Introduction
Motivations

- I come from computer science
- Why do I think linguistics is interesting?
- How do I look at languages?
Linguistics: what and why?

- Language: a mapping, thoughts $\leftrightarrow$ 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
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
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)

\[
\text{language} = \text{dialect} + \text{army} + \text{navy}
\]  

(1)

Fundamental theorem of language politics
Number of speakers

Size of language communities in the world

Robert Östling
Linguistics, mathematics and the Bible
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)
Syntax and morphology

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

<table>
<thead>
<tr>
<th>English</th>
<th>Mandarin</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP → V NP</td>
<td>VP → V NP</td>
</tr>
<tr>
<td>V → drink</td>
<td>V → he</td>
</tr>
<tr>
<td>N_{plu} → N s</td>
<td>N_{plu} → N</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
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)

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.
Linguistic typology

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

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

<table>
<thead>
<tr>
<th>Concept</th>
<th>English</th>
<th>Swedish</th>
<th>Mandarin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Father’s older brother</td>
<td>uncle</td>
<td>farbror</td>
<td>daye</td>
</tr>
<tr>
<td>Father’s younger brother</td>
<td>uncle</td>
<td>farbror</td>
<td>shu</td>
</tr>
<tr>
<td>Mother’s older brother</td>
<td>uncle</td>
<td>morbror</td>
<td>jiu</td>
</tr>
<tr>
<td>Mother’s younger brother</td>
<td>uncle</td>
<td>morbror</td>
<td>jiu</td>
</tr>
</tbody>
</table>
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
Data... from where?

- Just one problem, how to get this data?
  1. Lots and lots of manual work
  2. Automation
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 $L_i$
- $n \approx 8000$, $m \approx 1500$

\[
\begin{bmatrix}
L_1(x_1) & L_1(x_2) & \ldots & L_1(x_n) \\
L_2(x_1) & L_2(x_2) & \ldots & L_2(x_n) \\
\vdots & \vdots & \ddots & \vdots \\
L_m(x_1) & L_m(x_2) & \ldots & L_m(x_n)
\end{bmatrix}
\]

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

- Let $x_{\text{house}}$ be the sentences about houses.
- Which symbols are most overrepresented in $\{L_i(x_j) | x_j \in x_{\text{house}}\}$? 
  *Now we know the symbol for house in each language!*
- What about $x_{\text{small}}$ (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 $N_{\text{plu}}$ $\rightarrow$ N $\times$ or $N_{\text{plu}}$ $\rightarrow$ $\times$ N*
NP → Adj N or NP → N Adj?

ProtAdjective-Noun order
NP → Num N or NP → N Num?

Numeral-Noun order

Robert Östling
Linguistics, mathematics and the Bible
NP \rightarrow \text{RelCl} \ N \text{ or } NP \rightarrow N \text{ RelCl?}
VP → NP V or VP → V NP?

Object-Verb order
Testing theories

- If languages can be divided into left- and right-branching, these should be strongly correlated.
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?
Can we trust this?

Object-Verb

Number of languages

WALS says VO
WALS says None
WALS says OV

Robert Östling
Linguistics, mathematics and the Bible
Can we trust this?

ProtAdjective-Noun

WALS says NAdj
WALS says None
WALS says AdjN

Number of languages

Robert Östling
Linguistics, mathematics and the Bible
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
Variability within languages

“first” is less strict than “second” on average
Variability within languages

Word order entropy per adjective-like concept across languages
Syntax vs Semantics

Adjective-like concepts clustered by word order across languages
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*
Iconicity

Place of articulation for some concepts across sign languages
One factor: inherent plurality

Number of hands used per concept per language, non-plural concepts
One factor: inherent plurality

Number of hands used per concept per language, plural concepts
Summary

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

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

Robert Östling
Linguistics, mathematics and the Bible