Linguistics, mathematics and the Bible

Robert Östling

Department of Linguistics
Stockholm University
robert@ling.su.se

2018-08-29

Stockholm University

Introduction

Stockholm University

Motivations

- I come from computer science
- Why do I think linguistics is interesting?
- How do I look at languages?

Stockholm University

Linguistics: what and why?

- Language: a mapping, thoughts $\Longleftrightarrow$ sequences of symbols
- The world contains many such mappings, evolving through:
  - inheritance (learned from parents)
  - contact (influenced by surroundings)
  - innovations (adaptable to new situations)
- Constant pressure from cognitive, physiological and efficiency constraints

Stockholm University

Linguistics: why should I care?

- Language evolution is a stochastic process
- Roughly independent Markov chains
  (closer approximation on a small and remote island)
- Transitions in the process depend on:
  1. migration
  2. contact between peoples (trade, cultural influence, etc.)
  3. societal structure
  4. people being creative
  5. the human brain (and mouth, and ears)
- The world's languages make up a sample from this process
- **We can make inferences about the factors above**

Stockholm University

Languages today

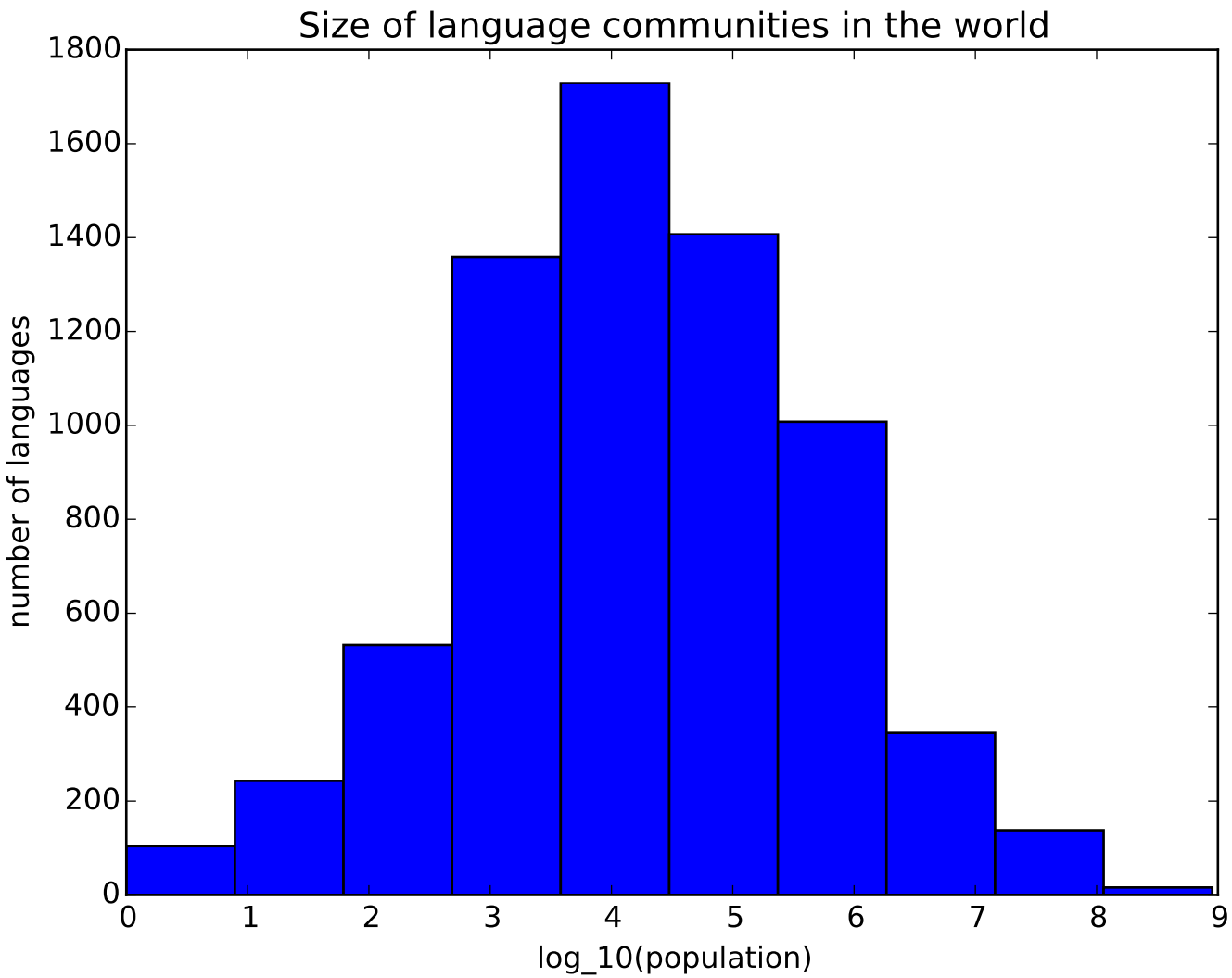
- About 7 000 languages exist today
- A human has an *idiolect*, a personal thought–symbol mapping
- A language is a cluster of mutually intelligible idiolects
- That is, $L_i \approx L_j \equiv L_i^{-1}(L_j(x)) \approx x$
- Somewhat subjective, see Eq. (1)

$$\boxed{\text{language} = \text{dialect} + \text{army} + \text{navy}} \tag{1}$$

Fundamental theorem of language politics

Stockholm University

Number of speakers



Stockholm University

Language structure

- Abstract mappings are not too useful
- Linguistics is about peeking into these mappings
- Multiple levels are used:
  - Morphology (word formation rules, $\approx$ linear)
    morphemes, smallest meaning-bearing unit
  - Syntax (sentence formation rules, $\approx$ hierarchical)
    words, largest non-hierarchical unit
    phrases, non-leaf nodes in the tree
  - Semantics (the "thoughts" link)

Stockholm University

Syntax and morphology

- Linguistic symbols (morphemes/words) form categories
- nouns (cat, rat, hat), number markers (cat**s**), etc.
- Syntax/morphology rules how categories are combined
- That is, they constrain the *range* of L_i
- Described formally e.g. as rewriting rules
- A language is associated with a set of such rules

English	Mandarin
$\mathrm{VP} \rightarrow \mathrm{V\ NP}$	$\mathrm{VP} \rightarrow \mathrm{V\ NP} \mid \mathrm{ba\ NP\ V}$
$\mathrm{V} \rightarrow \mathrm{drink}$	$\mathrm{V} \rightarrow \mathrm{he}$
$\mathrm{N_{plu}} \rightarrow \mathrm{N\ s}$	$\mathrm{N_{plu}} \rightarrow \mathrm{N}$
...	...

Stockholm University

Semantics

- Symbols (and their combination) map to "thoughts"
- Thoughts are fluid, English *cat* captures lots of variation
- A symbol maps to a volume in a semantic space
- *Kitten* maps to a subset of the *cat* space
- *Abyssinian* to another (partially overlapping) subset
  ... and also to other things from Ethiopia
- The structure of this mapping differs across languages
  (e.g., not all languages have kittens)

Source: Wikipedia

Stockholm University

Form

- The forms of linguistic symbols (morphemes) also evolve
- No coincidence *cat*, *Katze*, *katt* etc. sound similar
- In spoken languages, near-arbitrary
- Not so in signed languages!
- Still subject to physiological constraints
- Good source of evolutionary evidence

Stockholm University

Linguistic typology

- The study of linguistic variation
- Examples of interesting differences:
  - Does L use the rule $\mathrm{N}_{\mathrm{plu}} \to$ N x or $\mathrm{N}_{\mathrm{plu}} \to$ x N?
  - Does L use the rule NP $\to$ Adj N and/or NP $\to$ N Adj?
  - Is there any symbol used by L that maps to the semantic spaces of both *tree* and *fire*?

Stockholm University

Lexical typology

- How do the symbols of a language partition the semantic space?

Concept	English	Swedish	Mandarin
Father's older brother	uncle	farbror	daye
Father's younger brother	uncle	farbror	shu
Mother's older brother	uncle	morbror	jiu
Mother's younger brother	uncle	morbror	jiu

Stockholm University

Automated linguistics

- Define lots of features where languages do differ
- Describe each language in terms of its feature values
- Data $\rightarrow$ math $\rightarrow$???

Existing attempts

- Automated Similarity Judgement Program (ASJP)
  - words for 40 concepts in nearly all languages
  - mostly useful for studying evolution of form
  - and for checking which semantic concepts overlap in form
  - and for investigating cognitive/physiological constraints
- World Atlas of Language Structures (WALS), GramBank
  - syntax/morphology in lots of (up to a thousand) languages
  - evolution of syntax/morphology

Stockholm University

Data... from where?

- Just one problem, how to get this data?
  1. Lots and lots of manual work
  2. Automation

Stockholm University

Parallel texts

- Translated text, e.g. the Bible has 2000+ translations
- The text contains multiple sentences x_j
- Each sentence is encoded by multiple languages L_i
- $n \approx 8\ 000$, $m \approx 1500$

$$\begin{bmatrix} L_1(x_1) & L_1(x_2) & \dots & L_1(x_n) \\ L_2(x_1) & L_2(x_2) & \dots & L_2(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ L_m(x_1) & L_m(x_2) & \dots & L_m(x_n) \end{bmatrix}$$

There must be a good way to use this structure...

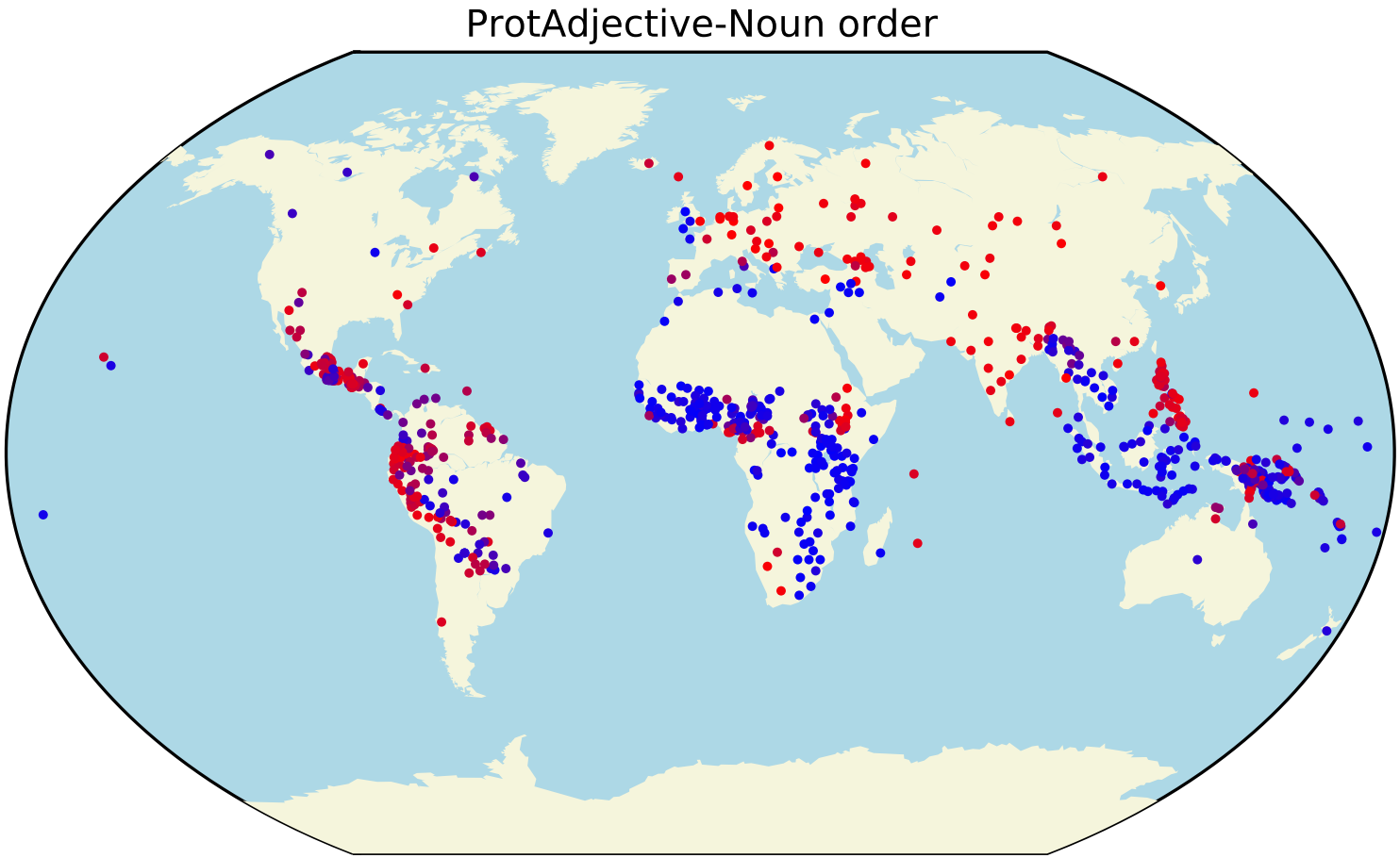
Stockholm University

Extracting data

- Let $\mathbf{x}_{\text{house}}$ be the sentences about houses
- Which symbols are most overrepresented in $\{L_i(x_j) | x_j \in \mathbf{x}_{\text{house}}\}$?
  Now we know the symbol for house in each language!
- What about $\mathbf{x}_{\text{small}}$ (referring to houses)?
  ...and for small!
- What is the order of these symbols?
  ...now we know if NP $\rightarrow$ N Adj or NP $\rightarrow$ Adj N
- Is something appended or prepended to the *house* symbol when multiple houses are referred to?
  ..now we know if $\text{N}_{\text{plu}} \rightarrow$ N x or $\text{N}_{\text{plu}} \rightarrow$ x N

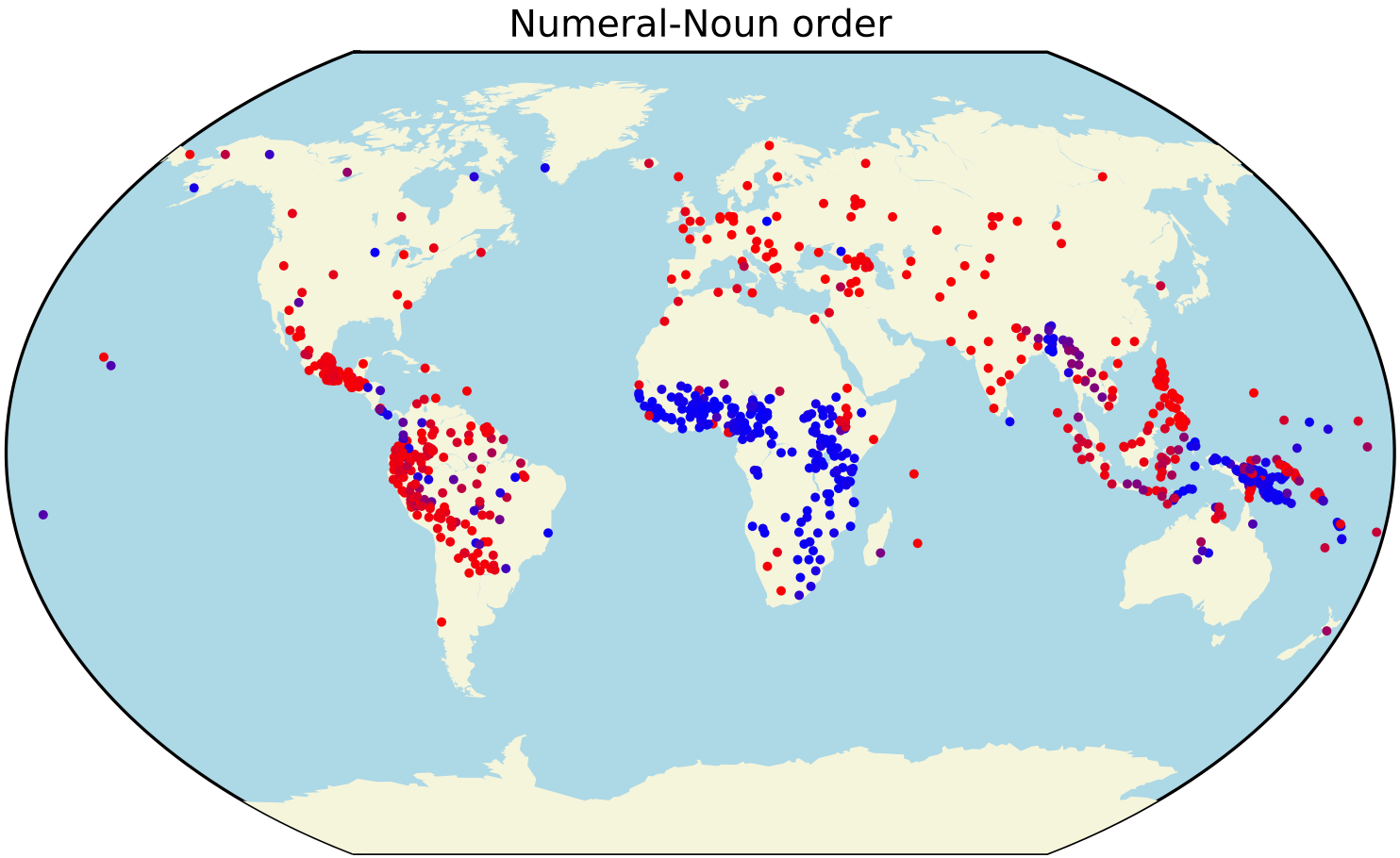
Stockholm University

NP → Adj N or NP → N Adj?



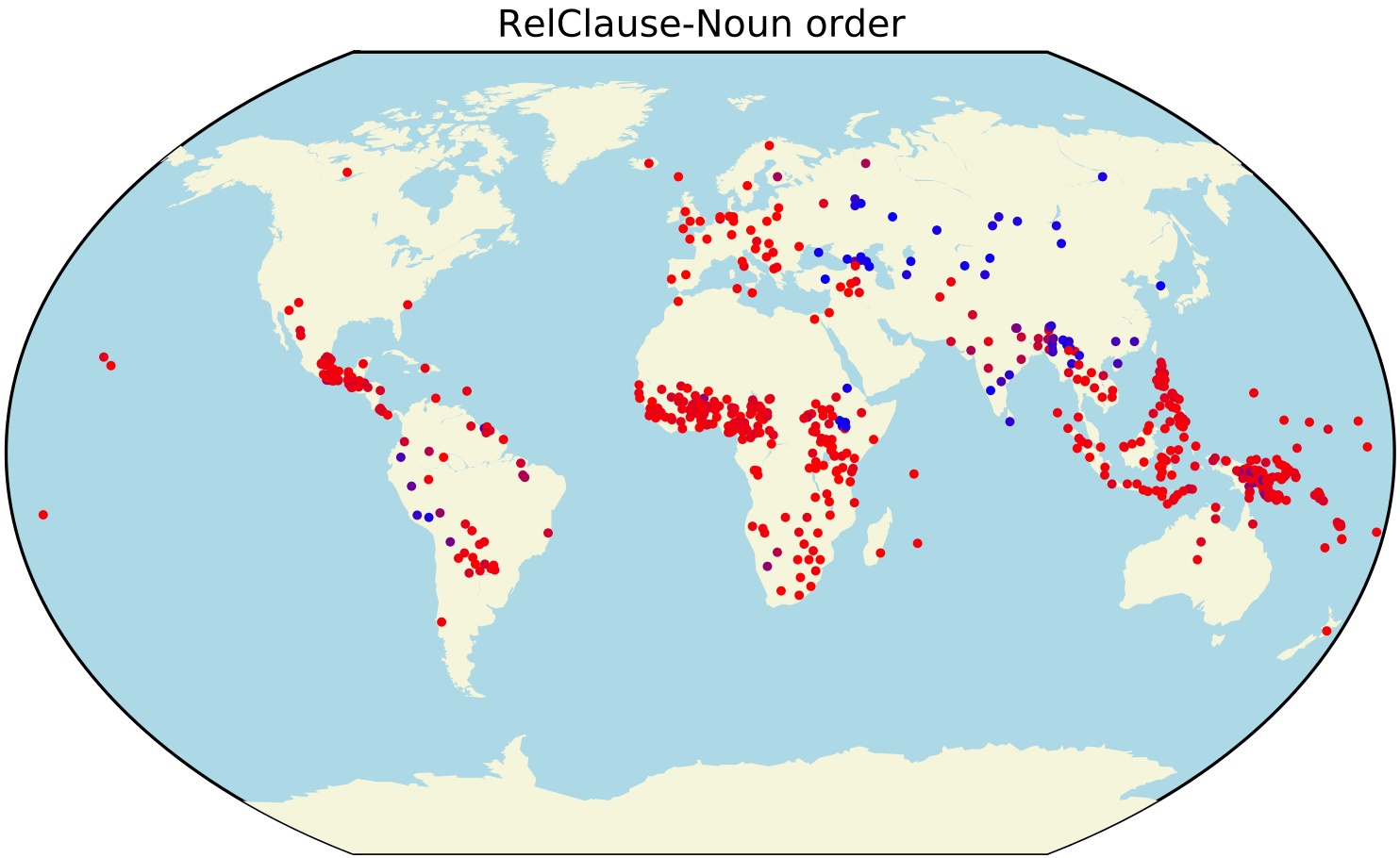
Stockholm University

NP → Num N or NP → N Num?



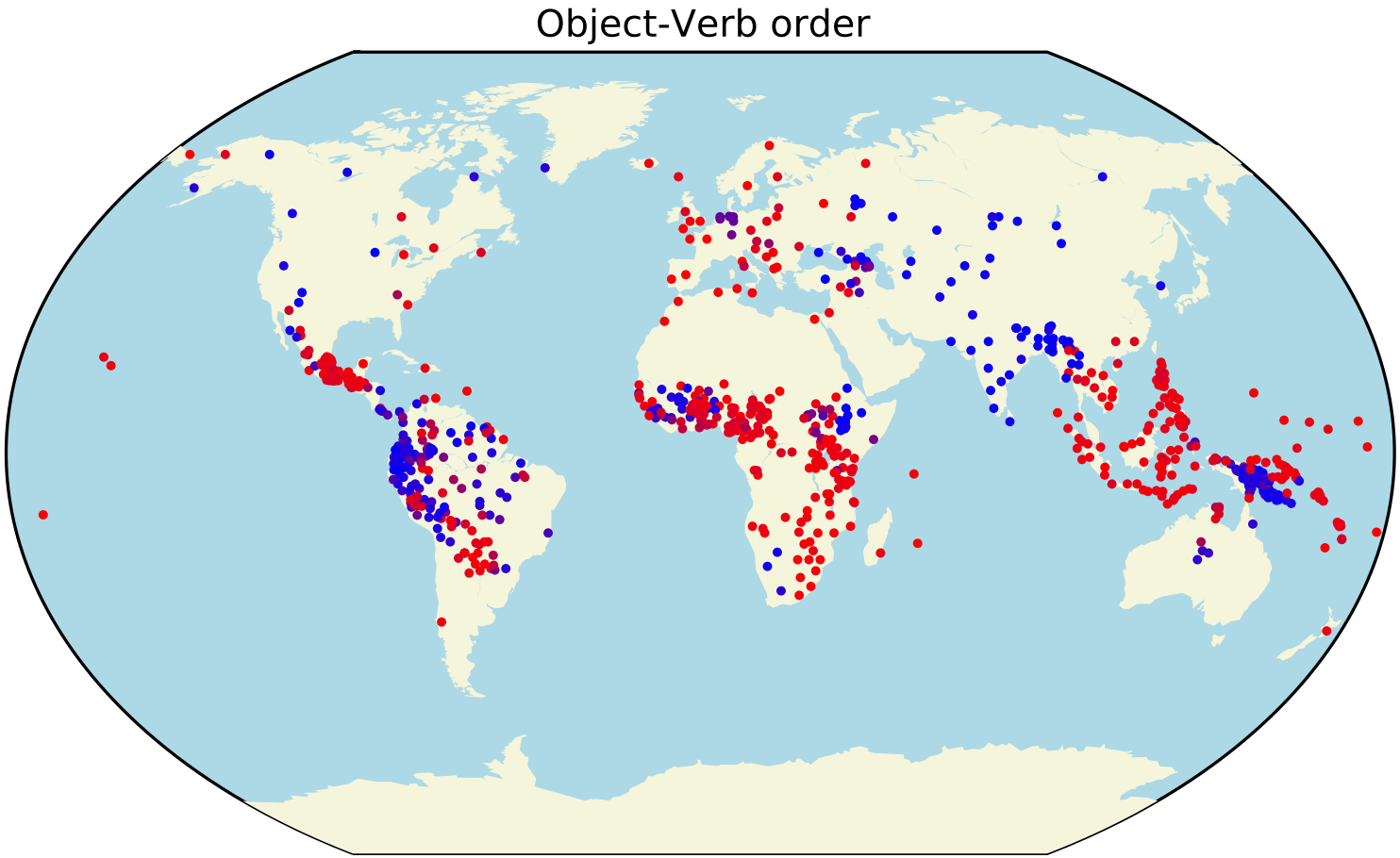
Stockholm University

NP → RelCl N or NP → N RelCl?



Stockholm University

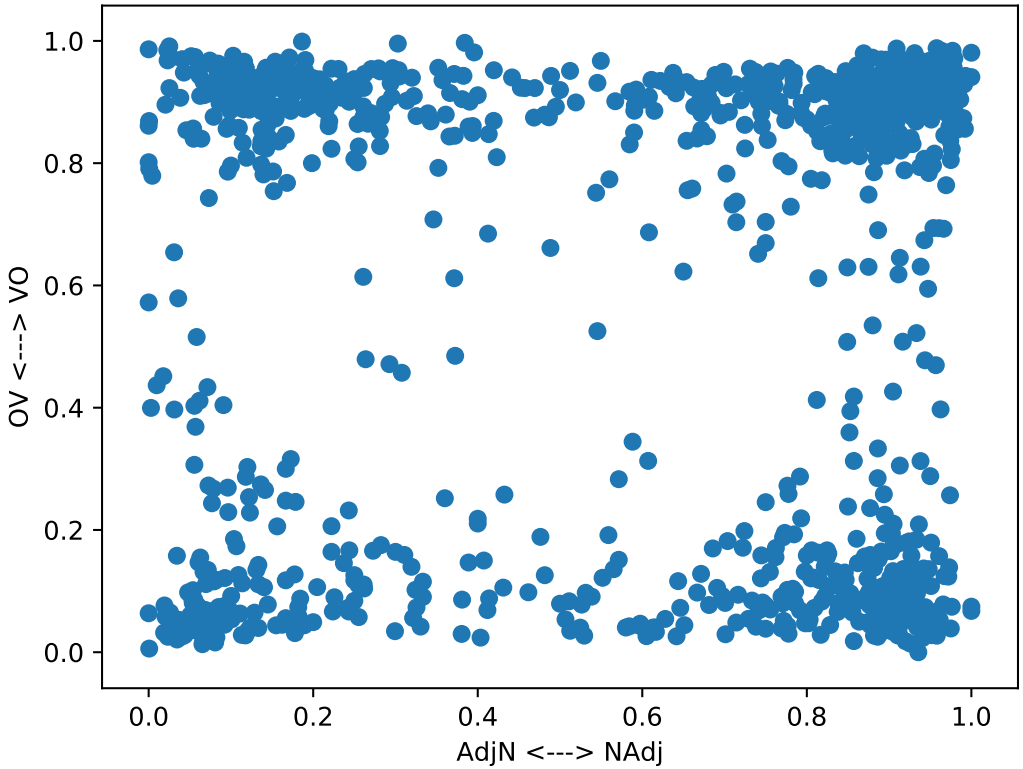
VP → NP V or VP → V NP?



Stockholm University

Testing theories

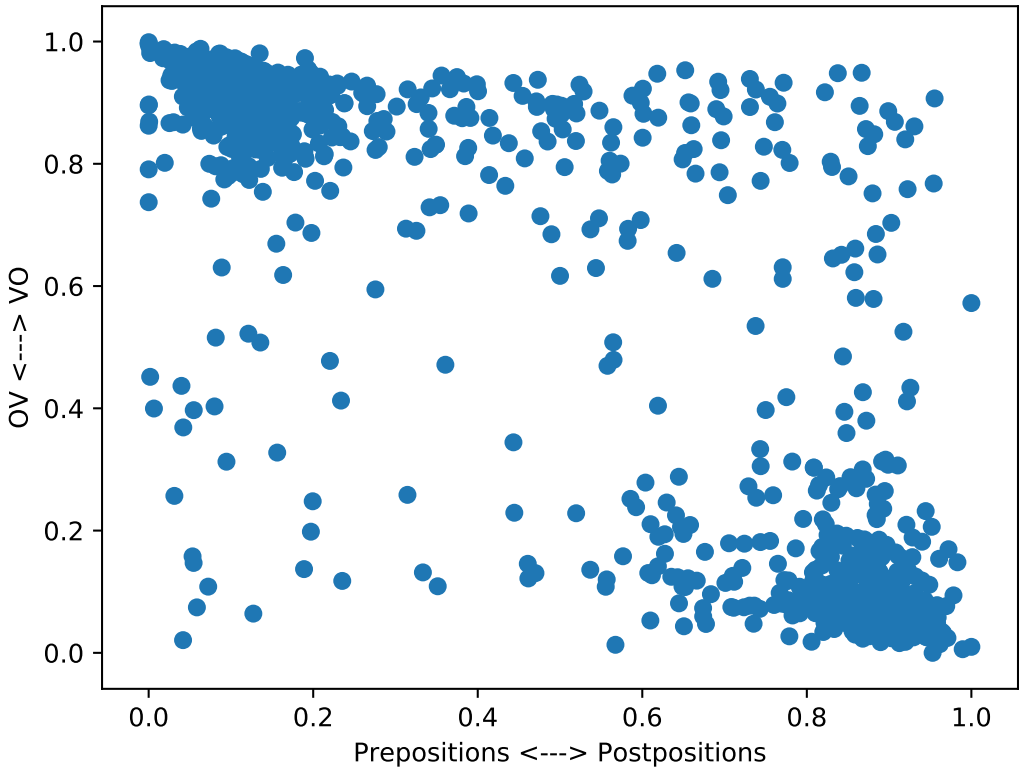
- If languages can be divided into left- and right-branching, these should be strongly correlated



Stockholm University

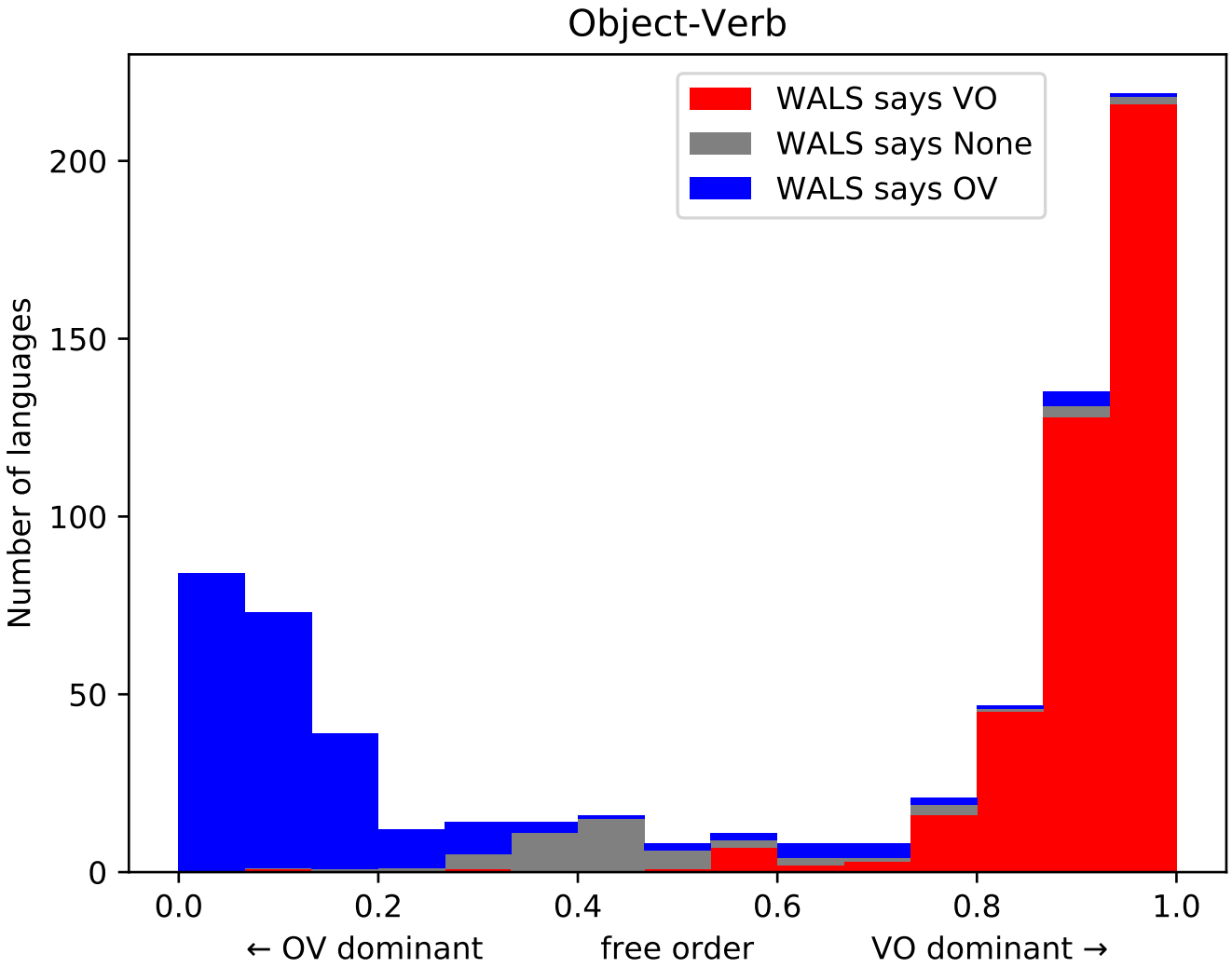
Testing theories

- YES: eating bread at home, bread home at eating
- NO: eating bread home at, bread at home eating
- Cognitive constraint, or historical accident?



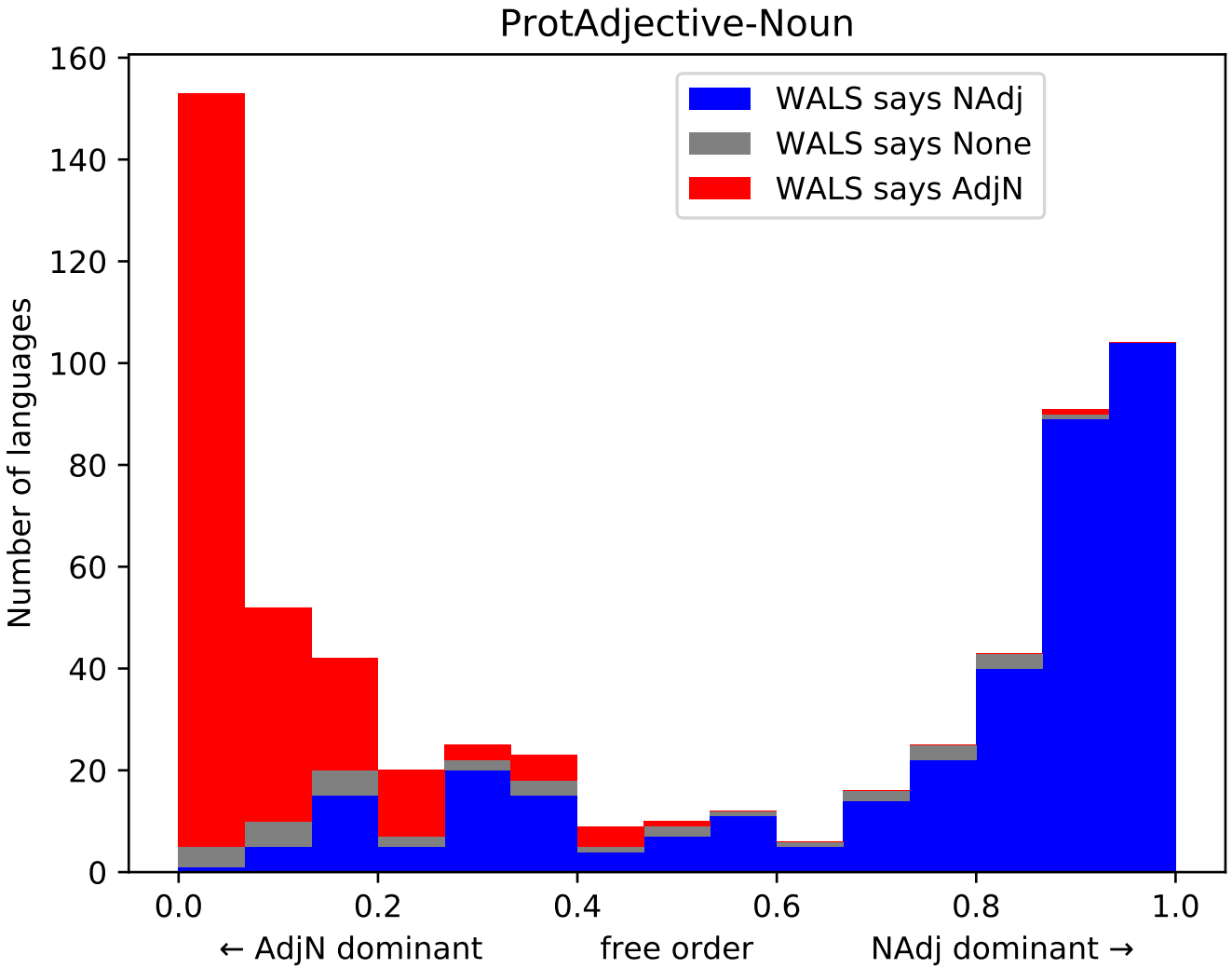
Stockholm University

Can we trust this?



Stockholm University

Can we trust this?



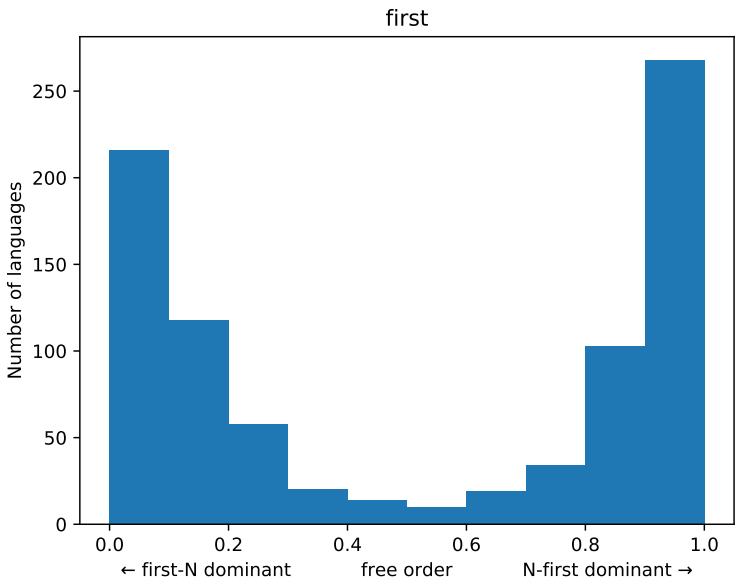
Stockholm University

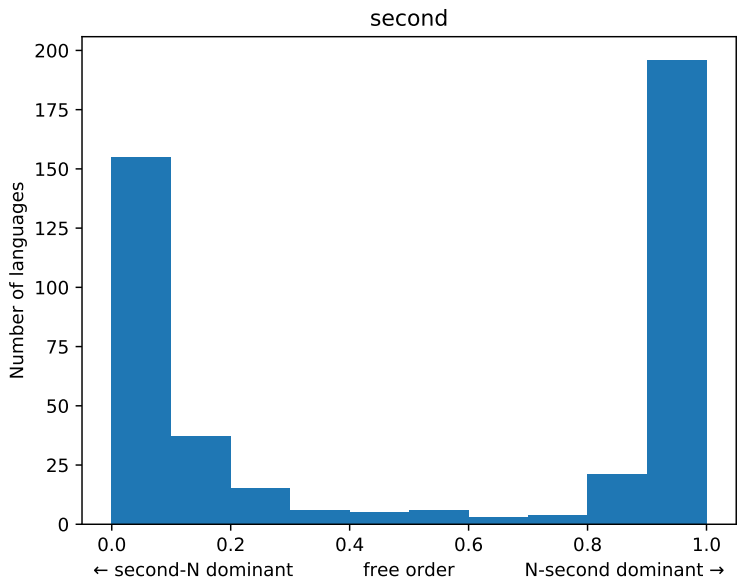
Explaining differences

- Sometimes the translation differs from the linguist's description
- Turns out adjectives are not a perfect syntactic category
- Individual words often behave different (French, Tok Pisin, ...)
- In some cases (Tagalog) there are complex interacting rules

Stockholm University

Variability within languages

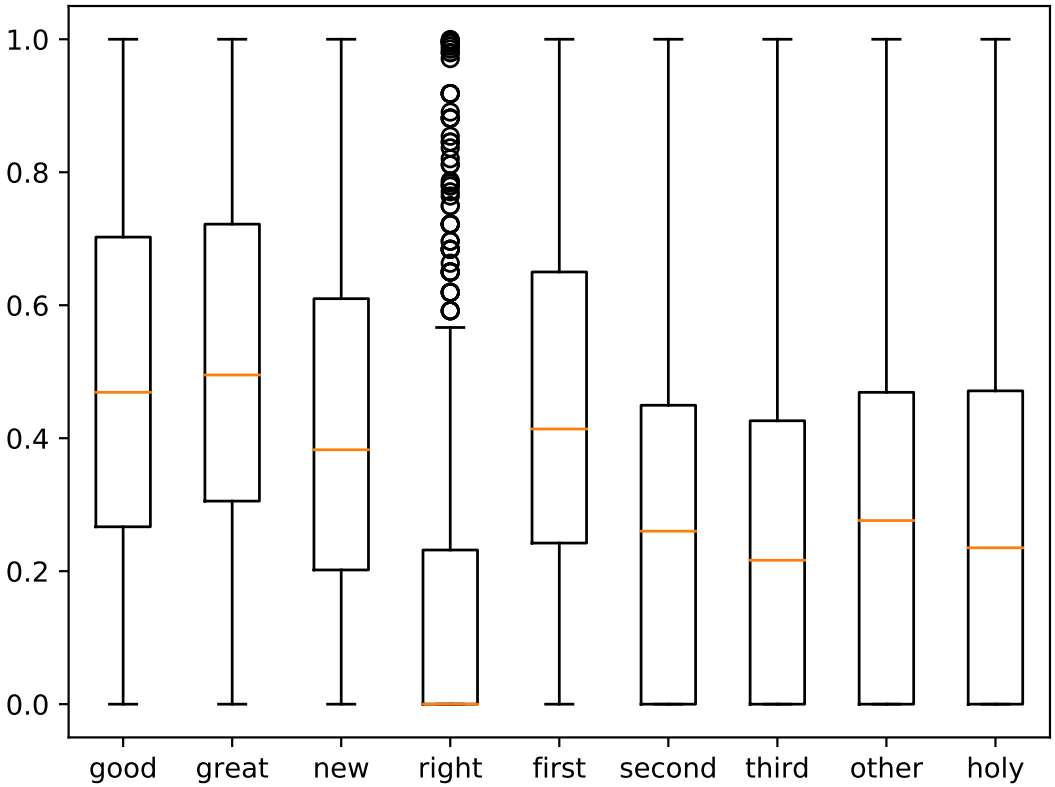




"first" is less strict than "second" on average

Stockholm University

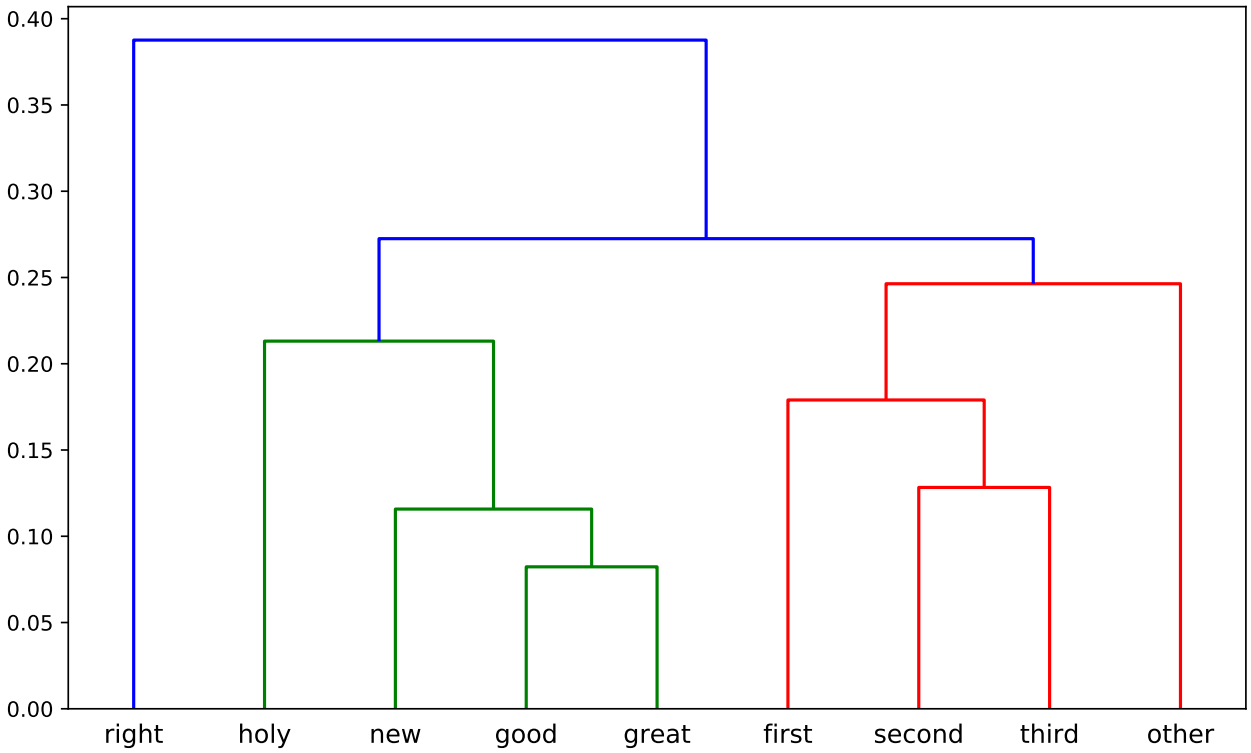
Variability within languages



Word order entropy per adjective-like concept across languages

Stockholm University

Syntax vs Semantics



Adjective-like concepts clustered by word order across languages

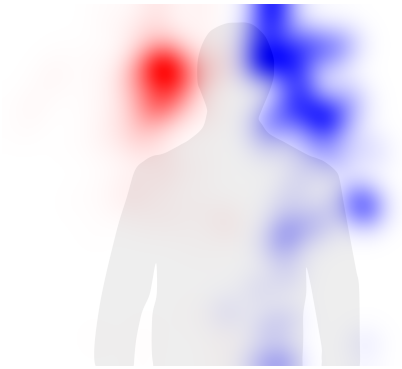
Stockholm University

Sign languages

- In many ways similar to spoken languages
- Balance of constraints is somewhat different
- *Iconicity* plays a larger role
- Ease of learning *vs* ease of use
  Example: *head*, *belly*, *foot*

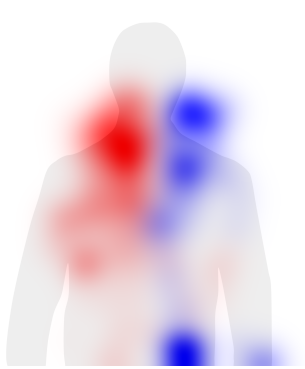
Stockholm University

Iconicity



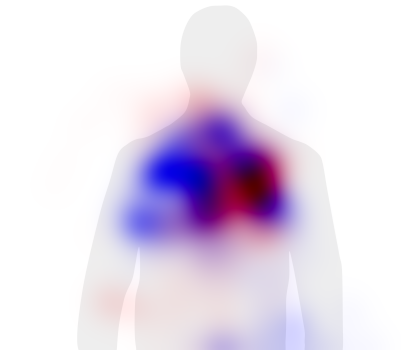
THINK (28 signs)

HEAR (25 signs)

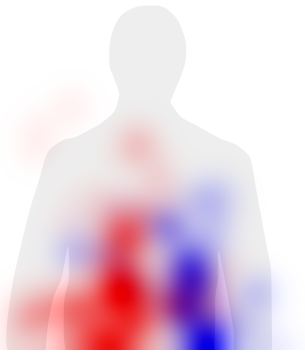


SAY (25 signs)

FOOD (25 signs)



LOVE (25 signs)

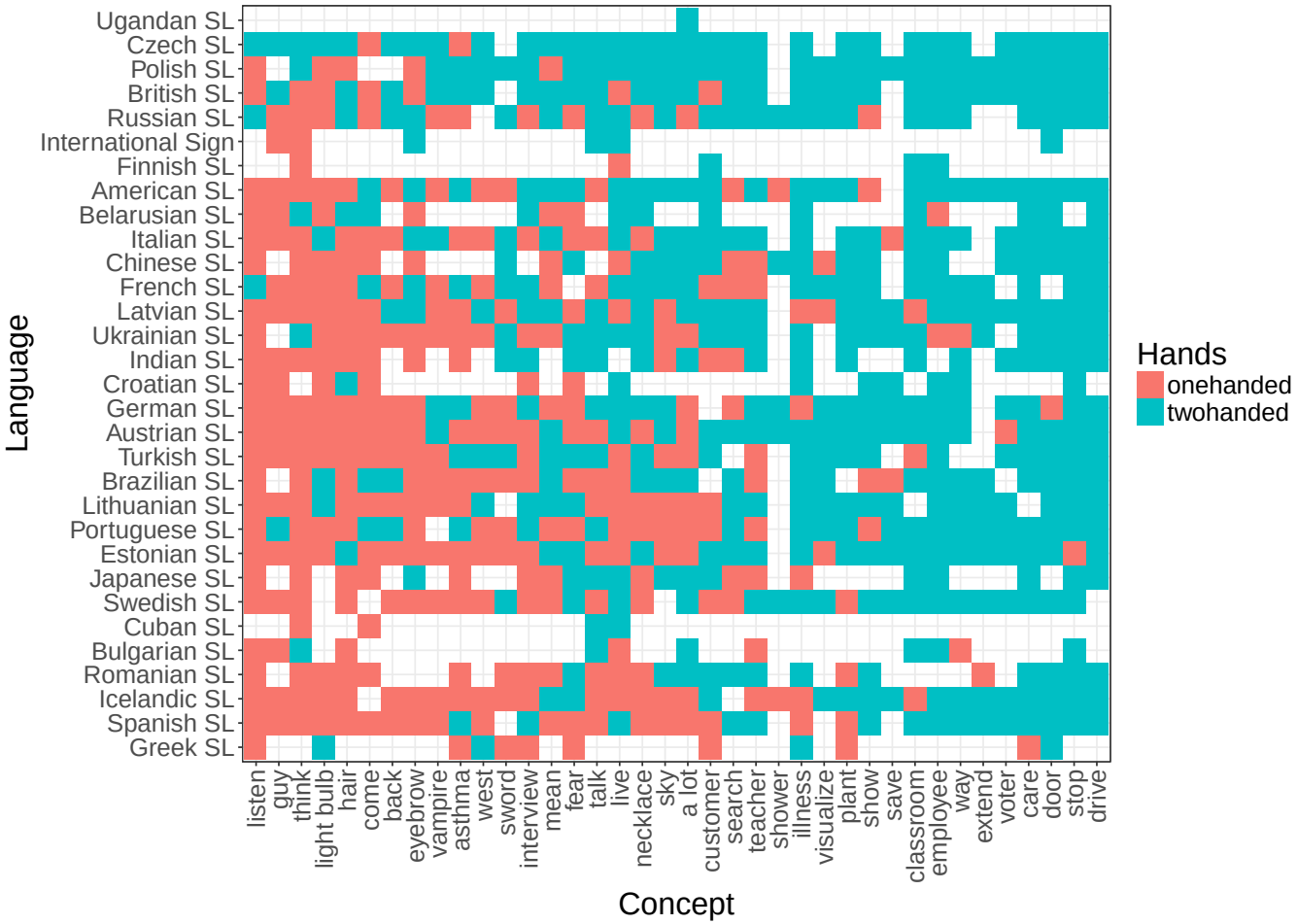


HUNGRY (24 signs)

Place of articulation for some concepts across sign languages

Stockholm University

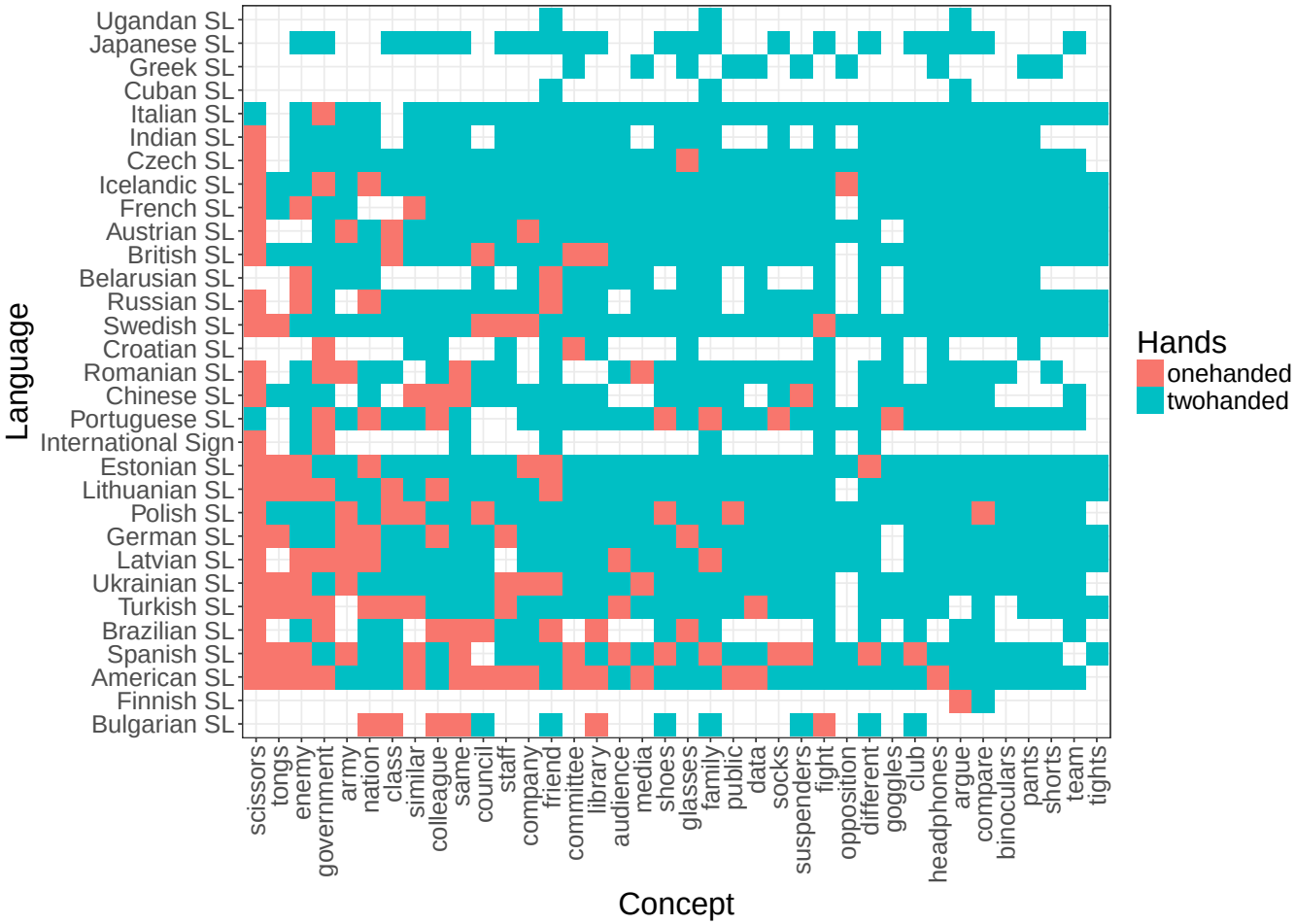
One factor: inherent plurality



Number of hands used per concept per language, non-plural concepts

Stockholm University

One factor: inherent plurality



Number of hands used per concept per language, plural concepts

Stockholm University

Summary

- Languages are fascinating
- There are tons of parameters you can extract
- The Bible is an excellent data source with structured variation
- Start calculating!

$$\begin{bmatrix} L_1(x_1) & L_1(x_2) & \dots & L_1(x_n) \\ L_2(x_1) & L_2(x_2) & \dots & L_2(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ L_m(x_1) & L_m(x_2) & \dots & L_m(x_n) \end{bmatrix}$$

Stockholm University