Parallel text typology

ETT2024 master class

Robert Östling
robert@ling.su.se

Institutionen för lingvistik
Stockholms universitet

2024-06-12

Stockholm University

Parallel texts

English	Likewise , every good tree bears good fruit ,
Swedish	Så bär varje gott träd bra frukt ,
German	So bringt jeder gute Baum gute Früchte ,
Chinese	照樣 ， 凡 好 樹 都 結 好 果實 ，

Plus over 1500 other languages!

Parallel texts

English	Likewise , every good tree bears good fruit ,
Swedish	Så bär varje gott träd bra frukt ,
German	So bringt jeder gute Baum gute Früchte ,
Chinese	照樣 ， 凡 好 樹 都 結 好 果實 ，

Plus over 1500 other languages!

How can we make the most out of this?

Alignment

- The core technology here is *word alignment*
- Translation-equivalent words (or better: morphemes) are linked across languages
- A difficult problem (I wrote a whole thesis on this!)
- What can we learn from word-aligned parallel texts?

Parallel texts

English	Likewise , every good$_1$ tree$_2$ bears$_3$ good$_4$ fruit$_5$,
Swedish	Så bär$_3$ varje gott$_1$ träd$_2$ bra$_4$ frukt$_5$,
German	So bringt$_3$ jeder gute$_1$ Baum$_2$ gute$_4$ Früchte$_5$,
Chinese	照樣 ， 凡 好$_1$ 樹$_2$ 都 結$_3$ 好$_4$ 果實$_5$ ，

Parallel texts

English	Likewise , every good$_1$ tree$_2$ bears$_3$ good$_4$ fruit$_5$,
Swedish	Så bär$_3$ varje gott$_1$ träd$_2$ bra$_4$ frukt$_5$,
German	So bringt$_3$ jeder gute$_1$ Baum$_2$ gute$_4$ Früchte$_5$,
Chinese	照樣 ， 凡 好$_1$ 樹$_2$ 都 結$_3$ 好$_4$ 果實$_5$ ，

Assume we know how to analyze English, we can directly infer:

- Order between 1 and 2: AdjN in all cases
- Order between 2, 4, 5: SVO vs VSO

Remember, this is for *one* transitive clause and *one* adjectival modification. We learn more by computing summary statistics over whole texts.

Summary tables

	VO	OV	NAdj	AdjN
English	624	28	23	637
Swedish	489	34	27	417
German	304	437	36	770

...

What are those few casse of unexpected OV or NAdj order? In most cases, evidence that the alignment procedure is not perfect.

Further conclusions

We can also gather the set of word forms for each English form:

English	Swedish	Chinese
good	god, gott, goda	好
fruit-∅.SG	frukt, frukten	果實
fruit-s.PL	frukter, frukterna	果實

What can we extrapolate from these and similar examples?

- Chinese nouns are not inflected for number (fruit = fruits)
- Swedish nouns mark number, but also something else (definiteness, which can be seen by the association with English *the*)
- Swedish adjectives are inflected (for gender and number)

A challenge

Dear GramBank people... or anybody?

Code a few languages by referring only to a Bible translation instead of a reference grammar.

- How much time did this take?
- How many features are (im)possible to figure out?
- What is the inter-annotator agreement?

Automation

Dear self

I know you all too well, you do not have the patience, focus or time to be a GramBank coder, so please figure out a way to do the above automatically.

- So far, we (Amanda, Bernhard, Östen, myself and others) have used simple heuristics and statistical patterns. Results will be presented soon.
- LLMs are becoming much better at general linguistics recently, to what extent can we now emulate the human analysis process?

Resources

OPUS

- Collection of public parallel corpora:
  https://opus.nlpl.eu/
- Search for translation equivalents:
  https://opus.nlpl.eu/legacy/lex.php

Bible-derived data

- Publication and repository:
  https://doi.org/10.1162/coli_a_00491
  https://zenodo.org/records/7506220
- For this class:
  http://robos.org/ett2024/