

Multilingual Corpora

Zdeněk Žabokrtský

📅 December 11, 2024



Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



unless otherwise stated

Linguistic typology

- in general, linguistic typology studies the ways in which languages vary
- e.g. researchers analyze similarities and differences in grammatical patterns across languages
- multilingual corpora – valuable resources for conducting such research in a systematic and data-driven way

A terminological note

- a **multilingual** corpus – any collection of texts in more than one language
- a **parallel** corpus – a collection of texts and their corresponding translations (possibly with document/sentence/word level alignment)
- a **comparable** corpus – something in between: texts from the same domains in multiple languages; not translations of each other, just thematically related
- a **code-switching** corpus – technically also multilingual, but something very different: a collection in which two or more languages alternate in a single text (possibly even in a single sentence)

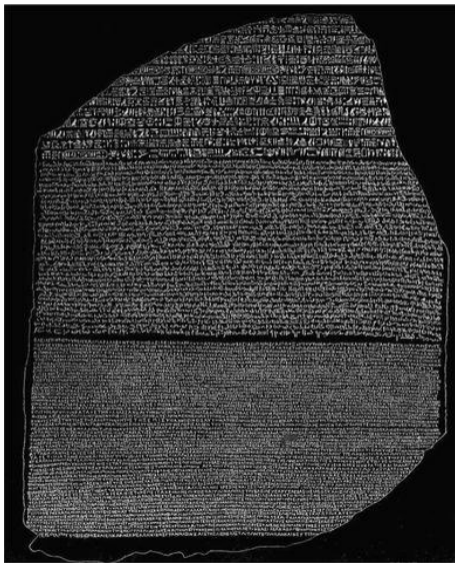
Parallel corpora

Other multilingual non-parallel corpora

Homework

Parallel corpora

The mandatory cliché introduction - the Rosetta Stone



Factors in design of parallel corpora

- which languages; bilingual vs. multilingual
- unidirectional vs. bidirectional
 - Why to consider translation direction? E.g. because of studying “translationese”
- included data sources
 - as in any other corpus: notoriously problematic notion of representativeness/balance
 - authentic vs. synthetic data
- provided alignment
 - document level
 - sentence level
 - word level
 - subword/morpheme level (very rare)
- additional markup: lemmas, POS, syntactic trees ...

The Bible Corpus

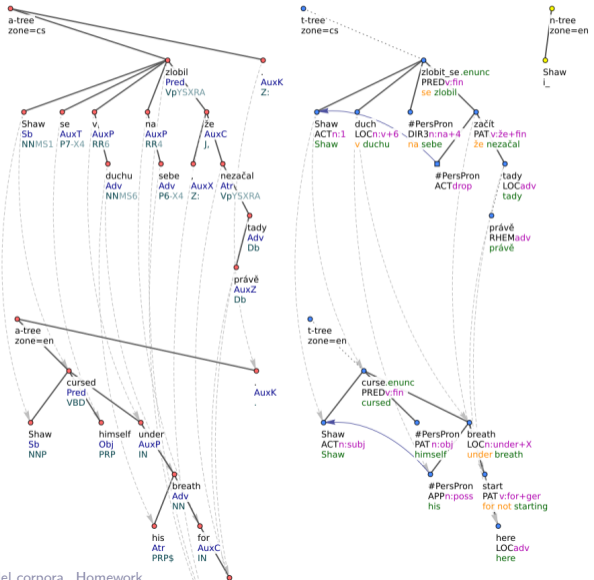
- translations of the Bible.
- easily aligned: verse-parallel alignments
- massively multilingual: 1,600 languages from 90 language families in the JHU version, 4,000 translations (McCarthy, 2020)
- 0.8 MW in the English Bible translation

- (Agic, Vulic, 2019)
- a complete crawl of jw.org (Jehovah's Witnesses)
- 300 languages, 100 kS per language pair, 1,4 GW in total

- 61 languages (in 2023), 5GW in total
- parallel search possible: <https://treq.korpus.cz>
- word-to-word alignment (manual alignment for the “core” part, the rest aligned automatically)

- a Czech-English parallel corpus
- collected at ÚFAL for MT purposes
- in 2.0: 61 M sentence pairs (0.7 GW in English, 0.6 GW in Czech)
- automatically parsed in the Prague Dependency Treebank style

CzEng, cont.



- proceedings of the European Parliament
- 21 European languages: Romanic (French, Italian, Spanish, Portuguese, Romanian), Germanic (English, Dutch, German, Danish, Swedish), Slavik (Bulgarian, Czech, Polish, Slovak, Slovene), Finni-Ugric (Finnish, Hungarian, Estonian), Baltic (Latvian, Lithuanian), and Greek.
- 60 MW/language

- The Acquis Communautaire – the total body of EU law
- JRC = Joint Research Center, a European Commission's science and knowledge service
- JRC-Acquis – a parallel corpus of EU legal texts from 1950 to now
- 22 languages: Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Maltese, Dutch, Polish, Portuguese, Romanian, Slovak, Slovenian and Swedish

The OPUS Corpus

- a collection of translated texts from the web and from many already existing parallel corpora
- in 2020:
 - 57 source corpora
 - 700 languages
 - 70,000 bitexts
- online search interface <https://opus.nlpl.eu/>

OPUS search interface

← → ↻ 🏠 <https://opus.nlpl.eu> 120% ☆ 🔍 Search

Home / Query / WordAlign / Wiki

[ada83] [bible] [bianet] [books] [CCAligned] [CCMatrix] [CAPES] [DGT] [DOGC] [ECB] [EhuHac] [EITB] [Elhuyar] [Ei
[finlex] [fiskmõ] [giga] [GNOME] [GlobalVoices] [hren] [infopankki] [JRC] [KDE4/doc] [liv4ever] [MBS] [memat] [Monten
[NC] [Ofis] [OO/OO3] [subs/16/18] [Opus100] [ParaCrawl] [ParCor] [PHP] [QED] [sardware] [SciELO] [SETIMES] [SF
[UNPC] [WikiMatrix] [Wikimedia] [Wikipedia] [WikiSource] [WMT,

ORPUS ... the open parallel corpus

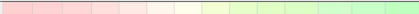
OPUS is a growing collection of translated texts from the web. In the OPUS project we try to convert and align free online data, to add linguistic annotation, and to provide the community with a publicly available parallel corpus. OPUS is based on open source products and the corpus is also delivered as an open content package. We used several tools to compile the current collection. All pre-processing is done automatically. No manual corrections have been carried out.

The OPUS collection is growing! Check this page from time to time to see new data arriving ... Contributions are very welcome! Please contact <opus-project@helsinki.fi >

Search & download resources: show all versions

Language resources: click on [tmx | moses | xces | lang-id] to download the data! (raw = untokenized, ud = parsed with universal dependencies, alg = word alignments and phrase tables)

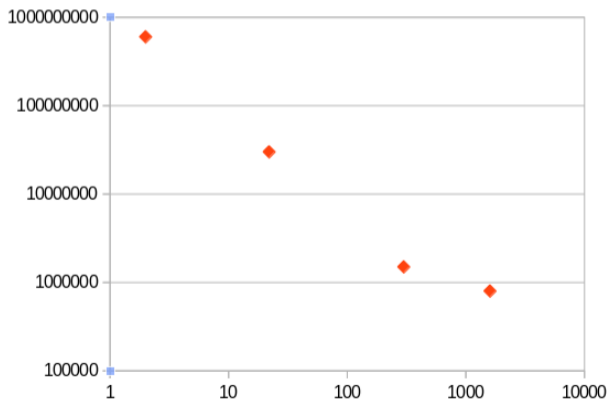
corpus	doc's	sent's	cs tokens	mn tokens	XCES/XML	raw	TMX	Moses	mono	raw ud	alg	dic	freq	other files
GNOME v1	649	0.2M	1.2M	1.2M	xces cs mn	cs mn	tmx	moses	cs mn	cs mn	alg smt		cs mn	sample
MultCCAligned v1.1	1	0.2M	138.7M	5.5M	xces cs mn	cs mn	tmx	moses	cs mn	cs mn			cs mn	sample
XLEnt v1.2	1	52.0k	0.1M	0.1M	xces cs mn	cs mn	tmx	moses	cs mn	cs mn			cs mn	sample
QED v2.0a	219	22.5k	0.3M	0.2M	xces cs mn	cs mn	tmx	moses	cs mn	cs mn	alg smt	dic	cs mn	sample
TED2020 v1	181	17.2k	0.3M	0.3M	xces cs mn	cs mn	tmx	moses	cs mn	cs mn	alg smt	dic	cs mn	sample
NeuLab-TedTalks v1	114	7.1k	0.1M	0.1M	xces cs mn	cs mn	tmx	moses	cs mn	cs mn			cs mn	sample
Mozilla-110n v1	1	1.0k	0.4M	7.5k	xces cs mn	cs mn			cs mn	cs mn			cs mn	sample
total	1166	0.5M	141.1M	7.5M	0.5M		0.3M	0.3M						

color: 
size (src+trg): 16.4k 32.8k 65.5k 0.1M 0.3M 0.5M 1.0M 2.1M 4.2M 8.4M 16.8M 33.6M 67.1M 134.2M

- 100 languages
- an English centric version – a selection of texts whose translations are available also in English
- 44 languages with at least 1M sentences (sentence pairs with English)
- 55M sentence pairs in total

Size or size?

- Obvious trade-off: number of languages vs. size of parallel data (words per lang)
- sample: CzEng, JRC-Acquis, JW300, Bible



Synthetic parallel corpora

- authentic data limited in size, need for data augmentation
- an old idea in Machine Translation: back-translation
 - you need huge training data for MT training from language A to language B
 - you take monolingual data in B and translate them back to A
 - and you use it as parallel training data for A-to-B direction
 - the trick: good quality of the B side; quality of the A side less important
 - possible modification: filter the back-translated data, i.e. keep only reasonably looking sentence pairs

Other multilingual non-parallel corpora

Common Crawl

- growing steadily: 200-300 TB/month
- some 160 languages recognized
- in total tens of PB (?)
- massive amount of comparable content can be expected, but no obvious alignment

Universal Dependencies

- 100+ languages
- a common annotation scheme – syntactic trees
- mostly tens to hundreds kW per language (max a few MW/language)

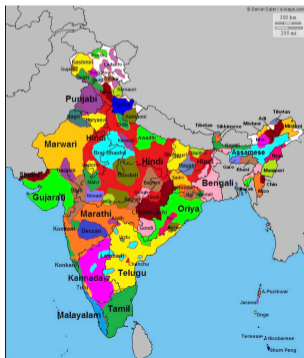
- created at ÚFAL
- coreference + UD
- coreference – a relation between two or more expressions (“mentions”) in a text that refer to the same entity (or event)
- 21 coreference-annotated datasets for 13 languages
- coreference annotation combined with (automatic or manual) dependency syntax annotation

Code switching corpora

- code switching - alternation between two or more languages in a single utterance
- distinct from other language-contact phenomena such as lexical borrowing or calques
- English-Hindi, English-Spanish, Turkish-German ...

HinDialect - a dialect-continuum corpus

- a collection of folksongs for 26 languages that form a dialect continuum in North India and nearby regions
- Angika, Awadhi, Baiga, Bengali, Bhadrawahi, Bhili, Bhojpuri, Braj, Bundeli, Chhattisgarhi, Garhwali, Gujarati, Haryanvi, Himachali, Hindi, Kanauji, Khadi Boli, Korku, Kumaoni, Magahi, Malvi, Marathi, Nimadi, Panjabi, Rajasthani, Sanskrit.
- see <https://lindat.cz>



Multilingual data and language universals

- language universals – statements that hold for all languages
- recall: absolute vs. statistical universals, implicational vs. unconditional universals

Time for an exercise

- given the data resources you learned about, can you verify/falsify previously proposed language universals? (e.g. Greeberg's universals)
- examples:
 - Languages with dominant VSO order are always prepositional.
 - If the nominal object always precedes the verb, then verb forms subordinate to the main verb also precede it.
 - If either the subject or object noun agrees with the verb in gender, then the adjective always agrees with the noun in gender.
 - Negation markers in sentences are almost always placed near the verb.
 - If a language has gender categories in the noun, it has gender categories in the pronoun.
 - The most common syllable structure is CV.
 - All languages have some way of forming questions.

Homework

Homework specification

- Download a parallel corpus (any).
- Find out how exactly the alignment is represented in that corpus.
- Find pairs of aligned sentences whose lengths (in terms of the number of words) differ substantially.
- Try to explain the main reasons behind the differences. Are they typologically grounded, or is it rather noise resulting from automatic processing, or ... ?
- Summarize and exemplify your findings in a 0.5-1 page report (PDF).
- Create directory hw3 in your git repo, and submit the report into it.
- Optional: you can submit also small data samples which you find relevant for your report.