Universität Potsdam

Creating a German discourse parsing corpus by transferring relations between languages

Johann Seltmann

Shallow Discourse Parsing

- Find *relations* between adjacent sentences or sentence parts, called *arguments*.
- Each relation has a *sense*.
- *Explicit* relations have a *connective*; a word that connects the two arguments. *Implicit* relations don't.

Problem

- expensive to annotate
- biggest corpora for German:
  - Potsdam Commentary Corpus (PCC): 2200 relations
  - TED Multilingual Discourse Bank: 3600 relations
- corpus for English: Penn Discourse Treebank (PDTB): 40k relations

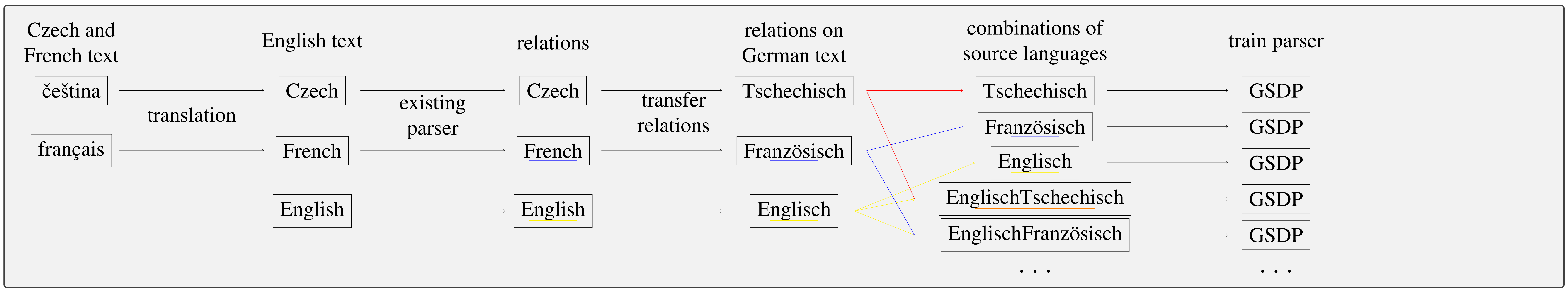
Example

Explicit relation with sense *Contingency.Condition*:

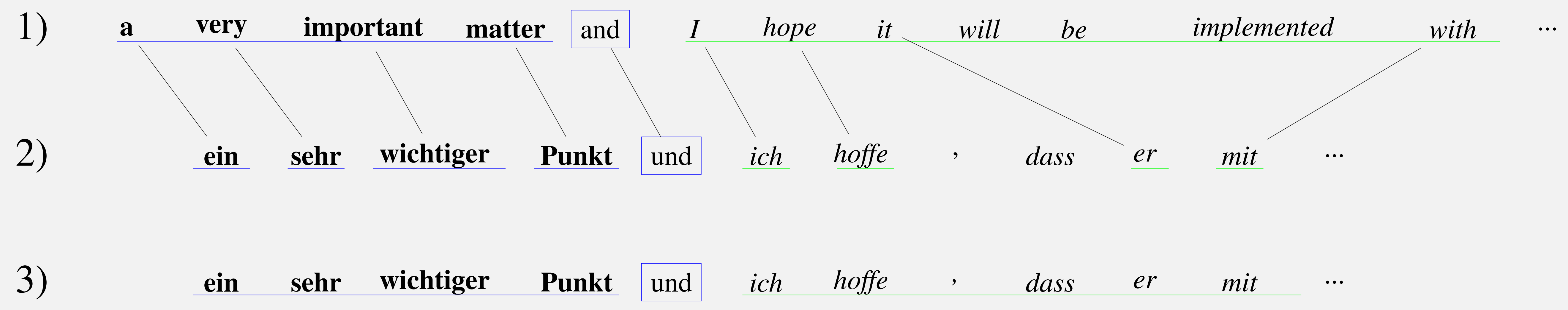
If, **at the end of this process, the Iranian fundamentalist regime has reinforced its influence in the region, [...]** then *that region will be further away from peace and the world will be facing a greater threat.*

Idea

- Translated texts should contain the same relations as the original.
- Take English parser trained on PDTB: Wang/Lan 2015
- Europarl corpus contains speeches from the European parliament, translated in 24 languages.
- Parse English text, transfer relations to German text



Transfer of relations along word-alignments


1) a very important matter and I hope it will be implemented with ...
2) ein sehr wichtiger Punkt und ich hoffe , dass er mit ...
3) ein sehr wichtiger Punkt und ich hoffe , dass er mit ...

Using back-translation

- Parsers work better on explicit relations, since connectives are often connected to specific senses.
- When there are multiple translated versions, some of them might include connectives, others might not.
- Idea from (Shi et.al., 2017): Use Moses for back-translation into English to get more data.
- Using the Czech and French sections of Europarl.
- Intersect corpora created from different languages to get more accurate annotations.

Corpora

- 12 different corpora
- examples: corpora created from English text, by combining English and Czech

	from_en	en_cs	PCC
#documents	56k	56k	176
relations per document	11.94	8.18	12.52
explicit relations per document	3.77	2.89	6.32

- difference in senses compared to PCC due to different kind of text and idiosyncracies of parser

Training

- train German Shallow Discourse Parser (GSDP) (Bourgonje&Stede, 2018)
- test on PCC and on held-out set

results:

- finds too few explicit relations on PCC
- low token-wise argument extraction accuracy
- weak on sense classification (F1: 11%-16% on PCC, depending on corpus)

Conclusion

- Needs better parser as base.
- Transfer works well, but need better heuristic for connectives.
- Using multiple languages improved some results.
- Larger size didn't offset lower quality.