Linking discourse relations and induction of bilingual discourse connective lexicons

Murathan Kurfalı [a], Sibel Özer [b,*], Deniz Zeyrek [b], Amália Mendes [c] and Giedrė Valūnaitė Oleškevičienė [d]

[a] *Linguistics Department, Stockholm University, Stockholm, Sweden*
E-mail: murathan.kurfali@ling.su.se
[b] *Cognitive Science Dept., Middle East Technical University, Ankara, Turkey*
E-mails: e159606@metu.edu.tr, dezeyrek@metu.edu.tr
[c] *Center of Linguistics, University of Lisbon, Lisbon, Portugal*
E-mail: amaliamendes@letras.ulisboa.pt
[d] *Institute of Humanities, Mykolas Romeris University, Vilnius, Lietuva*
E-mail: gvalunaite@mruni.eu

Abstract.
The single biggest obstacle in performing comprehensive cross-lingual discourse analysis is the scarcity of multilingual resources. The existing resources are overwhelmingly monolingual, compelling researchers to infer the discourse information in the target languages through error-prone automatic means. The current paper aims to provide more direct insight regarding the cross-lingual variations in discourse structures by offering an aligned version of a multilingual resource, namely TED-Multilingual Discourse Bank, which consists of independently annotated six Ted talks in seven different languages. It is shown that discourse relations in these languages can be automatically aligned with high accuracy, verified by the experiments on the manual alignments of three diverse languages. The resulting alignments have a great potential to reveal the divergences the target languages exhibit in local discourse relations, with respect to source text, as well as to lead to new resources, as exemplified by the induction of bilingual discourse connective lexicons.

Keywords: discourse relations, discourse connectives, lexicons of discourse connectives, linking discourse relations, parallel corpus

1. Introduction

Representing linguistic content in the form of linked data has recently become an active area of research in the field of Natural Language Processing. There has been a growing interest for linked data models and applications, leading to knowledge graphs, the wordnet, and dictionaries, to name a few. Following the TextLink project[1], there has been an effort to present discourse-level phenomena in the form of linked data, one of the most prominent of these being the Connective-Lex database [1]. Connective-Lex is a joint online database project, which currently hosts monolingual connective lexicons of ten different languages. It provides a web-based interface together with a cross-linguistically applicable XML schema and has the aim of extending the database to other languages. The entries in the Connective-Lex database provide information on discourse connectives (*but, once, although*) such as their orthography, syntactic category (coordinating conjunction, adverb, subordinating conjunction), and the senses they convey (contrast, temporal, concession).

TED-Multilingual Discourse Bank (TED-MDB), a resource created to capture the discourse-level properties of English TED talks and translations into multiple languages (European Portuguese, Lithuanian, German,

*Corresponding author. E-mail: e159606@metu.edu.tr.
[1] http://textlink.ii.metu.edu.tr/

Russian, Polish, and Turkish) offers an ideal domain to induce monolingual and bilingual discourse connective lexicons for a new set of languages. But, given its design characteristics, where each language set was annotated independently of the source language (see Section §3), this resource presents a challenge to the induction of discourse connective lexicons because the discourse relations are not aligned, thus they cannot be linked to each other. This, in turn, hinders any efforts of lexicon induction or cross-linguistic comparisons among the languages included in the corpus. To support further research, a discourse relation (DR) alignment task (or, discourse relation linking task) must be performed on TED-MDB (in this work, the term 'linking' is used interchangeably with the 'alignment' of discourse relations to refer to the true semantic relationship between a pair of discourse relations in the source text and the target text).

The present work aims to interlink the discourse-annotated data of TED-MDB and induce bilingual lexicons. The main contributions of the paper are: (1) to propose two alternative methods to align the discourse relation annotations of TED-MDB, one relying on traditional word alignments and the other one employing multilingual sentence embeddings. To the best of our knowledge, the latter method has not been investigated specifically for linking data in the context of discourse research, and not for the languages under consideration in the present work; (2) to automatically align the discourse relations annotated in TED-MDB enhancing the data structure of the corpus; and (3) to automatically induce new bilingual discourse connective lexicons for each TED-MDB language (Target Language-TL) and English (Source Language-SL), substantially increasing the number of available such bilingual lexicons[2].

The rest of the paper proceeds as follows: in the next section (§2), the main data source, TED-MDB is summarized along with the existing bilingual and multilingual discourse connective lexicons in the literature. §3 offers the description of the data linking task (alternatively referred to as the Discourse Relation (DR) Alignment task) by providing the details of the two proposed methods (§3.1 and §3.2). This section also provides an evaluation of the interlinked data as well as various issues and challenges confronted during the DR alignment task (§3.3). In §4, an overview of the discourse structures observed in TED-MDB is presented together with the statistics obtained from DR mappings. In §5, the bilingual lexicons induced from the DR-aligned data are described. The paper ends with a conclusion and some future directions for further research §6.

[2]All lexicons are publicly available at: http://metu-db.info/mdb/ted/resources.jsf

2. Background

2.1. TED Multilingual Discourse Bank

TED talks are prepared presentations given in English to a live audience. The audio/video recordings are made available online, together with English subtitles in a large set of languages, which are translated by volunteers and checked by experts. The subtitles ignore most dysfluencies, such as hesitations and filled pauses, although pragmatic discourse makers, such as *well*, are usually retained. The wide coverage of TED talks in terms of topics and translated languages make them an ideal source of data for parallel corpora and contrastive studies on a spoken genre.

The raw texts annotated in TED-MDB consist of English transcripts, and their translations into six different languages annotated in the Penn Discourse TreeBank (PDTB) style [2]. The talks were presented by native English speakers and cover different themes as listed in Table 1.

Table 1
The list of the TED talks annotated in TED-MDB [3]

ID	Author	Title
1927	Chris McKnett	The investment of logic for sustainability
1971	David Sengeh	The sore problem of prosthetic limbs
1976	Jeremy Kasdin	The flower-shaped starshade that might help us detect Earthlike planets
1978	Sarah Lewis	Embrace the near win
2009	Kitra Cahana	A glimpse of life on the road
2150	Dave Troy	Social maps that reveal a city's intersections and separations

In TED-MDB, discourse relations are identified as holding between two arguments, Arg1 and Arg2, taken as abstract objects [4] and are typically marked by a discourse connective, such as *and*, *because*, *however*. Arg2 is the text segment that is syntactically related to the discourse connective, Arg1 is the other text segment. TED-MDB has applied the 5 types of relations identified by the PDTB 2.0: Explicit, Implicit, Alterna-

tive Lexicalization (AltLex), Entity Relation (EntRel) and No Relation (NoRel).

A discourse relation is Explicit when a discourse connective makes the relation that holds between the two arguments salient, as in example 1. When there is no discourse connective that marks the relation, the relation is inferred from the context and the annotator inserts a connective (referred to as the 'implicit connective') that would make the inferred relation explicit, as in example 2. Discourse relations may be conveyed by lexical elements other than connectives. In those cases, it is not possible to insert an implicit connective because the context already contains elements that make the relation explicit, and the relation is annotated as an AltLex (example 3).

Discourse relations of the type Explicit, Implicit and AltLex are labelled with a sense chosen from the PDTB 3.0 hierarchy, such as Contingency:Cause:Result [5]. The format of the sense tags is such that, the first sense is referred to as the top-level or Level1 sense (e.g. Contingency). It shows the highest semantic category in the hierarchically organized semantic categories encompassing a set of subsenses. The sense tag lists the second level sense, or Level2 sense (Cause) of the top category, followed by the third level sense, or Level3 sense (Result), providing information about the full semantics of the relation.

Relations can also hold between entities, where one of the arguments provides additional information about an Entity introduced in the discourse in the other argument. These contexts are annotated as an Entity Relation, as illustrated in example 4. Finally, when no relation holds between the two adjacent segments, the relation is of the type NoRel (example 5). In TED-MDB, Explicit relations have been annotated intra and inter-sententially, while Implicit relations have only been annotated inter-sententially. The examples below are taken from the TED-MDB online; the connective is underlined, Arg1 is rendered in italics, and Arg2 in bold type; each example of the discourse relation, except EntRel and NoRel, is labelled with a sense.

1. *The world is changing in some really profound ways*, <u>and</u> **I worry that investors aren't paying enough attention to some of the biggest drivers of change, especially when it comes to sustainability**.
   [Explicit, Expansion:Conjunction] (English, TED Talk no. 1927)

2. *Os protésicos ainda usam processos convencionais , como a criação de moldes e gesso , para confecionar encaixes de próteses de um único material* . (implicit = por conseguinte) **Esses encaixes provocam uma quantidade intolerável de pressão nos membros de_ os pacientes , deixando -os com escaras e ferida**
   [Implicit, Contingency:Cause:Result] (Portuguese, TED Talk no. 1971)
   'Prosthetists still use conventional processes like molding and casting to create single-material prosthetic sockets. (implicit = consequently) Such sockets often leave intolerable amounts of pressure on the limbs of the patient, leaving them with pressure sores and blister'

3. *many of my early memories involved intricate daydreams where I would walk across borders, forage for berries, and meet all kinds of strange people living unconventional lives on the road.* <u>Years have passed</u>, but **many of the adventures I fantasized about as a child – traveling and weaving my way between worlds other than my own — have become realities through my work as a documentary photographer**
   [AltLex, Temporal:Precedence] (English, TED Talk no. 2009)

4. *I didn't understand how even one was going to hit the ten ring*. **The ten ring from the standard 75-yard distance, it looks as small as a matchstick tip held out at arm's length** [EntRel] (English, TED Talk no. 1978)

5. *They would, in fact, be part of a Sierra Leone where war and amputation were no longer a strategy for gaining power*. **As I watched people who I knew, loved ones, recover from this devastation, one thing that deeply troubled me was that many of the amputees in the country would not use their prostheses** [NoRel] (English, TED Talk no. 1971)

Additionally, the authors have included a new top-level sense called Hypophora, that applies in contexts where the speaker asks a question and immediately answers it with the purpose of creating dialogism and making the presentation livelier (example 6).

6. <u>Are</u> *investors, particularly institutional investors, engaged*? Well, **some are, and a few are really**

at the vanguard [AltLex, Hypophora] (English, TED Talk no. 1927)

During the annotation phase, each language was annotated simultaneously but independently of the original English texts to ensure that annotations capture the discourse structure of each translated language as independently as possible. This design criterion, coupled with variations in translation lead to different sets of relations annotated for each language. Table 2 provides the number and the percentage of each type of relation (Explicit, Implicit, AltLex, EntRel and NoRel) in each language.

2.2. Discourse Connective Lexicons

The last two decades have seen an upsurge in the development of discourse connective lexicons, such as DiMLex [6] and LexConn [7]. Researchers are also envisioning linking the existing lexicons [1]. However, the linking task poses certain challenges as various discourse connective lexicons vary in depth and detail of the information concerning discourse connectives. The *Spanish Diccionario de partículas discursivas del español* (DPDE – [8]) includes explicit information on discourse particles in Spanish but it excludes conjunctions and prepositions. The German resource *Handbuch der Konnektoren* [9, 10] contains discourse connective representations including their possible positions in a sentence, also the register and possible modifiers. Another problem faced while linking discourse connective lexicons is that discourse connectives are language specific and before linking the lexicons, researchers need to decide on the information that characterizes discourse connectives in different languages. In discourse-annotated corpora, the use of different annotation schemes such as PDTB, RST (Rhetorical Structure Theory) also pose more challenges for linking the information on discourse connectives.

Despite the challenges to the creation of discourse connective lexicons, and the difficulties posed by the TED-MDB data format, the uniform PDTB-style annotation of TED-MDB is a tremendous advantage. Finally, except for some recent attempts, multilingual discourse connective lexicons are few ([11] and [12]), and the field needs lexicons for more languages to enable various technology applications. The connective lexicons created in the present work are hoped to bring an added dimension to the existing lexicons.

3. Aligning the Discourse Relations of TED-MDB

The alignment of TED-MDB's independently created discourse relation annotations can be seen as a variant of the annotation projection task, where the aim is to transfer (manually or automatically), the annotated relations in one language to another through parallel corpora [13–15]. Annotation projection is used as a low-effort way of constructing linguistic resources in the target languages which is, otherwise, very costly. Hence, in annotation projection, the linguistic information is available only for one language. Being completely clueless about the target language, the projection methods are deemed successful to the extent that they mimic the annotations in the source language. In a corpus like TED-MDB, the annotations are available in the source language and the target language(s). While being highly useful in revealing the discourse-level information of specific languages, it is different from annotation projection, because in its present format, it is not known which source language relations correspond to the target language relations, or which source language relations are missing in the target language. Thus, to enhance the quality of TED-MDB data structure and to make it fully available for further cross-linguistic research, the discourse relations in English and the target languages must be aligned. To address this data linking challenge, cross-lingual variations among the discourse relations must be understood and handled carefully. These differences appear at several levels: Typically, the argument spans of the relations tend to vary across languages. For example, it may be that in one language, the abstract object interpretation of a text piece can be captured by annotating a longer span than what is annotated in the source language. Also, since both intra- and inter-sentential relations are annotated, a number of relations may be created over the same text piece (see example 7), which raises the need for finding the correct alignment among the annotated tokens with overlapping arguments. In example 7, inter-sentential relations conveyed by implicit English *because* and Turkish *çünkü* share argument spans with other relations of the same text signaled by English *also* and Turkish *de*, respectively (The last paragraph of the example is the retranslation of the target language (in this case, the Turkish relation) to English).

7. And these things are fundamental, of course, but *they're not enough.* (Implicit=because) **Investors**

Table 2
Distribution of discourse relation types in TED-MDB [3]

Language	Explicit	Implicit	AltLex	EntRel	NoRel	Total
English	290 (44%)	198 (30%)	46 (7%)	78 (12%)	49 (7%)	661
German	240 (43%)	214 (38%)	17 (3%)	59 (11%)	30 (5%)	560
Lithuanian	377 (50%)	246 (33%)	18 (2%)	79 (11%)	32 (4%)	752
Polish	218 (37,5%)	195 (33,5%)	11 (2%)	104 (18%)	52 (9%)	580
Portuguese	269 (43%)	256 (41%)	29 (5%)	38 (6%)	33 (5%)	625
Russian	237 (42%)	221 (39%)	20 (4%)	57 (10%)	30 (5%)	565
Turkish	315 (45%)	202 (29%)	60 (9%)	69 (10%)	51 (7%)	697
Total	1946	1532	201	484	277	4440

should also look at performance metrics in what we call ESG: environment, social and governance. [Explicit, Expansion:Conjunction] (English, TED Talk no. 1927)

Ve tabii ki bunların tümü gerekli, ama *yeterli değiller.* (Implicit=çünkü) **Yatırımcılar ÇSY diye adlandırdığımız üç faktörün performans metriklerine de bakmalılar: Çevre, sosyal ve yönetim** . [Explicit, Expansion:Conjunction] (Turkish)

'And of course all of these are necessary, but *not enough*.(Implicit=because) **Investors should also look at the performance metrics of three factors we call ESG: environment, social and governance.**'

The rest of the section presents two different approaches to respond to the above-mentioned challenges describing how discourse relation alignments are performed. The first approach is based on the traditional way of performing annotation projection through word alignments [13–15]. This approach requires large amounts of parallel data which makes it impractical for low-resource languages. The second approach employs multilingual sentence embeddings, which are shown to be effective for a related task, namely, parallel sentence mining even in the low resource scenario [16, 17].

3.1. Method I: DR Alignment through Word Alignments

In this approach, the discourse relations are mapped to each other through word alignments using a set of heuristics. In a pre-processing step, all the raw texts are sentence tokenized and aligned to remedy the variations in their format. Then, the texts are aligned at the word level through a statistical aligner (EFLOMAL [18]) to perform the discourse relation alignment in the final phase.

3.1.1. Sentence Alignment

Although TED-MDB is built upon the parallel corpus of TED talk subtitles, these texts are not explicitly aligned and exhibit differences to varying degrees in how they are structured. Therefore, firstly, all raw texts are normalized to a standard sentence-per-line format, where the paragraphs are separated.

Using NLTK's sentence tokenizer, the sentence segmentation procedure is performed. Then, all documents are aligned at the sentence level using the LF-aligner software[3], which is based on the hunalign algorithm [19]. The first attempt of aligning all seven languages to each other generated a number of mismatches due to the varying number of sentences in each translation, as listed in Table 3. Sentence alignments are especially crucial as any error in this step would be propagated through the pipeline. Therefore, all languages are aligned with English separately to maximize the alignment quality.

Table 3
Sentence counts in each talk of TED-MDB

TalkID	EN	DE	PL	LT	RU	PT	TR
Talk 1927	114	127	117	122	122	128	117
Talk 1971	27	26	30	31	26	28	28
Talk 1976	88	89	86	96	87	85	100
Talk 1978	82	81	95	88	85	83	83
Talk 2009	30	31	32	32	31	31	31
Talk 2150	44	58	58	45	65	57	62

3.1.2. Obtaining Word Alignments

Having aligned the raw texts with their English counterparts, the next step was to obtain word align-

[3]https://sourceforge.net/projects/aligner/

ments. However, the performance of word aligners heavily depend on the size of the parallel data and TED-MDB was too small to obtain reliable alignments. Therefore, for each language pair (i.e. English-Language X), separate model priors are developed through a custom parallel data using the model 3 of EFLOMAL[4] [18]. A parallel corpus is created for each language pair by concatenating the largest corpora of each language pair in the OPUS database [20]. All the corpora are obtained and processed using OpusTools[5] [21]. The data sizes of each corpora are listed in Table 4.

Table 4

The sizes of training sets used to train the word aligner for each English-Language X pair. The number refers to the sentences in one language.

Target Language	# of sentences
German	45,514,709
Lithuanian	4,915,547
Polish	52,800,073
Portuguese	48,663,333
Turkish	50,238,588
Russian	33,684,711

Word alignment is performed in both directions, resulting in two sets of alignments: *the forward alignments* include the alignments where the source language is set as English, and *the reverse alignments* involve word alignments in the reverse direction, where the source language is set to the non-English language. Yet, using alignments directly from either direction is reported to underperform [13, 15]; therefore, based on previous work, several symmetrization heuristics that combine forward and reverse alignments are explored:

- **Intersection**: keeps the alignments that exist in both directions. It is the most strict heuristic and leads to fewer but precise alignments.
- **Grow-diag**: Grow-diag expands on the intersection set by adding the diagonally neighbouring points.
- **Grow-diag-final**: Adds another step on grow-diag the heuristic, where the unaligned word pairs in grow-diag are aligned provided that those word pairs are in the union of the forward and reverse alignments.

[4]https://github.com/robertostling/eflomal
[5]https://github.com/Helsinki-NLP/OpusTools

3.1.3. Aligning Discourse Relations

In the last step, discourse relation alignment is performed using the word alignments. Due to the differences in the argument spans as well as translation effects (e.g. the omission of a connective present in the source language, or the addition of a connective not present in the source language), alignment cannot be straightforwardly performed by matching the relations, the words of which are found to be equivalent by the word aligner.

Discourse relation alignment is performed as follows: Given a relation in the source language, each component of that relation, i.e. the first and the second argument and (if any) the connective, are projected to the target language using the word alignments. As the initial check, it is made sure that more than half of the words in any part of the source relation is projected to the target text. Then, each relation in the target language is scored according to the overlap between their components and the projected spans. Connectives are given priority; if a target relation has a connective that perfectly matches the projected connective, then those relations are matched without further checking their arguments. For other relations, the target relation which has the highest score (i.e. in terms of the amount of overlap between the components of the target and projected relations) is selected as the aligned pair. However, particularly in cases where multiple relations are annotated over similar text spans (see example 7), the scores based on lexical overlap fail to be adequately discriminative. In those cases, the match between the target relation and the source relation is recorded as 1 if the senses match, 0 otherwise, and added to the score.

3.2. Method II: Alignment through Cross-lingual Sentence Embeddings

The second method utilizes the recent advancements in multilingual embeddings, where semantically similar linguistic units across languages are assigned similar representations, enabling a mapping with each other. The approach is built upon a previous study [22] that performed discourse relation alignments only for the English-Turkish pair in TED-MDB. This method starts with a pre-processing step which is similar to that of the first method (Section §3.1), the raw texts are sentence tokenized and aligned in the same manner.

For discourse relation alignment, all discourse relations in each bi-text unit are paired constructing DR matrices. Then, all pairs are assigned a composite score that reflects the agreement between the Level1

sense, Level2 sense and type of the matched pair as well as the semantic similarity between their text spans.

The overall score is scored as follows:

1. Firstly, the similarity score between the relation pairs are calculated. Pairs which do not have acceptable cosine similarity[6] are discarded. The semantic similarity is calculated as the cosine similarity between the LASER embeddings [23] of each relation. LASER supports 93 different languages and embeds sentences into a shared space where semantically similar sentences, regardless of their languages, are assigned similar representations. LASER provides two advantages, namely, it handles different languages in a very smooth way (as opposed to machine translation used in [22] which is a costly operation) and assigns contextual embeddings that capture the overall sentence meaning in a more compact way than the previously employed bag of words approach.
2. A score that reflects the SL DR - TL DR match (1 for match, 0 for mismatch) is added to the semantic similarity score. In a ranked manner, a match on Level1 sense is given a score of 1000, a match on Level2 sense is assigned 100, a match of Level2 sense is given 10, and 1 is assigned for DR type match. While the alignment algorithm gives the highest priority to Level1 sense matches, as there is no sense information for NoRels and EntRels, the DR type match also becomes prominent.
3. For each source relation, the target relation which yields the maximum score is marked as its alignment and the same procedure is repeatedly applied until no discourse relation pair is left in the matrices.

The whole procedure is exemplified on a sample sub-corpus of discourse relations given in example 4 consisting of three Explicit relations in two languages (EN, TR) signaled by (*but, as, and*) and (*ama 'but', gibi 'as', ve 'and'*), respectively. As the first step, all pairwise combinations of these relations are calculated, resulting in a (3x3) DR matrix as shown in Table 5. Then, following the scoring procedure, each pair is assigned a score. In example 4, while Turkish *Ama* matches English *But* in all four criteria and receives 1s, it matches *as* in two criteria, namely in Level1 sense and the DR type. Then, for each source relation (i.e. for each row), the target relation (i.e. the column) which has the maximum score is aligned (shown in bold Table 5). Brief explanations of the matched criteria are provided after example 4.

4. Years have passed, but many of the adventures I fantasized about as a child – traveling and weaving my way between worlds other than my own — have become realities through my work as a documentary photographer. But no other experience has felt as true to my childhood dreams as living amongst and documenting the lives of fellow wanderers across the United States. (English, TED Talk no. 2009)

   Yıllar geçti, ama çocuk olarak hayalini kurduğum birçok macera – benim dünyam dışındaki dünyalar arasında seyahat ederken ve yoluma dokunurken – bir belgesel fotorafçısı olarak işim aracıyla bunlar gerçek oldu. Ama hiçbir başka deneyim çocukluk rüyalarımı yaşayanlar arasında olmak kadar ve Birleşik Devlet boyunca gezgin arkadaşların arasında yaşamak kadar gerçek hissettirmedi. (Turkish, TED Talk no. 2009)

5. *English* :
   - DR-Explicit-Comparison.Concession.Arg2-as-denier-DC-**But**
   - DR-Explicit-Comparison.Similarity-DC-**as**
   - DR-Explicit-Expansion.Conjunction-DC-**and**

   Turkish:
   - DR-Explicit-Comparison.Concession.Arg2-as-denier-DC-**Ama**
   - DR-Explicit-Comparison.Similarity-DC-**kadar**
   - DR-Explicit-Expansion.Conjunction-DC-**ve**

Table 5

DR matrix for the sample corpus in example 4. The numbers refer to alignment scores based on sense/type agreement + semantic similarity (Args+Connectives)

	Ama	**kadar**	**ve**
But	**1111+1.71**	1001+1.56	0+1.62
and	0+1.49	0+1.05	**1101+1.75**
as	1001+1.55	**1101+1.71**	0 + 1.61

The examination of the preliminary results revealed the need for certain revisions. As mentioned before, it is common for more than one discourse relation to

[6] threshold of 0.6 is used during the experiments

hold between similar arguments ([24]) which leads to false discourse relation pairings. Semantic similarity between discourse connectives is also checked in addition to similarity between text spans. Second, due to structural differences among languages or translation choices, an AltLex in one language may map to an Explicit relation in another language. The alignment algorithm is unable to cover such cases as it works on sentence-aligned bitext units. In order to eliminate this pitfall, if a DR relation is not matched with a TL DR in its parallel unit, it is evaluated once more in the succeeding alignment unit.

3.3. Evaluation

In the literature, data linking quality is evaluated by using the standard precision, recall and the F-score metrics. Precision is the positive predictive value or the proportion of the assigned links that are true matches (also known as true positives). Sensitivity or recall is the proportion of the true matches that are correctly identified, and finally, accuracy is the proportion of the valid matches and non-matches that are correctly identified. F-score represents the performance of the method and it is the harmonic mean of precision and recall [24].

Data linking quality is dependent on the task domain and there is always a trade-off between precision and recall. Usually, when the number of non-matches is large in the data set, accuracy is not considered as a good measure. However, as the task at hand is aligning SL and TL annotations, accuracy should also be taken into consideration; providing information on the non-matching data pairs is as important as providing matching data. In annotation alignment, non-matching data offers valuable insights into linguistics, machine translation and in particular, into the assessment of the annotation quality.

The methods proposed in the current work are evaluated against the gold DR alignments of English-Lithuanian, English-European Portuguese and English-Turkish, which were provided by the TED-MDB annotators of the respective language. Gold alignments for other languages (English-German, English-Polish, English-Russian) are currently not available. In this section, for each method, the alignment performance of the two methods for each language pair will be presented using the metrics introduced before.

If a DR in the source language matches a DR in the target language and if this is captured by the alignment algorithm, it is counted as TP (True Positive). If the algorithm matches the DR of the source language to a false DR in the target language, it is counted as FP (False Positive). If the source DR does not have a match in the target language and if this non-matching relation is found by the algorithm, it is counted as TN (True Negative). In cases where the algorithm incorrectly matches a non-matching source DR to a target DR, it is counted as FN (False Negative). As the number and the set of relations differ from language to language, evaluation is done in each language and each aligning direction (SL > TL and vice versa). This evaluation method is preferred because only evaluating the DR pairs in one direction (e.g. SL > TL) would mean not considering TL DRs that have no matches in SL.

The evaluation results for both methods are given in Table 6 and 7. Overall, both methods yielded a good degree of and almost similar performance. In particular, Method I achieves a good degree of precision (Table 6), meaning that the links it finds have high probability to be a true match. However, the main difference arises at the point of recall and accuracy, because when compared to Method II (Table 7), Method I yielded more DRs that are left unaligned (False Negatives), missing a good number of existing alignments. The number of missed relations decrease as the symmetrization heuristics become less restrictive (grow-diag-final achieves the best recall for all language pairs); yet, the gain is minimal. A closer look at Method I's alignments revealed that some of the errors stem from the misaligned sentence pairs and were corrected through manual correction. Therefore, the second method stands out as the better alternative as it yields a higher performance as well as having a relatively simple pipeline with less dependencies.

Regardless of which method is used, the performance on the Lithuanian data is the lowest; that is, for Lithuanian, less uniformity is obtained with the Gold DR alignments. One of the possible reasons is that the total number of Lithuanian DRs to be matched (in total 752 DRs in six files) is more than the files of European Portuguese and Turkish DRs. More importantly, regardless of the language pair, there are several challenges to the DR alignment task. Translation itself, linguistic differences between SL and TL, or different annotation choices could be the cause of such differences. In those cases, both methods fail and performance decreases due to an increase either in False Positives (see 8) or False Negatives (see 6 and 7). An increase in those numbers affect all the performance metrics (precision, recall and accuracy). Here are some instances that led to performance drop:

Table 6
Method I (Alignment through Word Alignments) Quality metrics for each language. The first column refers to symmetrization heuristics, ranked from the most restrictive to least restrictive, as explained in sect. 3.1.2

Heuristics	Lang. Pair	TP	FN	FP	TN	Accuracy	Precision	Recall	F-Score
Intersect	EN LT	377	83	108	88	0.71	0.78	0.82	0.8
	LT EN	377	102	108	165	0.72	0.78	0.79	0.78
	EN PT	434	98	45	79	0.78	0.91	0.82	0.86
	PT EN	434	84	45	62	0.79	0.91	0.84	0.87
	EN TR	485	90	27	54	0.82	0.95	0.84	0.89
	TR EN	485	90	27	95	0.83	0.95	0.84	0.89
Grow-diag	EN LT	381	78	109	88	0.71	0.78	0.83	0.8
	LT EN	381	97	109	165	0.73	0.78	0.8	0.79
	EN PT	444	87	47	78	0.8	0.9	0.84	0.87
	PT EN	444	73	47	61	0.81	0.9	0.86	0.88
	EN TR	498	73	33	52	0.84	0.94	0.87	0.9
	TR EN	498	75	33	91	0.85	0.94	0.87	0.9
Grow-diag-final	EN LT	388	68	116	84	0.72	0.77	0.85	**0.81**
	LT EN	388	85	116	163	0.73	0.77	0.82	**0.79**
	EN PT	450	80	50	76	0.8	0.9	0.85	**0.87**
	PT EN	450	64	50	61	0.82	0.9	0.88	**0.89**
	EN TR	505	57	43	51	0.85	0.92	0.9	**0.91**
	TR EN	505	61	43	88	0.85	0.92	0.89	**0.91**

Table 7
Method II (Alignment through Cross-lingual Sentence Embeddings) Quality metrics calculated for each language

Lang. Pair	TP	FN	FP	TN	Accuracy	Precision	Recall	F-Score
EN-LT	470	14	114	58	0.8	0.8	0.97	0.88
LT-EN	470	23	114	145	0.82	0.8	0.95	0.87
EN-PT	518	17	53	68	0.89	0.91	0.97	0.94
PT-EN	518	6	53	48	0.9	0.91	0.99	0.95
EN-TR	533	20	64	39	0.87	0.89	0.96	0.93
TR-EN	533	24	64	76	0.87	0.89	0.96	0.92

Different argument spans are selected in language pairs: Translation differences often lead to different argument span annotations as in example 6. Here, since the text *now it occurred to me, as I thought about this* is translated as 'as I thought about this it occurred to me that ...' a longer Arg1 span had to be selected for the TL DRL.

6. Now *it occurred to me* , as **I thought about this**, why the archery coach told me at the end of that practice, out of earshot of his archers, that he and his colleagues never feel they can do enough for their team, never feel there are enough visualization techniques and posture drills to help them overcome those constant near wins. [Explicit, Temporal:Synchronous] (English, TED Talk no. 1978)

Bunun hakkında düşününce *neden okçuluk koçunun idmanın sonunda bana okçularının işitmeyeceği mesafeden, onun ve meslektaşlarının ekipleri için ne yapsalar yetmeyeceğini düşündüklerini, kazanmak üzere olmak konusunu aşmalarına yardımcı olması için yeterli gözünde canlandırma tekniği ve duruş eğitimi olmadığını söylediğini anlıyorum.* [Explicit, Temporal:Synchronous] (Turkish, TED Talk no. 1978)

'**When I think about this**, *I understand why the archery coach told me at the end of the training that from the distance that his archers wouldn't hear, what he and his colleagues thought was not*

enough for their team, that there was not enough visualization technique and posture training to help them overcome the issue of being about to win.'

Shifts in translation results in different realizations of discourse connectives. In example 7, even though the DR in the English sentence is aligned with the DR in Lithuanian, neither method could capture this link due to the different Arg2 annotations.

7. Now these initiatives create a more mobile workplace , and *they reduce our real estate footprint* , and **they yield savings of 23 million dollars in operating costs annually**, and avoid the emissions of a 100,000 metric tons of carbon . [Explicit, Expansion:Conjunction] (English, TED Talk no. 1927)

   To rezultatai šiandien – mobilesnės darbo vietos , mažinančios mūsų nekilnojamojo turto pėdsaką , o **tai leidžia sutaupyti 23 milijonus dolerių kasmetinių veiklos išlaidų ir sumažinti anglies dioksido išmetimą 100 000 metrinių tonų**. [Explicit, Contingency:Cause:Result] (Lithuanian, TED Talk no. 1927)

   'The result today is more mobile jobs that reduce our real estate footprint, saving 23 million dollar in annual operating costs and reducing carbon emissions by 100,000 metric tons.'

The argument spans of the SL DR set are only partially selected as an argument in the TL DR set. In example 8, the English DR is a non-matching data. However, both methods fail in this instance and match the DR with a TL DR as it shares a part of the Arg1 span.

8. *Good, you like it. I like it too. (Laughter) I like it because it pokes fun at both sides of the climate change issue.* I bet you can't guess which side I'm on. But **what I really like about it is that it reminds me of something Mark Twain said, which is, "Plan for the future, because that's where you're going to spend the rest of your life.** [Explicit, Expansion:Conjunction] (English, TED Talk no. 1927)

   Ótimo , vocês gostaram . Eu também gosto (Risos) Eu gosto porque faz troça de os dois lados de a questão de a alteração climática . *Aposto que não adivinham de que lado estou* . Mas **o que eu gosto em isto é que me lembra uma coisa que Mark Twain disse : " Planeia para o futuro , " porque é onde vais passar o resto de a tua vida.** [Explicit, Expansion:Conjunction] (European Portuguese, TED Talk no. 1927)

   'Great, you guys liked it. I like it too (Laughter) I like it because it makes fun of the two sides of the issue of climate change. I bet you don't guess which side I'm on. But what I like about this is that it reminds me of something that Mark Twain said: "Plan for the future," because that's where you're going to spend the rest of your life.'

4. Overview of the Discourse Structures of the TED-MDB languages

Parallel corpora have enabled a leap ahead in cross-linguistic investigations and in translation studies. Due to the scarcity of parallel corpus annotated for discourse relations on both sides, previous cross-lingual work is largely confined to a specific aspect of discourse, e.g. omission of discourse markers [25, 26], mostly using parallel data with manual annotations on only one side. However, thanks to the availability of discourse information on both ends and the DR alignments carried out in this work, TED-MDB enables studying the discourse of English and the translated texts in a comprehensive manner. To this end, in the rest of the section, a general overview of how the discourse structure of English and the TLs differ is outlined concentrating on two questions: (i) Do discourse relations exhibit differences in how they are realized (e.g. explicitly or implicitly) in different languages? (ii) How do the semantics of the relations that hold between the same text spans change cross-lingually? To answer these questions, the automatic alignments obtained from the second method (section 3.2) are used due to its higher performance. Therefore, the observations should be approached cautiously due to possible misalignments; yet, the high F-scores on capturing gold alignments (see Table 7) suggest that the reported results closely follow the gold distribution.

The following analysis is mainly confined with the descriptive analysis of the aforementioned points,

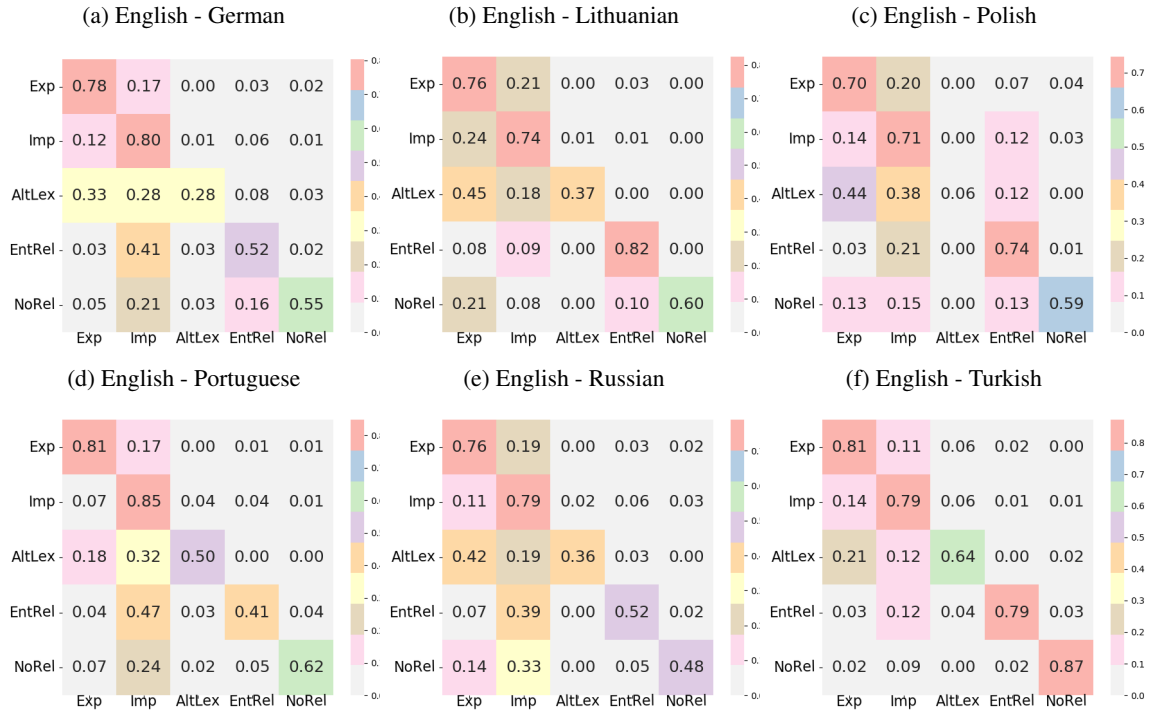

Fig. 1. Heatmap visualizations of the confusion matrices for relation type of the aligned discourse relations. Rows correspond to the English relations and columns denote target languages. The matrices are normalized row-wise where each cell denotes the percentage of English relations converted to the respective label in the target language.

leaving an in-depth linguistic investigation as a future work.

Cross-lingual Variation in Relation Types: In order to answer the first question, the relation types (Explicit, Implicit etc.) of each aligned relation are compared with each other in a pair-wise manner. Figure 1 shows the heat-map visualizations of the row-wise normalized confusion matrices for discourse relations in all language pairs. The rows represent the relations in English where each cell shows how often English relations are realized as the respective label on the X-axis. (e.g. the second cell of the first row of Figure 1a reads "17% of English explicit relations are realized implicitly in German.") Colors represent the density of agreements, where lighter colors visualize low agreement, getting redder as the agreement increases (a more detailed breakdown of the color-coding is provided in each figure). In a perfect match, only the diagonal cells would be red with the off-diagonal cells being complete white/gray.

According to Figures 1a to 1f, discourse relations vary greatly with respect to English annotations in terms of their types. On average, 555.5 of the English relations are aligned to each target language, and only 71.98% of them retained their type. Of the five types, the Implicit relations (77.95%) and Explicit relations are conserved most frequently, whereas the majority of the AltLex relations (61.71%) are converted into other relation types. The language-specific breakdown of these variations can be read in Figures 1a to 1f.

When all language pairs are considered, the top three conversions (from English to the target languages) are as follows: 33.33% of Altlex relations become Explicit; 27% of Entrels become Implicit and 17.37% Explicit relations become Implicit.

Of these three, English Entrels becoming Implicit are likely to stem from the annotators' interpretations, rather than language-specific factors, as 78.76% of these cases are annotated as implicit Expansion relations in the non-English language. Entrels and Implicits have been reported to be the most easily confused pairs even within the same language [27] as their distinction is very subtle. These two relations are semantically related to the extent that Entrel relations are exploited as implicit Expansion relations to increase the available training data in implicit discourse relation recognition task, yielding increases in overall performances. [28, 29].

Table 8
The sense distribution of the English relations that are implicitated (the left part) and those that are explicitated in the target language (the right part). AltLexes are included in the analysis.

	Implicitation				Explicitation			
	Expansion	Contingency	Comparison	Temporal	Expansion	Contingency	Comparison	Temporal
German	24	8	-	1	6	6	-	1
Lithuanian	29	6	2	2	15	13	2	4
Polish	32	2	3	3	2	9	-	2
Portuguese	31	6	2	-	3	7	-	1
Russian	23	4	1	-	3	7	1	-
Turkish	20	3	1	2	3	14	-	2

Finally, implicitation (the omission of a connective where there is a connective in the source language) is found to be the third common shift (or the second one, if Entrel to Implicit conversions are dismissed as being reasonably interchangeable) in relation types. Given that implicitation (and, its reverse, explicitation) are actively studied topics in discourse relations [30], the results of the current work can be used safely in future crosslinguistic investigations of implicitation (or explicitation). In all language pairs in TED-MDB, at least 10% of the English relations are found to be realized implicitly. These results raise a further question: are all explicit relations equally likely to be realized implicitly in the target language? Interestingly, implicitation dominantly occurs with Expansion relations (Table 8). The same is not true for explicitation, where Contingency relations are relatively more frequently explicitated than others on average, but they are far from being as dominant as the implicitated expansion relations (Table 8).

Cross-lingual Variation in Relation Sense: Unlike relation types, the senses of relations are found to be more stable across languages. On average, 85.53% of English relations retained their top-level sense in the target languages.[7] Comparison > Expansion seems to be the most frequent conversion (15.07%) followed by Temporal > Expansion (11.96%) cross-lingually.

When considered together with the higher level of variation in relation types, the consistency in relation senses may suggest that translators take liberty in adapting the source material into their languages; yet, naturally, these variations in form did not affect the semantics as the senses of the relations are mostly preserved.

[7]Only the relations annotated with a sense tag (i.e. Explicit, Implicit and AltLex) are considered.

5. Building Bilingual Discourse Connective Lexicons

In addition to enabling linguistic investigations of cross-lingual discourse structures, a parallel corpus aligned at the discourse level has a number of practical use cases, where building bilingual discourse connective (DC) lexicons is one of them. Bilingual DC lexicons document the relationships between discourse connectives over two languages. They are important resources as discourse connectives are shown to be challenging in both machine translation [31] and second language learning [30, 32] due to their varying degrees of ambiguity. However, the existing connective lexicons are overwhelmingly monolingual, where [11, 12, 33] are the only notable exceptions. Standard dictionaries or similar lexical resources (e.g. word alignment databases such as Treq [34] or OPUS[8]) often fall short of providing an exhaustive list of translations for connectives, let alone grouping them according to their semantics [11, 33]. However, bilingual DC lexicons compiled from resources where their contexts and usages are annotated (e.g. in the form of discourse relations) readily have access to such discourse-level information regarding connectives and can capture the complex mappings between them across languages. To this end, discourse relation alignments are exploited to build such lexicons for each English-language X pair. In the rest of the paper, the TED-MDB lexicons are introduced including their extraction procedure. Their coverage and limitations are also discussed.

5.1. Procedure

One way of compiling a bilingual lexicon involves interlinking existing monolingual connective lexicons

[8]http://opus.nlpl.eu/lex.php

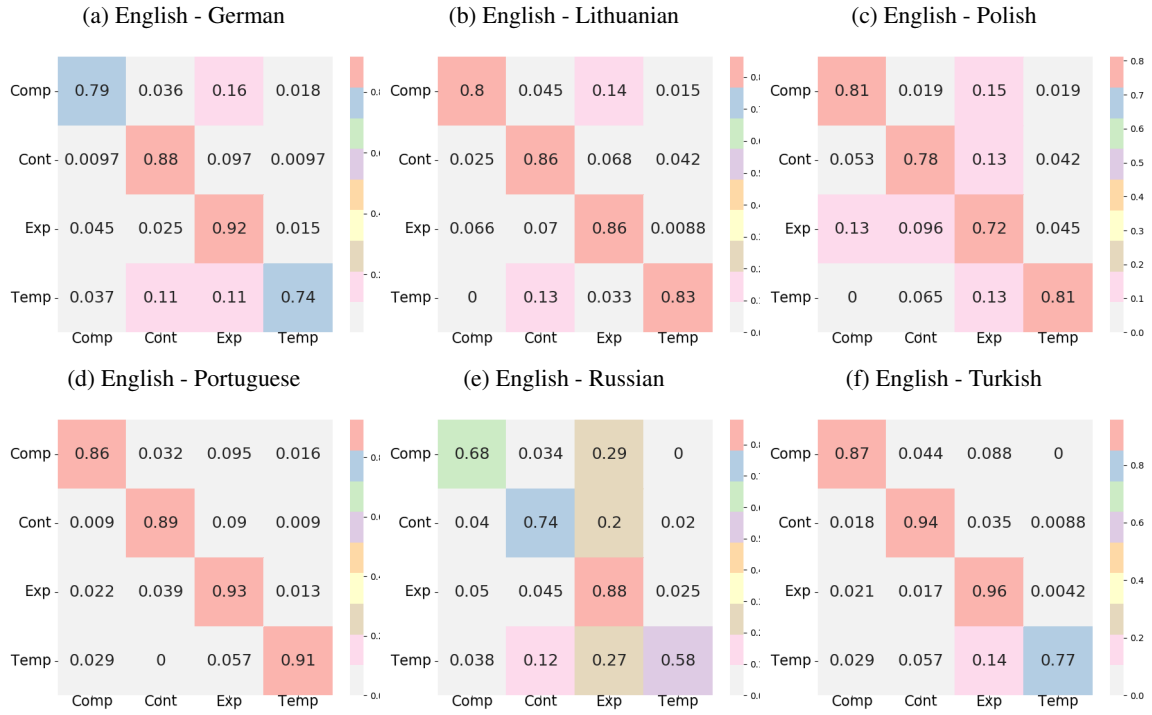

Fig. 2. Heatmap visualizations of the confusion matrices for the sense of aligned discourse relations. Rows correspond to the English relations and columns denote target languages. The matrices are normalized row-wise, where each cell denotes the percentage of English relations converted to the respective label in the target language.

by exploiting translation candidate tables calculated from large parallel corpora. To arrive at the bilingual DC lexicon, the translation candidates are filtered in a way that for each possible sense of the source connective, only those translations that can signal the same sense (determined by the DC lexicons of those particular languages) are kept [12]. In the current study, a more direct approach is adopted instead, where discourse relation alignments are exploited. This alleviates the need for other resources. The procedure mimics the extraction of monolingual lexicons from an annotated corpus, closely following [33]. Using the alignments, connectives in different languages are mapped with one another, provided that they exist in an aligned relation pair which conveys the same sense.

The selection of discourse connectives and the languages solely rely on the TED-MDB annotations.[9] The extraction of bilingual connective lexicons from aligned relations is straightforward as the more burdensome issues such as deciding which lexical items serve as discourse connectives or which sense they convey in a particular context have already been handled and implemented on the annotations. One limitation of working with TED-MDB is its size, which amounts to 255 Explicit relations on average (Table 2). To remedy this situation and extend the coverage of the lexicons, implicit connectives are also included, as in [33]. Specifically, the method consists of two steps, preceded by pre-processing:

0. In the pre-processing step, all alignments that include a non-Explicit or a non-Implicit relation in either side, as well as those mapping relations that are not annotated with exactly the same sense are filtered out.
1. For each connective in the source language, the list of its possible senses is compiled.
2. For each observed sense of each SL connective, translation equivalents are searched among the target language annotations using the relation alignments. Therefore, connective translations are provided (if any) separately for each sense. However, it is not uncommon for a matched connective pair to be polysemous between the same set of senses (e.g. the “in fact/na verdade” pair is found to signal both *Expansion:Instantiation*

[9] which is the only resource for most of those languages.

and *Expansion:Level-of-detail:Arg2-as-detail* in English and Portuguese, respectively), so sometimes, the same translations re-appear under different senses.

This procedure is applied in both directions for each language pair (of the form English-Language X). Again, the alignments obtained through the second method are used in the compilation of the lexicons.

5.2. Lexicons

The generated TED-MDB lexicons adopt a common structure. To repeat:

- **Connective**: Each lexicon entry is anchored to a connective. The connectives can be of any kind, single-word, multi-word or discontinuous (e.g. if...if). The connectives are not processed in anyway, except being lower-cased.
- **Dimlex link**: The TED-MDB annotations, therefore the TED-MDB lexicons, do not include any syntactic/orthographic information regarding connectives. In order to make that information available as well as creating a bridge between the bilingual and monolingual lexicons, each connective and its translations are accompanied with a URL to their connective-lex[10] entry.
- **Sense list:** The list of observed senses (according to the PDTB3 sense hierarchy) of the head connective in TED-MDB is provided in the body of each entry.
- **List of translation candidates**: The translation candidates in the target language are displayed under each observed sense. The candidates are guaranteed to have their own entry and can be accessed directly by clicking.
- **Example sentence**: To exemplify the context in which the connectives appear, each translation candidate is accompanied with an example relation pair from TED-MDB.

The statistics regarding each lexicon are provided in Table 9. As the entire lexicon induction phase is completely automatic, including the alignment of the discourse relations in the respective languages, the lexicons are bound to involve some errors. To evaluate the lexicons, firstly, the performance in aligning Explicit relations and Implicit relations is checked, as those relations constitute the basis of the lexicons (Table 10). In comparison to Table 7, these relation types turn out to be easier to align; in all directions, an average F-score of 93.8 is achieved. As a more direct evaluation, the lexicons generated from automatic alignments are compared against those from gold alignments that are available for three languages (LT, PT, TR). On average, 97.46% of the entries in the gold lexicons are also in the automatically generated lexicons of these languages, suggesting that the generated lexicons are of very high quality. Considering the typological variety in the evaluation languages (LT, PT, TR), it is safe to assume that the results are generalizable to other TED-MDB languages (DE, PL, RU).

Overall, through adopting a fully automatic pipeline, a number of high quality bilingual DC lexicons are generated. Considering the scarcity of such resources, the proposed lexicons are believed to be valuable additions to the cross-lingual studies. Furthermore, these lexicons can be easily verified and converted into gold standard by the discourse communities of the respective languages, which would, otherwise, require a great deal of manual labor.

6. Conclusion

In the current work, two methods for aligning discourse relations are proposed, one of them using word alignments and the other relying on distributional semantics. Due to the challenges specific to the current task, each method is tailored to the current context through a set of heuristics. Overall, the second method, which employs multilingual embeddings to align discourse relations, is favored over the more traditional first method, due to its higher performance as well as the latter's dependency on external resources (e.g. large parallel corpus, a sentence aligner), which may not be available for most of the language pairs.

The present paper has applied the data linking terminology to a different area of research, that is, to the alignment of discourse connective annotations in different languages. This has two promising results: First, an aligned multilingual corpora not only on the sentence level but also on the discourse level would enable many cross-linguistic studies to be performed, including machine translation, shallow discourse parsing, etc. Secondly, six bilingual discourse connective lexicons have been extracted purely contextually. These lexicons can be useful in many domains of information technology.

[10] http://connective-lex.info/

Table 9

Statistics regarding the generated lexicons. Exp and Imp columns refer to the number of connectives from Explicit and Implicit relations, respectively. The total number of connectives is calculated by counting explicit and implicit connectives separately (Total) and together (Unique). Min, Max and Avg columns correspond to the minimum, maximum and the average number of (i) senses per connective; (ii) translation equivalents available for each connective in the lexicons, respectively

Language	Connectives			Senses			Translations		
	Exp	Imp	Total (Unique)	Min	Max	Avg	Min	Max	Avg
English	26	26	52 (44)	1	3	1.25	1	6	1.79
German	29	20	49 (43)	1	3	1.24	1	8	1.90
English	27	32	59 (51)	1	5	1.20	1	9	2.27
Lithuanian	33	35	68 (59)	1	5	1.38	1	4	1.97
English	17	22	39 (33)	1	4	1.18	1	7	2.21
Polish	31	25	56 (51)	1	4	1.25	1	3	1.54
English	28	34	62 (53)	1	3	1.23	1	6	1.84
Portuguese	27	27	54 (44)	1	6	1.46	1	6	2.11
English	22	20	42 (35)	1	3	1.10	1	5	1.76
Russian	31	12	43 (43)	1	3	1.12	1	5	1.72
English	25	33	58 (48)	1	4	1.29	1	9	2.48
Turkish	38	40	78 (66)	1	5	1.44	1	4	1.85

Table 10

The performance of the method II on only implicit and explicit relations

Language Pair	TP	FN	FP	TN	Accuracy	Precision	Recall	F Score
en lt	321	8	84	51	0.8	0.79	0.98	0.87
lt en	321	0	34	0	0.9	0.9	1	0.95
en pt	367	10	36	50	0.9	0.91	0.97	0.94
pt en	367	0	14	0	0.96	0.96	1	0.98
en tr	348	12	50	37	0.86	0.87	0.97	0.92
tr en	348	0	22	0	0.94	0.94	1	0.97

Currently, English, the source language, is taken as the basis for all the bilingual dictionaries presented in this work. For the future, extending the bilingual lexicons to the multilingual level is planned; extracting the lexicons at a multilingual level would definitely provide a better perspective on the use of discourse connectives across multiple languages.

References

[1] M. Stede, T. Scheffler and A. Mendes, Connective-lex: A web-based multilingual lexical resource for connectives, *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics* (2019).

[2] R. Prasad, B. Webber and A. Joshi, Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation, *Computational Linguistics* **40**(4) (2014), 921–950.

[3] D. Zeyrek, A. Mendes, Y. Grishina, M. Kurfalı, S. Gibbon and M. Ogrodniczuk, TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style, *Language Resources and Evaluation* (2019), 1–27.

[4] N. Asher, *Reference to Abstract Objects in Discourse*, Kluwer, Dordrecht, 1993.

[5] B. Webber, R. Prasad, A. Lee and A. Joshi, The penn discourse treebank 3.0 annotation manual, *Philadelphia, University of Pennsylvania* (2019).

[6] T. Scheffler and M. Stede, Adding semantic relations to a large-coverage connective lexicon of German, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 1008–1013.

[7] C. Roze, L. Danlos and P. Muller, LEXCONN: a French lexicon of discourse connectives, *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics* (2012).

[8] A. Briz, S. Pons and J. Portolés, Diccionario de partículas discursivas del español, in: *El diccionario como puente entre las lenguas y culturas del mundo. Actas del II Congreso Internacional de Lexicografía Hispánica. Alicante, Biblioteca Virtual Cervantes*, 2008, pp. 217–227.

[9] R. Pasch, U. Brauße, E. Breindl and U.H. Waßner, *Handbuch der deutschen Konnektoren: linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen*

Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln), Vol. 2, Walter de Gruyter, 2003.

[10] E. Breindl, A. Volodina and U.H. Waßner, *Handbuch der deutschen Konnektoren 2: Semantik der deutschen Satzverknüpfer*, Vol. 13, Walter de Gruyter GmbH & Co KG, 2014.

[11] P. Bourgonje, Y. Grishina and M. Stede, Toward a bilingual lexical database on connectives: Exploiting a German/Italian parallel corpus, in: *Proceedings of the Fourth Italian Conference on Computational Linguistics–CLIC-IT*, 2017, pp. 53–58.

[12] L. Poláková, K. Rysová, M. Rysová and J. Mírovskỳ, GeCzLex: Lexicon of Czech and German Anaphoric Connectives, in: *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 1089–1096.

[13] Y. Versley, Discovery of Ambiguous and Unambiguous Discourse Connectives via Annotation Projection, *AEPC 2010* (2010), 83.

[14] J.J. Li, M. Carpuat and A. Nenkova, Cross-lingual Discourse Relation Analysis: A corpus study and a semi-supervised classification system, in: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 577–587.

[15] M. Laali, Inducing Discourse Resources Using Annotation Projection, PhD thesis, Concordia University, 2017.

[16] H. Schwenk, Filtering and Mining Parallel Data in a Joint Multilingual Space, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 228–234.

[17] M. Kurfalı and R. Östling, Noisy parallel corpus filtering through projected word embeddings, in: *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, 2019, pp. 277–281.

[18] R. Östling and J. Tiedemann, Efficient word alignment with markov chain monte carlo, *The Prague Bulletin of Mathematical Linguistics* **106**(1) (2016), 125–146.

[19] D. Varga, P. Halácsy, A. Kornai, V. Nagy, L. Németh and V. Trón, Parallel corpora for medium density languages, *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4* **292** (2007), 247.

[20] J. Tiedemann, Parallel Data, Tools and Interfaces in OPUS., in: *Lrec*, Vol. 2012, 2012, pp. 2214–2218.

[21] M. Aulamo, U. Sulubacak, S. Virpioja and J. Tiedemann, OpusTools and Parallel Corpus Diagnostics, in: *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, 2020, pp. 3782–3789. ISBN 979-10-95546-34-4. https://www.aclweb.org/anthology/2020.lrec-1.467.

[22] S. Özer and D. Zeyrek, An automatic discourse relation alignment experiment on TED-MDB, in: *Proceedings of the 2019 Workshop on Widening NLP*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 31–34.

[23] M. Artetxe and H. Schwenk, Massively Multilingual Sentence Smbeddings for Zero-shot Cross-lingual Transfer and beyond, *Transactions of the Association for Computational Linguistics* **7** (2019), 597–610.

[24] V. Pyatkin and B. Webber, Discourse Relations and Conjoined VPs: Automated Sense Recognition, in: *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, pp. 33–42.

[25] J. Hoek, S. Zufferey, J. Evers-Vermeul and T.J. Sanders, Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study, *Journal of Pragmatics* **121** (2017), 113–131.

[26] S. Zufferey, Discourse connectives across languages: factors influencing their explicit or implicit translation, *Languages in Contrast* **16**(2) (2016), 264–279.

[27] D. Zeyrek and M. Kurfalı, TDB 1.1: Extensions on Turkish discourse bank, in: *Proceedings of the 11th Linguistic Annotation Workshop*, 2017, pp. 76–81.

[28] J. Park and C. Cardie, Improving implicit discourse relation recognition through feature set optimization, in: *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2012, pp. 108–112.

[29] Y. Ji and J. Eisenstein, One vector is not enough: Entity-augmented distributed semantics for discourse relations, *Transactions of the Association for Computational Linguistics* **3** (2015), 329–344.

[30] S. Zufferey, W. Mak, L. Degand and T. Sanders, Advanced learners' comprehension of discourse connectives: The role of L1 transfer across on-line and off-line tasks, *Second Language Research* **31**(3) (2015), 389–411.

[31] T. Meyer, A. Popescu-Belis, N. Hajlaoui and A. Gesmundo, Machine translation of labeled discourse connectives, in: *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, 2012.

[32] M. Wetzel, S. Zufferey and P. Gygax, Second Language Acquisition and the mastery of discourse connectives: Assessing the factors that hinder L2-learners from mastering French connectives, *Languages* **5**(3) (2020), 35.

[33] M. Kurfalı, S. Ozer, D. Zeyrek and A. Mendes, TED-MDB Lexicons: TrEnConnLex, PtEnConnLex, in: *Proceedings of the First Workshop on Computational Approaches to Discourse*, 2020, pp. 148–153.

[34] M. Škrabal and M. Vavřín, The translation equivalents database (treq) as a lexicographer's aid, in: *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, 2017, pp. 124–137.

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

[36] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado and J. Dean, Distributed representations of words and phrases and their compositionality, in: *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[37] P. Christen and K. Goiser, Quality and complexity measures for data linkage and deduplication, in: *Quality measures in data mining*, Springer, 2007, pp. 127–151.

[38] M. Dupont and S. Zufferey, Methodological issues in the use of directional parallel corpora: A case study of English and French concessive connectives, *International journal of corpus linguistics* **22**(2) (2017), 270–297.