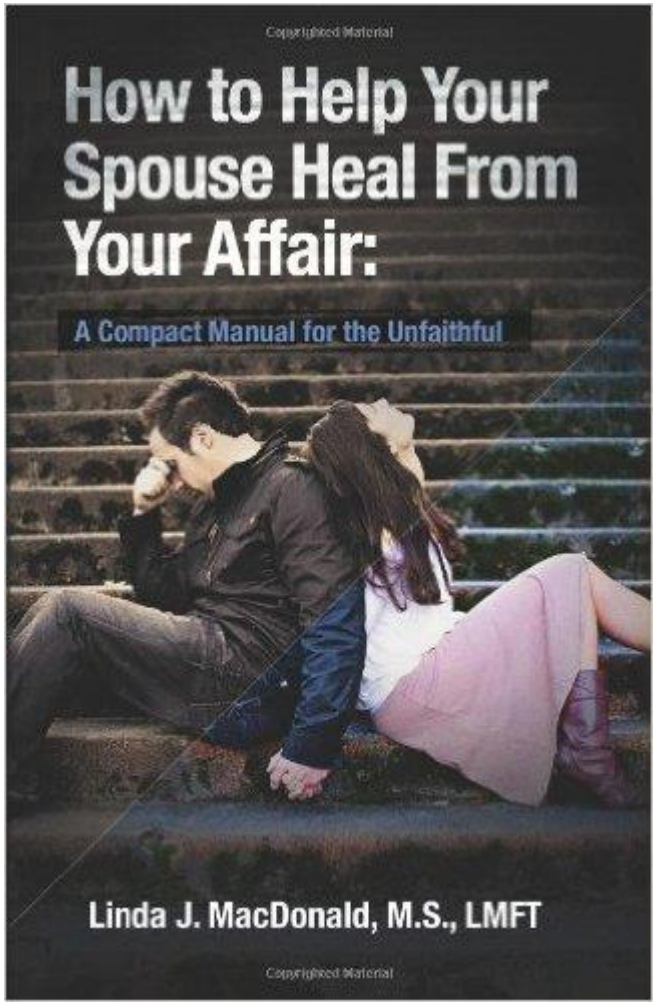
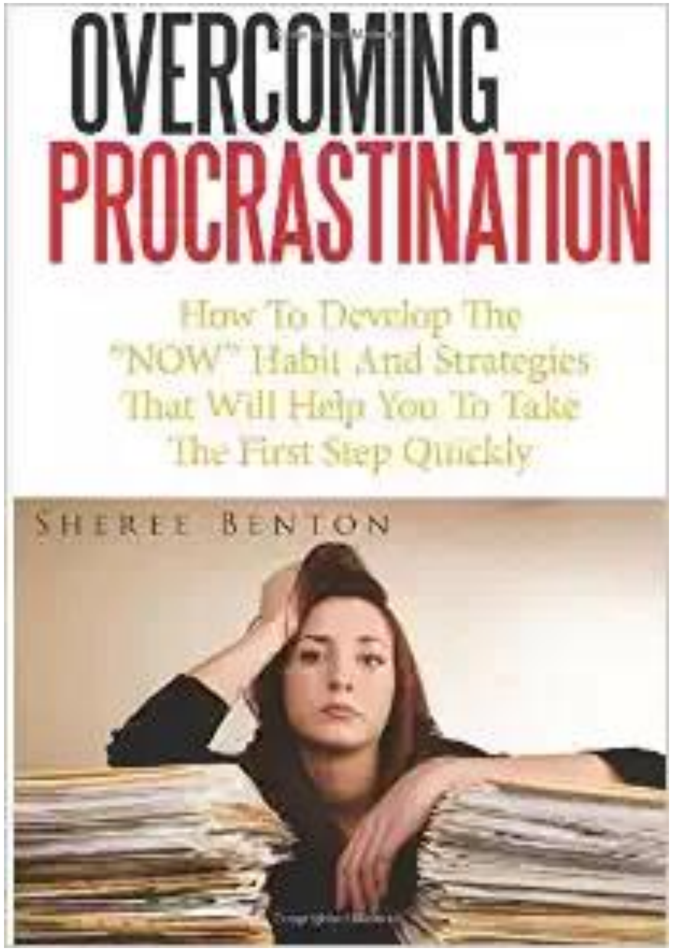
Help + (to) Infinitive in twenty geographic varieties of web-based English: A Bayesian mixed-model approach

Natalia Levshina

Leipzig University

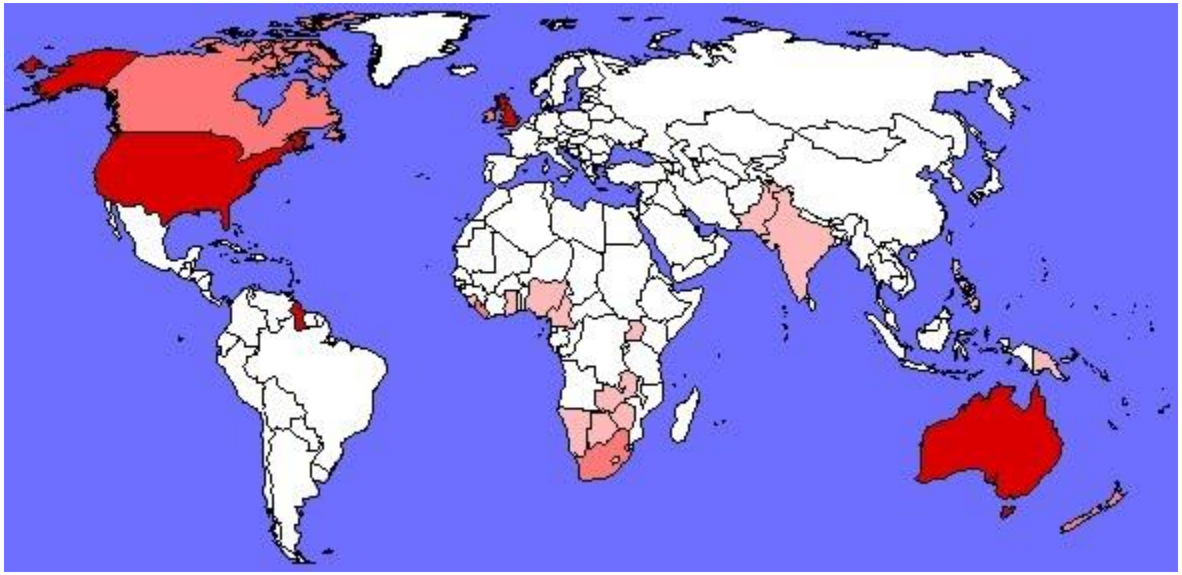

Copyrighted Material
How to Help Your Spouse Heal From Your Affair:
A Compact Manual for the Unfaithful
Linda J. MacDonald, M.S., LMFT
Copyrighted Material

???


OVERCOMING
PROCRASTINATION
How To Develop The "NOW" Habit And Strategies That Will Help You To Take The First Step Quickly
SHEREE BENTON
Copyrighted Material

Web-based English from 20 countries

http://corpus.byu.edu/glowbe/



Australia (AU), Bangladesh (BD), Canada (CA), Great Britain (GB), Ghana (GH), Hong Kong (HK), India (IN), Ireland (IE), Jamaica (JM), Kenya (KE), Malaysia (MY), Nigeria (NG), New Zealand (NZ), Pakistan (PK), the Philippines (PH), Singapore (SG), South Africa (ZA), Sri Lanka (LK), Tanzania (TZ), the USA (US)

Outline

- Previous research
- Data
- Variables
- Bayesian models
- Is there variation?

Lohmann 2011: *help* in BrE

- Complexity (Rohdenburg 1996; Hawkins 2004):
  - the greater the distance between *help* and Inf, the more difficult it is to recognize the latter as a component of the construction. Therefore, the speaker is more likely to mark it with *to*.
  - *I **helped** the cat that killed the rat (…) **to escape** from the dog.*
- Avoidance of identity (*horror aequi*) (Rohdenburg 2003):
  - *I want to help you do it* (+) vs. *to help you to do it* (-)
  - The effect of *horror aequi* wanes with the increasing number of intervening words between *help* and Inf.

Lohmann 2011 (cont.)

- Iconicity (e.g. Haiman 1983):
  - the Helper is less involved = greater conceptual distance between events > greater formal distance (*to*)
    - *This trick will help you to lose twenty kilos in two weeks.*
    - *Mary helped John cook the dinner. She cut the vegetables.*
  - Animate Helpers somewhat increase the chances of *Ø*-Inf because they show a potentially greater involvement in the effected event (?)
- Ø-Inf is more common after *helping* and *helps* than after *help* (cf. inflectional islands, Newman & Rice 2005)
- Implicit Helpees (e.g. *This will help to understand this problem*) increase the chances of *to*-Inf.

Variation

- Diachronic variation: ongoing auxiliarization of *help* and gradual disappearance of *to* (Mair 2006)
- Regional variation: *help* + Ø-Inf is more frequent in AmE than in BrE (e.g. McEnery & Xiao 2005)
- Register variation: *help* + Ø-Inf is preferred in less formal and spoken discourse (e.g. Lind 1983)

Main research question

- Is there regional (national) variation in the preference of the infinitive form?
- OK, but what is variation?

3 main aspects of variation (1)

- Variation of probabilistic constraints in time or space
    - Szmrecsnanyi (To appear): We can posit (probabilistic) grammar change if the stochastic effect of language-internal predictor variables varies as a function of real time. Here: the effect varies across language varieties.

3 main aspects of variation (2)

- Variation of baseline proportions of variants (when controlling for the environment variables)
  - Szmrecsnanyi (To appear): frequencies are not a reliable indicator of language change/variation. A change in frequencies over time can be due to changing frequencies of environment variables that trigger one or the other variant. Example: a dip in the frequency of ’s-genitive in 1650 – 1850 is explained by a decrease in the overall frequency of animate possessors in Genitive constructions.
  - Still, if we control for the environment variables (e.g. by taking the intercept values), a comparison of baseline frequencies is not uninteresting. After all, frequency can cause conservation, formal reduction, etc. (e.g. Bybee 2007).

3 main aspects of variation (3)

- Variation of constructional habitats, such as the frequency of animate possessors in the Genitive alternation:
  - spurious (e.g. poorly balanced corpora)
  - scientifically interesting (i.e. culture)
- Why bother? Non-spurious differences in habitat may trigger differences in probabilistic constraints.
  - Example: development of Dutch causative *doen*, which expresses direct causation. The construction has been in decline since the 18th century because interpersonal direct causation represented by *doen* became marginal due to liberalization of social relationships and change in our idealized cognitive models of authority and relationships between men and women (Verhagen 2000). Overall, the meaning of *doen* and its constructional network have changed, and so have its relationships with near-synonyms.

Outline

- Previous research
- Data
- Variables
- Bayesian models
- Interpretation

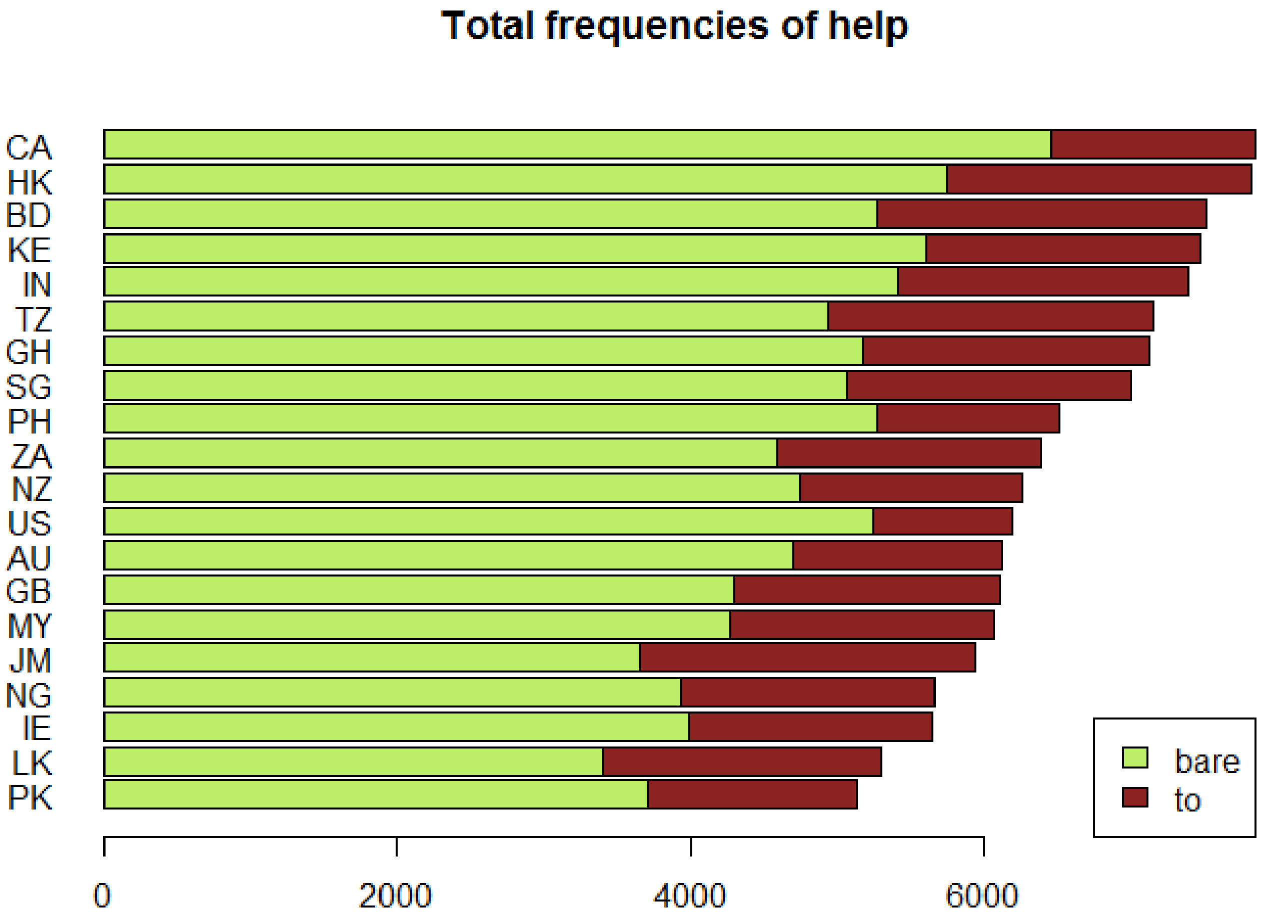
Corpus

- Global corpus of Web-based English (e.g. Davies and Fuchs 2015): national varieties in web-based communication
- Pros:
  - 1.9 bn tokens
  - 20 English-speaking countries
  - blogs (usually informal) and other ‘general’ (usually formal) sources
- Cons:
  - A lot of rubbish and duplicates
  - Website country != speakers’ variety
  - the blog/general distinction is problematic (see below)

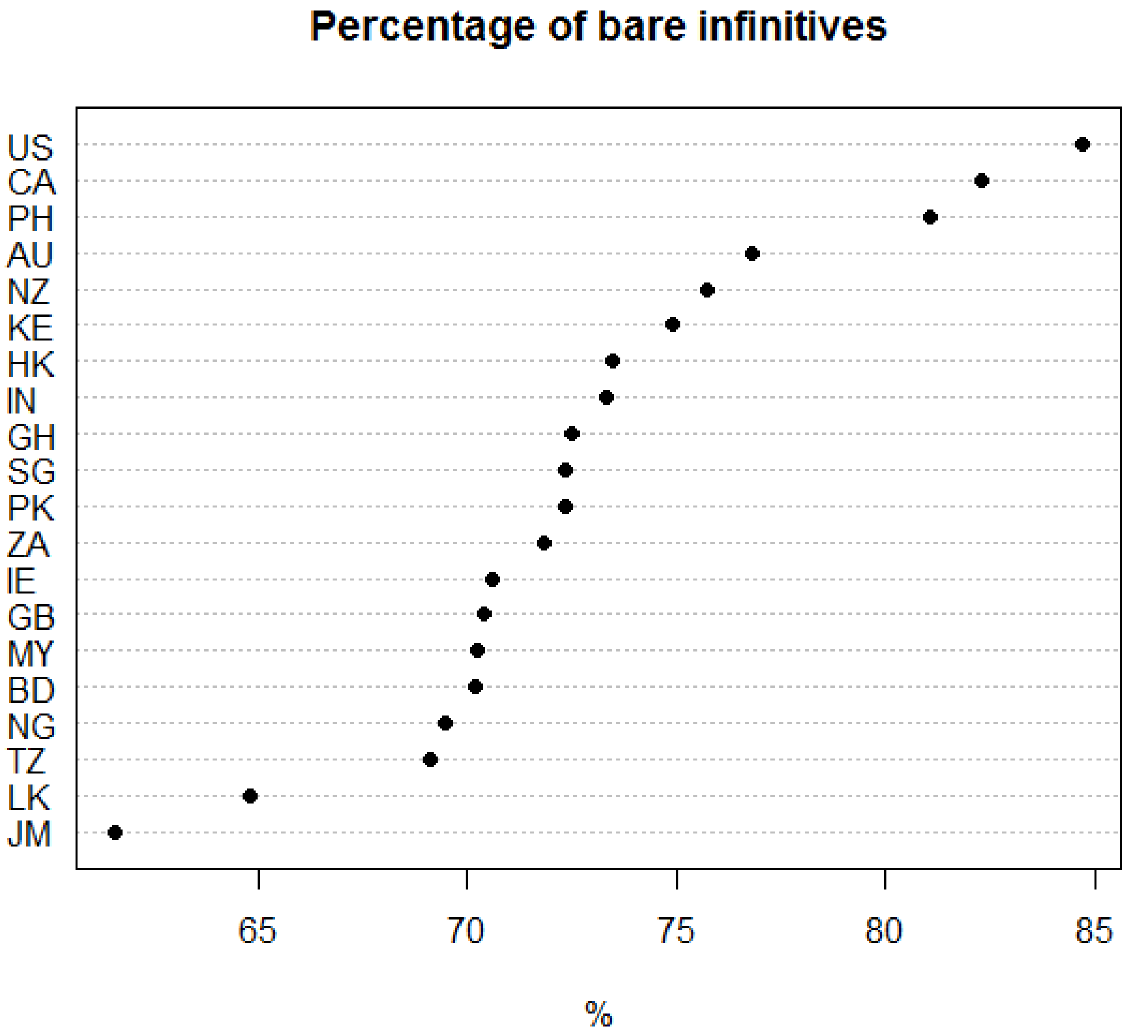
Data set

- Each country: a sample of 9M tokens from blogs and an equal sample from the ‘general’ subcorpus
- Extracted all instances of *help* followed by an Inf, without subordinate conjunctions, finite verb forms, subject pronouns (*I*, *he*, *she*, *we*, *they*), etc. between *help* and Inf
- Precision: 93%, recall: 86%
- Cleaning up: long distance between *help* and Inf, some prepositions, punctuation, *help desk*, *help me please*, remaining duplicates, etc.
- Appr. 130,000 observations from 20 countries

Absolute frequencies of both constructions



Relative frequencies



Outline

- Previous research
- Data
- Variables
- Bayesian models
- Is there variation?

Variables (from previous research)

- V1 form (according to the POS tagging):
  - *help*$_{PRES}$, *help*$_{INF}$, *helps*, *helped*$_{PAST}$, *helped*$_{PPART}$, *helping*
- Horror aequi
  - *to + help* vs. all other contexts
- Linguistic distance between *help* and Inf, in words
  - *Meditation will help* ***you immensely*** *to discover your true self*. (2)
- The presence or absence of the Helpee NP
  - *This helps* ***you*** *(to) relax* vs. *This helps (to) relax*.

Variables (new)

- Transitive or intransitive Inf (DO) – Stanford Parser
  - *I helped him write* ***an application*** vs. *I helped him escape*.
- Collostructional attraction between *help* and (*to*) Inf…
- Register…

Collostructional attraction

	Verb X	**Other verbs**
help + *to/Ø* Verb	*a*	*c*
Other constructions	*b*	*d*

Retrieved separately for each subcorpus!

Collostructional measures

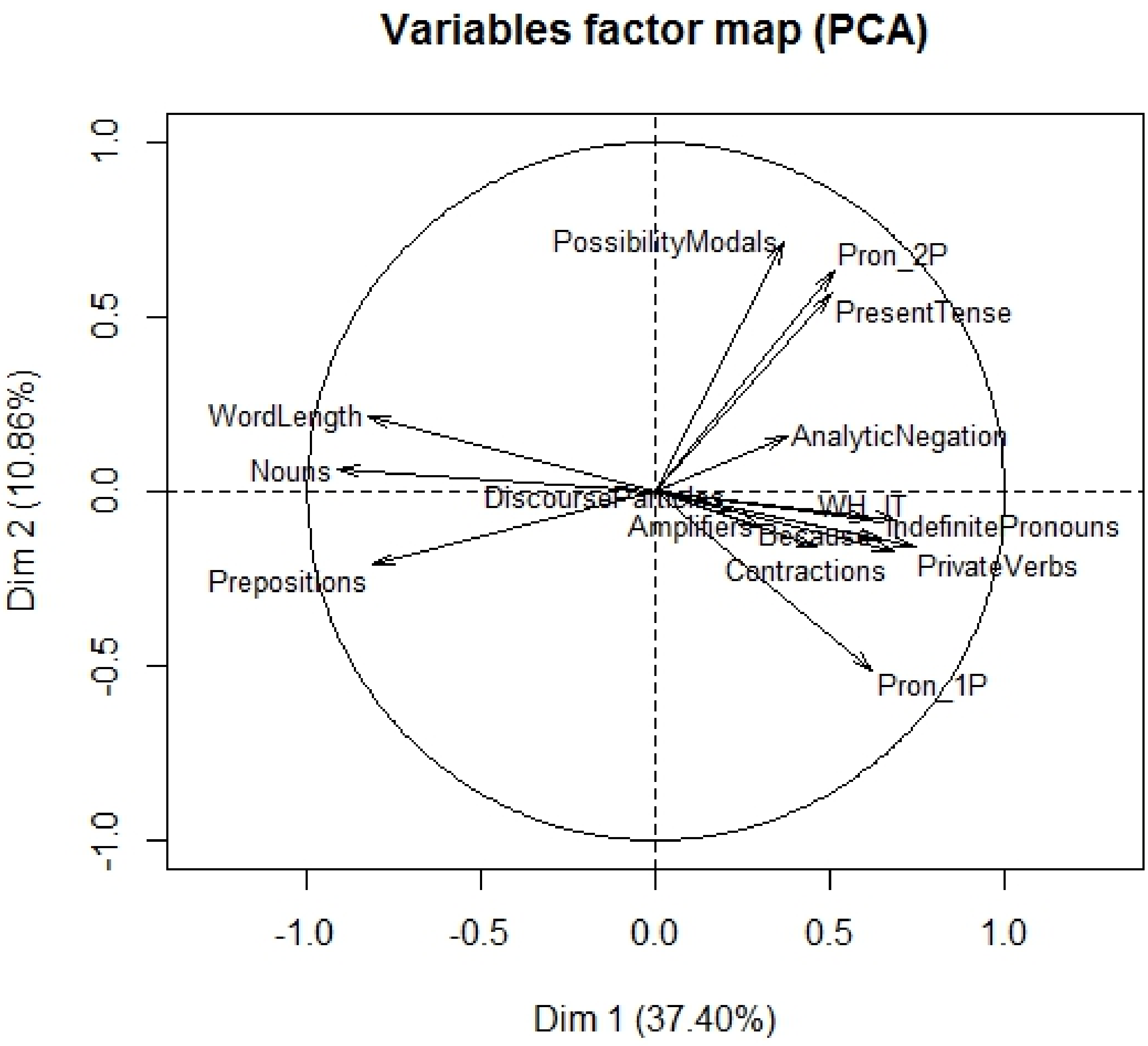
- Attraction = $a/(a + c)$ (Schmid 2000)
- Reliance = $a/(a + b)$ (Schmid 2000)
- Odds ratio = $a*d/b*c$
- Minimum sensitivity = min (Attraction, Reliance) (Pedersen & Bruce 1996)
- ΔP with construction as cue (Ellis 2006)
- ΔP with verb as cue (Ellis 2006)
- Collostructional Strength = log-p of FET (Stefanowitsch & Gries 2005, but see criticisms in Schmid & Kuechenhoff 2013) works the best!

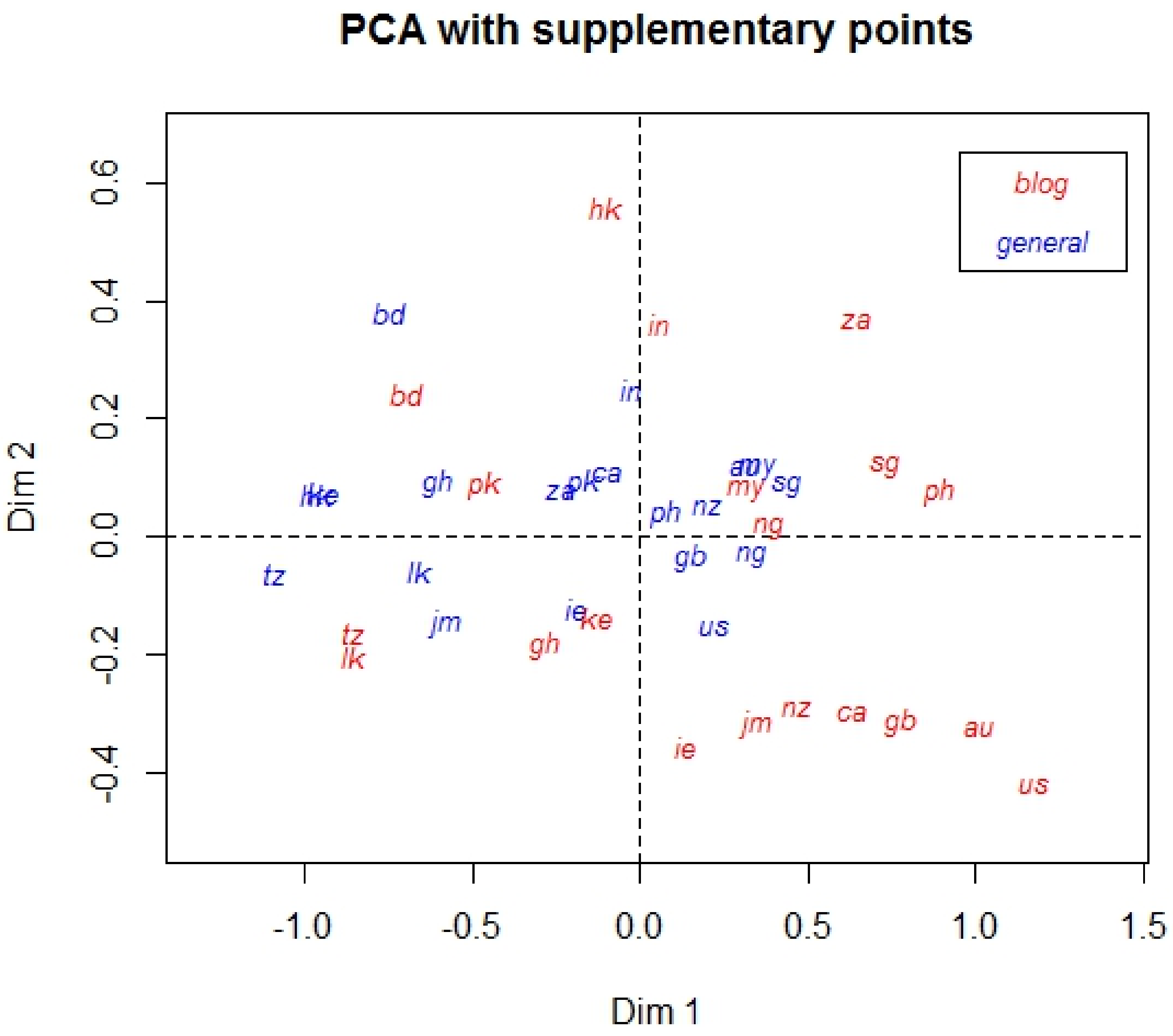
Most attracted verbs in US blogs

- *help* + (*to*)…
    - *understand*
    - *keep*
    - *achieve*
    - *shape*
    - *build*
    - *get*
    - *prevent*
    - *create*
    - *navigate*
    - *reduce*

Register

- Blog::general = Informal::formal? An empirical question!
- Biber’s (1988) Multidimensional Analysis: Dim 1 “Informational vs. involved production”: highly informational, edited, careful vs. affective, involved, produced in real-time
    - Features:
        - informational: longer words, nouns, prepositions
        - involved: 1st pp, 2nd pp, discourse particles, private verbs, contractions, present tense, amplifiers, etc.
    - 32500 files from all 20 varieties (1000 and more tokens)
    - PCA


Variables factor map (PCA)
PossibilityModals
Pron_2P
PresentTense
WordLength
AnalyticNegation
Nouns
DiscourseParticles
WH_IT
IndefinitePronouns
Amplifiers
Because
PrivateVerbs
Contractions
Prepositions
Pron_1P
Dim 2 (10.86%)
Dim 1 (37.40%)
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0


PCA with supplementary points
blog
general
Dim 1
Dim 2
-1.0
-0.5
0.0
0.5
1.0
1.5
-0.4
-0.2
0.0
0.2
0.4
0.6
hk
bd
in
za
bd
in
sg
ph
au
my
my
sg
lk
ke
gh
pk
za
pk
ca
ph
nz
ng
tz
lk
gb
ng
ie
ke
jm
us
tz
lk
gh
jm
nz
ca
gb
au
ie
us

Average word length as a proxy of Dim 1

- About 50% of all files in the data set contain less than 1,000 tokens.
- Approximation of Dim 1: Average word length in a text. Very strong correlation between word length and Dim 1 ($r = -0.82$, $p < 0.0001$), also across all countries.

Outline

- Previous research
- Data
- Variables
- Bayesian models
- Is there variation?

Why Bayesian?

- Frequentist statistics (Pearson, Fisher, etc.):
  - The *p*-value is used to decide if the null hypothesis of no effect, no difference, etc. can be rejected.
  - A dichotomous decision (significant or not significant).
  - We lose a lot of information!
- Bayesian statistics:
  - We can both test the null hypothesis AND obtain directly the probability that a variable has an effect in a particular direction.
  - More information
  - No *p*-value-hacking

The notorious problem of priors

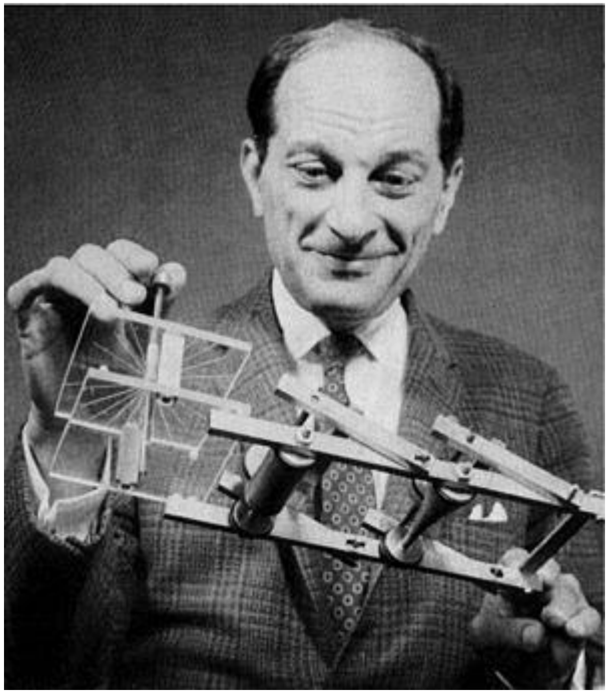
- In Bayesian statistics, the probability of a variable having an effect in a given direction (so-called posterior) is computed on the basis of two sources of information:
  - Prior probability (or simply prior)
  - Likelihood based on data
- For many people, using priors sounds arbitrary and subjective.
- Here, I use non-informative priors. Only the data play a role! No worries about subjectivity.

20 country-specific models

- Logistic mixed models
    - the response: bare or *to*-infinitive
    - fixed effects (the contextual variables)
    - The second verb (Inf) and corpus file ID as random effects

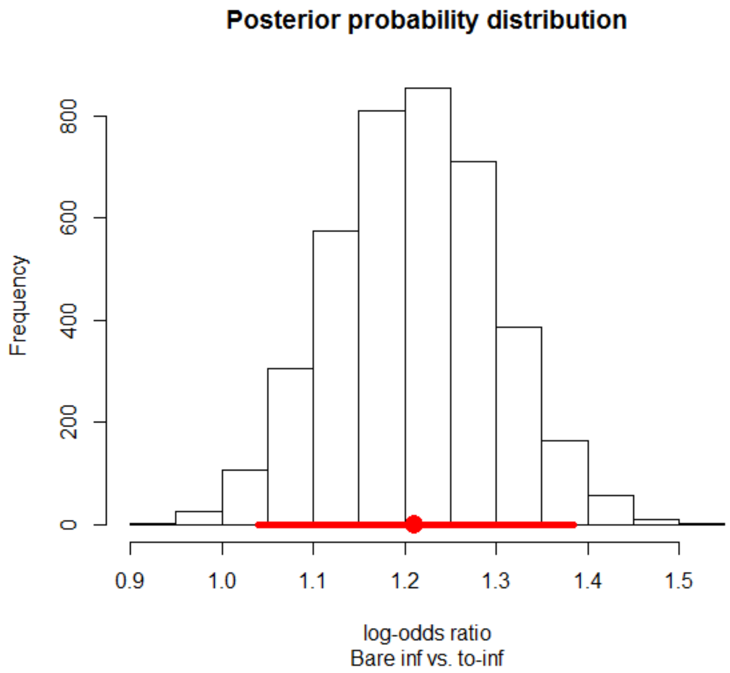
- Stan software for Bayesian statistics via R package rstan

Why “Stan”?



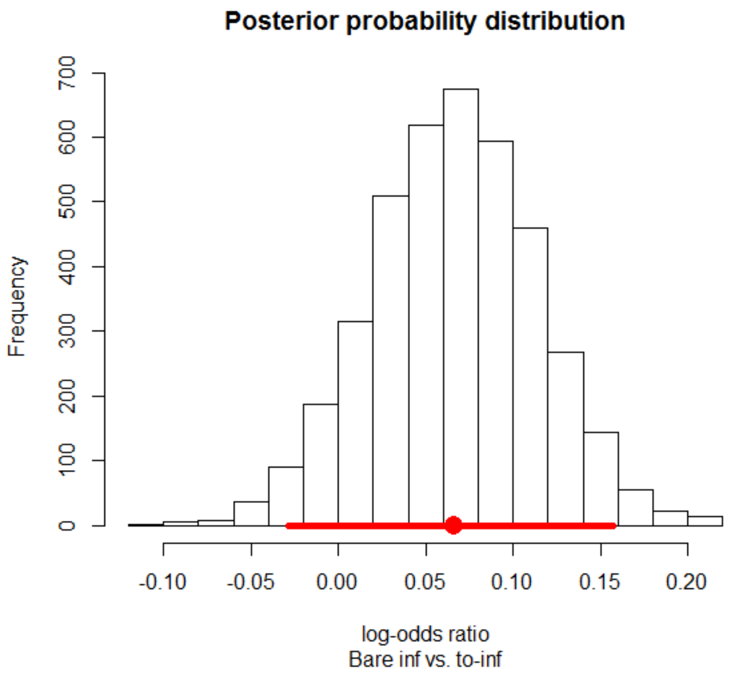
Stanisław Ulam, pioneer of the Monte Carlo method

Example 1: effect of *to help* on the choice of Inf in Australian subcorpus



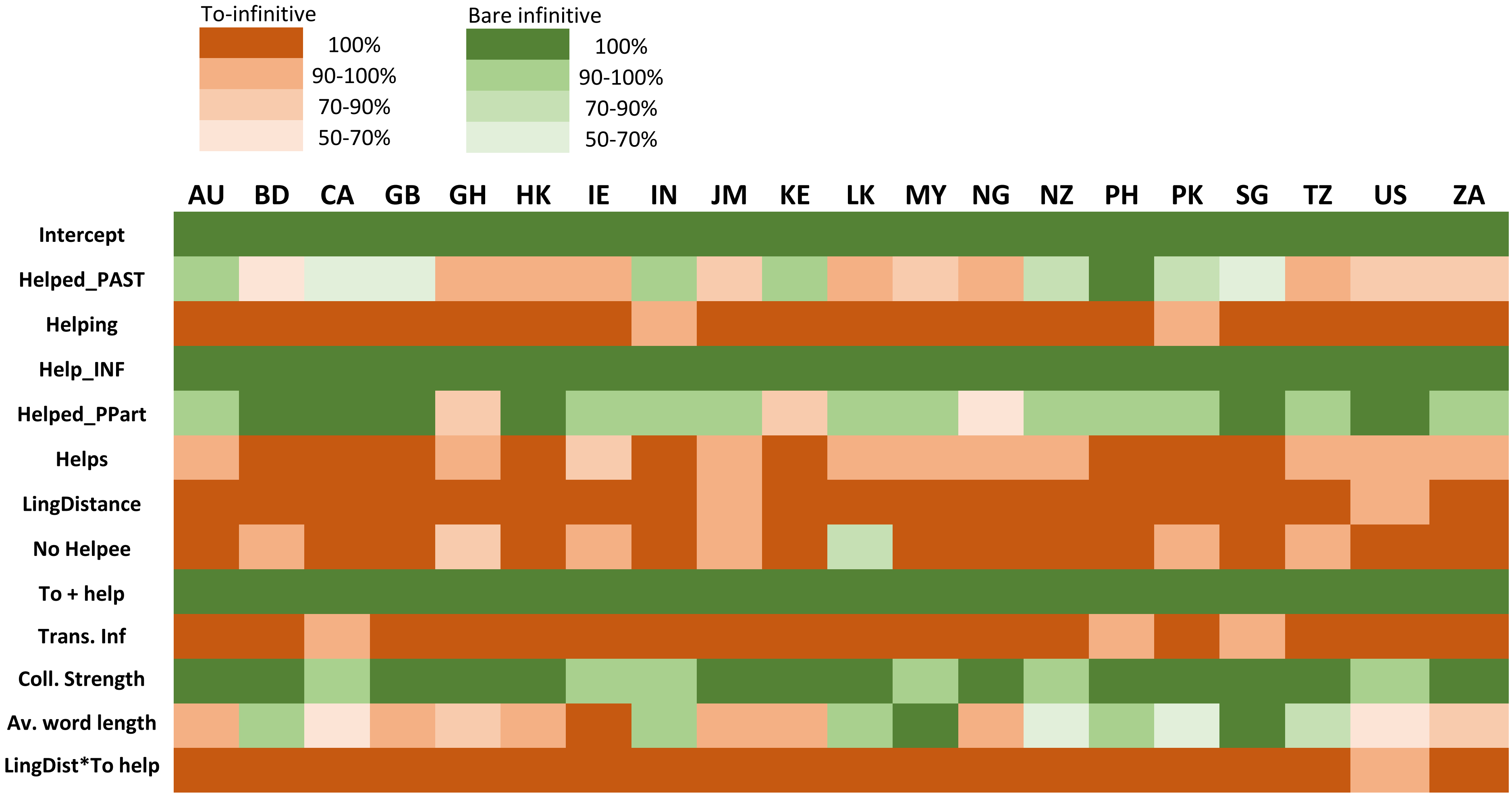
The probability that the effect is positive (favouring Ø-inf) is 100%

Example 2: effect of average word length in Bangladesh subcorpus

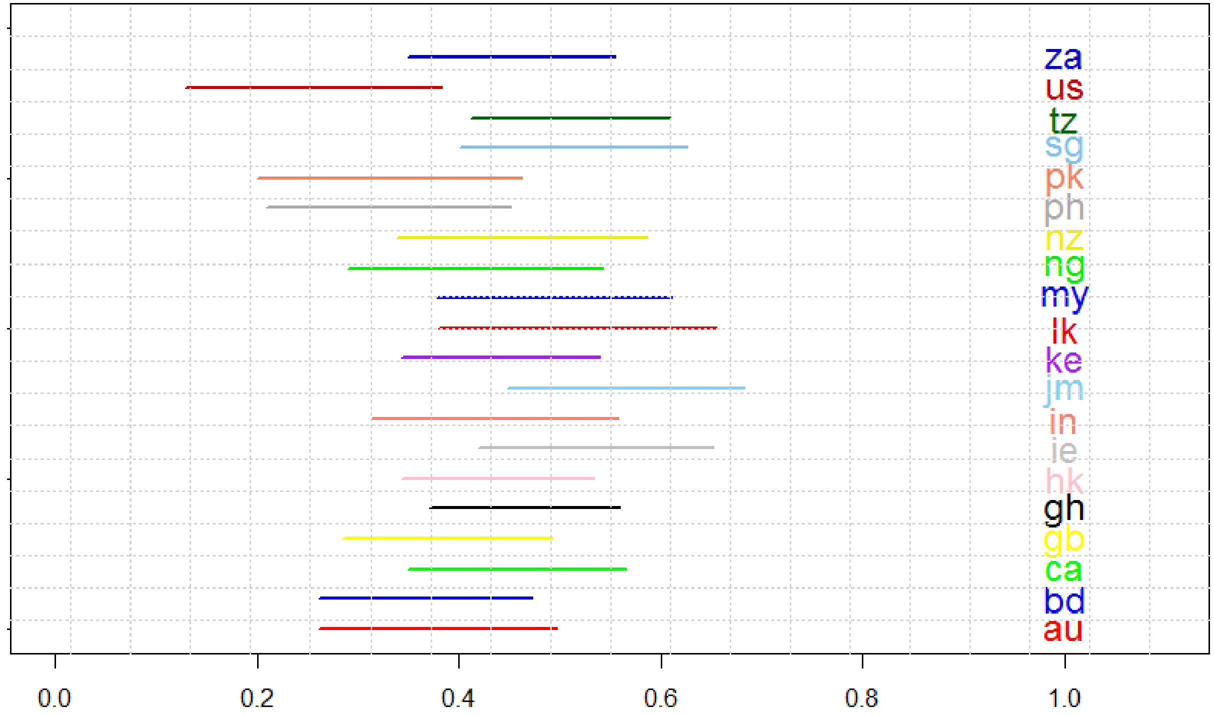


The probability that the effect is positive (favouring Ø-inf) is 91.8%

Posteriors of predictor effects



Overlapping 95% credible intervals: example (interaction term)



log-odds

Overlapping 95% credible intervals

Parameter	Lack of overlap (out of 190 pairs)
Collostructional Strength	0
ING form (helping)	3
Interaction LingDistance*To help	4
Transitive Infinitive	6
To help	13
LingDistance	14
3rd person SG form (helps)	18
Implicit helpee	36
Average word length (register)	46

Least overlap of 95% CI

- Ghana
- USA
- Jamaica
- Sri Lanka
- India
- Bangladesh

Greatest overlap of 95% CI

- Malaysia
- New Zealand
- South Africa
- Australia
- Great Britain
- Philippines

Outline

- Previous research
- Data
- Variables
- Bayesian models
- Is there variation?

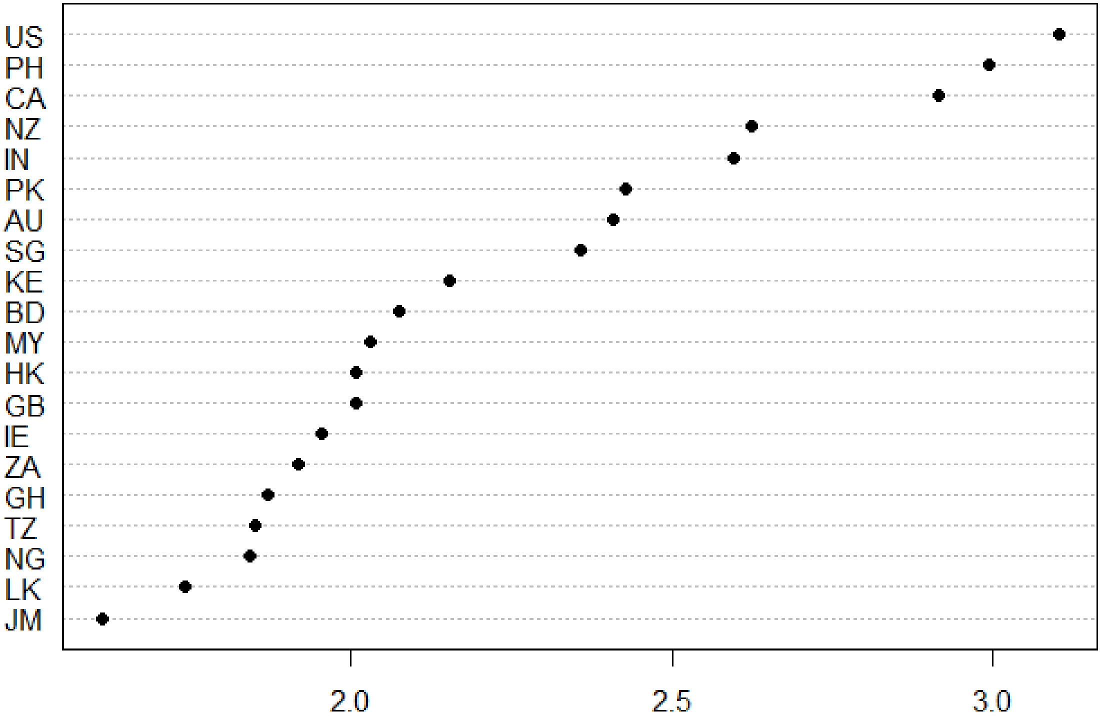
Variation of probabilistic constraints

- Very little
- Pragmatics (register, ellipsis) > Form

Variation of baseline proportions of variants

- Intercept values (sum contrasts, centred variables), which represent the log odds of Ø/*to*-infinitive for an abstract average situation (all variables are controlled for).
- Results in little difference from simple relative frequencies (ρ = 0.86, p < 0.001). => There is little variation in the frequencies of the environment variables.
    - This may also explain the lack of substantial differences in the constraints.

Intercept values (bare vs. *to*)

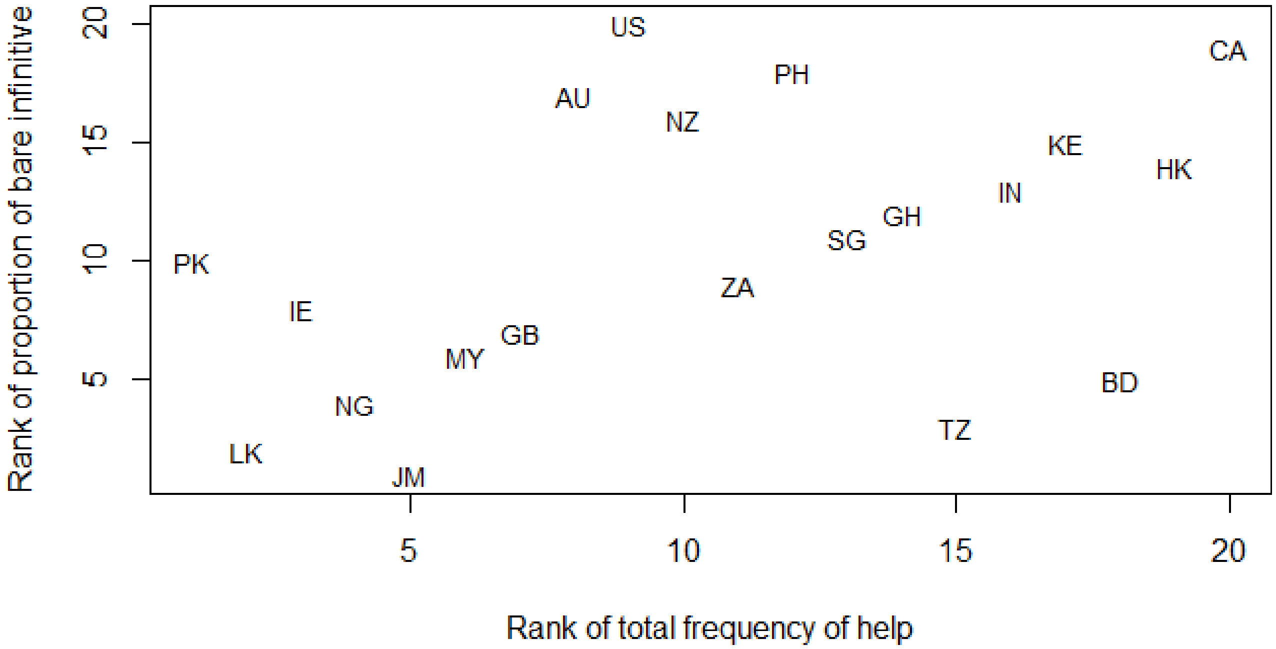


Influence of substrate and creoles/pidgins?

- Jamaican creole continuum
  - Yuh **help fi** keep di peace (Mek Wi Laugh & Talk: An Anthology of Jamaican Poems by Donna Hart 2014)
- Nigerian English: educated NigE speakers tend to overgeneralize the infinitive (e.g. *he made her to do it*). Interestingly, less educated speakers tend to omit the infinitive (e.g. *I won go Amerika*), which is typical of Nigerian Pidgin (Taiwo 2012).

Different stages of auxiliarization of *help* (Mair 2006)?

- Total frequency of *help* vs. proportion of Ø-inf: Spearman $\rho = 0.43$, $p = 0.06$
- Economy (e.g. Haiman 1983; Haspelmath 2008)



Summary

- The lectal grammars as sets of probabilistic constraints are rather similar.
- At the same time, there’s some minor variation. While formal constraints are rather stable, pragmatic ones are more variable.
- Although all dialects “prefer” Ø-inf, there’re cross-lectal differences in relative frequencies and baseline odds of the variants, as represented by the intercepts. These differences can be at least partly explained by the Principle of Economy and possibly language contact and sociolinguistic situation.

Thanks!

Data set and R and Stan code: soon on GitHub.

Slides: www.natalialevshina.com/presentations.html

Email: natalia.levshina@uni-leipzig.de

References

Biber, Douglas. 1988. *Variation across speech and writing.* Cambridge: Cambridge University Press.

Bybee, Joan. 2007. *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Ellis, Nick. 2006. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1). 1–24.

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

Haiman, John. 1983. Iconic and economic motivation. *Language* 59(4). 781–819.

Haspelmath, Martin. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19(1). 1–33.

Hawkins, John. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Lohmann, Arne. 2011. Help vs. help to - a multifactorial, mixed-effects account of infinitive marker omission. *English Language and Linguistics* 15(3). 499-521.

Lind, Age. 1983. The variant forms of help to/help 0. *English Studies* 64. 263–275.

Mair, Christian. 2006. *Twentieth-Century English: History, variation and standardization*. Cambridge: Cambridge University Press.

McEnery, Anthony & Zhonghua Xiao. 2005. HELP or HELP to: What do corpora have to say? *English Studies* 86(2). 161–87.

Pedersen, Ted & Rebecca Bruce. 1996. What to infer from a description. Technical Report 96-CSE-04, Southern Methodist University, Dallas, TX.

Rice, Sally & John Newman. 2005. Inflectional islands. Paper presented at the International Cognitive Linguistics Conference, Seoul, South Korea.

Rohdenburg, Gunther. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2). 149–182.

Rohdenburg, Gunther. 2003. *Horror aequi* and cognitive complexity as factors determining the use of interrogative clause linkers. In Gunther Rohdenburg & Britta Mondorf (Eds.), *Determinants of Grammatical Variation in English*, 205–250. Berlin/New York: Mouton de Gruyter.

Schmid, Hans-Jorg. 2000. *English Abstract Nouns as Conceptual Shells. From corpus to cognition*. Berlin/New York: Mouton de Gruyter.

Schmid, Hans-Jorg & Helmut Kuchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3). 531–577.

Stefanowitsch, Anatol & Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1). 1–43.

Szmrecsanyi, Benedikt. To appear. About text frequencies in historical linguistics: disentangling environmental and grammatical change" *Corpus Linguistics and Linguistic Theory* (special issue, ed. by Martin Hilpert & Hubert Cuyckens).

Rotimi Taiwo. (2012). Nigerian English. In Bernd Kortmann & Kerstin Lunkenheimer (Eds.) *The Mouton World Atlas of Variation in English*, 410 - 416. Berlin/New York: Mouton de Gruyter.

Verhagen., Arie. 2000. Interpreting Usage: Construing the History of Dutch Causal Verbs. In Michael Barlow and Suzanne Kemmer (Eds.), *Usage-Based Models of Language*, 261-286.. Stanford, CA: CSLI Publications.