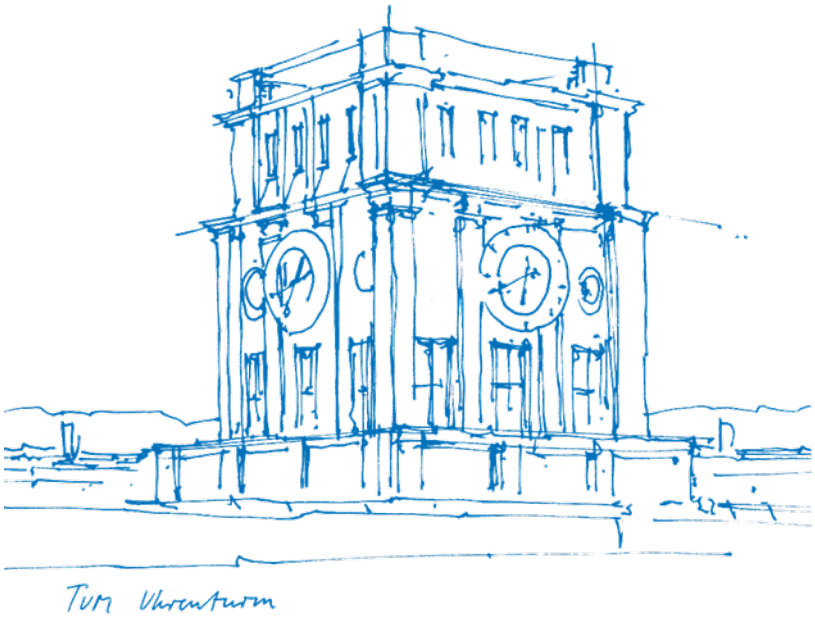
First online workshop on NLP tools for language communities

Shu Okabe & Alexander Fraser

Data Analytics & Statistics
Technische Universität München
shu.okabe@tum.de

4th & 6th February 2025



About the workshop

Schedule:

1. Presentation of the parallel sentence mining tool
   Around 40 minutes of presentation + Q&A
2. Presentation of NLP tools and tasks
   Around 40 minutes of presentation + Q&A

Goals of the workshop:

- Present NLP tools for language communities and language activists
- Survey and understand the needs of the community side

Linked to the ERC Proof of Concept Grant to create tools for language activists

Frequent topics from the online form

- NLP models and tasks for low-resource languages: Machine Translation, Large Language Models, Conversational AI, Automatic Speech Recognition, evaluation, ...
- NLP models for specific languages: how to adapt existing tools to another language?
- Data access, collection, and pre-processing
- How to work with language communities?

Parallel Sentence Mining for Low-Resource Languages

First online workshop on NLP tools for language communities

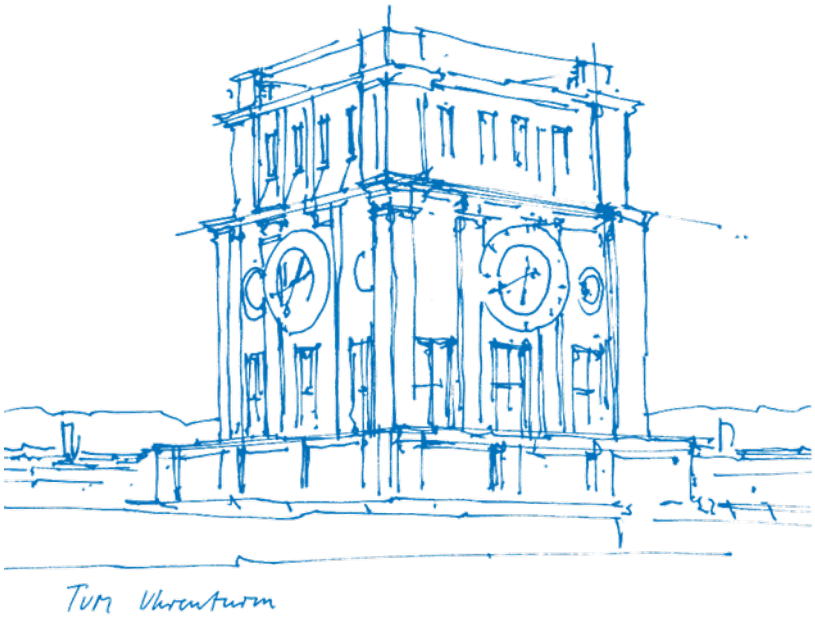
Shu Okabe & Alexander Fraser

Data Analytics & Statistics

Technische Universität München

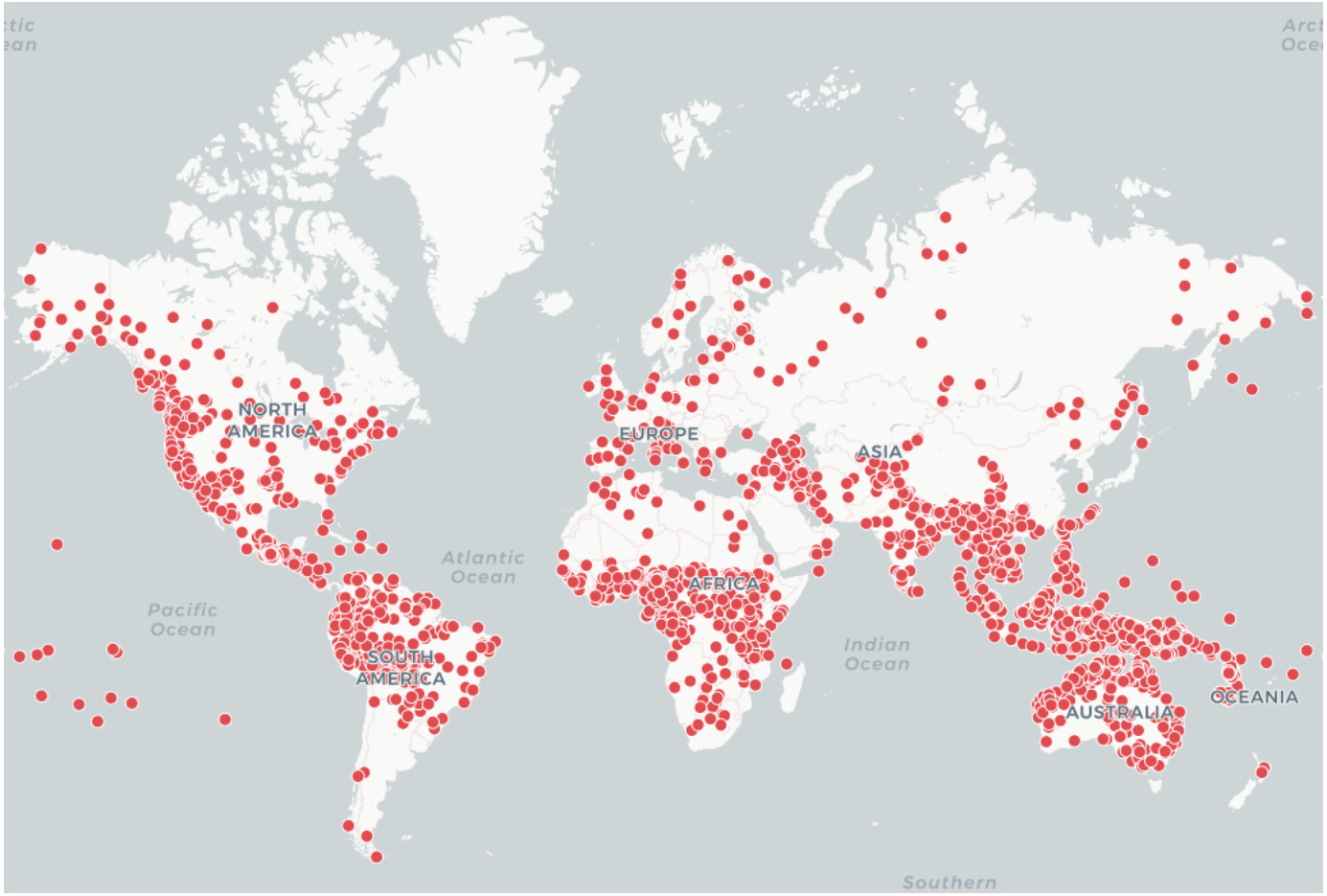
shu.okabe@tum.de

4th & 6th February 2025



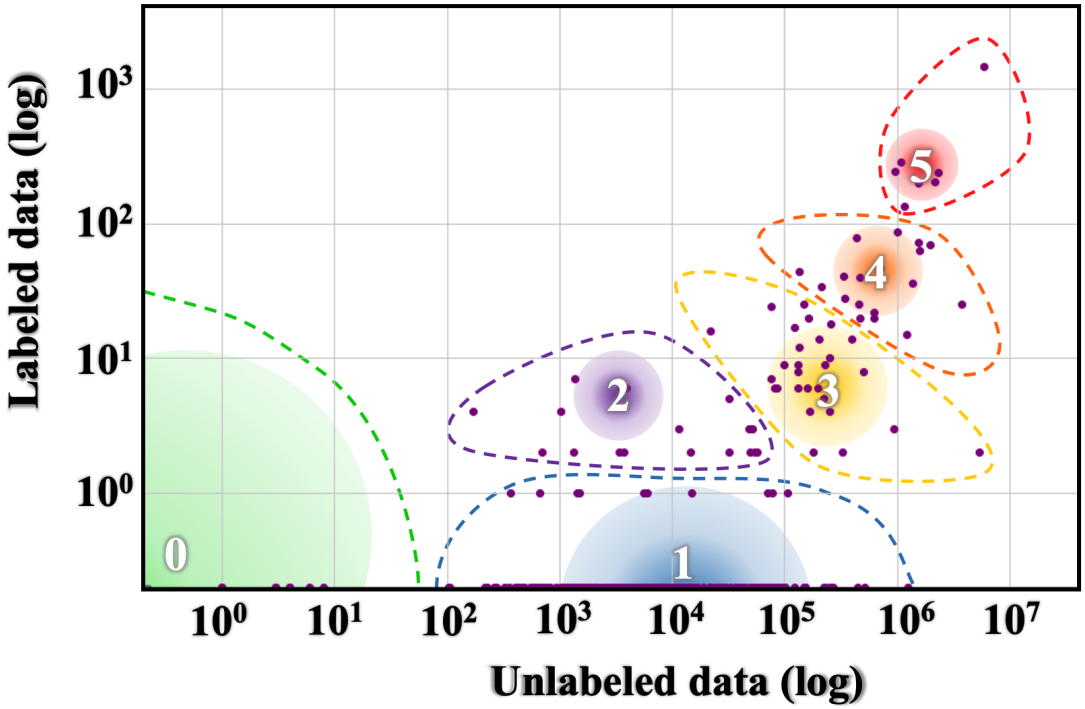
Introduction
Language endangerment & NLP

Language vitality in the world
3,170 languages are currently endangered



From Ethnologue (Eberhard et al., 2024)

Language taxonomy in the NLP community



From (Joshi et al., 2020)

0. The Left-Behinds
1. The Scraping-Bys
2. The Hopefuls
3. The Rising Stars
4. The Underdogs
5. The Winners

Unlabelled data: Wikipedia pages
Labelled data: datasets in LDC catalog and the ELRA Map

NLP community & Language diversity recently

- **ACL 2022 theme track**: *Language Diversity: from Low-Resource to Endangered Languages*
- **ComputEL workshops** (The Use of Computational Methods in the Study of Endangered Languages): co-located with either the International Conference of Language Documentation and Conservation (ICLDC) or *ACL
- Dedicated **workshops**, co-located with *ACL, NLP or ML conferences: African NLP, AmericasNLP, SIGTURK 2024 Workshop, Special Interest Group on Under-resourced Languages (SIGUL), Slavic NLP, VarDial, . . .
- **Shared tasks**: e.g., Low-Resource Indic Language Translation (WMT 2023, 2024), Translation into Low-Resource Languages of Spain (WMT 2024)
- **Collaboration** between NLP and language community: Masakhane, AmericasNLP

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)

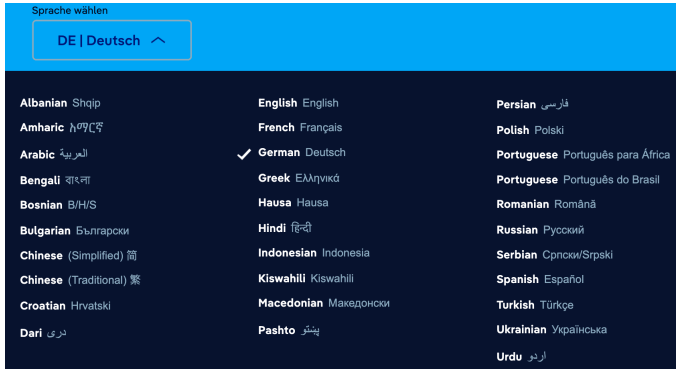
→ Valuable resource for downstream NLP tasks such as Machine Translation

Where can we find parallel sentences?

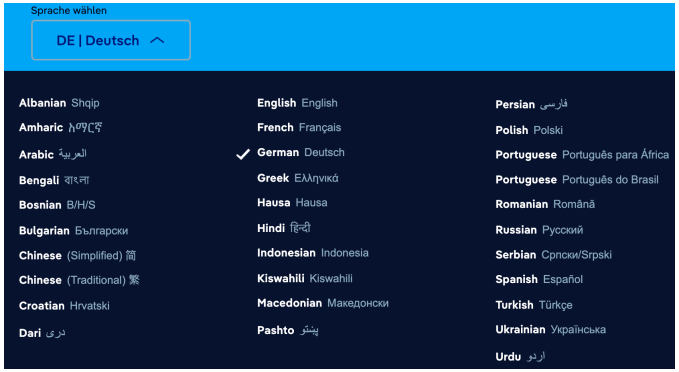
The BBC is in multiple languages

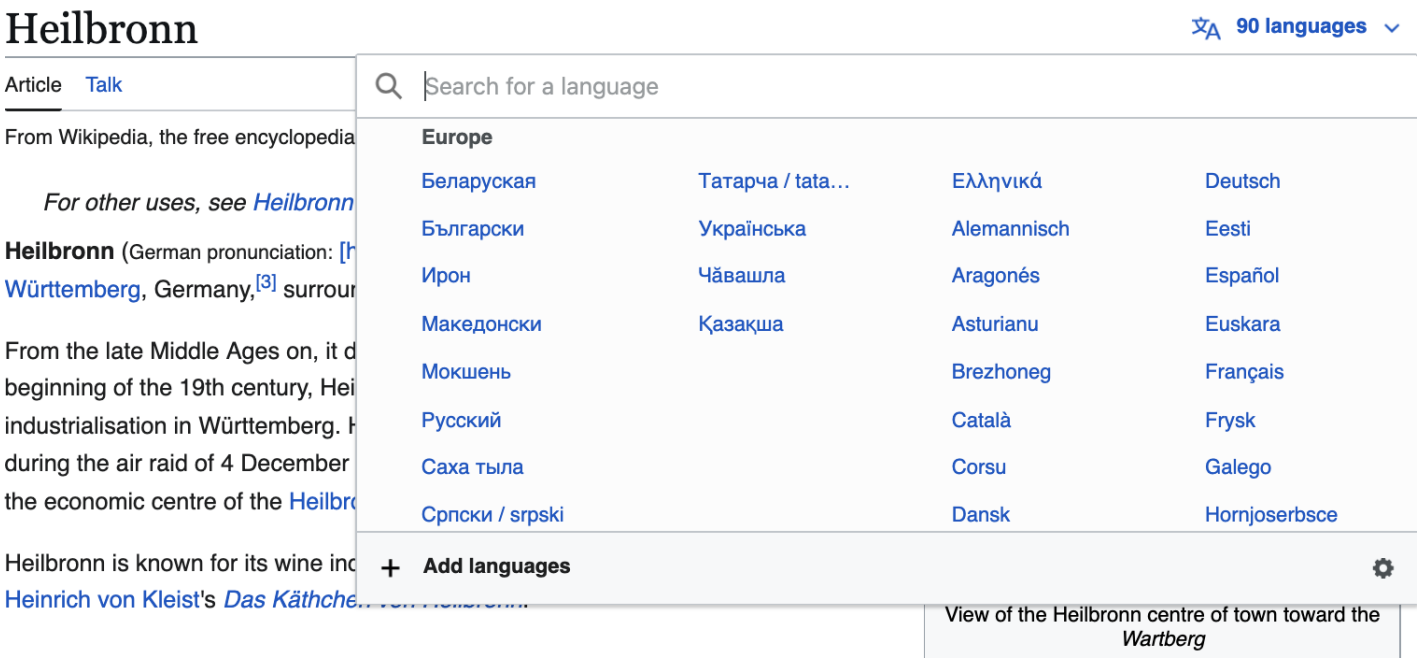
Read the BBC In your own language

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



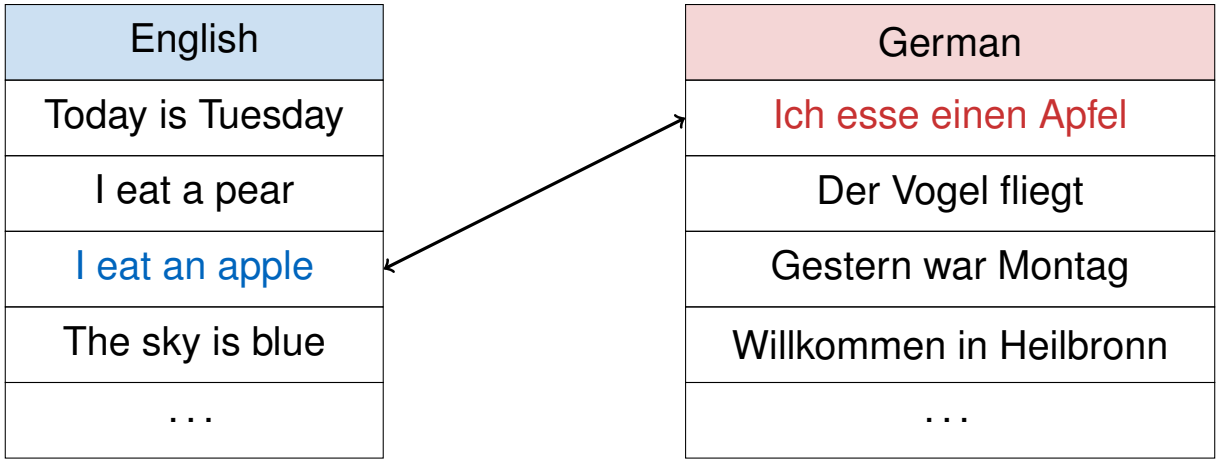
Where can we find parallel sentences?



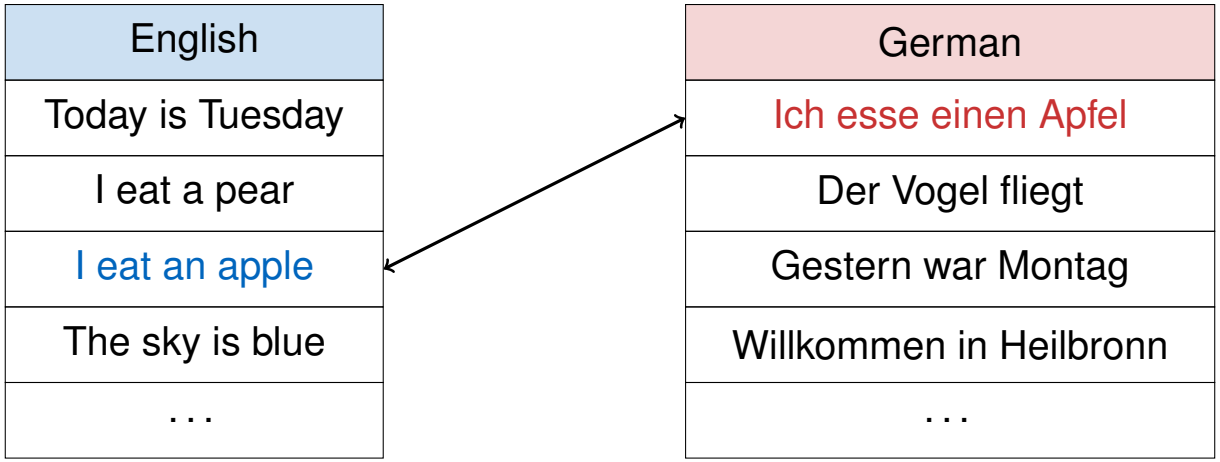


Parallel sentence mining

Extracting parallel sentences from two monolingual corpora



Parallel sentence mining
Extracting parallel sentences from two monolingual corpora



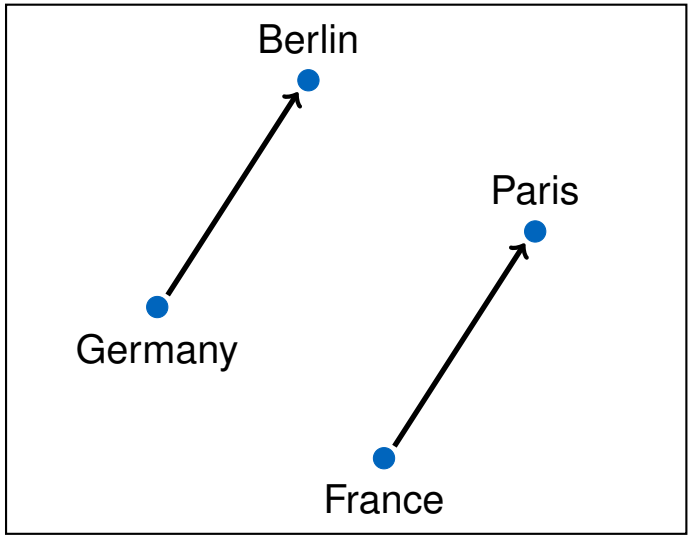
Challenges:

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

Mining pipeline

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

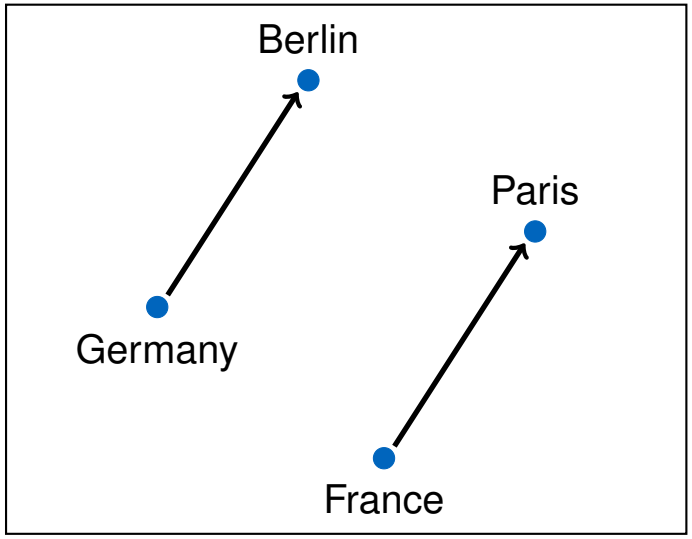
$$v(Berlin) - v(Germany) + v(France) \approx v(Paris)$$



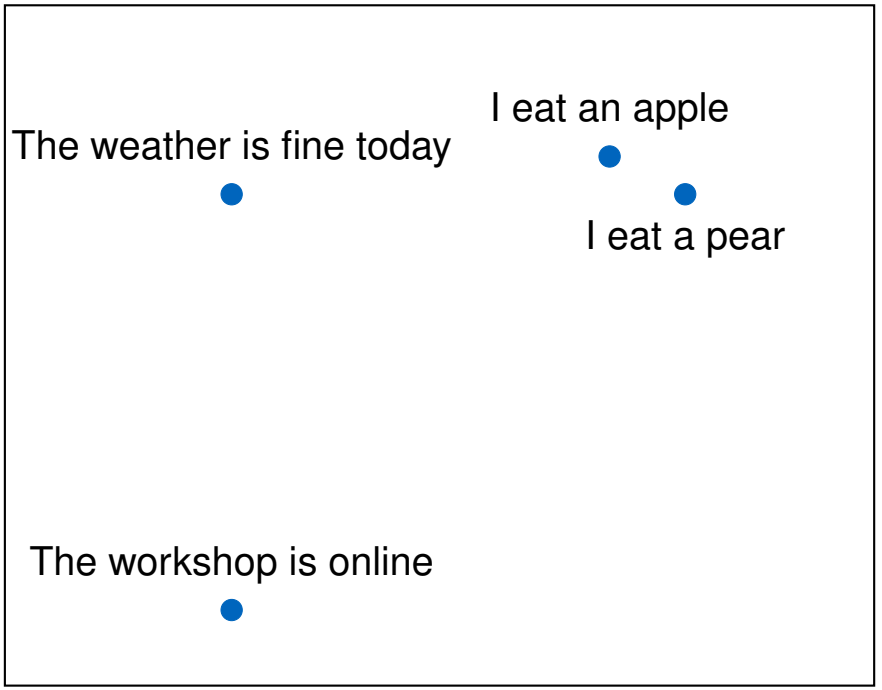
Word embeddings, from (Mikolov et al., 2013)

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

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

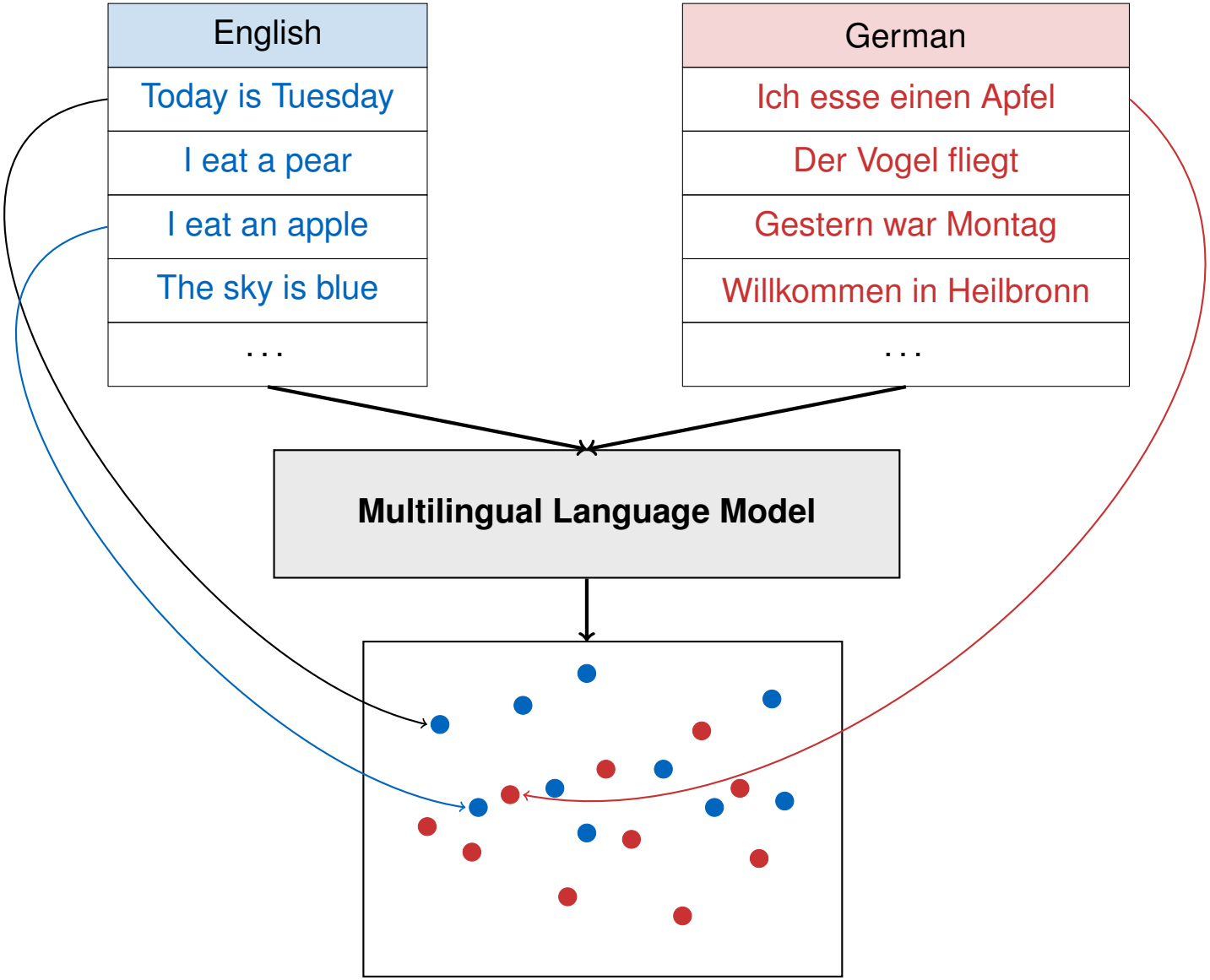


Word embeddings, from (Mikolov et al., 2013)

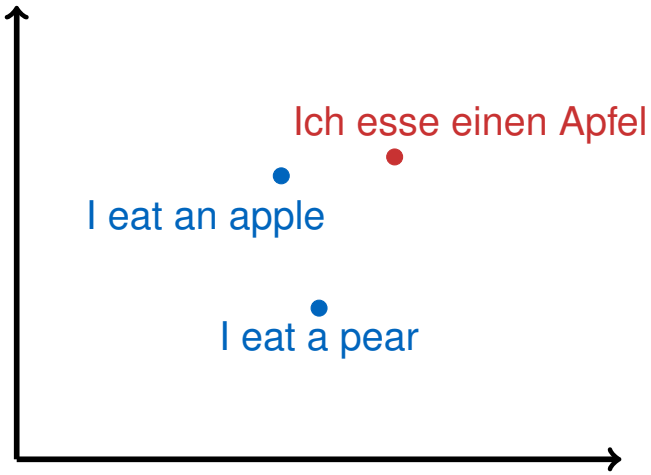


Sentence embeddings

Step 1: From monolingual sentences to vectors

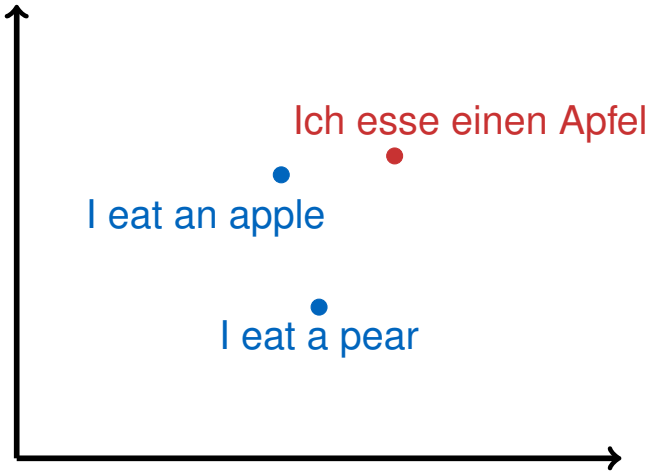


Step 2: Compute similarity between sentences



Multilingual sentence embeddings

Step 2: Compute similarity between sentences



Multilingual sentence embeddings

Similarity between sentences (vectors)

- Pipeline from (Hangya & Fraser, 2019)
- Compute a similarity score based on cosine similarity (using Faiss (Johnson et al., 2019))

sim(I eat an apple, Ich esse einen Apfel) $>$ sim(I eat a pear, Ich esse einen Apfel)

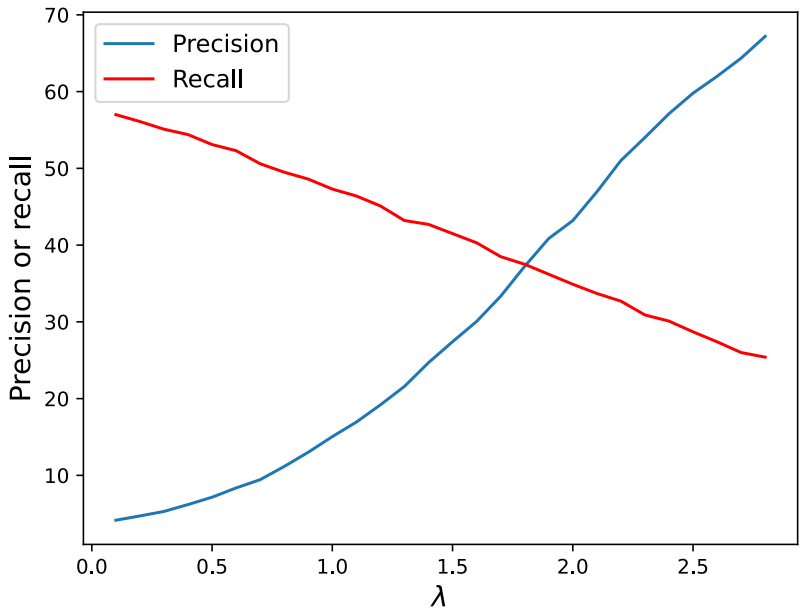
Step 3: Filtering sentence pairs
Setting a similarity threshold

Source sentence	Closest target sentence	Similarity score
EN0001	DE3721	0.34
EN0002	DE8460	0.92
EN0003	DE1298	-0.05
EN0004	DE5943	0.21

Step 3: Filtering sentence pairs
Setting a similarity threshold

Source sentence	Closest target sentence	Similarity score
EN0001	DE3721	0.34
EN0002	DE8460	0.92
EN0003	DE1298	-0.05
EN0004	DE5943	0.21

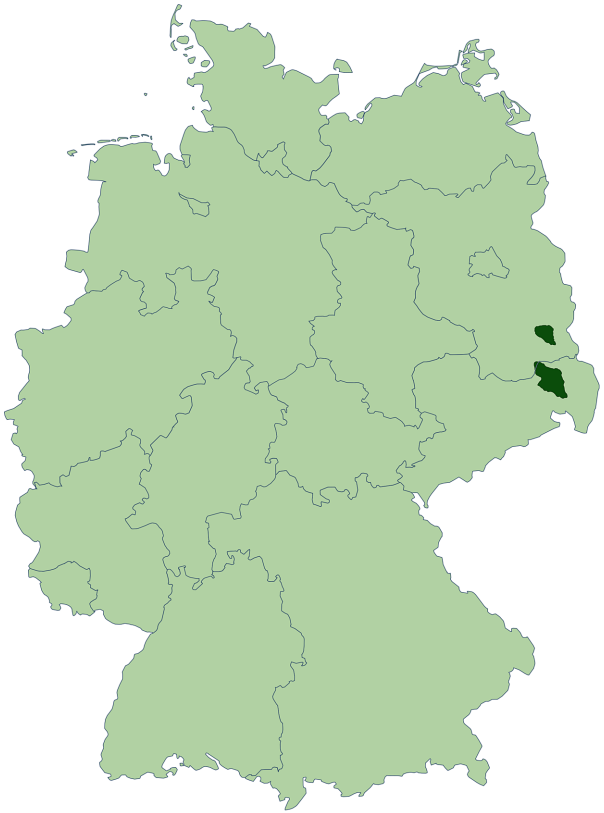
Trade-off: how cautious should we be?



Case study on Sorbian languages

Case study on Upper and Lower Sorbian

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

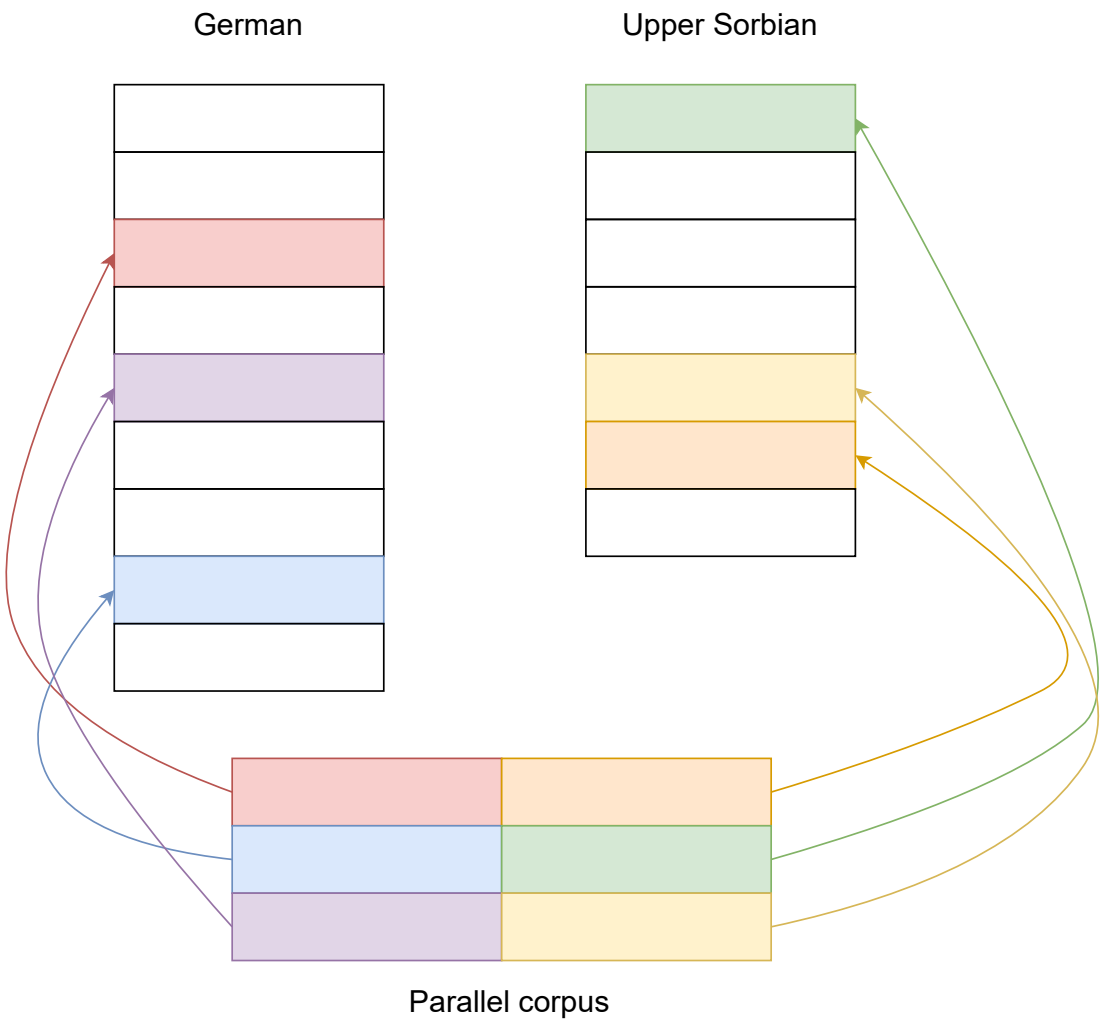


- Previous cooperation with non-profits (e.g., WMT Shared Tasks):
  the Witaj Sprachzentrum (Witaj Language Centre) and the Sorbian Institute

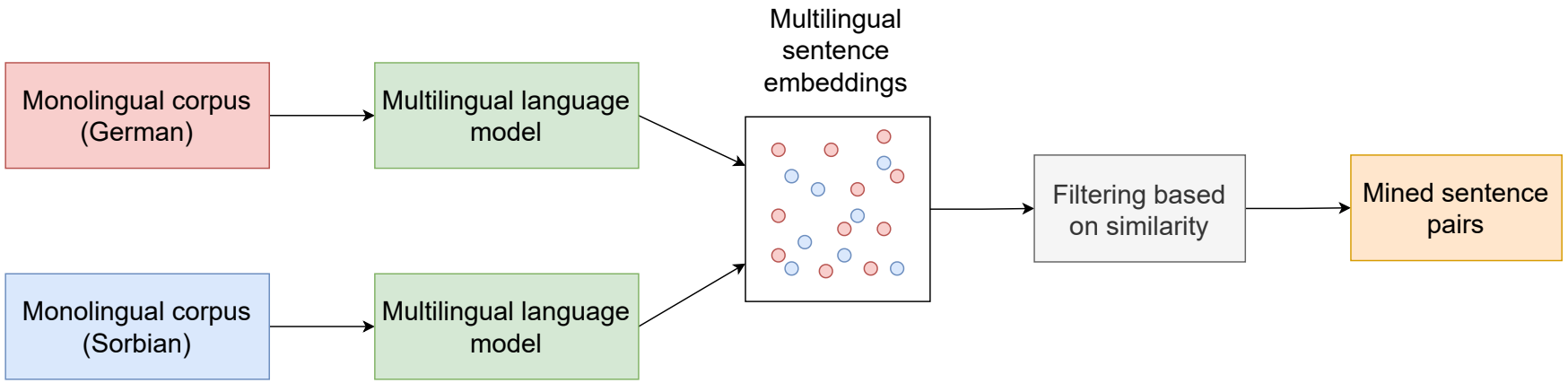
Parallel sentence mining for Sorbian languages

- To what extent can some knowledge of a low-resource language (Upper Sorbian) help another low-resource language (Lower Sorbian)?
- How much data is needed to expect a significant improvement directly (Upper Sorbian) and indirectly (Lower Sorbian)?
- How useful is it to use additional data from languages of the same family?

Experimental methodology: corpus creation
Injecting parallel sentences in monolingual corpora



Mining pipeline



Baseline multilingual language models

Sentence embeddings from averaged word embeddings

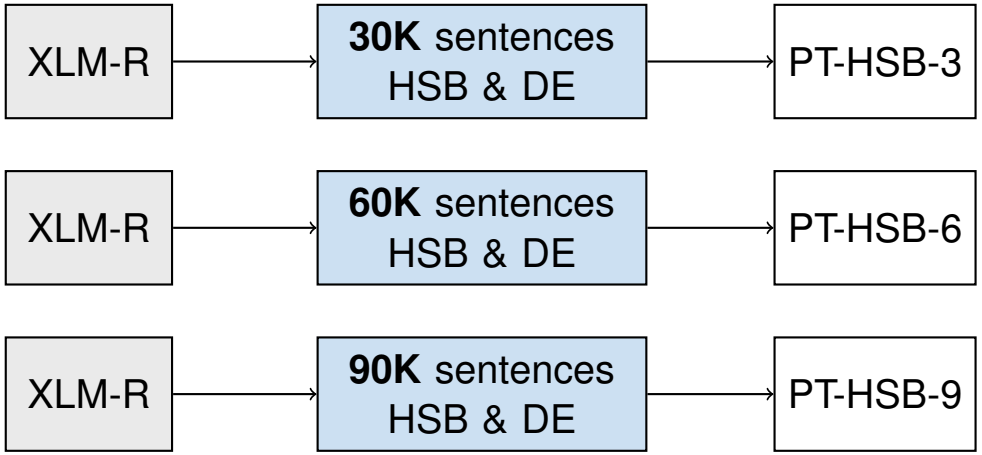
Off-the-shelf baseline multilingual models:

- 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)

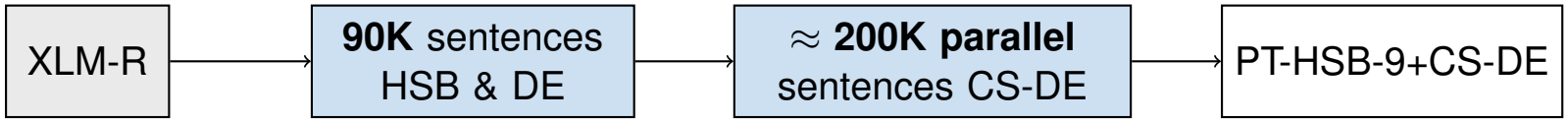
	XLM-R (base)	Glot500-m
Number of covered languages	100	511
Czech & Polish?	✓	✓
Upper Sorbian?	✗	✓
Lower Sorbian?	✗	✗

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

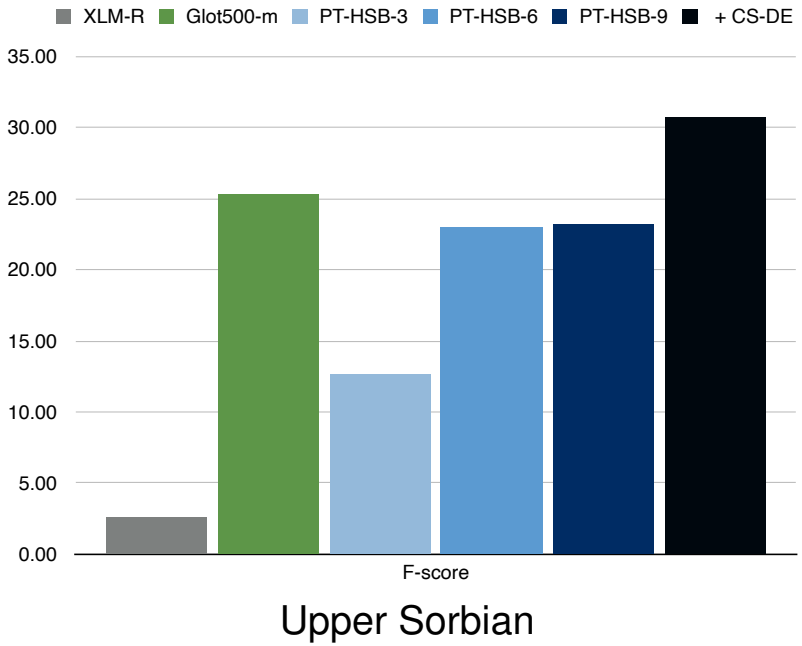


- Using German-Czech **parallel** sentences



Mining results for Upper and Lower Sorbian

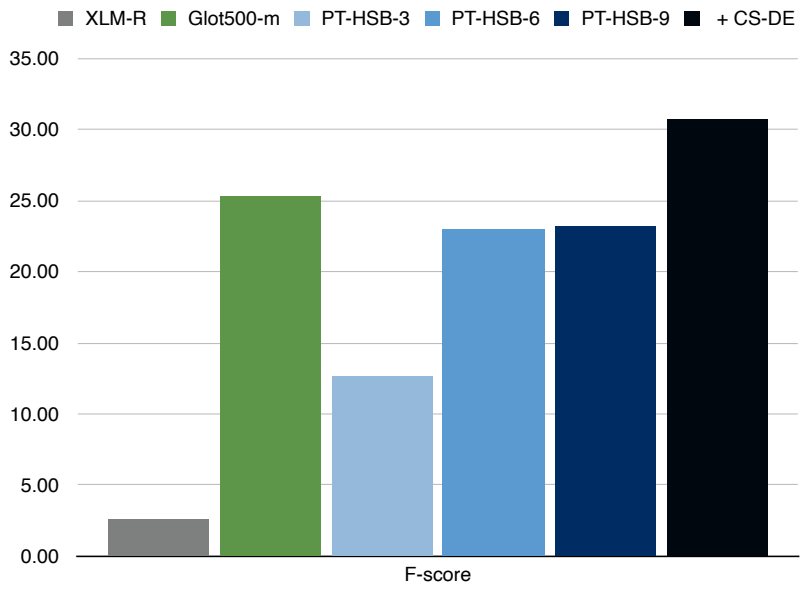
→ Measuring how well the tool can mine parallel sentences



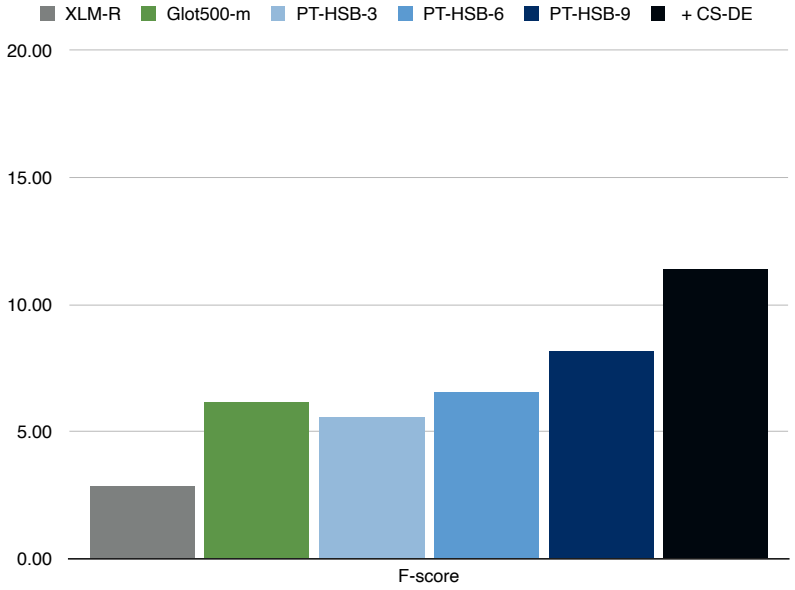
Upper Sorbian

Mining results for Upper and Lower Sorbian

→ Measuring how well the tool can mine parallel sentences



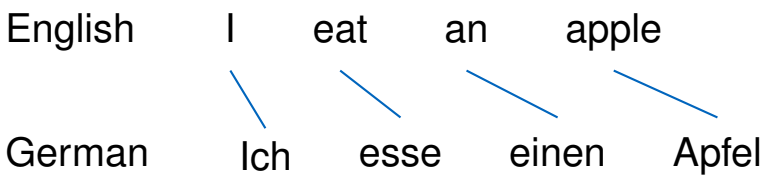
Upper Sorbian



Lower Sorbian

Unsupervised alignment post-processing

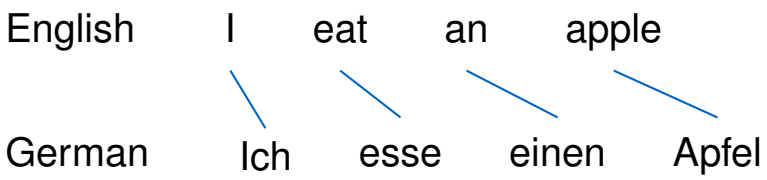
Additional filtering of mined sentences based on alignment proportions



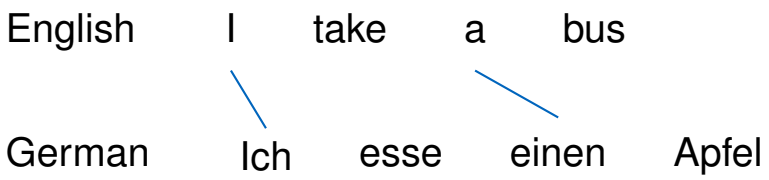
4 alignment links
(alignment score: 100%)

Unsupervised alignment post-processing

Additional filtering of mined sentences based on alignment proportions



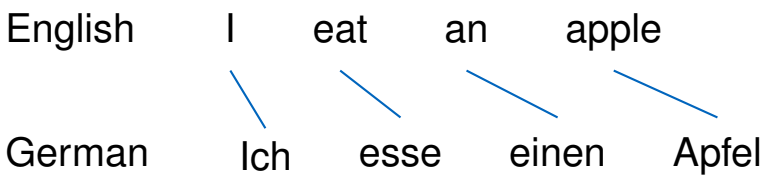
4 alignment links
(alignment score: 100%)



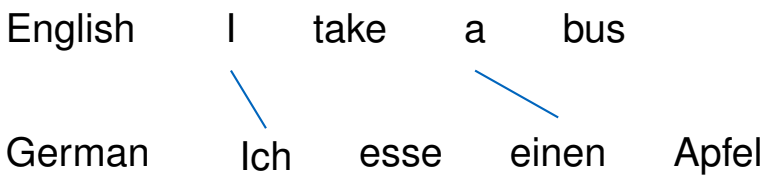
Only 2 alignment links
(alignment score: 50%)

Unsupervised alignment post-processing

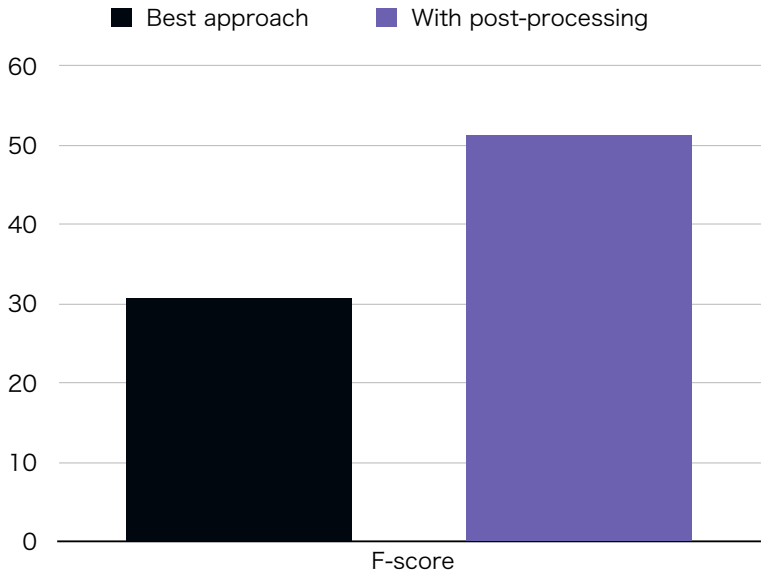
Additional filtering of mined sentences based on alignment proportions



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`

Limitations in the current benchmark

Upper Sorbian-German and Lower Sorbian-German language pairs are an ‘ideal’ case

- Related to two better-resourced languages: Czech and Polish
  - Extensive parallel data for both Czech-German and Polish-German
  - Both Czech and Polish are (better) supported by state-of-the-art language models
- Latin script for both languages (same script)
- Sorbian languages and German are both Indo-European languages
- (Partial) support in existing multilingual language models
- Data availability for both Sorbian languages

→ How well can it be applied to other languages?

Next steps: with you

- Are you interested in such a mining tool? or more interested in Machine Translation directly?
- How should the tool be?
  - Better take into account technical constraints?
  - Dedicated session for practical tool use?
- What should we prioritise?
  - So far, no language-specific processing has been used: adapting to each language?
  - Or, improve overall support of languages?

Challenges:

- Data availability: comparable monolingual corpora
- Model support: is the language (or related languages) supported by an existing multilingual model?
- Language distance: how close are the languages in the pair?

Thank you for your attention!