NPFL150

Multilingual Corpora

Zdeněk Žabokrtský

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

Linguistic typology

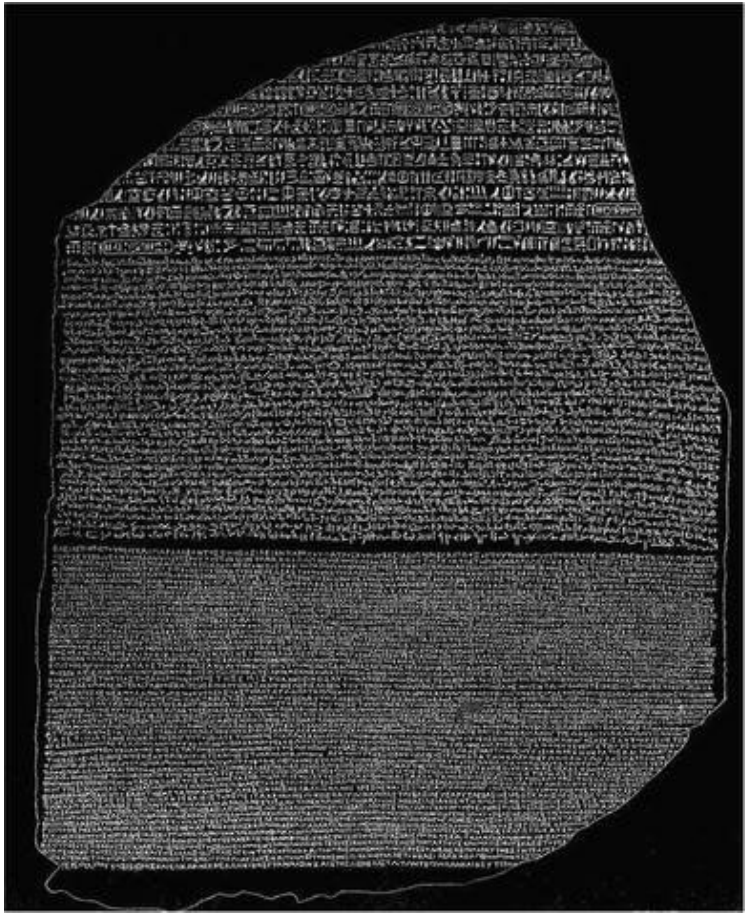
- in general, linguistic typology studies the ways in which languages vary
- e.g. researchers analyze similarities and differences in grammatical patterns across languages
- multilingual corpora - valuable resources for conducting such research in a systematic and data-driven way

Different kinds of multilingual content

- 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)

Selected parallel corpora

The mandatory cliché introduction



The Rosetta Stone

- a granodiorite stele, during the reign of Ptolemy V of Egypt, 196 BCE
- the same (administrative) text in three scripts
    - Hieroglyphic (used for formal religious texts)
    - Demotic (everyday Egyptian script)
    - Ancient Greek (administrative language of the Ptolemaic rulers)
- found in 1799 by French soldiers during Napoleon's campaign in Egypt
- because scholars already understood Ancient Greek, they could use it to decode the meaning of the hieroglyphs (1822, Champollion)

Basic factors in designing parallel corpora

- how many languages
  - bilingual vs. multilingual
- translation direction
  - unidirectional vs. bidirectional
  - (Why to consider translation direction? E.g. because of studying “translationese”)
- authentic vs. synthetic data
- • provided alignment
- • document level

Typical sources of parallel texts

- International organizations often publish documents in several official languages (e.g. UN, EU)
- Government’s documents - in countries with multilingual policy (Kanada, Switzerland, India)
- Religious texts
- Literary translations
- Movie subtitles
- Software localization, technical documentation
- News, international media
- Webcrawling & mining parallel sentences, etc.

Note: like with all other types of corpora, the notorious problem of representative/balanced mixture of sources

Annotation in parallel corpora

- Specific for parallel corpora: **alignment** - linking corresponding units of text across two or more languages
    - Document-level alignment (basically always)
    - Sentence-level alignment
    - Word-level alignment

- Other possibilities like with all other corpora
    - Lemmatization, POS tagging, syntactic parsing…

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

JW300

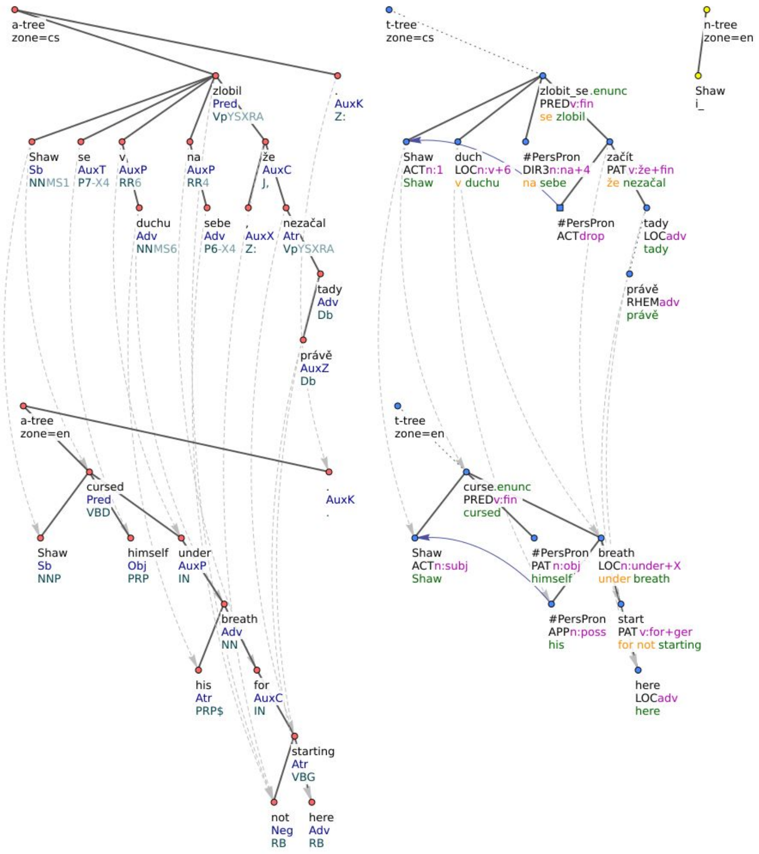
- compiled by (Agic, Vulic, 2019)
- a complete crawl of jw.org (Jehovah's Witnesses)
- 300 languages, 100 kS per language pair, 1,4 GW in total

InterCorp

- a multilingual parallel corpus maintained by the Czech National Corpus
- 61 languages (47 tagged and lemmatized)
- 5 GW in total
- version InterCorp v16ud - makes use of parts of the Universal Dependencies schema
- parallel search possible: https://treq.korpus.cz

CzEng

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



EuroParl

- proceedings of the European Parliament
- EU’s languages - mostly IE, but still quite diverse
  - Romance (French, Italian, Spanish, Portuguese, Romanian)
  - Germanic (English, Dutch, German, Danish, Swedish)
  - Slavic (Bulgarian, Czech, Polish, Slovak, Slovene)
  - Baltic (Latvian, Lithuanian)
  - Finno-Ugric (Finnish, Hungarian, Estonian)
  - Greek
- 60 MW/language

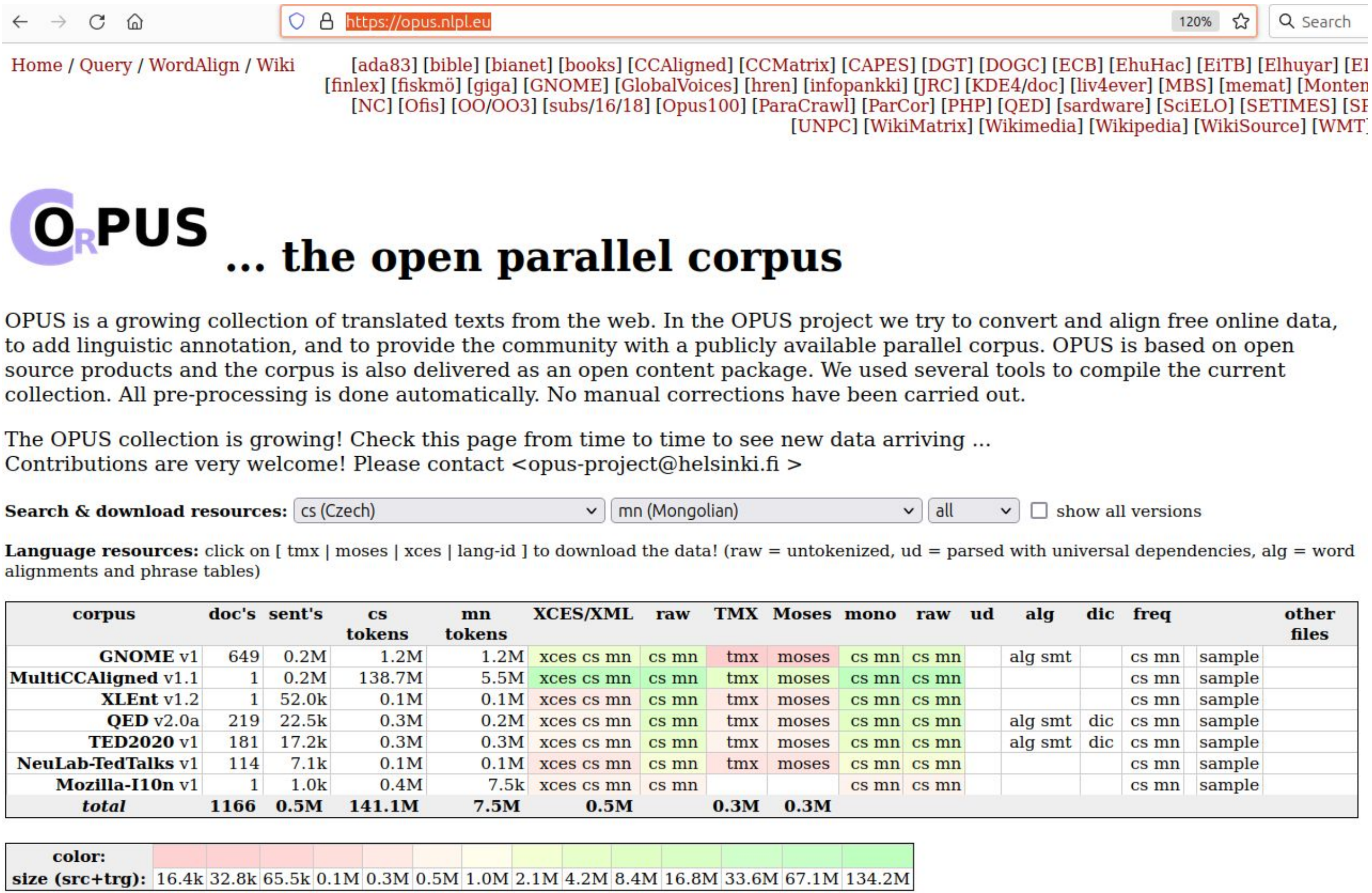
JRC-Acquis

- 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

The OPUS Corpus

- a huge collection of translated texts from the web and from many already existing parallel corpora
- in 2025:
  - 1005 languages
  - 1213 corpora
  - 60G sentence pairs
- online search interface https://opus.nlpl.eu/

OPUS search interface



https://opus.nlpl.eu

Home / Query / WordAlign / Wiki

[ada83] [bible] [bianet] [books] [CCAligned] [CCMatrix] [CAPES] [DGT] [DOGC] [ECB] [EhuHac] [EiTB] [Elhuyar] [EI
[finlex] [fiskmö] [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]

OPUS ... 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: cs (Czech) | mn (Mongolian) | all | □ show all versions

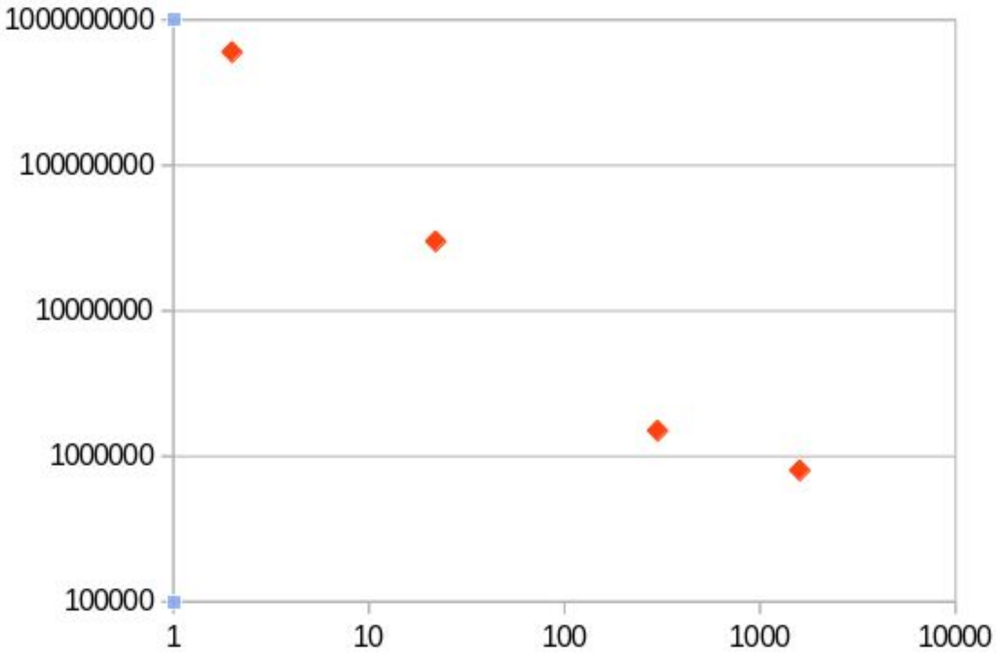
Language resources: click on [tmx | moses | xces | lang-id] to download the data! (raw = untokenized, ud = parsed with universal dependencies, alg = word alignments and phrase tables)

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	
MultiCCAligned 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-I10n 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

A more general glimpse: size or size?

- An obvious tradeoff: either many languages, or many sentence pairs
- Four selected datapoints: 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
-

Some other multilingual corpora, now non-parallel ones

Common Crawl

- growing steadily: 200-300 TB/month
- some 160 languages recognized
- 300 billion pages spanning 18 years
- 3-5 billion pages added each month
- massive amount of comparable content can be expected, but no obvious alignment

Universal Dependencies

- 150+ languages
- a common annotation scheme – POS and other morphological categories, lemmatization, syntactic trees
- mostly tens to hundreds kW per language (max a few MW/language)

Code-switching corpora

- Code-switching - when a speaker alternates between two or more languages, dialects, or language varieties within a conversation, sentence, or even a single utterance

- distinct from other language-contact phenomena such as lexical borrowing or calques

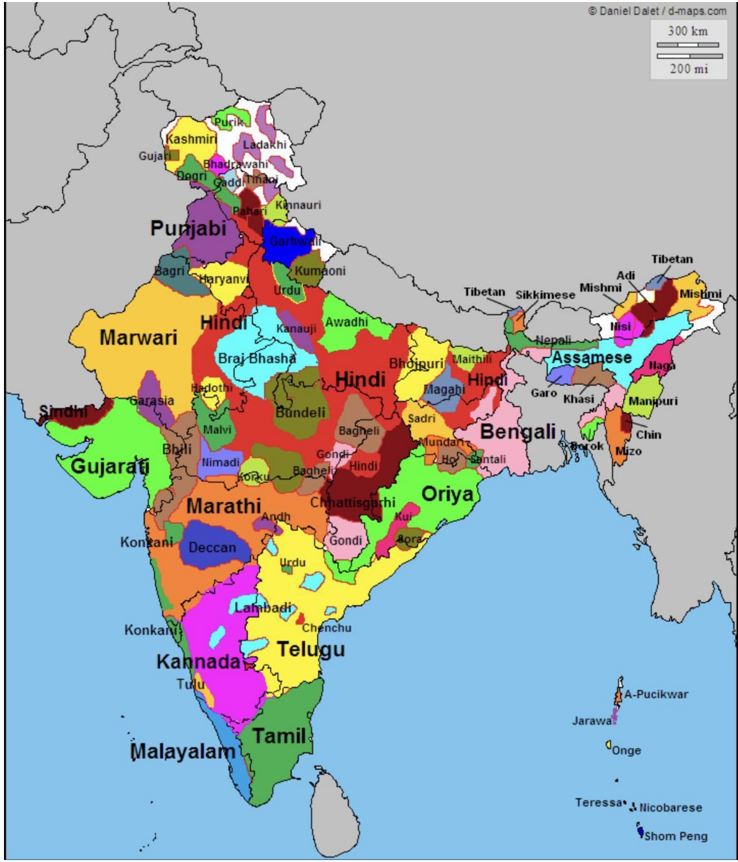
- English-Hindi, English-Spanish, Turkish-German …

HinDialect - a dialect-continuum corpus

- A dialect continuum is - phenomenon in which neighboring dialects of a language (or related languages) are mutually intelligible, but dialects at the extremes of the continuum are not.

- 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.



Multilingual data and language universals

- language universals – statements that hold for all languages
- recall: absolute vs. statistical universals
- implicational vs. unconditional universals
- given the data resources you learned about, can you verify/falsify previously proposed language universals? (e.g. Greeberg’s universals)

Multilingual data and language universals, cont.

- 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 assignment

HW3 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.