Cross-linguistic differences in discourse marking: A case study of German-English texts

Frances Yung, Merel Scholman, Vera Demberg

Introduction

- Studies on the translation of discourse connectives (DCs) often rely on corpus analyses (e.g. Becher 2011)
- Manual analysis can only focus on limited samples and a subset of DCs → need more comprehensive insights.
- We explore the use of **NLP tools** to analyze the translation of a large sample and many different DCs.
- Our research questions:
  1. Are NLP tools useful for the study of discourse relations?
  2. How are discourse connectives translated from English to German, and from German to English?
- We first consider the pooled **translation equivalents**. Our next step is to consider translation direction.

Data

Europarl Direct Corpus (Cartoni and Meyer 2012): written proceedings of the European Parliament and their translation.

- 18 English-to-German texts (170K English tokens)
- 15 German-to-English texts (95K German tokens)
- Sentences were aligned cross-lingually

Methodology

- Identify English and German discourse connectives (discourse and non-discourse usage) and annotate their senses using language-specific **automatic shallow discourse parsers**
- Align words of each sentence pair cross-lingually using an **automatic word alignment model**.
- Tools: Knaebel 2021, Bourgonje 2021, Dou and Neubig 2021
- Refine with heuristics (e.g. "damit...,..zu.." is not a DC)

Tool performance

- **8776 German and 7995 English DCs** were identified

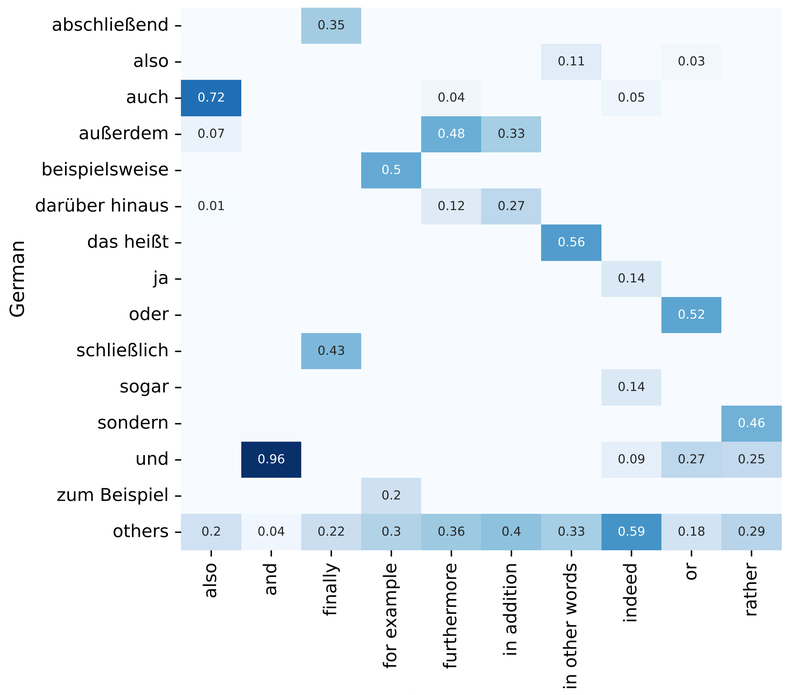
	German	English
aligned to DCs of the other language	5494	5593
aligned to other non-DC words	1542	1517
not aligned to any words	1740	885

- Accuracy based on **manual analysis of 400 instances**

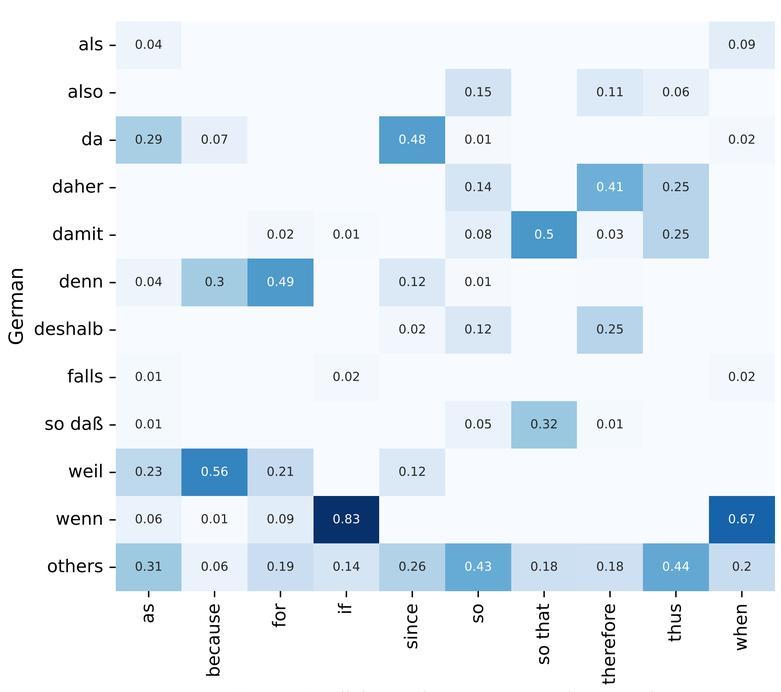
Parser accuracy	German	English
DC identification	83%	85%
DC sense annotation	90%	84%

Aligment results	Accuracy
word-to-word alignment	89%
DC-align. w/o sense	78%
DC-align. with sense	52%

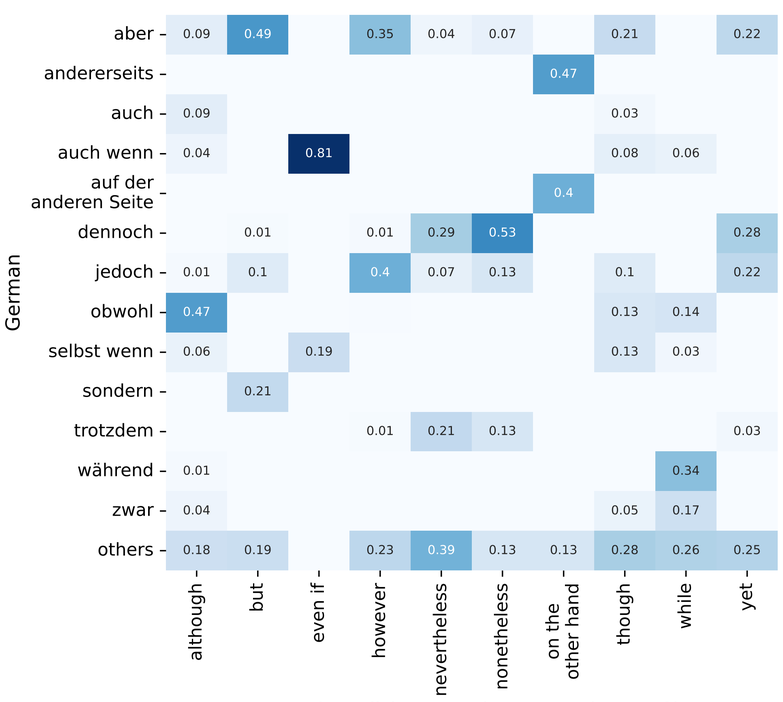
Alignment of English connectives to German words



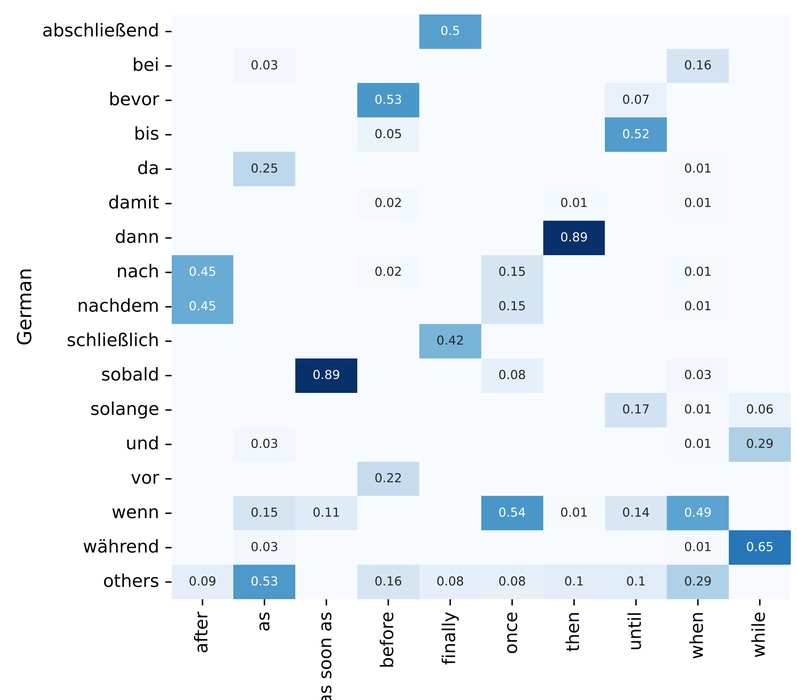
Top 10 English expansion conn. (n=3311)



Top 10 English contingency conn. (n=1422)



Top 10 English comparison conn. (n=1448)



Top 10 English temporal conn. (n=584)

- Finer-grained differences in meaning get translated (*because* = *weil/denn*; *since* = *da* but to a lesser extent *weil/denn*)
- Connective choice is influenced by various other aspects:
  - Style choices (*nevertheless/nonetheless* = *dennoch* rather than *trotzdem*)
    Nonetheless, these are important criteria for the EU.
    Dennoch ist das ein wichtiges Kriterium für die EU.
  - Lexical elements in context (*not+until* = *nicht+bis* or *wenn*)
    The process can not take place until the EU...
    Das Prozeß kann erst dann stattfinden, wenn der EU...
  - Nominalization in the arguments (*before* = *bevor* or *vor*+nominalization)
    Before the vote started...
    Vor der Abstimmung...

Conclusion

- Automatic annotations are useful to **recognize the overall patterns** of cross-lingual differences in DC usage.
- Specific samples can be identified for **detailed manual analysis**.
- Our next step is to analyze the patterns of explicitation and implicitation of DCs in relation of the explicity of other discourse relations in context.