Shallow Discourse Parsing for Under-Resourced Languages: Combining Machine Translation and Annotation Projection

Henny Sluyter-Gäthje, Peter Bourgonje, Manfred Stede
Universität Potsdam, Applied Computational Linguistics
Potsdam, Germany
firstname.lastname@uni-potsdam.de

Abstract

Shallow Discourse Parsing (SDP), the identification of coherence relations between text spans, relies on large amounts of training data, which so far exists only for English - any other language is in this respect an under-resourced one. For those languages where machine translation from English is available with reasonable quality, MT in conjunction with annotation projection can be an option for producing an SDP resource. In our study, we translate the English Penn Discourse TreeBank into German and experiment with various methods of annotation projection to arrive at the German counterpart of the PDTB. We describe the key characteristics of the corpus as well as some typical sources of errors encountered during its creation. Then we evaluate the GermanPDTB by training components for selected sub-tasks of discourse parsing on this silver data and compare performance to the same components when trained on the gold, original PDTB corpus.

Keywords: machine translation, annotation projection, discourse parsing

1. Introduction

Texts are not a random collection of sentences: they are texts because they convey a certain sense of coherence. The uncovering of the coherence relations holding a text together is referred to as the task of *discourse parsing*. Like many other tasks based on automatically parsing input text in the larger field of Natural Language Processing, procedures often rely on the availability of training data annotated for the type of information to be extracted. In the case of coherence relations, such annotations are notoriously difficult and time-consuming to obtain, and inter-annotator agreement rates are lower than for many other tasks. As a result, the amount of available training data is comparatively small, especially for languages other than English (see Section 2.).

In this paper, we present a corpus annotated for discourse relations obtained through automatically translating an existing English corpus (the Penn Discourse TreeBank, henceforth: PDTB, (Prasad et al., 2008)), and using word alignment to project the English annotations on the German target text. The result is the GermanPDTB, a German corpus annotated for shallow discourse relations in the (financial) news domain. We provide details on the method used to create this corpus, sum up the key characteristics and use the GermanPDTB to enrich a pre-existing German connective lexicon. In addition, we provide an extrinsic evaluation of the corpus using components of a German discourse parser and compare performance of selected (sub-)tasks on GermanPDTB to the original English PDTB.

The rest of this paper is structured as follows: Section 2. lists similar corpora for German and other languages. Section 3. briefly describes the different coherence annotation types in the original PDTB and consequently the GermanPDTB. Section 4. explains our method of constructing the GermanPDTB and Section 5. explains the manual corrections done on this automatically produced output. Section 6.1. provides an intrinsic evaluation of the German Machine Translation output and the quality and possible sources of error for the annotation projection, and Section 6.2. provides the extrinsic evaluation, using the GermanPDTB as training data for a discourse parser. Finally, Section 7. sums up our main findings and points to future work.

2. Related Work

Our starting point for the GermanPDTB is the original English PDTB in its 2.0 version (Prasad et al., 2008). More specifically, we use the subset also used in the 2016 CoNLL shared task on discourse parsing (Xue et al., 2016). The PDTB is by far the largest corpus annotated for coherence relations, with over 1m words and over 40k annotated relations in its 2.0 version. Other corpora annotated for coherence relations are considerably smaller, and also distributed over different frameworks, most notably Rhetorical Structure Theory (Mann and Thompson, 1988) and Segmented Discourse Representation Theory (Asher and Lascarides, 2005). We refer the reader to Zeldes et al. (2019) for an overview of corpora for different languages and frameworks.

For German, our language of interest, to date the largest annotated corpus is the Potsdam Commentary Corpus (henceforth: PCC, (Bourgonje and Stede, 2020)) and a smaller corpus exists as a discourse annotation layer over parts of the TÜBA-D/Z corpus (Versley and Gastel, 2012). The PCC contains 2,208 relations, annotated according to the guidelines used for the PDTB2 (Prasad et al., 2008). Because of the much larger size of the PDTB, in our experiments we hope to collect considerably more instances of discourse relations in German.

Through our method, exploiting machine translation and annotation projection, we will extract silver data in the sense that the resulting annotations cannot be guaranteed to be correct (i.e., they are not all checked by a human, though the next sections describe heuristics for quality assurance), but because of the much larger size of the PDTB, we end up with many more instances of relations in German, and from a slightly different domain, with the PCC representing

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) , pages 1037–1043
Marseille, 11–16 May 2020

the news editorial/commentary domain, and PDTB articles representing the financial news domain.

The procedure of annotation projection has been used in the context of coherence relations before, but remained restricted to explicit discourse connectives, for example to create or extend discourse lexicons and disambiguate connectives (English-French (Laali and Kosseim, 2014), English-Chinese (Zhou et al., 2012) and German-Italian (Bourgonje et al., 2017)), to compile a metric to score machine translation output (English-Arabic (Hajlaoui and Popescu-Belis, 2013)) or to create a corpus annotated with discourse markers to train a parser (English-German (Versley, 2010) and English-French (Laali, 2017)). The novelty of our work lies in using the procedure for entire coherence relations, as opposed to restricting it to connectives. In contrast to Versley (2010) and Laali and Kosseim (2014) who use existing parallel corpora for which they automatically annotated the English side, we create a parallel corpus by machine-translating the manually annotated PDTB. After machine translation, we rely on word alignments produced with GIZA++ (Och and Ney, 2003) that are post-processed using some heuristics implemented in the Moses statistical machine translation system (Koehn et al., 2007). In addition, we perform an automatic corpus analysis to examine how different types of annotation interdepend, and accordingly we compile rules for the projection process.

3. Annotation Structure

In the PDTB framework, coherence annotations are divided into five different relation types.

(1) Explicit relations consist of an overtly realised discourse connective (such as *because, although, if*) and two arguments; one external argument (Arg1) and one internal argument (Arg2), the latter being syntactically integrated with the discourse connective. Finally, they contain a relation sense, to be selected from the PDTB sense hierarchy (see Prasad et al. (2008)). Arg1 and Arg2 are referred to as such because this reflects the unmarked order of the arguments, but the reverse order can occur as well. Explicit relations make up ~43% of all relations (see also Table 3).

(2) Implicit relations consist of two arguments (Arg1 and Arg2) only, because an overtly realised connective was considered redundant by the author; in a sequence like *"Mary broke her leg. She could not attend the festival the next day."*, a causal relation can easily be inferred without an explicit connective. Corresponding to the PDTB annotation guidelines, for implicit relations the annotators specified the connective that could be inserted between the two arguments (but crucially is not present in the text). Finally, they equally contain a relation sense. Implicit relations make up ~38% of all relations.

(3-5) If between two adjacent segments (typically sentences), neither an explicit nor an implicit relation could be assigned, the annotator furthermore had the option to choose between the remaining three types of entity relation (EntRel), alternative lexicalisation (AltLex) or no relation (NoRel). EntRel cases (~12% of all relations) are those where no particular relation from the PDTB sense hierarchy could be assigned, but the two segments speak of the same entities. As such, they only contain two arguments (and no – explicit or implicit – connective). AltLex cases (~2% of all relations) are those where the relation was explicitly expressed through something other than a discourse connective. Discourse connectives are seen as a closed class (though different theories and frameworks disagree on specifics), and a typical alternative lexicalisation would be *At that time*, expressing a Temporal.Synchronous relation sense. Finally, NoRel cases (~0.6% of all relations) are those where no relation between two adjacent segments could be established by the annotator. We adopt this scheme and attempt to project any relation (except NoRel) found in the PDTB onto the GermanPDTB.

4. Method

The creation of the GermanPDTB can be decomposed into several steps, explained in more detail in the following subsections. First, we need to create a parallel, sentence-aligned corpus, comprising the raw, English text of the original PDTB on the one hand, and the raw, German text of the GermanPDTB in-the-making on the other hand. Second, we extract word alignments from the parallel sentences. Third, we establish a set of heuristics based on the different annotation types present in the PDTB. In the process, we extend an already existing German lexicon of connectives: DiMLex (Stede, 2002). First introduced in 1998, this lexicon has been extended and refined over the last 20 years, resulting in a relatively exhaustive and stable lexicon of German discourse connectives. Still, in the process of creating the GermanPDTB, we found several items we consider candidate entries for the lexicon.

4.1. Creation of the Parallel Corpus

We use machine translation to produce a parallel corpus. We considered and tested five different systems – Google Translate[1], DeepL[2], Bing[3], Edinburgh's Neural Machine Translation system (Sennrich et al., 2016) and Moses (Koehn et al., 2007) – by translating the English side of a parallel news corpus (Tiedemann, 2012) and scoring the translation using BLEU (Papineni et al., 2002). Google Translate and DeepL produced the best translations with BLEU scores of 26.6 and 28.07 respectively, so we proceeded with these two systems and translated the English raw text of the PDTB. Though the BLEU scores are not particularly good, we determined by manual inspection that the translations can generally be considered good enough for creating the corpus. Next, we performed a separate manual evaluation on a subset of 50 sentences, following the approach proposed by Popovic et al. (2013). Since the translations for these 50 sentences were of equal quality for both systems, we determined for which one a direct alignment (German-English) retrieved more explicit connectives. As this was the case for the DeepL translation, we continued to work with this system.

4.2. Alignment Heuristics

Having obtained parallel English-German sentences, we proceeded with extracting word alignment using GIZA++.

[1] https://translate.google.com/
[2] https://www.deepl.com/en/translator
[3] https://www.bing.com/translator

First experiments with direct alignments were not promising and we encountered similar issues as those reported by Laali (2017). We therefore applied additional alignment heuristics implemented in Moses (in which GIZA++ is executed using IBM Model 4), similar to Laali (2017) who used the intersection and the grow-diag, and to Versley (2010) who used the intersection and the grow-diag-final heuristics. We experimented with six alignment versions in total. To evaluate these, we extracted the aligned German connective candidates and matched them against DiMLex to see how many are found.

All six versions build on the direct (English to German) and the inverse alignment (German to English). The intersection only contains alignment points that appear in the direct and the inverse alignment, while the union contains all alignment points from both alignments. The four remaining heuristics augment the intersection with alignment points from the union in various ways, as proposed by Och and Ney (2003). In the grow heuristic, word pairs neighbouring already aligned word pairs are aligned if they occur in the union. The grow-diag heuristic extends the notion of neighbouring words and is therefore less restrictive. In the grow-diag-final heuristic, a final step is added in which remaining word pairs get aligned if one word is not yet aligned and the word pair is aligned in the union. The grow-diag-final-and implements a more restrictive final step in which a word pair is only checked against the union if both words are not yet aligned. In short, the grow method is the most restrictive, followed by the grow-diag, the grow-diag-final-and and the grow-diag-final method.

The more restrictive a heuristic is, the more precise it is, but the fewer connectives are found in total. The results for the six different methods are presented in Table 1. We decided to favour precision over recall, and as the intersection version is the most precise, we use this heuristic as default alignment for projecting the connectives. For projecting the arguments, all alignment versions are used and a majority vote is retrieved. The same applies if in the intersection version a discourse marker is aligned to "NULL" or if the aligned word is not found in DiMLex.

4.3. Extension of the Lexicon

Using our heuristic of choice as described above, we manually analysed a subset of the projected (German) explicit connectives that were not found in DiMLex, allowing us to find sources of error in alignment/projection. The majority of cases that emerged from this manually-analysed subset evolve around modifiers for discourse markers. In the PDTB, modifiers (for example temporal modifications and focus particles as in *shortly thereafter* and *especially if*) are annotated as part of the explicit connective. In this respect, DiMLex has a more strict definition and includes the head of the explicit connective only, while regarding the modifier as an optional element (some of those are focus particles whose combination with connectives is restricted, which is also recorded in the DiMLex entries). To be able to reliably evaluate the explicit connectives we only annotate the "pure" form in the GermanPDTB, i.e., we iterate over all (German) words that are aligned to the (English) explicit connective and only annotate the ones matching an entry in DiMLex. After this step, some explicit connectives were found to be correctly aligned yet not present in DiMLex. For such cases, inspired by Meyer and Webber (2013), we extracted all explicit discourse markers from the PDTB, translated them with DeepL, checked them against DiMLex and discussed the ones not yet present. This resulted in 17 candidates that can be considered as additions to DiMLex.

4.4. Projection

We project the annotations from the English to the German side of the parallel corpus sentence-wise and relation-wise (several discourse relations can be annotated for one sentence). To not rely on the word alignments alone, we conducted an automatic analysis of the PDTB and compiled rules for the projection of the arguments and for the projection of the relations that are not explicit. For example, if an argument spans a whole sentence, the projection is possibly based on sentence alignment alone, and no word alignments are needed. Since the position of the arguments depends on the position of the connective, we start the projection by checking if a connective is present in the English sentence. If this is the case, we retrieve the alignment and check the word(s) to which the English connective is(are) aligned against DiMLex. If the result set is empty, we retrieve alternative alignments (see Section 4.2.) and look up these alignments. If this results in a non-empty set, we proceed with this instead. If none of the alternative alignment procedures resulted in a German word or phrase present in DiMLex, we extract all n-grams in a window with a size of five tokens around each connective and check if any is present in DiMLex with a matching relation sense. If this resulted in an empty result set too, we manually annotate the connective.

For implicit and AltLex relations, we want to assign the relation sense to the appropriate word or phrase (the alternative lexicalisation in the AltLex case, typically the first word of Arg2 in the implicit case), but in this case we cannot check the alignments against DiMLex. So for these relation types, we check whether the tags can be transferred using a rule (e.g. is annotated to the first word of a sentence); if they cannot, we retrieve a majority vote on the tag's position with all six alignment versions. EntRel relations are always annotated to the first word of the second argument, therefore the projection is included in the argument projection process.

The arguments, for all relation types, are projected in the following way; if the argument is continuous, a majority vote is compiled for the start and the end of the argument. Otherwise, we split the argument into continuous parts and retrieve the majority vote for the start and the end of each part.

Furthermore, to correctly annotate the implicit relations we created an English-German mapping for the connectives to be inserted. We further PoS-tagged the German raw text using MarMoT (Müller et al., 2013) to be able to present the GermanPDTB enriched with the same information as the PDTB.

Heuristic/Category	% of explicits not found	Total number explicits	Null alignments
Intersection	**8.7**	16401	1817
Grow	10.2	16901	1043
Grow-diag	11.5	17428	579
Grow-diag-final-and	12.0	17634	372
Grow-diag-final	13.5	18059	24
Union	14.8	**18354**	**23**

Table 1: Performance for the explicit connective projection for the different alignment heuristics sorted by restrictiveness.

Word	Count
if	17
and	14
however	14
while	13
meanwhile	12
also	8

Table 2: Most frequent explicit connectives that were annotated as implicit relations.

5. Manual Correction

After going through the procedures outlined above, 2.7% of the explicit connectives needed to be manually corrected, mainly due to four reasons.

1. The connective was correctly annotated but is not in DiMLex. These cases were discussed and four of them were considered candidates for DiMLex, resulting in 21 candidate entries for DiMLex, in total.

2. The connective was present in the sentence, but not found through alignment or by looking for it around the projected position. This was mostly due to the window size of five being too narrow in some cases (further increasing the window size however led to the inclusion of too much noise).

3. The connective was not present in the translation. Explicit relations in some language A can tend to be expressed more often implicitly in language B. This zero-translation case is a known problem in the literature (Meyer and Webber, 2013) .

4. No discourse relation is expressed in the translation mostly because a part of the translation is missing.

Table 2 displays the most frequent explicit connectives that were expressed through an implicit relation in German. In total, 164 explicit relations turned implicit. *if* was often not translated, mostly when it was at the beginning of a sentence, and *however* was mostly omitted when inserted within the sentence. Examples of both cases are provided below.

EN If, by that time, the network reaches 14 million homes, the contract will be renewed for five more years.

DE Erreicht das Netz bis zu diesem Zeitpunkt 14 Millionen Haushalte, wird der Vertrag um weitere fünf Jahre verlängert.

EN Few small neighborhood businesses, however, can afford such protection, even in collaboration with other local merchants.

DE Nur wenige kleine Unternehmen in der Nachbarschaft können sich einen solchen Schutz leisten, auch in Zusammenarbeit mit anderen lokalen Händlern.

An example of an incorrect alignment due to translation error is shown below, where the phrase *but that won't be enough* is missing in the German target text.

EN Mr. Koch already has announced he will drop 3,200 jobs from the city payroll, *but that won't be enough*.

DE Koch hat bereits angekündigt, dass er 3.200 Stellen von der Lohnliste der Stadt streichen wird.

6. Evaluation

We provide two types of evaluation. The intrinsic evaluation (first subsection) focuses on the output of the translation and projection procedures, and discusses frequent sources of errors. The extrinsic evaluation (second subsection) describes experiments using the obtained silver data for discourse parsing, i.e., it evaluates the quality of the output with regard to usability as training data for the parsing task.

6.1. Intrinsic Evaluation

Some key characteristics of the original PDTB and the GermanPDTB are summarised in Table 3. Due to the manual correction of the explicit connectives, there are fewer discourse relations in total in the GermanPDTB. There are also fewer explicit but more implicit relations; the number of AltLex and EntRel relations stays the same. The PDTB has more unique discourse markers than the GermanPDTB. This is most likely due to modifiers, which in the PDTB are part of the connective, but not so in DiMLex (see Section 4.3.). For comparison, we also extracted the heads of the connective in a naive way by only considering the last word. This method fails for some connectives, e.g., for *on the contrary*. Since the GermanPDTB contains more unique "naive" heads than the PDTB, this compensates for the difference in ambiguity of discourse connectives. In other words, when considering the full connective, German seems to be more ambiguous, because it expresses roughly the same number of relations/senses with fewer unique connectives. However, when looking at the head, the situation is reversed. We have not quantified in how many cases our naive way of extracting the head results in wrong heads being extracted, though.

	PDTB	GermanPDTB
Total relations	39,319	39,311
Explicit relations	16,888	16,670
Implicit relations	15,369	15,533
EntRels	4783	4,783
AltLexes	602	602
Unique connectives (in explicit relations)	232	185
Unique connectives (naive heads)	91	168
Arg1 token length (average)	18.13	17.91
Arg2 token length (average)	16.86	16.58

Table 3: Key characteristics of original PDTB and GermanPDTB.

To further evaluate our corpus we manually examined 150 discourse relations. Based on this, we distinguish between four kinds of errors:

1. Punctuation errors: A punctuation mark is not included in the annotation even though it is included in the PDTB relation, or vice versa.

2. Minor word errors: One word is not included in the annotation even though it is included in the PDTB relation, or vice versa.

3. Severe word errors: More than one word is not included in the annotation even though it is included in the PDTB relation, or vice versa.

4. Connective errors: The connective is wrongly annotated.

In the manually examined set, 141 out of the 150 relations were accurate (94%). In three cases (2%) there were punctuation errors; furthermore there were four minor (3%) and two severe (1%) word errors. Only severe word errors render an annotation useless, so based on this manual, intrinsic evaluation, we can conclude that 99% of cases are usable for our purposes. With the set of relations under investigation being very small though (150 relations, which is $<1\%$ of all relations), a larger sample size would provide a more reliable perspective.

6.2. Extrinsic Evaluation

Having established the quality of the annotations by looking at the relations themselves, we now turn to a more use-case driven evaluation. Specifically, we use individual components of a German discourse parser currently under development to assess the suitability of the obtained data for the tasks of connective disambiguation and argument extraction. To put performance into perspective, we compare performance of these components to their performance on the original, English PDTB data. All scores reported on are the result of 10-fold cross-validation.

6.2.1. Connective Disambiguation

First, we establish the quality of the projected data with regard to the task of connective disambiguation. To exemplify the task, consider the sentences in (1) and (2).

(1) A small but significant effect.

(2) Lucy had very little contact with the folks outside her cubicle day, but she found it suitable and she liked it that way.

While the *but* in (1) simply coordinates two noun phrases (with the noun elided in the first NP), the *but* in (2) indicates a relation between two propositions, and puts them in a contrastive relation. Our task entails binary classification, classifying candidates as having either sentential (as in (1)) or discourse (as in (2)) reading. Using the classifier described in (Bourgonje and Stede, 2018), on the GermanPDTB data, we get a binary f1-score of **94.04**. When using the same classifier on the English PDTB, Bourgonje and Stede (2018) report a very similar binary f1-score of 93.64. Comparing this, in turn, to the English competition, we note that the overall winning system of the 2016 CoNLL shared task on discourse parsing (Oepen et al., 2016) reports an f1-score of 91.79 for the sub-task of connective disambiguation. The system with the highest score for this sub-task in that same competition, however, achieved an f1-score of 98.38 (Li et al., 2016).
We suspect the difference in performance to be due to language-specifics, similar to those reported in Section 5., where German in some cases tends to implicit realisation, whereas English uses an explicit form. Further investigation would be needed to find the root cause of the 0.4 point difference in f1-score, but we consider the fact that scores are relatively close together a confirmation of generally good quality which we observed from manual evaluation in Section 6.1.

6.2.2. Argument Extraction

The second component on which we evaluate the GermanPDTB is argument extraction. In the PDTB framework, each coherence relation has two arguments which are put in some kind of relation to each other. Consider the sentence in (3).

(3) powerful political pressures may convince the Conservative government to keep its so-called golden share, *which limits any individual holding to 15%*, until **the restriction expires on Dec. 31, 1990**
(from the PDTB2.0: WSJ 0745)

The first argument is in italics and the second argument in bold face. The task of argument extraction is to decide upon the scope of both arguments and to extract (in the optimal case) the exact token span that makes up the argument. As

our discourse parser for German only works for explicit relations so far, the scores reported here for argument extraction are based on the 16,670 explicit relations in the GermanPDTB only. We use the approach described in (Bourgonje and Stede, 2019) and we follow their evaluation metric, which for every argument measures the token overlap between the actual and the predicted argument in the sense that every token that truly belongs to the argument and is classified as such results in a true positive; every token that does not belong to the argument and is classified as such results in a false positive; and every token that belongs to the argument and is not classified as such results in a false negative. Scores are averaged over 10 cross-validation runs, and use the connective annotation from the GermanPDTB directly, instead of using the classifier to predict the presence of connectives. When using the classifiers in combination with heuristics, we get an f1-score of **62.45** for Arg1 spans and **81.33** for Arg2 spans. The corresponding numbers for English reported by Bourgonje and Stede (2019) are 59.35 and 88.63, meaning that interestingly, Arg1 spans are easier to detect in the GermanPDTB, while Arg2 spans are more difficult to detect, compared to the original PDTB. We refer the reader to (Bourgonje and Stede, 2019), Section 5 for more details on how this compares to other competitors.

Upon manual investigation, we found that for both argument types (Arg1 and Arg2), attribution was a frequent source of error. The heuristics described in (Bourgonje and Stede, 2019) were devised based on the PCC, which consists of news commentary and contains very few cases of attribution. The PDTB contains such cases much more frequently (Prasad et al., 2006), and the token span expressing the attribution is typically left out of the annotated argument, but is included by the heuristics. This is supported by the lower precision and higher recall for both Arg1 spans and Arg2 spans (59.08 (precision), 66.25 (recall) and 78.79 (precision), 84.04 (recall), respectively). This, however, impacts both the German and English processing, and does not explain the difference in performance between the two. We leave further investigation into the cause for this difference to future work.

7. Conclusion & Future Work

We demonstrate how a large discourse-annotated corpus can be created by machine-translating the original English Penn Discourse TreeBank and exploiting word alignments to project the annotations over the English text onto the translated – in our case, German – text.[4] We discuss the procedure used to obtain the corpus and evaluate it by manually establishing the quality of the annotations on the German text for a small subset of the corpus. Additionally, in this process, we identify 21 candidates that we consider potentially valuable additions to DiMLex (a German connective lexicon). For an extrinsic evaluation, we use the obtained corpus as training data for selected sub-tasks (connective classification and argument extraction) of the larger task of discourse parsing and compare the obtained results to the same architectures trained on the original English, obtaining similar results for the two sub-tasks under investigation.

[4]The release of the data via LDC is currently in preparation.

Another important piece of future work is the further extrinsic evaluation of the corpus using a German discourse parser currently under development. Once this component is available for the parser, we plan to use the GermanPDTB for the sub-task of sense classification (the next step after connective classification and argument extraction in a typical pipeline setup). In addition, we plan to establish whether or not individual components trained on the GermanPDTB improve performance when evaluating on a gold corpus, the Potsdam Commentary Corpus.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 323949969. We would like to thank the anonymous reviewers for their helpful comments on an earlier version of this manuscript.

8. Bibliographical References

Asher, N. and Lascarides, A. (2005). *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press.

Bourgonje, P. and Stede, M. (2018). Identifying Explicit Discourse Connectives in German. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 327–331, Melbourne, Australia. Association for Computational Linguistics.

Bourgonje, P. and Stede, M. (2019). Explicit Discourse Argument Extraction for German. In *Proceedings of the 21st International Conference on Text, Speech and Dialogue*, Ljubljana, Slovenia.

Bourgonje, P. and Stede, M. (2020). The Potsdam Commentary Corpus 2.2: Extending Annotations for Shallow Discourse Parsing. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Paris, France, May. European Language Resources Association (ELRA).

Bourgonje, P., Grishina, Y., and Stede, M. (2017). Toward a bilingual lexical database on connectives: Exploiting a German/Italian parallel corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics*, Rome, Italy, December.

Hajlaoui, N. and Popescu-Belis, A. (2013). Assessing the Accuracy of Discourse Connective Translations: Validation of an Automatic Metric. In *14th International Conference on Intelligent Text Processing and Computational Linguistics*. Springer.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Laali, M. and Kosseim, L. (2014). Inducing Discourse Connectives from Parallel Texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 610–619, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Laali, M. (2017). *Inducing Discourse Resources Using Annotation Projection*. Ph.D. thesis, Concordia University.

Li, Z., Zhao, H., Pang, C., Wang, L., and Wang, H. (2016). A Constituent Syntactic Parse Tree Based Discourse Parser. In *Proceedings of the CoNLL-16 shared task*, pages 60–64, Berlin, Germany, August. Association for Computational Linguistics.

Mann, W. and Thompson, S. (1988). Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text*, 8:243–281.

Meyer, T. and Webber, B. (2013). Implicitation of Discourse Connectives in (Machine) Translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26.

Müller, T., Schmid, H., and Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Oepen, S., Read, J., Scheffler, T., Sidarenka, U., Stede, M., Velldal, E., and Øvrelid, L. (2016). OPT: Oslo–Potsdam–Teesside—Pipelining Rules, Rankers, and Classifier Ensembles for Shallow Discourse Parsing. In *Proceedings of the CONLL 2016 Shared Task*, Berlin.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Popovic, M., Avramidis, E., Burchardt, A., Hunsicker, S., Schmeier, S., Tscherwinka, C., Vilar, D., and Uszkoreit, H. (2013). Learning from human judgments of machine translation output. In *Proceedings of the MT Summit XIV. Machine Translation Summit (MT-Summit-2013)*, Nice, France, September. The European Association for Machine Translation.

Prasad, R., Dinesh, N., Lee, A., Joshi, A., and Webber, B. (2006). Annotating Attribution in the Penn Discourse TreeBank. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, pages 31–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Sennrich, R., Haddow, B., and Birch, A. (2016). Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany, August. Association for Computational Linguistics.

Stede, M. (2002). DiMLex: A Lexical Approach to Discourse Markers. In Lenci A. et al., editors, *Exploring the Lexicon - Theory and Computation*. Edizioni dell'Orso, Alessandria.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Versley, Y. and Gastel, A. (2012). Linguistic Tests for Discourse Relations in the TüBa-D/Z corpus of written German. *Dialogue and Discourse*, 4(2):142–173.

Versley, Y. (2010). Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In L. Ahrenberg, et al., editors, *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, pages 83–92. Northern European Association for Language Technology (NEALT).

Xue, N., Ng, H. T., Pradhan, S., Rutherford, A., Webber, B., Wang, C., and Wang, H. (2016). CoNLL 2016 Shared Task on Multilingual Shallow Discourse Parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany, August. Association for Computational Linguistics.

Zeldes, A., Das, D., Maziero, E. G., Antonio, J., and Iruskieta, M. (2019). The DISRPT 2019 Shared Task on Elementary Discourse Unit Segmentation and Connective Detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 97–104, Minneapolis, MN, June. Association for Computational Linguistics.

Zhou, L., Gao, W., Li, B., Wei, Z., and Wong, K.-F. (2012). Cross-Lingual Identification of Ambiguous Discourse Connectives for Resource-Poor Language. In *Proceedings of COLING 2012: Posters*, pages 1409–1418, Mumbai, India, December. The COLING 2012 Organizing Committee.