Bilingual Sentence Mining for Low-Resource Languages A Case Study on Upper and Lower Sorbian

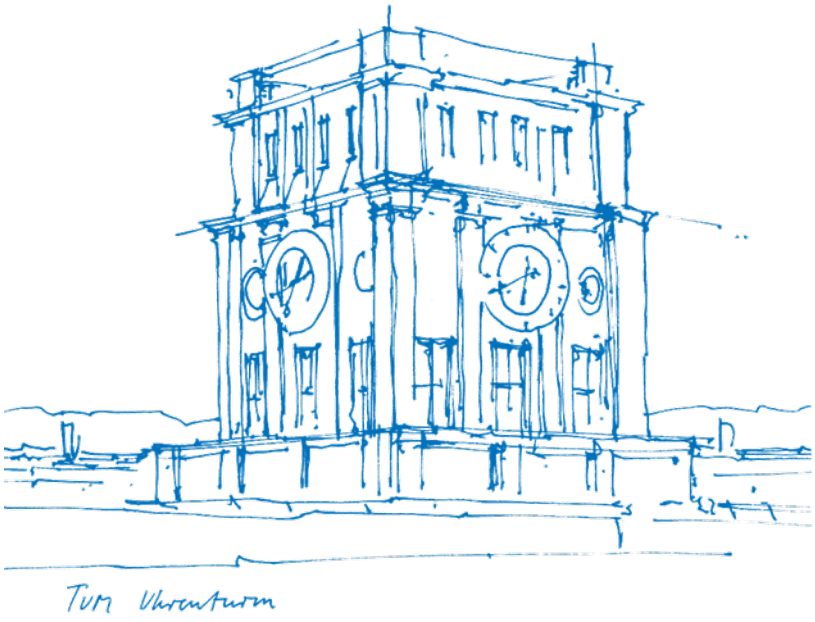
Shu Okabe & Alexander Fraser

Data Analytics & Statistics

Technische Universität München

shu.okabe@tum.de

4th March 2025



Parallel sentence mining

Parallel sentences and parallel corpora

English	German
Please rise, then, for this minute's silence	Ich bitte Sie, sich zu einer Schweigeminute zu erheben
(The House rose and observed a minute's silence)	(Das Parlament erhebt sich zu einer Schweigeminute)
Madam President, on a point of order	Frau Präsidentin, zur Geschäftsordnung
This is all in accordance with the principles that we have always upheld	All dies entspricht den Grundsätzen, die wir stets verteidigt haben

Adapted from the Europarl corpus (European parliament)

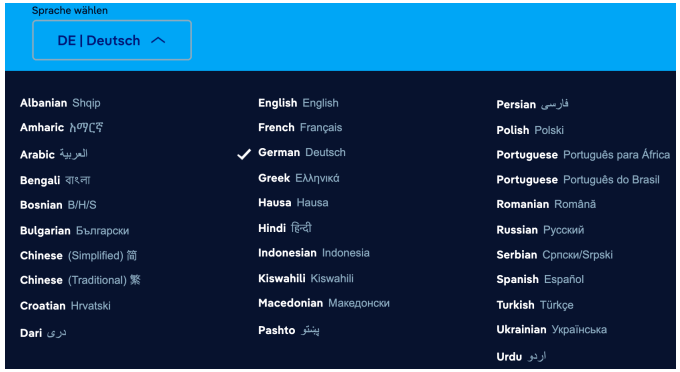
→ Valuable resource for downstream NLP tasks such as Machine Translation

Where can we find parallel sentences?

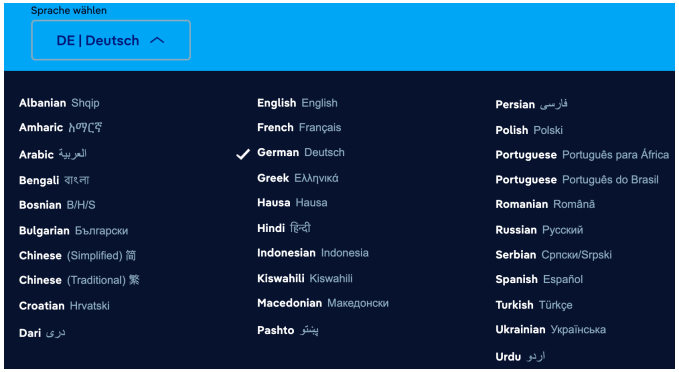
The BBC is in multiple languages

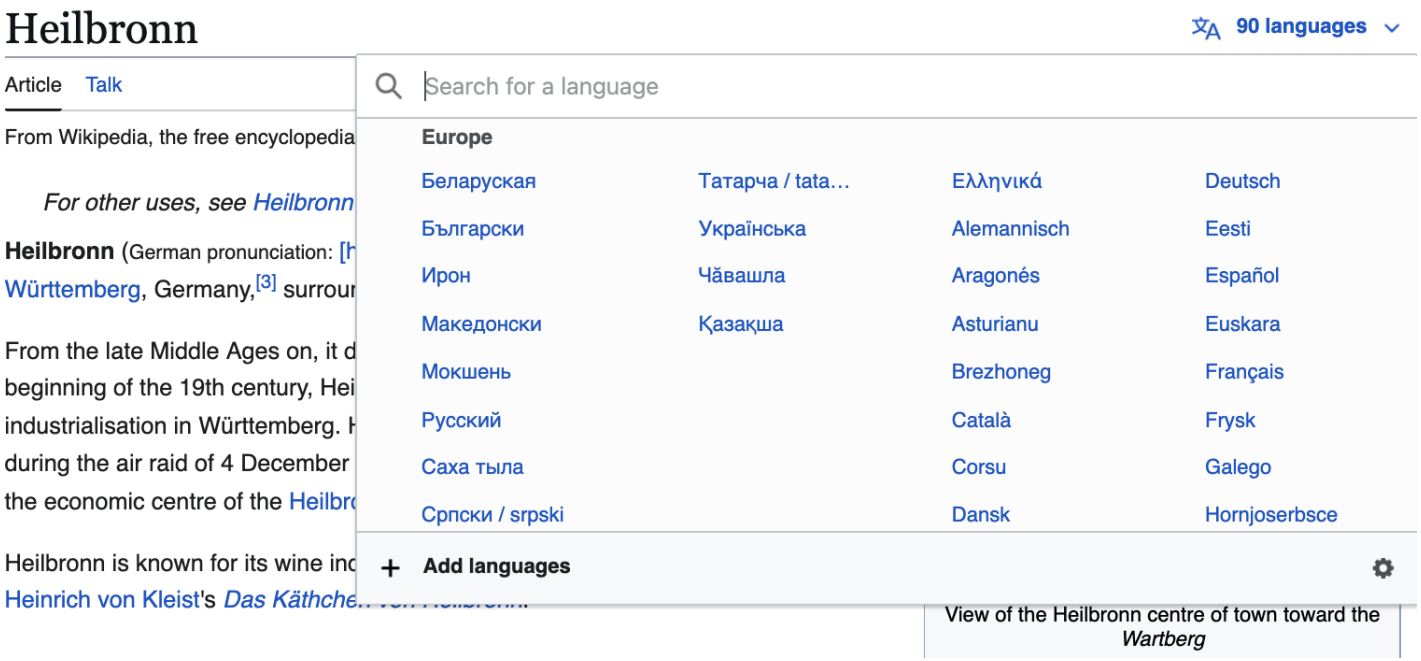
Read the BBC In your own language

Oduu Afaan Oromootiin
Amharic ዜና በአማርኛ
Arabic عربي
Azeri AZƏRBAYCAN
Bangla বাংলা
Burmese မြန်မာ
Chinese 中文网
French AFRIQUE
Hausa HAUSA
Hindi हिन्दी
Gaelic NAIDHEACHDAN
Gujarati ગુજરાતીમાં સમાચાર
Igbo AKỤKỌ N'IGBO
Indonesian INDONESIA
Japanese 日本語
Kinyarwanda GAHUZA
Kirundi KIRUNDI
Korean 한국어
Kyrgyz Кыргыз
Marathi मराठी
Nepali नेपाली
Noticias para hispanoparlantes
Pashto پښتو
Persian فارسی
Pidgin
Portuguese BRASIL
Punjabi ਪੰਜਾਬੀ ਖ਼ਬਰਾਂ
Russian НА РУССКОМ
Serbian NA SRPSKOM
Sinhala සිංහල
Somali SOMALI
Swahili HABARI KWA KISWAHILI
Tamil தமிழில் செய்திகள்
Telugu తెలుగు వార్తలు
Thai ข่าวภาษาไทย
Tigrinya ዜና ብትግርኛ
Turkish TÜRKÇE
Ukrainian УКРАЇНСЬКА
Urdu اردو
Uzbek O'ZBEK
Vietnamese TIẾNG VIỆT
Welsh NEWYDDION
Yoruba ÌRÒYÌN NÍ YORÙBÁ



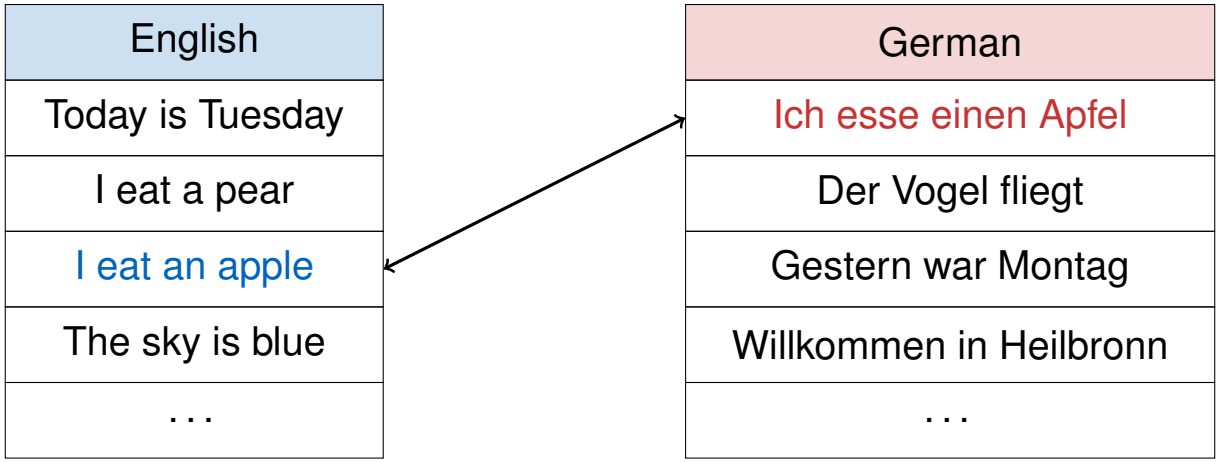
Where can we find parallel sentences?





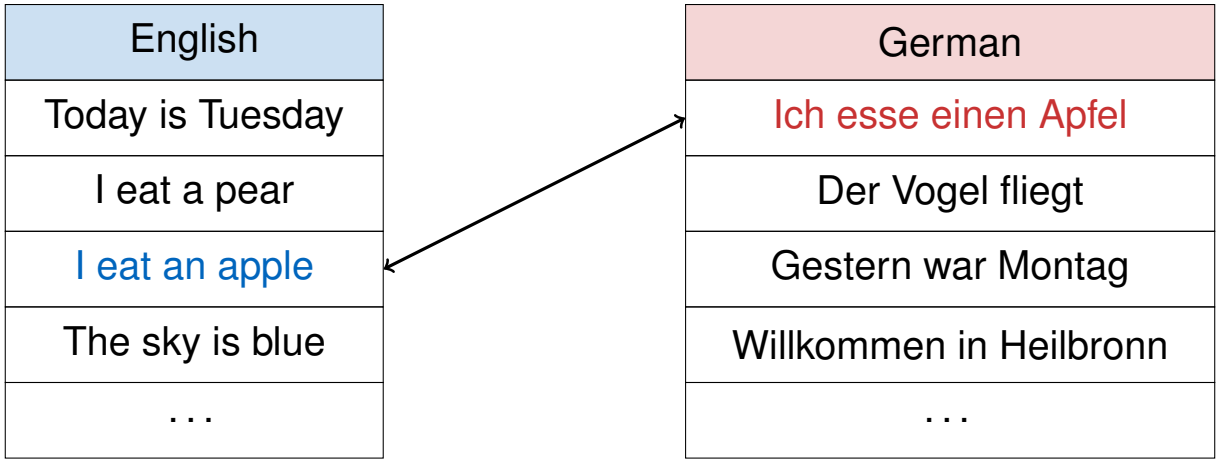
Parallel sentence mining

Extracting parallel sentences from two monolingual corpora



Parallel sentence mining

Extracting parallel sentences from two monolingual corpora

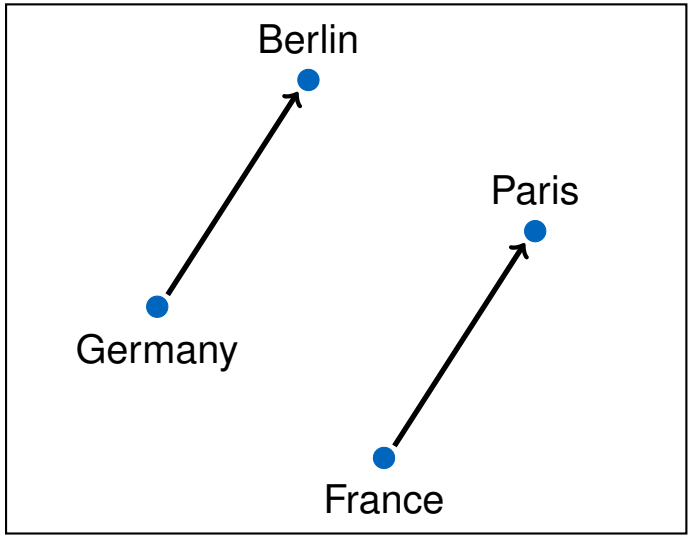


Challenges:

- No guarantee that a sentence has a matching counterpart in the other corpus
- Differences between the two languages

Word and sentence embeddings
Representing words (and sentences) as vectors

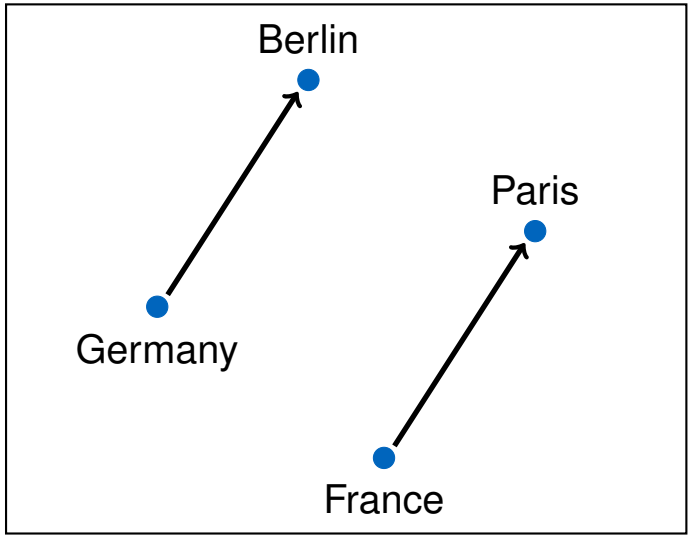
$$v(Berlin) - v(Germany) + v(France) \approx v(Paris)$$



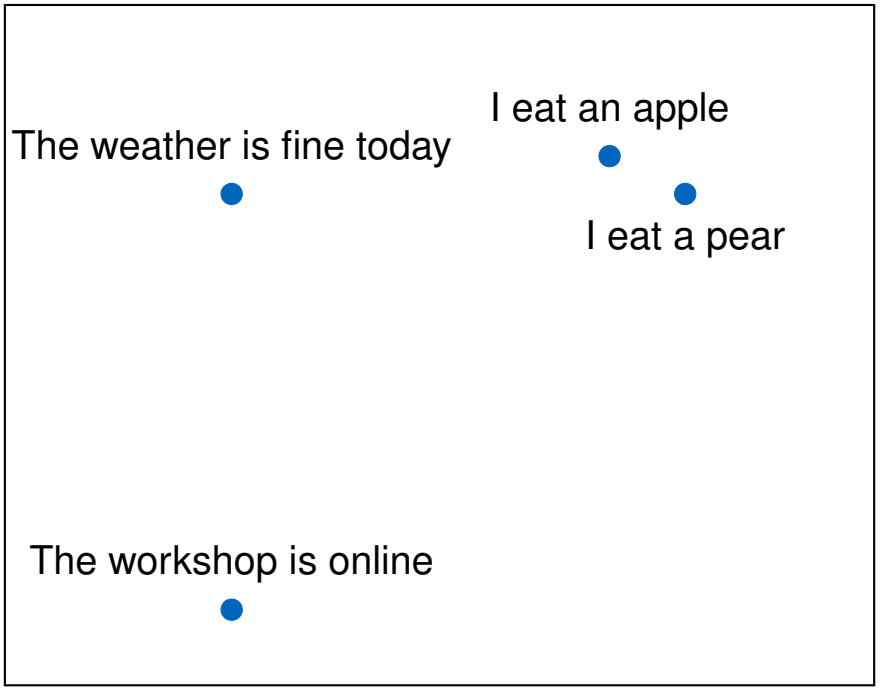
Word embeddings, from (Mikolov et al., 2013)

Word and sentence embeddings
Representing words (and sentences) as vectors

$$v(Berlin) - v(Germany) + v(France) \approx v(Paris)$$



Word embeddings, from (Mikolov et al., 2013)

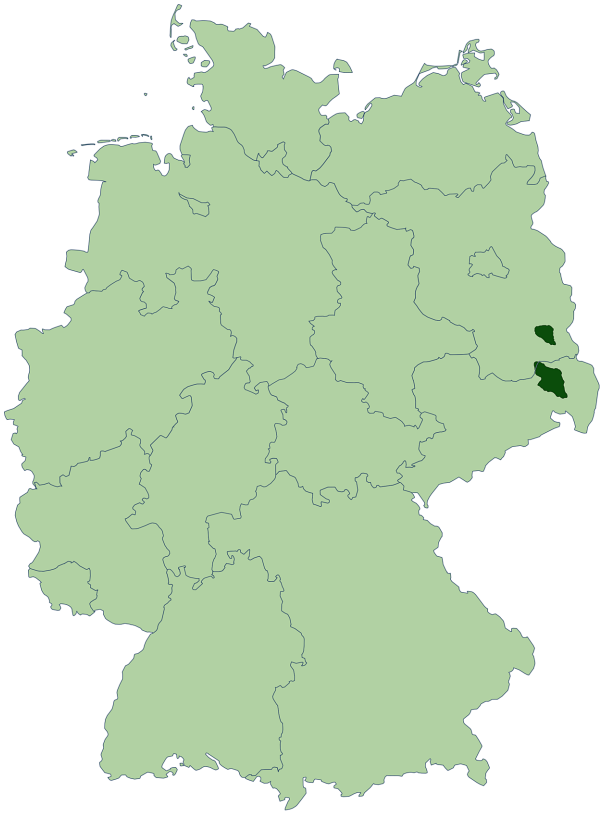


Sentence embeddings

Case study on Sorbian languages

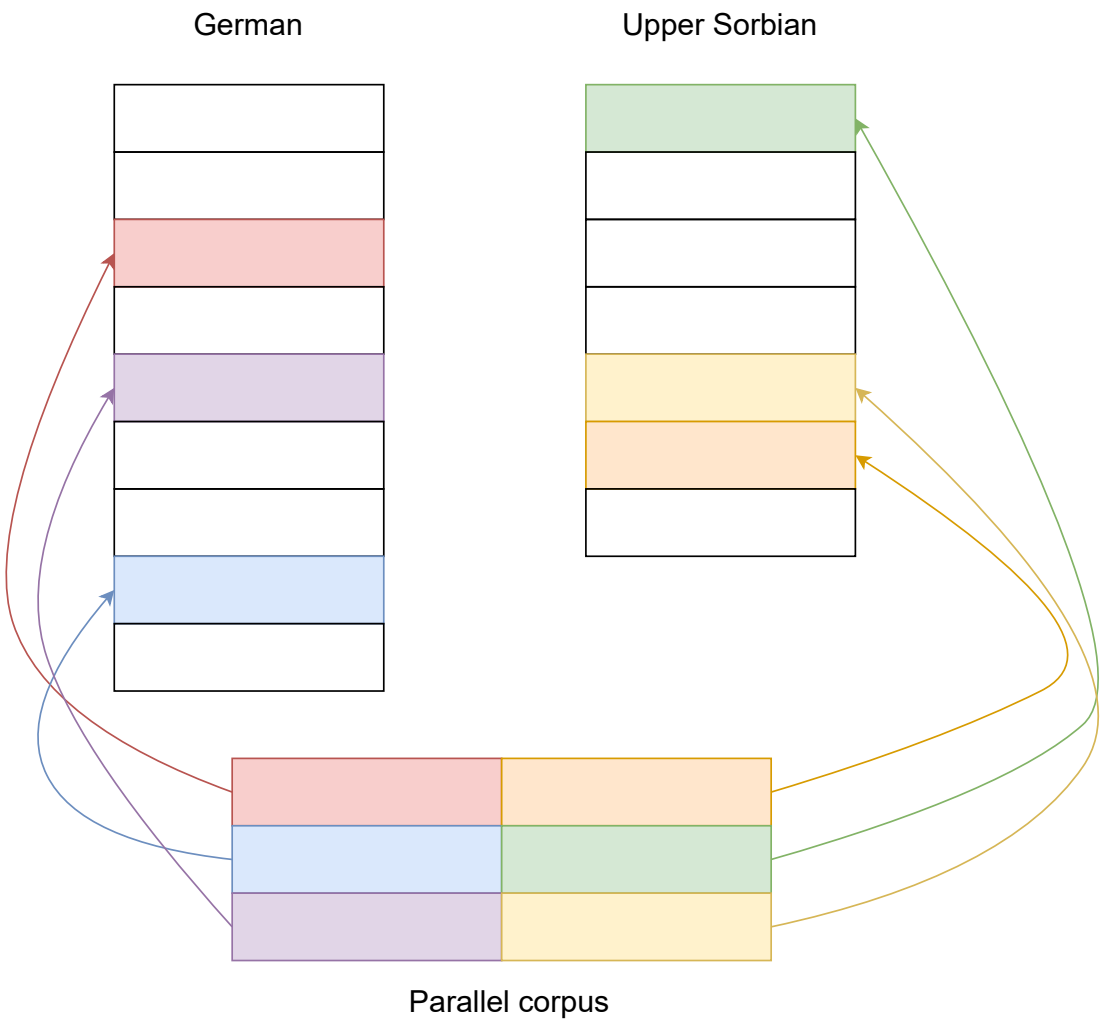
Case study on Upper and Lower Sorbian

Two endangered West Slavic languages (ISO codes: `hsb` and `dsb`)

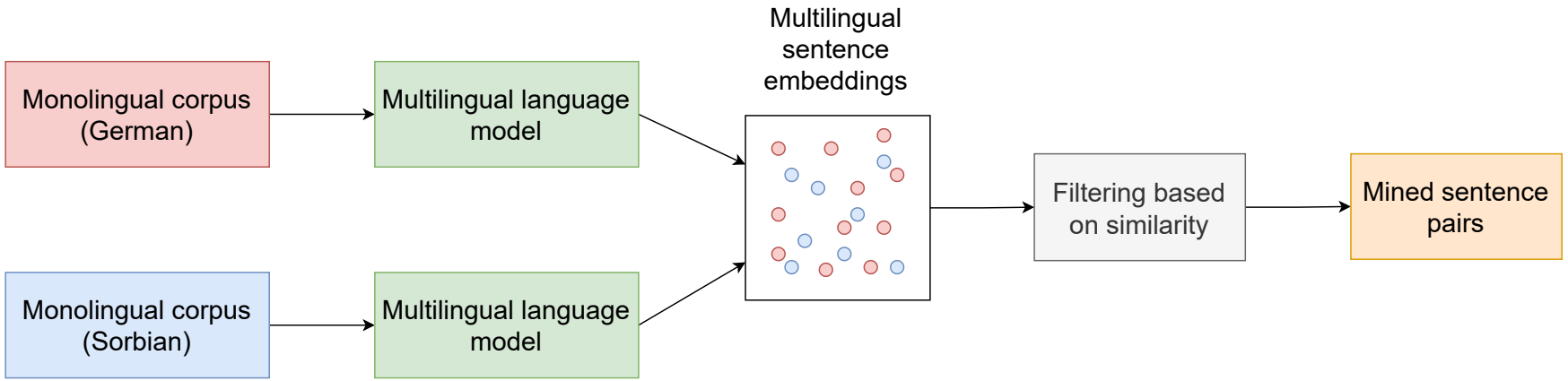


- Previous cooperation with non-profits (e.g., WMT Shared Tasks):
  the Witaj Sprachzentrum (Witaj Language Centre) and the Sorbian Institute

Experimental methodology: corpus creation
Injecting parallel sentences in monolingual corpora



Mining pipeline



Baseline multilingual language models

Sentence embeddings from averaged word embeddings

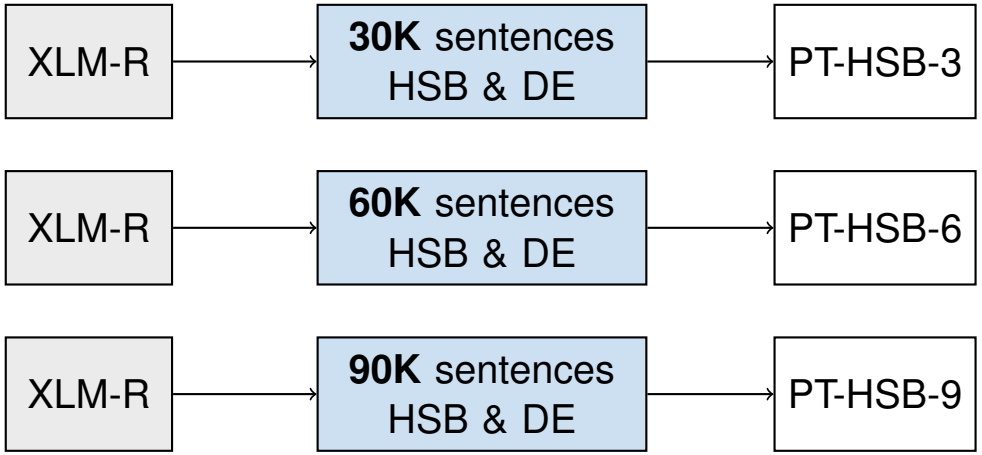
Off-the-shelf baseline multilingual models:

- XLM-R (base): competitive multilingual language model (Conneau et al., 2020)
- Glot500-m: extension of XLM-R to low-resourced languages (Imani et al., 2023)

	XLM-R (base)	Glot500-m
Number of covered languages	100	511
Czech & Polish?	✓	✓
Upper Sorbian?	✗	✓
Lower Sorbian?	✗	✗

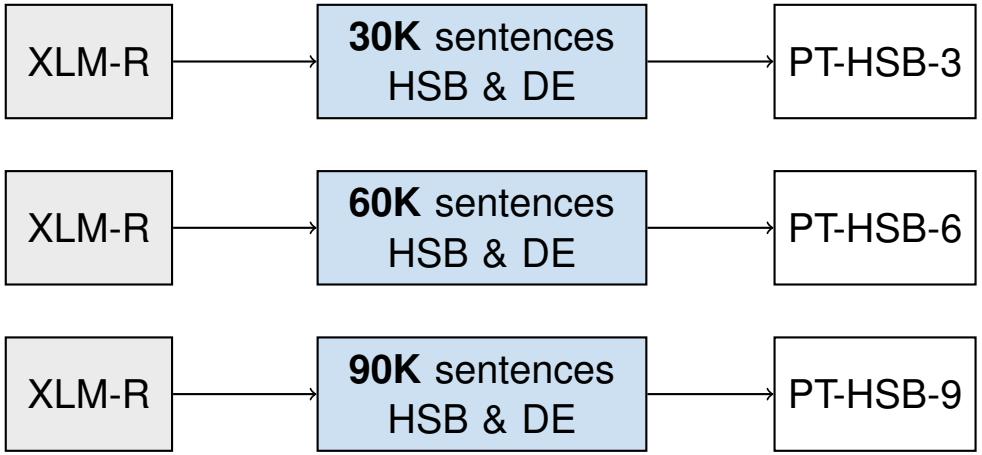
Pre-trained language models
Leveraging available data for Upper Sorbian in XLM-R

- Using German and Upper Sorbian **monolingual** texts
  varying the number of Upper Sorbian sentences

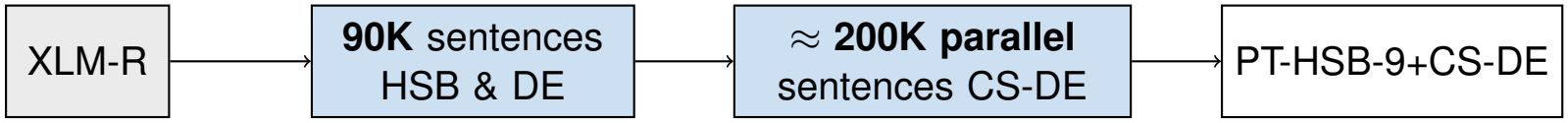


Pre-trained language models
Leveraging available data for Upper Sorbian in XLM-R

- Using German and Upper Sorbian **monolingual** texts
  varying the number of Upper Sorbian sentences

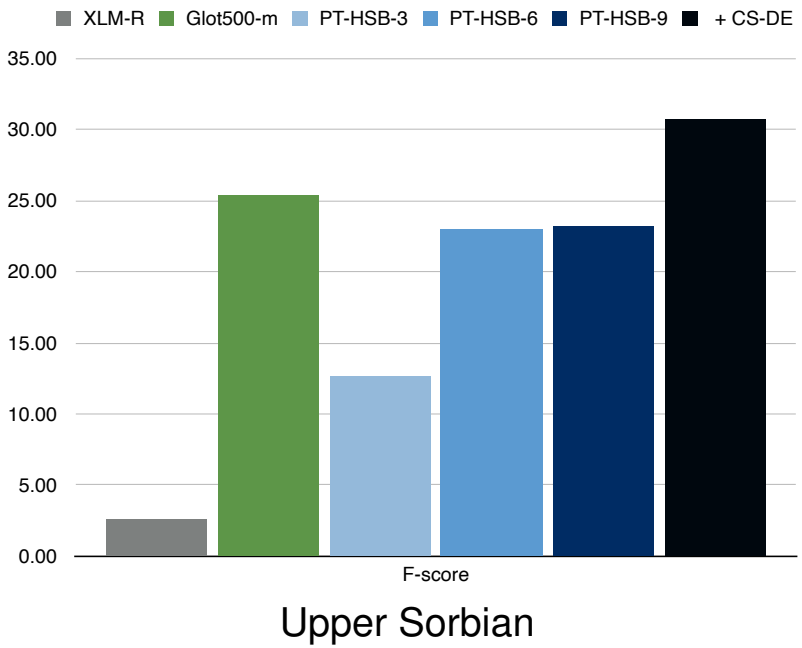


- Using German-Czech **parallel** sentences



Mining results for Upper and Lower Sorbian

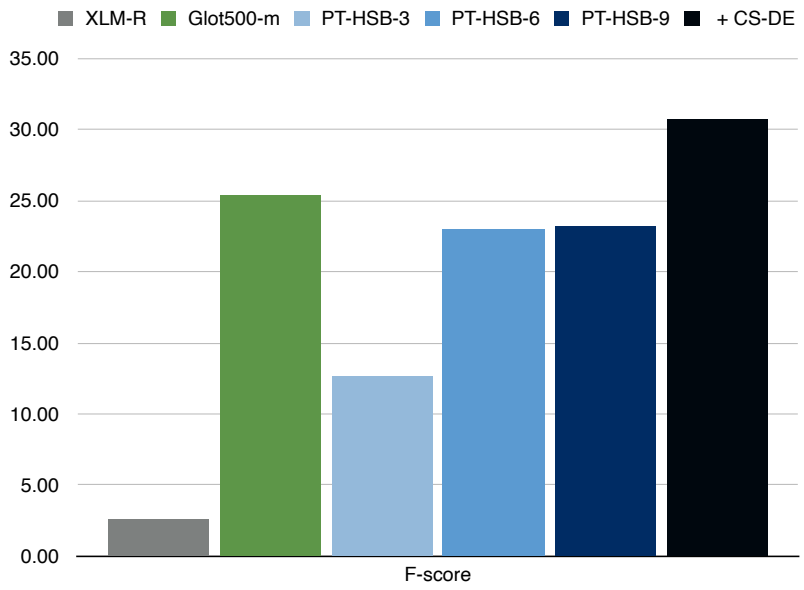
→ Measuring how well the tool can mine parallel sentences



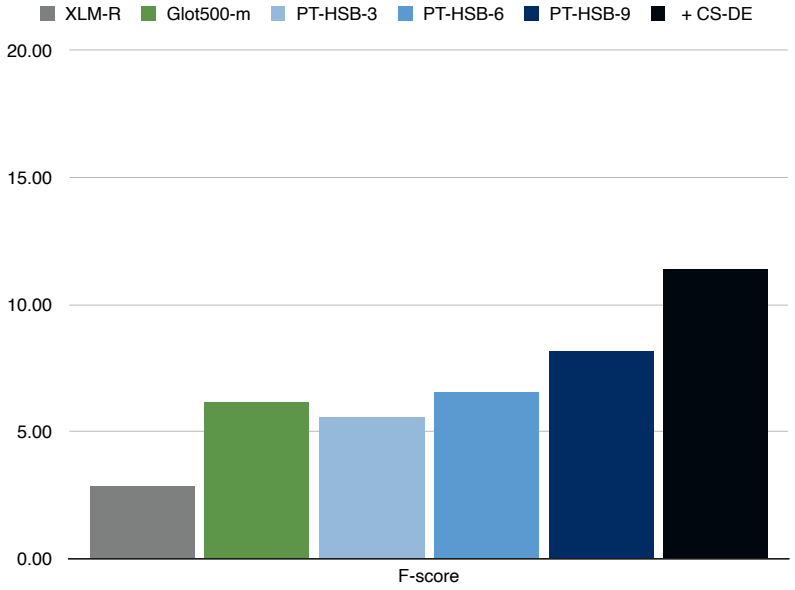
Upper Sorbian

Mining results for Upper and Lower Sorbian

→ Measuring how well the tool can mine parallel sentences



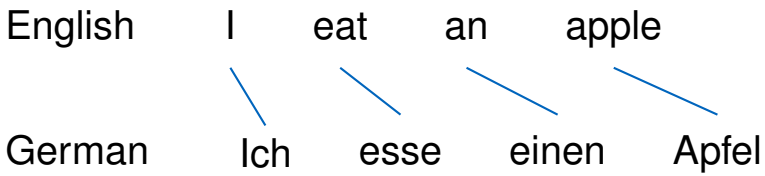
Upper Sorbian



Lower Sorbian

Unsupervised alignment post-processing

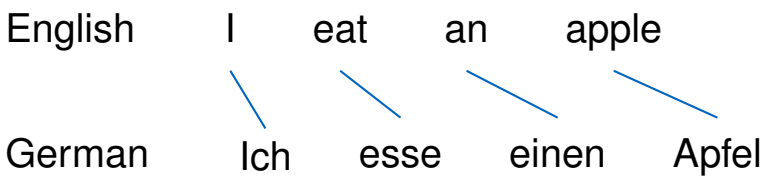
Additional filtering of mined sentences based on alignment proportions



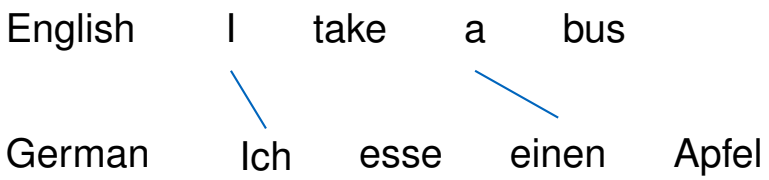
4 alignment links
(alignment score: 100%)

Unsupervised alignment post-processing

Additional filtering of mined sentences based on alignment proportions



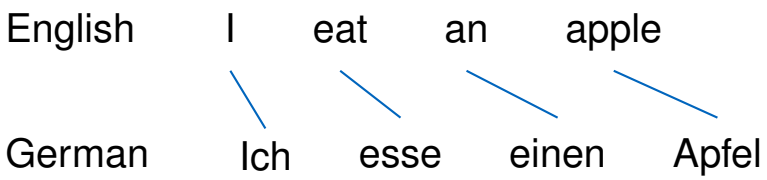
4 alignment links
(alignment score: 100%)



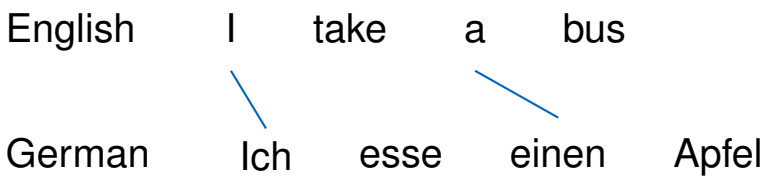
Only 2 alignment links
(alignment score: 50%)

Unsupervised alignment post-processing

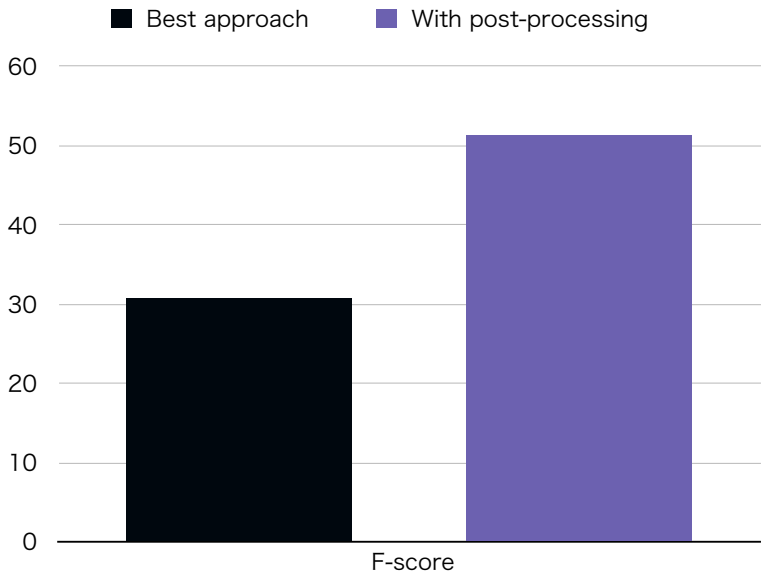
Additional filtering of mined sentences based on alignment proportions



4 alignment links
(alignment score: 100%)



Only 2 alignment links
(alignment score: 50%)



For Upper Sorbian

Qualitative analysis of mined sentence pairs

model	language	sentence
-	Upper Sorbian	Wón namjetuje moderěrowanu diskusiju wo tym.
XLM-R	German	Sie rechen das Laub der Laubbäume. *They rake the leaves of the deciduous trees.*
Best	German	Er schlägt eine moderierende Diskussion darüber an. *He proposes a moderated discussion about this.*

Qualitative analysis of mined sentence pairs

model	language	sentence
-	Upper Sorbian	Wón namjetuje moderěrowanu diskusiju wo tym.
XLM-R	German	Sie rechen das Laub der Laubbäume. *They rake the leaves of the deciduous trees.*
Best	German	Er schlägt eine moderierende Diskussion darüber an. *He proposes a moderated discussion about this.*

Upper Sorbian	Kocorowy oratorij „Serbski kwas“ zaklinči po něhdźe dźesać lětach zaso, a to tutu njedźelu, 15. julija, w 17 hodź.
German	Das große Finale von „Die Bachelorette“ läuft am Mittwoch, den 9. Dezember, um 20.15 Uhr bei RTL.

Dates in red (Sunday, 15th of July and Wednesday, 9th of December) and times in blue.

Conclusion

- Sentence mining pipeline with multilingual language models
- Parallel sentence mining for two endangered low-resource languages: Upper and Lower Sorbian
- Pre-training on the language is essential to start to have a decent mining quality
  Relying on related languages helps but is not enough
- Alignment post-processing to improve mining quality
- Benchmark to evaluate parallel sentence mining tool:
  `https://github.com/shuokabe/PaSeMiLL/tree/main/data`

Workshop organisation

- This work was part of the ERC Proof of Concept Grant to create tools for language activists
- Organisation of an online workshop in February 2025
  NLP tools for language communities
- Ongoing collaboration, extension to other languages (language pairs)
- In case you are interested, you can join our Google Group for future updates!

Thank you for your attention!