Creating a German discourse parsing corpus by transferring relations between languages

Johann Seltmann

1 Introduction

Shallow Discourse Parsing is the task of finding relations between adjacent segments of text, to understand the structure of the text. It is called "shallow" since it does not look for bigger (tree-like) structures in the text. The two connected segments are referred to as *Arguments*. The arguments can be connected by a word or a phrase signifying the relation, which is called a *Connective*. In that case, the relation is called *explicit*, otherwise *implicit*. One can also differentiate other relation types, but I limited the corpus I created to these two. Each relation also has a *sense*, which signifies of what kind the relation is.

Example 1.1. An example of a discourse relation on a sentence in the Europarl corpus:

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.*

The sense of this relation is Contingency.Condition, meaning the first argument lays out the condition under which the situation described in the second argument will come to pass. I marked the first argument with a blue underline and boldface and the second one with a green underline and italics. I will keep it like that in the further examples. The boxes mark the connective, which in this case is not continuous.

A major problem for shallow discourse parsing is that any data for it has to be expensively annotated by humans qualified for that task. As a result, there are not many large datasets available. For English, the PDTB [12] contains more than 40000 relations. This is large enough to train parsers on it. For German, on the other hand, the largest existing corpora are the Potsdam Commentary Corpus (PCC) [3] and the TED Multilingual Discourse Bank [21], which contain only about 2200 and 3600 relations, respectively. This limited data presents a challenge in training discourse parsers for German and other languages with small or no corpora.

One approach to generating more data for shallow discourse parsing relies on the assumption that translations will maintain the discourse structure of the overall text, i.e. if there is a relation between two text segments in one language,

that relation will also exist between the translations of these segments. This idea has been used in different works that attempt to generate more data for discourse parsing.

In this project, I combined existing approaches in order to generate a shallow discourse parsing corpus for German.

This report consists of the following parts

- An overview of existing approaches to generating data for discourse parsing using translation.
- The method I used to create several versions of a corpus for German.
- An analysis of these corpora, comparing them to the PCC.
- An analysis of the individual parts of the creation process.
- Experiments using an existing parser for German, testing the trained parser on the PCC, in an attempt to measure the quality of the created corpus.

The code for the project is available at https://github.com/sejo95/GeDisCo.git.

2 Related Work

2.1 Finding implicit relations through back-translation

The PDTB still contains relatively little data. That especially hurts implicit sense classification, since that is harder than explicit sense classification, which can rely on the connectives. To solve that problem, Shi et al. use back-translation to automatically find more instances of implicit relations. They use a sentence-aligned English-French corpus and automatically back-translate the French text into English.

The idea is that a human translator will sometimes insert an explicit connective in a relation when translating (or conversely, drop the connective, thus turning an explicit relation implicit). They use a statistical translation system which is likely to preserve these inserted connectives. The back-translated text will then contain explicit relations, which are implicit in the original English

They then run a discourse parser, which was trained on the PDTB, on the back-translated text. Since parsers work better on explicit relations than implicit ones, the parser does more reliably find these relations. They then transfer the senses of these implicit relations onto the original English text if

- two adjacent sentences were also identified as being arguments of a relation by the parser run on the original English and
- there is a corresponding explicit relation in the back-translated text.

Essentially, they use the back-translation to improve the quality of the implicit relation senses in the English dataset they create.

To test their approach, they train a neural classifier on their created data and test it on the implicit relations in the PDTB test set. It achieved a 15-class accuracy of 45.5%.

In [14], they test their approach using different languages for the back-translations: They use the French, Czech, and German sections of the Europarl corpus [6] to create back-translations. In addition to identifying specific features arising with each source language, they also take majority votes between the different back-translations to create more accurate corpora. According to their experiments, they get the best results by combining the relations from two back-translations.

2.2 Identifying discourse connectives using inter-language alignment

Shi et al.'s approach is only feasible when there is an adequate parser for explicit relations in a language. In order to get more data for training such a parser, Laali and Kosseim automatically annotate discourse connectives in French texts. To do that, they use the English-French section of the Europarl corpus.

- In a first step, they identify possible discourse connectives (candidates) in a French text using a dictionary of French discourse connectives.
- Then, they identify discourse connectives in the English text using their ClaC parser [9].
- They use Moses to identify word alignments between the French and English texts.
- Then, for each French candidate connective, they differentiate:
  - if it was aligned with an English discourse connective marked by the ClaC parser, it is included in the corpus;
  - if it is aligned to something else than a connective, it is included, but marked as Non-discourse usage (NDU);
  - if it is not aligned to anything, it is dropped.

They compared their corpora against small gold sets, which they annotate. They also train a classifier for discourse connectives on their corpora and test them on the French Discourse Treebank (FDTB, [17]). These experiments generally yield a high precision and a low recall, indicating that this approach may identify too few explicit discourse connectives.

The GIZA++ [10] aligner, which is part of the Moses system, produces two alignments: one from French to English and one from English to French. They experiment with different variants of combining the two alignments. They conclude that the intersection between the two alignments works best for projecting discourse connectives, due to its high precision.

2.3 Translating the PDTB

Sluyter-Gäthje et al. tackle the problem of creating a German discourse parsing corpus by translating the text of the PDTB into German. They then use word-alignments produced by GIZA++to project the relations onto the German text. For explicit relations, they use the intersection alignment but include additional heuristics to find the connective in the German text if it is there. For implicit relations, conversely, they try to identify, if there is a connective in the German text.

That way, they create a corpus containing circa 39000 relations. In a manual analysis of 150 relations, only about 6% of these showed some kind of error. In addition, they train two components of the (forthcoming) GermanShallowDiscourseParser (GSDP)[1], specifically the Connective Classifier [1] and the Explicit Argument Extractor [2], and test them on a held-out part of their corpus. They achieve a 94% binary F1-score for the connective classifier, 62.45% for the extraction of Arg1 and 81.33% for Arg2.

This approach has the advantage that it can rely on gold standard annotations as a basis for the transfer of relations. A disadvantage is that it is still limited to the size of the PDTB.

3 Building the German corpora

3.1 Creating English discourse relations

The corpus is based on the Europarl corpus [6], in the pre-tokenized version from Opus [2] [18]. The Europarl corpus contains speeches from the European parliament, which have been translated into the official languages of the European Union. Like Shi et al., I use the English, French, and Czech sections of the corpus as bases of the corpora. This selection has some sense since these languages come from different language families which should increase the chance of having an explicit discourse connective in one of them. The files each contain one debate with several speakers; I split them into one file for each speech in order to have one continuous text in each file. I also remove files that do not occur in all three languages.

Example 3.1. A sentence from the Europarl corpus in its human-translated versions, which I will use as a running example:

EN: This is a very important matter and I hope it will be implemented with the same zeal with which we implement our own regulations affecting our own fishing communities .

CS: To je velice důležitá záležitost a doufám , že bude provedena se stejnou horlivostí , s jakou provádíme své vlastní předpisy týkající se našich vlastních rybolovných komunit .

[1] https://github.com/PeterBourgonje/GermanShallowDiscourseParser
[2] http://opus.nlpl.eu/Europarl.php

FR: C' est un point très important et j' espère que cette interdiction sera appliquée avec le même zèle que celui avec lequel nous mettons en œuvre nos réglementations qui affectent nos propres communautés de pêche .

DE: Das ist ein sehr wichtiger Punkt und ich hoffe , dass er mit dem gleichen Eifer umgesetzt wird , mit dem wir unsere eigenen Vorschriften für unsere eigenen Fischereigemeinschaften umsetzen .

Also like Shi et al., I use the Moses statistical machine translation system [7] to translate the French and Czech texts into English. We will refer to these translations as *translated French* and *translated Czech* texts. The reason for using Moses rather than a neural translation system is that this is more likely to preserve the explicit connectives. The fluency and accuracy of the translation, on the other hand, are of less concern, since the final corpus will only contain the (human-translated) German text. I use pre-trained versions of Moses, which were trained on the Europarl corpus, i.e. we are working on the training data, which should improve translation quality.

Example 3.2. Translated Czech and French texts:

CS: this is a very important issue and i hope it will be implemented with the same zeal with which we implement our own regulations affecting our own fishing communities .

FR: c' is a very important point and j' hope it will be implemented with the same zeal with which we implement our own regulations affecting our own fishing communities .

I use the parser by Wang and Lan to extract discourse relations from the the original English and translated texts. This parser was originally trained on the PDTB 2.0 corpus [12]. I chose this parser in part because it performed relatively well on the CoNLL 2015 Shared Task [20] and in part because it is publicly available and ready-to-use. Being trained for the CoNLL task, it uses the specific input format for the task, which is a json format that contains syntax information in addition to the plain text. In order to transfer the input to that format, I used a script from the UNITN Penn Discourse Treebank Discourse Parser, which in turn uses the Berkley Parser [11] and the Stanford parser [4].

Example 3.3. Example of a relation found by the parser on English sentence:

Expansion.Conjunction:

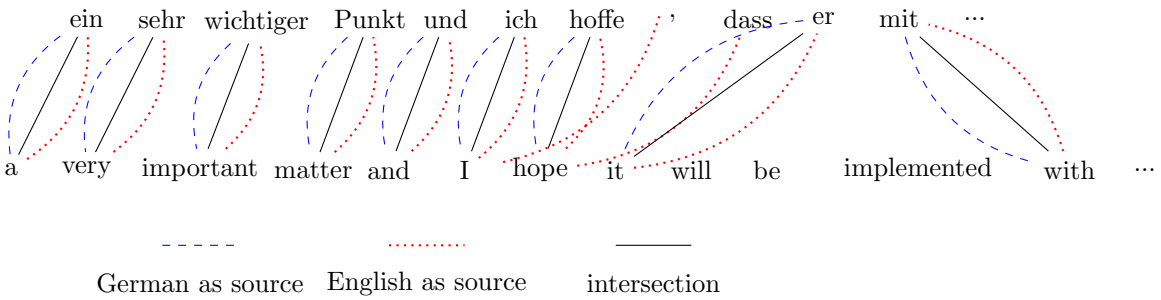
This is a very important matter and *I hope it will be implemented with the same zeal with which we implement our own regulations affecting our own fishing communities* .

3.2 Transferring relations to German text

In order to transfer the relations onto the German text, I created word alignments between the English, translated French, and translated Czech texts, respectively, and the German texts using the Giza++ alignment tool [10], specifically using the mgiza implementation. This creates two alignments, one with

German as source and one with German as target language. In order to have one definite alignment these two have to be combined, for which there are several methods. For this, I chose the *intersection* method, where two words are considered to be aligned with each other only if they are aligned with each other in both alignment directions. This method has the highest precision in identifying connectives, which is why it is also used by Laali and Kosseim.

Example 3.4. Example of an alignment found between English and German text:

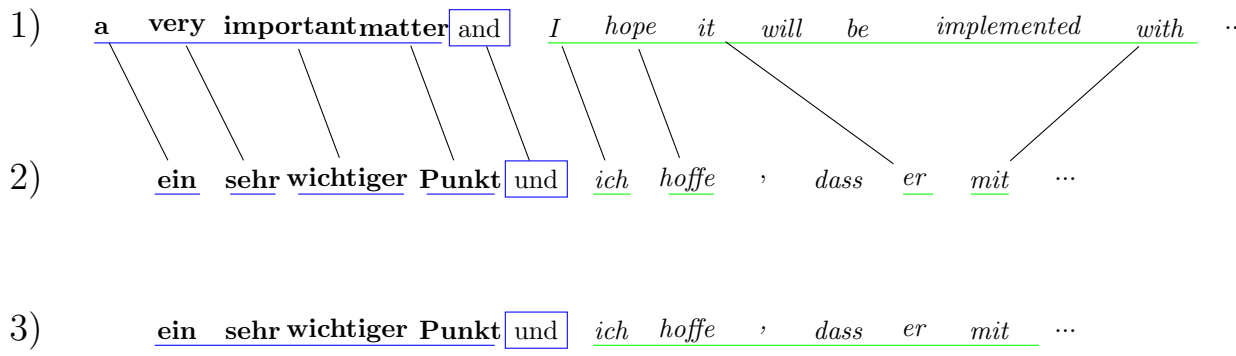


Alignments for words that are aligned to words outside of the text are not shown.

Then the process of transferring a relation is done by replacing the word indices in the connective and the arguments of the parsed relation with indices of the aligned German words. In addition to that, each transferred argument is made to be one continuous span in order to include words that may have been skipped by the alignment.

If the original relation is explicit, the words aligned with the connective are compared to the DimLex dataset [16], which contains German discourse connectives. If the aligned words form a German discourse connective whose sense matches the assigned sense of the relation, the relation is kept as explicit, otherwise the connective is ignored and the relation is taken to be an implicit one. On the other hand, if the original relation is implicit, we also check if one of the arguments ends or begins with a connective and if that connective corresponds to the sense with which the relation was annotated. In that case, the relation is taken as an explicit one.

Example 3.5. Example of the transfer of a found relation:



1) the relation found by the parser

2) transfer to the German words aligned with the English words in the relation

3) inclusion of all words in a continuous span into each argument

Since the Wang/Lan parser uses PDTB-2 senses, but DimLex is annotated with PDTB-3 senses, there are some senses in DimLex that do not occur in the parser output For the transfer, we map these senses to their lower levels, e.g. "Expansion.Level-of-detail" to "Expansion". We then assume senses to match if the sense assigned by the parser is also an "Expansion", regardless of what higher level is then assigned.

3.3 Senses

The senses in the created corpora are the senses output by the Wang/Lan parser. They are largely PDTB 2 senses, except for EntRel, which is treated as a sense for implicit relations, rather than a relation type by itself. Table 1 shows the senses occurring in the created corpora in comparison to the PCC. Table 5 contains a numerical overview about which senses occur how often in the different corpora. Since the parser output does not contain sense information relating to the function of the arguments of the relations, I ignore these in the PCC as well, e.g. "Contingency.Condition.Arg1-as-cond" and "Contingency.Condition.Arg2-as-cond" are both mapped to "Contingency.Condition".

3.4 Combining transferred relations from languages

Shi et al. also create additional corpora by taking majority votes between the relations transferred from different languages. Similarly to that, I create combined corpora from the transferred relation by taking an intersection of two sets of relations. In that, two relations are judged to be the same if they have the same sense and there is at least as 50% overlap between the arguments of the two relations. The first and second arguments of the new relation are created

created corpora	PCC
Comparison.Concession	Comparison.Concession
Comparison.Contrast	Comparison.Contrast
	Comparison.Similarity
Contingency.Cause.Reason	Contingency.Cause.Reason
Contingency.Cause.Result	Contingency.Cause.Result
Contingency.Condition	Contingency.Condition
	Contingency.Purpose
	Contingency.Negative-condition
EntRel	EntRel
Expansion.Conjunction	Expansion.Conjunction
Expansion.Exception	Expansion.Exception
Expansion.Instantiation	Expansion.Instantiation
Expansion.Restatement	
Expansion.Alternative	
Expansion.Alternative.Chosen alternative	
	Expansion.Level-of-detail
	Expansion.Equivalence
	Expansion.Disjunction
	Expansion.Substitution
	Expansion.Manner
Temporal.Asynchronous.Precedence	Temporal.Asynchronous.Precedence
Temporal.Asynchronous.Succession	Temporal.Asynchronous.Succession
Temporal.Synchrony	Temporal.Synchronous
	NoRel

Table 1: Senses occurring in the created corpora compared to PCC

as the union of the first (or respectively second) arguments of the original relations. If one of the original relations was explicit, its connective then becomes the connective of the new relation. Otherwise, the new relation is implicit.

3.5 Corpus formats

The repository also contains code to transfer the corpus to two different formats used in Shallow Discourse Parsing:

- The json format used in the CoNLL 2015 Shared Task. I use the Standford parser as part of this.
- The xml format used by the PCC. The PCC differentiates between internal and external arguments where internal arguments are more closely connected to the connective of the relation. In this transfer, I judge an argument to be internal if it contains at least parts of the connective while the other argument contains none. This case (with overlaps between the connective and an argument) can happen because of the alignments and because of the union over the arguments when combining relation from two languages. In all other cases, I simply set the first argument as the internal one. This format is also used as input format for the GSDP.

  In addition, the PCC also contains syntax trees for its text. I create these parses using the Berkeley Neural Parser [5].

4 Analysis

In order to understand the resulting corpora, I performed three analyses:

- an internal analysis in which I analyze various properties of the corpora
- an external analysis by training a model on the corpora and testing it against the PCC
- as analysis of some of the components in the corpus creation process

4.1 Created corpora

I created the following corpora:

- from only one language: *from_en*, *from_cs*, *from_fr*
- as intersection of the above three corpora: *en_cs*, *en_fr*, *cs_fr*, and *cs_fr_en*, this being an intersection of cs_fr with from_en
- For the experiments with the GSDP, I also created corpora that differ from the other corpora in one aspect of the creation process, in order to better understand that aspect.

	from_en	from_fr	from_cs	en_cs	en_fr	cs_fr	cs_fr_en	en_cs_one_word	PCC
#documents	56499	56085	56088	55803	55701	55287	55040	55803	176
#words	14.4 M	14.2 M	14.3 M	14.1 M	14.1 M	13.9 M	13.9 M	14.1 M	33 k
#relations	674663	675808	687031	456558	180465	174214	141427	499753	2204
relations per document	11.94	12.05	12.25	8.18	3.24	3.15	2.57	8.96	12.52
relations per 100 words	4.69	4.75	4.82	3.23	1.28	1.2	1.02	3.5	6.63

Table 2: General data of the corpora

- *en_cs_one_word*: a corpus combining from_en and from_cs, which uses a less strict criterion for the intersection of two relations, namely that their senses match and that they have at least one word overlap for each argument
- *from_en_drop*: a corpus based on from_en, where a percentage of relations are dropped if the senses of these relations are overrepresented in from_en compared to the PCC.
- *from_en_arg1int* and *from_en_arg2int*: Since the GSDP uses the format of the PCC as input format, the relations need to be transferred to that format. These two corpora differ from the normal from_en by also setting the first (or second, respectively) relation argument as internal, rather than using the heuristic described in section 3.5.
- *from_en_impl_sents*: corpus that only contains implicit relations, which span between two sentences.
- *from_en_simple_drop*: from_en contains some relation senses much more often than the PCC. In this corpus, I dropped relations with those senses with a certain probability, thus creating a more balanced corpus.

4.2 Internal analysis

4.2.1 General data

Table 2 shows a basic overview of the size of the corpora and the number of relations in them. The number of documents in each corpus differs slightly, as I removed files where a step of the creation process (mostly the parsing) led to an error.

The number of relations is also shown relative to the number of documents and relative to the number of words. For the corpora built from one language, the number of found relations is similar to that in the PCC. However, when using the intersection of relations from different languages, this number decreases, especially in the cases, where the version from French is involved. This indicates that there is much less of an overlap between the from_fr version on the on hand and from_en and from_cs on the other hand.

	from_en	from_fr	from_cs	en_cs	en_fr	cs_fr	cs_fr_en	en_cs_one_word	PCC
implicit relations per document	8.17	11.1	8.71	5.29	2.17	2.18	1.66	5.64	5.14
explicit relations per document	3.77	0.95	3.53	2.89	1.07	0.97	0.91	3.32	6.32
percentage of implicit relations between two sentences	77.2	35.8	76.6	78.8	42.6	42.9	42.9	76.2	100
not between prev. explicit	51.4	40.2	47.6	49.6	14.0	12.9	0.0	52.5	

Table 3: Data on number of explicit vs. implicit relations in each corpus. The row "not between prev. explicit" counts the fraction of implicit relations *not* spanning two sentences which used to be explicit.

	from_en		from_fr		from_cs	
	before	after	before	after	before	after
implicit relations per document	6.94	8.17	7.67	11.1	7.59	8.71
explicit relations per document	5.14	3.77	4.81	0.95	4.89	3.53

Table 4: Number of explicit and implicit relations in the parsed texts, before transferring to German

Also note that the number of relations relative to the length of the text (*per 100 words*) is lower for from_en, from_fr, and from_cs than in the PCC. This is presumably due to the difference in domain.

4.2.2 Explicit vs. Implicit relations

The created corpora have a higher number of implicit relations than the PCC, but a lower number of explicit relations (see table 3). Table 4 shows the reasons for this. It shows the number of implicit and explicit relations of the parsed English and translated Czech and French texts before and after the transfer to the German text. For implicit relations, the numbers before the transfer are already higher than for the PCC, which is presumably due to the Wang/Lan parser, which finds much more implicit relations than there actually are (see section 4.4.2). During the transfer, a larger number of explicit relations then becomes implicit. This is of course part of the intention of parsing on one language (as with Shi et al.), to find more implicit relations in order to better train models for this. However, in this case, the comparison to the PCC suggests, that the numbers of implicit and explicit relations were already close to the actual number. Of course, this might also be due to the differing domains and text lengths. It also does not say anything about the quality of the parses on

the English and translated texts.

Also notable here is that for from_fr, there is only a small number of explicit relations, and many more implicit relations. During combination with from_en and from_cz, most of these implicit relations are "filtered out", which indicates that these relations do not correspond to any explicit relations in these corpora.

Table 3 also shows information on what percentage of implicit relations cover two sentences (*inter-sentential*). Since the transfer to German text can turn an explicit relation into an implicit one, some of the resulting implicit relations are intra-sentential. In addition to that, an incorrect word-alignment in the transfer can also cause an implicit relation to not exactly span between two sentences. The values in the table do account for that to some degree, by also counting a relation as being inter-sentential if the arguments do not contain all words in the respective sentence or do contain one word of an adjacent sentence.

For from_en and from_cs about one quarter of all implicit relations does not span between two sentences. Of that quarter, about half were originally explicit relations. Since the Europarl texts are sentence-aligned, the other half has to be caused by problems with the word-alignment in the transfer.

4.2.3 Relation Senses

Table 5 shows the distribution of relation senses in the corpora. It is largely similar between the created corpora, however, it is markedly different from the one in the PCC. To some degree this is a result of the difference between PDTB-2 and PDTB-3 senses. For example, the sense Expansion.Restatement occurs relatively often in the created corpora but does not exist in the PCC. On the other hand, a number of senses occurs much more often (or much rarer) in the created corpora than in the PCC, e.g. Contingency.Cause.Reason and Contingency.Cause.Result occur about twice as often as in the PCC.

This difference is partially caused by the Wang/Lan parser, which seems to often output Contingency.Cause.Reason/Result as sense for implicit relations. However, it is likely also influenced by the difference in domain between the texts, e.g. a parliamentary speech might be more likely to give reasons for and results of statements than a newspaper article.

4.3 Experiments

I trained the GermanShallowDiscourseParser on our corpora and test them on the PCC. I split the from_en corpus in to train set with $\approx$ 45000 documents and a test set with $\approx$ 5000 documents. Since training takes a long time on that dataset, I created smaller versions for the other corpora, containing 4500 train and 500 test files. We will refer to these sets as *from_en (small)* etc.

4.3.1 The GermanShallowDiscourseParser

The GSDP consists of four components:

- The Connective Classifier [1] extracts connectives from a given text.

		from_en	from_fr	from_cs	en_cs	en_fr	cs_fr	cs_fr_en	en_cs_one_word	PCC
Comparison	Concession	0.9	0.8	1.0	0.8	0.5	0.5	0.5	0.9	14.2
	Contrast	7.6	6.9	7.1	8.5	7.8	7.3	8.4	8.6	3.3
	Similarity									0.1
Contingency	Cause.Reason	27.3	29.9	28.7	29.4	33.3	34.4	33.8	28.7	13.7
	Cause.Result	14.7	14.5	15.1	14.8	14.6	14.9	14.6	14.3	7.0
	Condition	2.8	2.4	2.1	2.2	1.0	0.8	0.8	2.5	4.5
	other									0.1
EntRel		5.7	4.6	4.9	4.2	5.1	4.6	4.5	3.9	0.3
Expansion	Alternative	0.4	0.4	0.4	0.4	0.1	0.1	0.1	0.4	
	Conjunction	20.2	19.5	19.4	21.2	17.7	17.0	18.9	22.3	25.8
	Instantiation	0.7	0.7	0.7	0.8	0.8	0.8	0.9	0.7	5.7
	Restatement	15.0	15.5	15.8	13.7	17.3	17.8	16.2	13.0	
	other	0.2	0.2	0.2	0.2	0	0	0	0.2	13.5
None										1.6
Temporal	Asynchronous. Precedence	1.1	1.4	1.4	1.0	0.4	0.7	0.4	0.1	3.8
	Asynchronous. Succession	0.4	0.4	0.4	0.3	0.1	0.1	0.1	0.4	0.8
	Synchronous/ Synchrony	3.1	2.8	3.0	2.6	1.0	0.9	0.9	3.0	2.0

Table 5: Percentages of relation senses in each corpus. Empty fields indicate that that sense does not occur on the created corpora. "Other" combines several sub-senses which did not occurred only rarely or not at all in the created corpora.

- The Explicit Argument Extractor [2] then identifies the arguments belonging to the connectives found by the Connective Classifier.
- The Explicit Sense Classifier then finds the senses for the extracted relations. It also contains functionality to overwrite the classified sense, if it does not match with the senses assigned to the connective of the relation in the DimLex lexicon. Since the senses in the created corpora do not match the ones in DimLex, I disabled that functionality.
- The Implicit Sense Classifier then does the same for sentence pairs in order to find implicit relations between them.

All four components extract BERT features and syntactic information from the text and then train classifiers on these features.

4.3.2 CoNLL 2016 partial match scorer

I adapted the partial match scorer from the CoNLL 2016 task for the evaluation of the parser output. In order to identify a parsed relation with a gold annotated relation in the test set, the full scorer demands that the relations exactly overlap in both arguments. However, the trained parser is not accurate enough to have this complete overlap.

The partial scorer, on the other hand, only demands the overlap of the arguments reaches a certain F1-score to identify two relations to each other. I set this score to 30% here. In addition, I add functionality that parsed and gold relations that have not been aligned to each other with that 30% overlap can be aligned with each other if they have one word in common in each argument.

After producing aligned relations, the scorer then calculates precision, recall, and F1-score for how many connectives and arguments fulfill the 30% condition. It also calculates precision and recall of the sense classification.

4.3.3 Argument extraction results

Table 6 shows the results of the argument extraction. Specifically, it shows for all Arg1 and Arg2 which percentage of them could be aligned with an Arg1 and Arg2 from the PCC gold data, respectively.

One notable thing about these results is that they differ very little between the corpora on which the parser was trained, with only few outliers, like the small from_en corpus. The precision is (in most cases) larger than the recall, presumably because the parser finds much fewer explicit relations than are contained in the PCC.

Conjunctive Arg1 & Arg2 simply means, that the alignment simply looks if there is another Argument, irrespective if it is an Arg1 or Arg2. These values are correspondingly higher. The results of from_en_arg1int and from_en_arg2int show that this is mostly due to overlap with arguments from other relations, rather than the parser having problems in identifying the different arguments.

trained on	Arg1			Arg2			Conjunctive Arg1 & Arg2		
	prec	rec	f1	prec	rec	f1	prec	rec	f1
from_en (complete)	0.85	0.81	0.83	0.85	0.79	0.82	0.96	0.87	0.91
from_en (small)	0.77	0.79	0.78	0.77	0.80	0.79	0.88	0.91	0.89
from_fr (small)	0.86	0.80	0.83	0.85	0.79	0.82	0.96	0.86	0.91
from_cs (small)	0.81	0.79	0.80	0.82	0.80	0.81	0.93	0.88	0.91
en_cs (small)	0.86	0.80	0.83	0.85	0.78	0.81	0.96	0.86	0.91
en_fr (small)	0.86	0.79	0.82	0.85	0.78	0.81	0.96	0.86	0.91
from_en_simple_drop (small)	0.83	0.79	0.81	0.83	0.79	0.81	0.95	0.88	0.91
en_cs_one_in_common (small)	0.80	0.80	0.80	0.80	0.80	0.80	0.91	0.90	0.91
from_en_arg1int (small)	0.77	0.79	0.78	0.77	0.79	0.78	0.89	0.91	0.90
from_en_arg2int (small)	0.80	0.80	0.80	0.79	0.80	0.80	0.91	0.90	0.91
from_en (small) on Europarl test set	0.77	0.84	0.80	0.77	0.85	0.81	0.85	0.94	0.90

Table 6: Overall argument extraction results on the PCC

trained on	Arg1			Arg2			Conjunctive Arg1 & Arg2		
	prec	rec	f1	prec	rec	f1	prec	rec	f1
from_en (complete)	0.37	0.11	0.17	0.78	0.39	0.52	1	0.08	0.15
from_en (small)	0.34	0.18	0.23	0.72	0.52	0.60	0.93	0.18	0.30
from_fr (small)	0.36	0.10	0.16	0.78	0.41	0.54	1	0.07	0.13
from_cs (small)	0.36	0.14	0.20	0.77	0.46	0.57	0.99	0.12	0.22
en_cs (small)	0.42	0.13	0.20	0.78	0.40	0.53	1	0.08	0.15
en_fr (small)	0.35	0.10	0.15	0.82	0.42	0.55	1	0.06	0.11
from_en_simple_drop (small)	0.52	0.23	0.32	0.76	0.45	0.56	0.99	0.21	0.35
from_en (small) on Europarl test set	0.32	0.34	0.33	0.46	0.52	0.49	0.47	0.53	0.50

Table 7: Argument extraction results for explicit relations

Interestingly, the results on the the held-out test set from from_en are slightly worse than the results for the PCC. This suggest that there is some general problem in using this data to train the parser.

This becomes more clear when looking at table 7, which shows the same metrics, but only for explicit relations. This happens presumably because the partial scorer can align the argument of an explicit relation to that of an implicit relation. Since this is not possible here, the results are much worse.

There are, however, a few noteworthy results in table 7. The precision is much higher than the recall, which happens because the parser only identifies a fraction of the explicit relations. Conversely, on the Europarl test set, the recall is higher, since for that set, the parser finds more relations.

The trained parser is also better at identifying Arg2 than Arg1. This is also noted by Sluyter-Gäthje et al., meaning this is not directly a result of the data. Interestingly, the parser performs much better when trained on

trained on	Arg1			Arg2			Conjunctive Arg1 & Arg2		
	prec	rec	f1	prec	rec	f1	prec	rec	f1
from_en (complete)	0.59	0.95	0.72	0.59	0.95	0.72	0.60	1	0.75
from_en (small)	0.59	0.94	0.73	0.59	0.94	0.73	0.61	1	0.75
from_en_simple_drop (small)	0.61	0.93	0.73	0.60	0.93	0.73	0.63	1	0.77
from_en_impl_sents (small)	0.60	0.91	0.73	0.60	0.91	0.73	0.63	1	0.77
from_en (small) on Europarl test set	0.76	0.89	0.82	0.76	0.88	0.82	0.81	0.96	0.87

Table 8: Argument extraction results for implicit relations

	explicit	implicit	total
contained in PCC	1112	905	2017
from_en (complete)	171	1757	1928
from_en (small)	525	1726	2251
from_cs (small)	329	1750	2079
from_fr (small)	145	1754	1899
en_cs (small)	170	1743	1913
en_fr (small)	147	1754	1901
cs_fr (small)	170	1743	1913
cs_fr_en (small)	155	1749	1904
en_cs_one_in_common (small)	429	1723	2152
from_en_simple_drop (small)	357	1670	2027
contained in Europarl test set	1760	4071	5831
from_en (small)	1755	4521	6276

Table 9: Number of relations contained in corpora and found by parser

from_en_simple_drop. This indicates that the transfer between languages might be a better approach for some relation senses than for others.

For the implicit relations (table 8), the results are almost the same for the different corpora. Specifically, the the recall ist close to one (exactly one in the case of Conjunctive Arg1 & Arg2) while the precision is only about 50%. This is because the originally parsed text contained an implicit relation for almost all sentence pairs. The GSDP trained on this data then does that, too, which means that most implicit relations in the PCC will be caught. For the Europarl test data, the recall is lower, because the GSDP only finds implicit relations between sentences, which means it misses implicit relations that do not span between two sentences. From_en_impl_sents leads to a lower recall (since it return fewer relations), but does not increase the precision.

Table 9 shows the number of relations found by the parsed, compared to the number of relations actually contained in the PCC and in the Europarl test set. For the Europarl test set, the numbers of the parser are close to the ones actually contained in the corpus. However, on the PCC, the parser finds much

fewer explicit relations than there are in the corpus and much more implicit ones.

This shows the difference in the number of explicit and implicit relations found is mostly a result of the data; not of the parser.

Of note is also that the number of found explicit relations varies somewhat between the training corpora. Especially surprising is that the parser finds more relations when trained on from_en (small) than on from_en (complete), which is also the reason for the slightly better performance in the explicit argument extraction. This shows that the larger size of the generated corpora does not necessarily outweigh the lower quality.

4.3.4 Tokenwise argument extraction

In addition to the above values, I also use the CoNLL2016 partial scorer to determine precision and recall on a token-wise base. So rather than simply identifying if the parser predicted a relation where there is a gold relation, this counts the words in the arguments of predicted and gold relations, respectively, and calculates the overlap between the two. For this, I ignore predicted relations, that do not have a corresponding gold relation, and vice versa. This is to get a better idea of how the parser performs in finding the correct argument spans to a given relation, rather than in identifying, if a relation is there after all. I only look at these values for explicit relations, since the GSDP only takes whole sentences as arguments to relations.

Table 10 shows the results for this. Overall, the results for Arg2 are (again) better than for Arg1. Generally, the precision is higher than the recall, indicating that the parser marks too few words as being part of an argument. Notably, the parser performs worse on the Europarl test set than on the PCC, which means that the performance here is to a large degree influenced by the parser, rather than the data it is trained on.

Interestingly, the corpora created by combining two languages perform better for Arg1 than the others, as does from_en_simple_drop. This, again, suggests that the transfer works better for specific senses.

I also used the test functionality of the GSDP to get tokenwise values for the Europarl test set, which can be compared to the results reported by Sluyter-Gäthje et al.. The results are worse than for theirs, achieving an f1-score of 84% for the Connective Classifier, 16% for Arg1 and 42% for Arg2. One reason for this is are the intra-sentential implicit relations, which the GSDP cannot find. But even considering this, the results are bad.

4.3.5 Sense classification

Since the senses of the produced Corpora and of the PCC do not exactly match, I test two versions of the parsed data: One with all sense levels and one where only the first sense level is considered.

Table 11 shows the average results for the sense-classification. The results are bad and do not differ much between corpora. The results for from_en_simple_drop

trained on	Arg1			Arg2			Conjunctive Arg1 & Arg2		
	prec	rec	f1	prec	rec	f1	prec	rec	f1
from_en (complete)	0.53	0.27	0.36	0.73	0.68	0.70	0.42	0.31	0.36
from_en (small)	0.47	0.26	0.33	0.67	0.59	0.63	0.35	0.26	0.30
from_fr (small)	0.44	0.28	0.34	0.77	0.65	0.71	0.38	0.31	0.34
from_cs (small)	0.46	0.27	0.34	0.68	0.61	0.65	0.41	0.33	0.37
en_fr (small)	0.49	0.31	0.38	0.81	0.69	0.75	0.36	0.29	0.32
en_cs (small)	0.52	0.38	0.44	0.74	0.64	0.69	0.64	0.51	0.57
cs_fr (small)	0.52	0.38	0.44	0.74	0.64	0.69	0.38	0.32	0.35
cs_fr_en (small)	0.51	0.33	0.40	0.74	0.68	0.71	0.38	0.31	0.34
from_en_simple_drop (small)	0.55	0.44	0.49	0.69	0.62	0.65	0.49	0.43	0.46
from_en (small) on Europarl test set	0.42	0.37	0.39	0.61	0.50	0.54	0.43	0.36	0.39

Table 10: Tokenwise relation extraction accuracy for explicit relations

	precision	recall	F1
from_en (complete)	0.12	0.11	0.11
from_en (small)	0.13	0.13	0.13
from_cs (small)	0.13	0.12	0.13
from_fr (small)	0.13	0.11	0.12
en_cs (small)	0.13	0.12	0.12
from_en_simple_drop (small)	0.17	0.16	0.16
from_en_impl_sents (small)	0.12	0.11	0.12
from_en (small) on Europarl test set	0.30	0.33	0.32

Table 11: Average accuracy of sense classification. The results of the other corpora are not shown, since they are the same. These values also include gold relations that were not found and parsed relations that do not exist in the gold data.

	precision	recall	F1
from_en (complete)	0.33	0.30	0.32
from_en (small)	0.32	0.33	0.32
from_cs (small)	0.32	0.31	0.32
from_fr (small)	0.33	0.30	0.31
en_cs (small)	0.32	0.29	0.30
from_en_simple_drop (small)	0.34	0.32	0.33
from_en (small) on Europarl test set	0.43	0.47	0.45

Table 12: Average accuracy of sense classification, looking only at first sense level

are slightly better, showing that the different distribution of senses has some influence on the performance. A more refined approach tackling this problem might yield slightly better results. Looking at the Europarl test set, again, shows results that are comparatively better but not very good.

Table 12 shows the results for only sense level one. These results are (as expected) better than for level three, but still not good. The advantage of from_en_simple_drop vanishes almost completely, which makes sense since most sense-levels one occurs in both the PCC and the created corpora. It also shows that the parser has at least to some degree learned to differentiate between the different (coarser) senses.

4.3.6 Parsing results for individual senses

Table 13 shows the accuracy for individual senses on the PCC for from_en_5k and from_en_simple_drop. 13 of the 24 senses are either never predicted or do not occur in the PCC. Of the remaining, only four (Contingency.Cause.Reason, Contingency.Cause.Result, Contingency.Condition, and Expansion.Conjunction), have f1-scores at or above the average. Two senses have high precision because the parser only assigns them rarely to a relation.

There does not seem to be a general pattern to theses values: Contingency.Cause.Reason and Result are among the relations occurring most often in the corpora. Contingency.Condition, on the other hand occurs much less in the corpora than in the PCC and has good results anyway. A reason for this might be that some senses are more often explicit, rather than implicit, which would make it easier to identify them. However, the Contingency.Cause relations are often implicit (see table 15), and still have better results than other senses.

When looking at from_en_simple_drop, there is a similarly unclear picture: The results for many senses are improved or at least remain similar, however,the results for Contingency.Condition and Comparison.Contrast actually drop. This suggest that this engineering of the corpus to be more similar to the PCC might too simplistic. On the other hand, some differences between the PCC and Europarl are to be expected, since they are from (slightly) different domains.

	from_en_5k			from_en_simple_drop		
	prec	rec	f1	prec	rec	f1
Comparison.Concession	0.50	0.01	0.01	0.13	0.00	0.01
Comparison.Contrast	0.05	0.15	0.07	0.04	0.08	0.06
Comparison.Similarity	1	0	0	1	0	0
Contingency.Cause.Reason	0.17	0.45	0.25	0.19	0.38	0.25
Contingency.Cause.Result	0.10	0.14	0.12	0.15	0.28	0.20
Contingency.Condition	0.42	0.17	0.24	0.25	0.11	0.15
Contingency.Negative-Condition	1	0	0	1	0	0
Contingency.Purpose	1	0	0	1	0	0
EntRel	0.06	0.20	0.09	0.06	0.31	0.10
Expansion.Alternative	0	1	0	0	1	0
Expansion.Alternative.Chosen alternative	0	1	0	0	1	0
Expansion.Conjunction	0.24	0.16	0.20	0.34	0.25	0.28
Expansion.Disjunction	1	0	0	1	0	0
Expansion.Equivalence	1	0	0	1	0	0
Expansion.Exception	1	0	0	1	0	0
Expansion.Instantiation	0.25	0.02	0.04	0.27	0.08	0.12
Expansion.Level-of-detail	1	0	0	1	0	0
Expansion.Manner	1	0	0	1	0	0
Expansion.Restatement	0	1	0	0	1	0
Expansion.Substitution	1	0	0	1	0	0
NoRel	1	0	0	1	0	0
Temporal.Asynchronous.Precedence	0.08	0.01	0.02	0.13	0.02	0.04
Temporal.Asynchronous.Succession	0	0	0	0.07	0.06	0.06
Temporal.Synchrony	0.07	0.02	0.03	0.25	0.05	0.08

Table 13: Results for individual senses senses on the PCC. Senses that either did not occur in the PCC or were never predicted are greyed out.

4.3.7 Manual error analysis

I also did a manual error analysis on two parsed documents from the PCC, where the parser was trained on from_en. Out of 33 relations, 12 were completely missing, 10 of which were explicit. Three relations identified by the parser were not present in the gold data. Of the relations which were correctly identified (ignoring incorrect argument spans), 15 had the wrong sense, 6 had the correct one.

This analysis fits largely with the analysis of the relations produced by the Wang/Lan parser (see table 14). There are more missing relations and fewer correct or incorrectly identified relations. This is mostly because the trained parser misses most explicit relations.

4.4 Analysis of corpus creation

In order to better understand the earlier results, I analyzed the different parts of the corpus creation process.

4.4.1 Translation

For the back-translation to English, I measured the corpus-level BLEU-scores (ignoring case). For Czech it is 0.85 and for French 0.63. Looking at the French translations (e.g., example 3.2), the main problem seems to be that they still contain French particles, like "j́" and "ć".

If we look at a case-sensitive BLEU-scores, (0.69 for Czech and 0.60 for French), the Czech one is similar to the scores of the translations used by Shi et al., while the French one is still lower. That indicates that the implicit relations in the from_cz corpus might be of comparable quality to the ones found in their work, although there are of course other influences on that.

4.4.2 Parsing on English and translated texts

In order to better understand parser performance, I manually analyzed a small number of relations found by the Wang/Lan parser on three different texts. Table 14 shows the results.

For the English texts, the parser found 33 relations. Out of these, only thirteen had the correct sense. Five relations were identified completely incorrectly, three (all of them explicit) were missed. However, the largest group of errors are relations where the sense was identified incorrectly. The same is true for the parses on the translated French texts. I did not analyze the parser performance on the translated Czech texts because of the large similarity between the from_cz and from_en corpora.

This weakness on the side of the original parser is presumably the main reason for the weak performance of the parser trained on the created corpora.

Correctly identifying the argument spans of the relations seems to work well (for English and the translated Czech), except that the parser sometimes misses one or two tokens at the beginning or ends of arguments.

	total	correct	first level sense correct	wrong sense	incorrect	missing implicit	missing explicit
English	33	13	3	11	5	0	3
translated French	32	10	1	13	8	1	5

Table 14: Manually analyzed relations

4.4.3 Transfer to German text

For the transfer of connectives and argument spans to the German text, I manually analyzed the relations for from_en and from_cs in one document. Out of a total of 31 relations, 18 were transferred correctly. 12 had a minor problem of missing a word at the beginning or end of an argument span. This was mostly due to different word orders in English and German. There was a bigger problem in only one relation, where Arg2 was reduced to one word in the transfer. Judging from this small sample, the transfer of relations seems to work well.

However, there is the issue of explicit relations becoming implicit. Table 15 shows for each sense in the from_en corpus, what percentage of its occurrences change their relation type between explicit and implicit. Most relations (when parsed on English) occur largely in one of the types. Senses that have a relatively large percentage of each explicit and implicit relations are Contingency.Cause.Reason/Result, Expansion.Conjunction, and Expansion.Instantiation. The first three are also among the most common senses.

In no case does a large number of implicit relations become explicit. However, for the senses that originally are mostly explicit, are large percentage of the relations becomes implicit. The distribution of relation types in the PCC (see table 3) shows that this is not due to a different distribution of explicit vs. implicit relations in German. Next to the parser, this is another weak point of the corpus creation process.

4.4.4 Combination of languages

In order to study the combination of languages, I manually analyzed one document from from_en and from_cs and compared how the relations matched up. The from_cs document contained 15 relations, from_en 16. Of these, 10 could be aligned with each other. However, for three relation pairs which were correct matches, the overlap was not high enough to match them together.

Because of that, I analyzed the percentage of aligned relations between the from_en and from_cs corpora for different alignment thresholds (see table 16). For the 50% overlap plus matching sense condition used in the corpus creation, about two thirds of the relations could be aligned with other relations. It is not possible to tell from this, whether two relations aligned with each other are actually the same relation. However, I also looked at the number of relations, that could be aligned with more than one relation in the other corpus (duplicates). Since that number is very low for most possible conditions, I assume that the match is usually correct.

	ex to ex	ex to im	im to ex	im to im	total number of relations
total	0.29	0.14	0.03	0.55	674663
Comparison.Concession	0.75	0.25	0	0	5867
Comparison.Contrast	0.73	0.25	0	0.02	50983
Contingency.Cause.Reason	0.09	0.02	0.06	0.83	184464
Contingency.Cause.Result	0.22	0.10	0.06	0.62	98933
Contingency.Condition	0.64	0.36	0	0	18668
EntRel	0	0	0	1	38526
Expansion.Alternative	0.69	0.31	0	0	2675
Expansion.Alternative.Chosen alternative	0.57	0.43	0	0	1073
Expansion.Conjunction	0.57	0.28	0.01	0.14	136256
Expansion.Exception	0.4	0.6	0	0	20
Expansion.Instantiation	0.34	0.49	0.0	0.17	4977
Expansion.Restatement	0.01	0.03	0.03	0.94	101467
Temporal.Asynchronous.Precedence	0.56	0.44	0	0	7242
Temporal.Asynchronous.Succession	0.52	0.48	0	0	2841
Temporal.Synchrony	0.61	0.39	0	0	20641

Table 15: Percentage of relations changing from explicit to implicit and vice versa for from_en. Rows sum to 100% (except for rounding errors)

The percentage of aligned relations grows when we only demand that there be one word in common per argument; over 90% when we put no restrictions on the sense. However, in that case, the number of duplicates also rather high. Therefore, I decided to also add a corpus created from from_en and from_cs in the experiments, which demanded that the arguments of the relations have on word in common and have the same sense.

5 Conclusion

The created corpora showed relatively poor performance, overall. The analysis revealed several weaknesses in the creation process that can explain that and some conclusions for future work in this area:

- The parser used to parse the English text produced too many incorrect relations. Using either a better parser or directly using gold annotations (like Sluyter-Gäthje et al.) would lead to much better results.
- The transfer of relations to a different languages works relatively well, at least for English and German, which are, of course, closely related to each other. However, there needs to be a better heuristic (like the one used by Sluyter-Gäthje et al.) to find discourse connectives. Otherwise, the transfer will result in too many implicit relations.

	from_en		from_cs	
	aligned	duplicates	aligned	duplicates
50% overlap	0.68	0.008	0.66	0.008
30% overlap	0.71	0.011	0.70	0.011
20% overlap	0.72	0.013	0.70	0.013
10% overlap	0.73	0.017	0.71	0.016
one in common ignore sense	0.94	0.098	0.91	0.095
one in common and sense	0.73	0.02	0.71	0.019
one in common and sense level 2	0.78	0.029	0.76	0.0278
one in common and sense level 1	0.78	0.029	0.76	0.0278

Table 16: Percentages of relations from from_en and from_cs that were aligned with relations from the other corpus according to different thresholds. For the "x% overlap" thresholds, the sense also has to be the same. The column "duplicates" gives the fraction of relations which could be aligned to one or more other relations.

- Using information from several languages did lead to a slightly higher performance on some sub-tasks of Shallow Discourse Parsing (like Argument Extraction, when looking at tokenwise accuracy). However, it did not generally lead to better results and, therefore, any attempt to improve the corpus creation should be focussed on the other issues, rather than the language combination.
- When using different languages, the translation quality is crucial to obtaining useful results.
- Comparing the performance on the complete and small versions of the from_en corpus does not show a clear picture: In some tasks, the larger corpus lead to better results, in some it did not, and in others, the small corpus even produced better results. Overall, the larger size did not offset the low quality of the corpus.
- Another issue is the mismatch between the annotation standards used for different corpora. Examples for that are the different senses between the PDTB and the PCC and the existence of intra-sentential implicit relations in the created corpora. This is not necessarily a problem when it comes to training discourse parsing models, but it makes it harder to test the quality of the corpora.
- Engineering the corpus to be more similar to the PCC (e.g. by dropping some relations) does improve the results by a bit, so improvements in this area might help overall. One could argue that this means "gaming

the system", instead of learning from the actual data, but I think it is justified since the PCC consists of gold data and the created corpora do not.

Overall, I conclude that while the approach of transferring languages can work, my particular attempt mostly failed.

References

[1] P. Bourgonje and M. Stede. Identifying explicit discourse connectives in German. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 327–331, Melbourne, Australia, 2018. Association for Computational Linguistics.

[2] P. Bourgonje and M. Stede. Explicit Discourse Argument Extraction for German. In *Proceedings of the 21st International Conference on Text, Speech and Dialogue*, Ljubljana, Slovenia, 2019. URL `https://link.springer.com/chapter/10.1007/978-3-030-27947-9_3`.

[3] P. Bourgonje and M. Stede. The Potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1061–1066, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://www.aclweb.org/anthology/2020.lrec-1.133`.

[4] D. Chen and C. Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1082. URL `https://www.aclweb.org/anthology/D14-1082`.

[5] N. Kitaev and D. Klein. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[6] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

[7] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P07-2045`.

[8] M. Laali and L. Kosseim. Improving discourse relation projection to build discourse annotated corpora. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 407–416, Varna, Bulgaria, Sept. 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_054. URL https://doi.org/10.26615/978-954-452-049-6_054.

[9] M. Laali, A. Cianflone, and L. Kosseim. The CLaC discourse parser at CoNLL-2016. In *Proceedings of the CoNLL-16 shared task*, pages 92–99, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-2013. URL https://www.aclweb.org/anthology/K16-2013.

[10] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[11] S. Petrov, L. Barrett, R. Thibaux, and D. Klein. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia, July 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220230. URL https://www.aclweb.org/anthology/P06-1055.

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

[13] W. Shi, F. Yung, R. Rubino, and V. Demberg. Using explicit discourse connectives in translation for implicit discourse relation classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 484–495, Taipei, Taiwan, Nov. 2017. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-1049.

[14] W. Shi, F. Yung, and V. Demberg. Acquiring annotated data with cross-lingual explicitation for implicit discourse relation classification. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 12–21, Minneapolis, MN, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2703. URL https://www.aclweb.org/anthology/W19-2703.

[15] H. Sluyter-Gäthje, P. Bourgonje, and M. Stede. Shallow discourse parsing for under-resourced languages: Combining machine translation and annotation projection. In *Proceedings of the 12th Language Resources and*

Evaluation Conference, pages 1044–1050, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.131.

[16] M. Stede. DiMLex: A lexical approach to discourse markers. In *Exploring the Lexicon - Theory and Computation*. Edizioni dell'Orso, Alessandria, 2002.

[17] J. Steinlin, M. Colinet, and L. Danlos. FDTB1: Repérage des connecteurs de discours en corpus. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 34–40, Caen, France, June 2015. ATALA. URL https://www.aclweb.org/anthology/2015.jeptalnrecital-court.6.

[18] J. Tiedemann. Parallel data, tools and interfaces in opus. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

[19] J. Wang and M. Lan. A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 17–24, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/K15-2002. URL https://www.aclweb.org/anthology/K15-2002.

[20] N. Xue, H. T. Ng, S. Pradhan, R. Prasad, C. Bryant, and A. Rutherford. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/K15-2001. URL https://www.aclweb.org/anthology/K15-2001.

[21] D. Zeyrek, A. Mendes, Y. Grishina, M. Kurfali, S. Gibbon, and M. Ogrodniczuk. Ted multilingual discourse bank (ted-mdb): a parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–38, 2019.