory-based language processing*. Studies in natural language processing. Cambridge, UK ; New York: Cambridge University Press.

Davies, M. and R. Fuchs (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide 36*(1), 1–28.

de Marneffe, M.-C., S. Grimm, I. Arnon, S. Kirby, and J. Bresnan (2012, January). A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes 27*(1), 25–61.

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.

References III

Gahl, S. and S. Garnsey (2004). Knowledge of Grammar, Knowledge of Usage: Syntactic Probabilities Affect Pronunciation Variation. *Language 80*, 748–775.

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

Grafmiller, J. (2014, November). Variation in English genitives across modality and genres. *English Language and Linguistics 18*(03), 471–496.

Greenbaum, S. (1991). ICE: the International Corpus of English. *English Today 7*(04), 3.

Kachru, B. B. (Ed.) (1992). *The Other tongue: English across cultures* (2nd ed ed.). English in the global context. Urbana: University of Illinois Press.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Philadelphia Press.

References IV

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.

Mesthrie, R. and R. M. Bhatt (2008). *World Englishes: the study of new linguistic varieties*. Key topics in sociolinguistics. Cambridge, UK ; New York: Cambridge University Press.

O'Reilly, R. C., D. Wyatte, S. Herd, B. Mingus, and D. J. Jilk (2013). Recurrent Processing during Object Recognition. *Frontiers in Psychology 4*.

Schneider, E. (2003). The dynamics of New Englishes: From identity construction to dialect birth. *Language 79*(2), 233–281.

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

References V

Shin, N. L. (2014). Grammatical complexification in Spanish in New York: 3sg pronoun expression and verbal ambiguity. *Language Variation and Change 26*, 1–28.

Szmrecsanyi, B., J. Grafmiller, B. Heller, and M. Röthlisberger. Around the world in three alternations: modeling syntactic variation in varieties of English around the globe. *English World-Wide 37*(2).

Theijssen, D., L. ten Bosch, L. Boves, B. Cranen, and H. van Halteren (2013, January). Choosing alternatives: Using Bayesian Networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory 9*(2), 227–262.

Trudgill, P. (2011). *Sociolinguistic typology : social determinants of linguistic complexity*. Oxford, New York: Oxford University Press.

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.