# Linking Discourse-level information and induction of bilingual discourse connective lexicons

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

[a] *Cognitive Science Dept., Middle East Technical University, Ankara, Turkey*
*E-mails: sibel.ozer@metu.edu.tr, dezeyrek@metu.edu.tr*
[b] *Linguistics Department, Stockholm University, Stockholm, Sweden*
*E-mail: murathan.kurfali@ling.su.se*
[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-level information in the target languages through error-prone automatic means. The current paper aims to provide more direct insight into the cross-lingual variations in discourse structures by linking the annotated relations of the TED-Multilingual Discourse Bank, which consists of independently annotated six Ted talks in seven different languages. It is shown that the linguistic labels over the relations annotated in the texts of these languages can be automatically linked with high accuracy, as verified against the semi-automatically linked relations of three diverse languages. The resulting corpus has a great potential to reveal the divergences the languages exhibit in local discourse relations, with respect to the source text, as well as leading 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 con-

*Corresponding author. E-mail: sibel.ozer@metu.edu.tr
†Equal contribution.
[1]http://textlink.ii.metu.edu.tr/

junction), and the senses they convey (contrast, temporal, concession).

TED-Multilingual Discourse Bank (TED-MDB) is a corpus annotated for discourse relations of English TED talks and translations into multiple languages (European Portuguese, Lithuanian, German, Russian, Polish, and Turkish) created during the lifespan of the TextLink project, when most of the languages involved in the project did not have a discourse connective lexicon. Thus, no discourse connective lexicons were utilized during the development of TED-MDB. The teams took the members of syntactic classes such as subordinating and coordinating conjunctions and adverbials as a starting point to determine the set of explicit discourse connectives in each language. Each team was also allowed to specify discourse connectives that go beyond the syntactic classes..

TED-MDB offers an ideal domain to induce monolingual and bilingual discourse connective lexicons for a new set of languages. But this resource presents a challenge to the induction of discourse connective lexicons because of the following reasons: (i) discourse relations in each language are annotated blind to the annotations performed on other languages, (ii) discourse relations, naturally, exhibit differences across languages, which could hinder any efforts of cross-linguistic comparisons or induction of new resources such as bilingual lexicons among the languages included in the corpus. Given that TED-MDB annotates discourse connectives, their binary arguments, and the discourse senses conveyed by discourse connectives (see §2.1), the discourse relations exhibit variations across languages on several levels, e.g. argument spans differ due to translation effects, language-specific facts, or the annotators' methodological choices. Thus, to support further research, a relation linking task[2] must be performed on TED-MDB, which involves the linking of labels over annotated relations across languages.

The main contributions of the paper are: (1) to introduce two alternative methods to link the relation labels in 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 neither been investigated for the relation linking of a multilingual discourse corpus, nor for the languages under consideration in the present work; (2) to present a newer version of TED-MDB with the linked labels over each text in the corpus, thus enhancing the data structure of the corpus; (3) to present a comprehensive overview of the discourse structures across TED-MDB languages and (4) 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 discourse connective bilingual lexicons.[3]

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 describes the data linking task, highlighting its challenges in §3.1, followed by the details of the two proposed methods in §3.2 and §3.3. This section also provides an evaluation of the interlinked data as well as various issues and challenges confronted during the linking task (§3.4). In §4, an overview of the discourse structures observed in TED-MDB is presented together with the statistics obtained from relation mappings. In §5, the bilingual lexicons, which link connectives across languages based on the meaning, i.e. sense, they convey, induced from the linked data are described. The paper ends with a conclusion and some future directions for further research in §6.

## 2. TED Multilingual Discourse Bank and the Linguistic Labels Over the Texts

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 The talks were presented by native

[2] Throughout the text, the general term of relation linking is adopted, instead of discourse relation linking, as our method also links EntRels or NoRels across languages, which are not discourse relations by definition.

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

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 |

### 2.1. Annotation Scheme

The texts are annotated in the Penn Discourse TreeBank (PDTB) style [4], where discourse relations that hold between two arguments of a discourse connective (Arg1 and Arg2) are identified. Discourse relations may be explicit, typically marked by a discourse connective, such as *and*, *because*, *however*, though there also exist implicit discourse relations and alternative lexicalizations, as shown in the annotation scheme in Table 2.

A discourse relation is Explicit when a discourse connective makes the relation that holds between the two arguments salient, as in Example 1.[4]

(1) *The world is changing in some really profound ways*, and **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)

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.

(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'

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 Alternative Lexicalization, or AltLex (Example 3).[5]

(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.* 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**
[AltLex, Temporal:Precedence] (English, TED Talk no. 2009)

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 [2]. 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)

[4]The examples are taken from the TED-MDB. In all the examples, discourse 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.

[5]TED-MDB does not annotate the AltLex-C cases, which PDTB 3.0 annotates.

Fig. 1. PDTB 3.0 sense hierarchy [2]

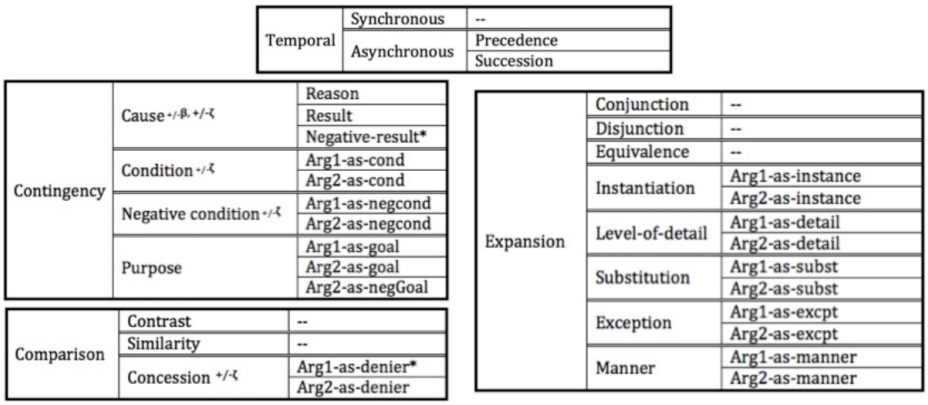



Table 2
TED-MDB Annotation Scheme

| Relation type | Relation anchor | Arguments | Sense |
|---|---|---|---|
| Explicit | Overt discourse connective | Arg1, Arg2 | Yes |
| Implicit | Inferred discourse connective | Arg1, Arg2 | Yes |
| Alternative Lexicalization (AltLex) | Alternative way of expressing the DR | Arg1, Arg2 | Yes |
| Entity Relation (EntRel) | None | Arg1, Arg2 | No |
| No Relation (NoRel) | None | Arg1, Arg2 | No |

of the top category, followed by the third level sense, or Level3 sense (Result), providing information about the full semantics of the relation. The complete sense hierarchy is provided in Figure 1.

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 sentences, the relation is of the type NoRel (Example 5). [6]

(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, a new top-level sense called Hypophora was introduced, which 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) Are *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 lead to different sets of relations annotated for each language. Ta-

[6] EntRels and NoRels are annotated within paragraphs and between sentences. The annotators are told to annotate a pair of adjacent sentences as No Relation if an implicit relation that relates them cannot be inferred.

ble 3 provides the number and the percentage of each type of relation (Explicit, Implicit, AltLex, EntRel and NoRel) in each language.

In order to test the reliability of the annotations, 25% of the whole corpus (i.e. two TED talks per language) are annotated by an independent annotator, using the annotation schema and following the annotation principles summarized above. The inter-annotator agreement (IAA) is performed on two levels following [5]: (i) whether or not annotators spotted a relation between same two discourse units, (ii) whether or not spotted relation is of the same kind (type and sense-wise). The agreement on relation spotting is measured via F-score, whereas the type and sense agreement on the spotted relations is measured via simple ratio agreement and Cohen's Kappa. The IAA on both levels is found to be at a good standard (> 70) which suggests the reliability of the annotations.

### 2.2. *Discourse Connective Lexicons*

The last decade has seen an upsurge in the development of discourse connective lexicons, including LexConn [6], LiCo[7], DiMLex [8], CzeDLex [9], LDM-PT [10] among others. 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 – [11]) includes explicit information on discourse particles in Spanish but it excludes conjunctions and prepositions. The German resource *Handbuch der Konnektoren* [12, 13] contains discourse connective representations including their possible positions in a sentence, also the register and possible modifiers. In discourse-annotated corpora, the use of different annotation schemes such as PDTB, Rhetorical Structure Theory (RST) [14] also poses challenges for linking the information on discourse connectives.

Despite the challenges in 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 ([15] and [16]), 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. Linking the Annotated Relations in TED-MDB

The outcome of the main task of the present work, i.e., relation linking, will identify and link the discourse labels over the texts of different languages that correspond to each other. This means enabling access to the discourse labels over texts in different languages on the level of format, as well as enabling easy access to the discourse structures of different languages, by means of the reference to existing labels [17]. Thus, the main task of the current paper will not only support the induction of a multilingual discourse connective lexicon, but it will also enable immediate access to different datasets within TED-MDB.

Such a task can be seen as a variant of the annotation projection task, where the aim is to transfer (manually or automatically), the annotated discourse relations in one language to another through parallel corpora [18–20]. Yet, despite certain similarities, they noticeably differ from each other, because in annotation projection, the linguistic information is available only for one language. Hence, being completely clueless about the target language, the projection method can be deemed successful to the extent that the projected relations mimic the original ones and cannot be punished for missing the discourse relations on the target side. However, in our case, the linguistic information is available for both sides and instead of an uninformed projection of the source text discourse relations, one should decide if there is a corresponding discourse relation on the target text, which is not a straightforward task. In the rest of this section, the challenges surrounding the task undertaken, the methods to address these challenges and their evaluation is provided, respectively.

### 3.1. *The Challenges*

In order to link two sets of relations, cross-lingual variations among the relations must be understood and handled carefully. The challenge could appear at several levels: Typically, the argument spans of the relations tend to vary across languages and Example 7 illustrates such an instance. Here, the Arg2 span to the English connective *and*, and the Arg1 span to the Turkish suffixal connective *-up* are supposed to match, but a difficulty arises due to the dropped object *it* that refers to *The Hubble Space Telescope* in the Turkish relation.

(7) ... *we take the Hubble Space Telescope* and **we turn it around** [Explicit, Expansion:Conjunction]

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

| Language | Explicit | Implicit | AltLex | EntRel | NoRel | Total |
|---|---|---|---|---|---|---|
| English | 289 (40%) | 254 (36%) | 46 (6%) | 78 (11%) | 49 (7%) | 716 |
| German | 240 (43%) | 214 (38%) | 17 (3%) | 59 (11%) | 30 (5%) | 560 |
| Lithuanian | 377 (46%) | 315 (38%) | 18 (2%) | 79 (10%) | 32 (4%) | 821 |
| Polish | 218 (37,5%) | 195 (33,5%) | 11 (2%) | 104 (18%) | 52 (9%) | 580 |
| Portuguese | 269 (40%) | 311 (46%) | 29 (4%) | 38 (6%) | 33 (5%) | 680 |
| Russian | 237 (42%) | 221 (39%) | 20 (4%) | 57 (10%) | 30 (5%) | 565 |
| Turkish | 315 (41%) | 264 (35%) | 60 (8%) | 70 (9%) | 51 (7%) | 760 |
| Total | 1945 | 1774 | 201 | 485 | 277 | 4682 |

(English, TED talk 1976)

Hubble Uzay Teleskobu'nu *tut*up **döndür-düğümüzü** ... [Explicit, Expansion:Conjunction] (Turkish, TED Talk 1976)

Secondly, since both intra- and inter-sentential relations are annotated, a number of relations may be created over the same text span, which raises the additional need to find the correct relation among those with overlapping arguments.

### 3.2. Method I: Linking Relations through Word Alignments

In the firs approach, the relations annotated over the texts of different languages are linked to each other through word alignments. In a pre-processing step, all the raw texts are sentence-tokenized and aligned to remedy the variations in their original format. Then, the texts are aligned at the word level by a statistical aligner (Eflomal [21]), and the relations are linked. The details are presented below.

#### 3.2.1. Sentence Alignment

Although TED-MDB is built upon the parallel corpora of TED talk subtitles, the texts on which relation annotations are created were not aligned, causing problems for relation linking. To alleviate problems, firstly, all raw texts are normalized to a standard sentence-per-line format, and paragraphs are separated. Using NLTK's sentence tokenizer, a sentence segmentation procedure is performed; then, using the LF-aligner software[7], based on the hunalign algorithm [22], a sentence alignment procedure that aligns the relations of all seven languages is carried out. This initial attempt generated a number of mismatches due to the varying number of sentences in each translation, as listed in Table 4. Since any error in this step would be propagated through the pipeline, we settled on aligning each language separately with English to maximize the alignment quality and the linking quality which would take place later in the pipeline.

Table 4
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 |

#### 3.2.2. Obtaining Word Alignments

Having aligned the raw texts with their English counterparts, the next step was to obtain word alignments. However, the performance of word aligners heavily depends 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 by using the model 3 of EFLOMAL[8] [21]. Custom parallel corpora are created for each language pair by concatenating the largest corpus of each language pair available in the OPUS database [23]. All the corpora are obtained and processed using OpusTools[9] [24]. The data sizes of each corpus are listed in Table 5.

Word alignment is performed in both directions, resulting in two sets of alignments: *the forward alignments* include the alignments where the source lan-

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

[8] https://github.com/robertostling/eflomal

[9] https://github.com/Helsinki-NLP/OpusTools

Table 5
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 |

guage is set as English, and *the reverse alignments* involve word alignments where the source language is set as the non-English language. Yet, using alignments directly from either direction is reported to underperform [18, 20]; 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.

#### 3.2.3. Linking the Relations

In the last step, the labels over the relations of English texts are linked to the labels over the texts of target languages using the word alignments. Due to the differences in the argument spans as discussed in§3.1, linking cannot be straightforwardly performed by matching the relations whose words are found to be equivalent by the word aligner. Hence, relation linking is performed as follows: Given a relation in the source text English, the labels over that relation, i.e. Arg1, Arg2, and the discourse connective (if there is any), are projected to the target text using the word alignments. As an 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 text is scored on the basis of the overlap between its components and the components of the projected relation. Discourse 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 relation and the projected relation) is selected as the linked pair. However, particularly in cases where multiple relations are annotated over similar text spans, 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 is added to the score (also see§3.3.2).

### 3.3. Method II: Linking Relations through Cross-lingual Sentence Embeddings

The second method utilizes the modern, language agnostic sentence encoders which are capable of assigning similar representations to the semantically similar linguistic units across languages. The method is a continuation of a previous study [25] that performed relation linking only for the English-Turkish pair in TED-MDB. It starts with a pre-processing step which is similar to that of the first method, i.e., the raw texts are sentence-tokenized and aligned in the manner already described. For relation linking, all relations in each bitext unit [10] are paired constructing relation matrices. Then, for all pairs with a semantic similarity over a certain threshold learned in the training phrase (§3.3.1), a composite score is calculated. This score not only reflects the agreement on all three sense levels and the relation type of the matched pair (if they have no match in their Level1 sense, relation type match is discarded), but also the semantic similarity between the text segments (Arg1+connective(if available)+Arg2). The semantic similarity is calculated as the cosine similarity between the LASER embeddings [27] of each relation's text segments.

#### 3.3.1. Adjusting the Semantic Similarity Threshold

Unlike the first method, the second method involves a training phase, namely, the learning of a semantic threshold parameter. To learn this parameter, a training is performed for language pairs involving the source language and three target languages, Turkish, Portuguese, and Lithuanian (EN-TR, EN-PT, EN-LT). The relation labels over English texts were automatically matched with those on the texts of these languages, and the performance of the automatic pro-

[10] Bitext is a pair of source and target sentences which have a degree of (partial or full) translation equivalences.[26]

cess was checked by the teams and wrong matches were manually corrected. In the training phase, further performance evaluation was done using this manually checked data. Throughout the paper, we refer to this data as manually-corrected data or semi-automatically linked data.

For training, six English files are split into the train and test datasets considering the overall relation counts in the English texts. As the data size is low, to eliminate over fitting, the data is evenly split into train and test sets.[11] Also, to have a representative training set, four talks are set aside as the training set whereas the other two are used as the test set.[12] Using semantic threshold values starting from 0 to 0.95 and incriminating by 0.05, the algorithm is repeated. The optimum threshold value that yields the best F-score on average for all three language pairs is selected and validated in the test set, and later applied to other language pairs.

Figure 2 shows the effect of the semantic threshold on the performance according to the evaluation metrics. For better readability, the figures start from 0.35. However, the performance is found to be stable between 0 and 0.55 across languages. The effect of threshold starts to become visible around 0.6 for all languages. Even though maximum performance is observed at 0.7 for Portuguese and at 0.65 for Turkish, the performance after the 0.6 threshold shows a rapid decrease for Lithuanian. Between the threshold values 0 to 0.55, the F-score is 0.88 on average for all three language pairs. However, keeping the parameter at this level causes false positives to increase. Due to no or little control of this parameter, the model relies on the similarity of two relations only at the sense levels and relation types. This reliance results in linking English relations with wrong target relations. This can be seen in example 8, where the DR anchored by *and* in the English sentence is falsely linked to the DR *ve* 'and' in Turkish, as their senses and relation types match.

(8) When we think about mapping cities , we tend to think about roads and streets and buildings , and the settlement narrative that led to their creation , or you might think about the bold vision of an urban designer , but there 's other ways to think *about mapping cities* and **how they got to be made** . [Explicit, Expansion:Conjunction] (English, TED Talk no. 2150)

Şehirlerin haritalarını oluşturmayı düşündüğümüzde yollar, sokaklar, caddeler, binalar ve şehirlerin oluşumuna yol açan yerleşim hikayeleri aklımıza gelir. Ya da bir kentsel tasarımcının cesur vizyonunu düşünebilirsiniz. Ancak, şehirlerin haritalarını oluşturmayı *düşünmenin* ve **yapmanın** başka yolları da var. [Explicit, Expansion:Conjunction] (Turkish, TED Talk no. 2150)

[11] The exact relation-wise train:test data ratio is 52:48.

[12] Specifically, talks with ids of 1971, 1978, 2009 and 2150 are used as the training data.

### 3.3.2. *Linking Relations*

1. As stated before, the similarity score between the relation pairs are calculated considering their text spans (Arg1+discourse connective/DC (if available)+Arg2). Pairs which do not exceed the similarity threshold learned in the previous step are discarded.
2. The semantic similarity score is combined with another score that reflects the *semantic* match between the relation pairs in the respective languages. That is, 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 Level3 sense is given 10, and 1 is assigned for the type match (Explicit, implicit etc.). While the algorithm gives the highest priority to Level1 sense matches, as there is no sense information for NoRels and EntRels, the type match also becomes prominent.
3. For each source relation, the target relation which yields the maximum score is marked as its linked pair and the same procedure is repeatedly applied until no relation pair is left in the matrices.

The whole procedure is exemplified on a sample sub-corpus of discourse relations given in example 9 consisting of three Explicit relations in two languages (EN, TR) signaled by (*but, as, and*) and (*ama 'but', kadar '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 6. Then, following the scoring procedure, each pair is assigned a score. For example, 9 contains separate discourse relations anchored by three explicit discourse connectives, *but, as, and*. On the Turkish side, while the connective label *Ama* (first column) matches the English connective label *But* in relation type and relation sense in all levels, the labels corresponding to

(a) English - Lithuanian (b) English - Portuguese (c) English - Turkish

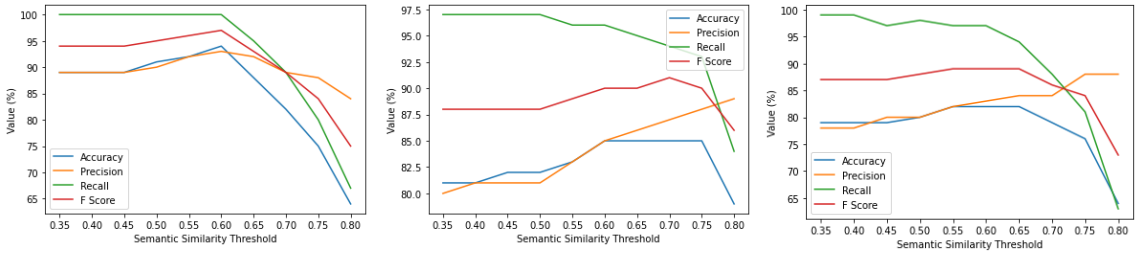



Fig. 2. The change of evaluation metrics (Accuracy, Precision, Recall and F-Score) at different levels of semantic threshold values. Although the threshold is searched between 0 and 0.95 at increments of 0.05, to achieve better visualization, only the values between 0.3 and 0.85 are provided.

*as* match in Level1 sense and relation type. The third English relation conveyed by *and* has no match with 'Ama' at any sense levels. For this reason, relation type match is not considered between 'Ama' and 'and'. Then, each source relation (i.e. each row) is linked to the target relation (i.e. each column) that has the maximum score (shown in bold Table 6).

(9) 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)

(10) *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 6

DR matrix for the sample corpus in example 9. The numbers refer to the scores based on sense/type agreement + semantic similarity of segments (Arg1+Conn(If available)+Arg2)

| | **Ama** | **kadar** | **ve** |
|---|---|---|---|
| **But** | **1111+0.85** | 1001+0.79 | 0+0.72 |
| **and** | 0+0.69 | 0+0.71 | **1111+0.75** |
| **as** | 1001+0.8 | **1111+0.85** | 0 + 0.77 |

The examination of preliminary results revealed the need for certain revisions. As mentioned before, it is common for more than one discourse relation to hold between similar arguments ([28]), which could lead to false relation linking if only the arguments are linked. So, in addition to the similarity between argument spans, the semantic similarity between discourse connectives is also checked. Second, an AltLex in one language may be converted into an Explicit discourse relation in another language. The linking algorithm is unable to cover such cases as it works on sentence-aligned bi-text units. In order to eliminate this pitfall, if a relation is not matched with a relation in the target language in its parallel unit, it is evaluated once more in the succeeding alignment unit.

### 3.4. Evaluation

In the literature, data linking quality is evaluated by using the standard precision, recall and F-score met-

rics. 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 [5, 28].

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 linking the relations in two languages, accuracy should also be taken into consideration; providing information on the non-matching data pairs is as important as providing matching data. In linking annotated labels such as ours, 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 manually corrected links between English texts and Lithuanian, European Portuguese and Turkish texts.[13] The linking performance of the proposed methods are measured for each direction, e.g. Lithuanian-to-English and English-to-Lithuanian, as the number and the set of relations differ from language to language. This evaluation method is preferred because only evaluating the relation pairs in one direction would mean not considering the relations in one language that have no matches in the other.

The evaluation results for both methods are given in Table 7. Overall, both methods yielded a good degree of performance. In particular, Method I achieves a good degree of precision, meaning that the links it finds have a 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, Method I yielded more relations that are left unlinked (False Negatives), missing a good number of existing links. The number of missed relation links decreases 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 performance revealed that some of the errors stem from the misaligned sentence pairs. 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, Lithuanian texts displayed less uniformity with English texts. One of the possible reasons is that the total number of Lithuanian relations to be matched (in total 752 relations in six files) is more than the relations in the European Portuguese and Turkish files. In such cases, both methods fail and performance decreases due to an increase either in False Positives (see 13) or False Negatives (see 11 and 12). An increase in those numbers affect all the performance metrics (precision, recall and accuracy). In the following, we report some instances that led to performance drop. A detailed analysis of linguistic reasons, methodological choices of the annotators, or translation decisions that possibly lead to such cases are left for further research.

*Different argument spans are selected in language pairs*: In example 11, 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 target language relation.

(11) 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 düşünün**ce *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)

*Different realizations of discourse connectives*. In example 12, even though the DR in the English sentence is linked with the DR in Lithuanian, neither method could capture this link due to the different Arg2 annotations.

[13] Unfortunately, relation labels on English texts and those on the remaining languages did not go through a manual correction procedure.

Table 7

Method I (Linking through Word Alignments) and Method II (Linking through Cross-lingual Sentence Embeddings) Quality metrics for each language obtained in two test files selected in 3.3.1. The first three parts refers to the results of the first method grouped by the symmetrization heuristics, ranked from the most restrictive to least restrictive, as explained in sect. 3.2.2

| Method | Lang. Pair | TP | FN | FP | TN | Accuracy | Precision | Recall | F-Score |
|---|---|---|---|---|---|---|---|---|---|
| Method I Intersect | EN LT | 245 | 34 | 26 | 41 | 0.83 | 0.9 | 0.88 | 0.89 |
| | LT EN | 245 | 36 | 26 | 63 | 0.83 | 0.9 | 0.87 | 0.89 |
| | EN PT | 160 | 51 | 96 | 39 | 0.58 | 0.62 | 0.76 | 0.69 |
| | PT EN | 160 | 51 | 96 | 22 | 0.55 | 0.62 | 0.76 | 0.69 |
| | EN TR | 255 | 44 | 13 | 34 | 0.84 | 0.95 | 0.85 | 0.9 |
| | TR EN | 255 | 49 | 13 | 49 | 0.83 | 0.95 | 0.84 | 0.89 |
| Method I Grow-diag | EN LT | 250 | 31 | 23 | 42 | 0.84 | 0.92 | 0.89 | 0.9 |
| | LT EN | 250 | 33 | 23 | 64 | 0.85 | 0.92 | 0.88 | 0.9 |
| | EN PT | 165 | 44 | 99 | 38 | 0.59 | 0.62 | 0.79 | 0.7 |
| | PT EN | 165 | 45 | 99 | 20 | 0.56 | 0.62 | 0.79 | 0.7 |
| | EN TR | 265 | 33 | 14 | 34 | 0.86 | 0.95 | 0.89 | 0.92 |
| | TR EN | 4265 | 39 | 14 | 48 | 0.86 | 0.95 | 0.87 | 0.91 |
| Method I Grow-diag-final | EN LT | 254 | 26 | 25 | 41 | 0.85 | 0.91 | 0.91 | **0.91** |
| | LT EN | 254 | 29 | 25 | 62 | 0.85 | 0.91 | 0.9 | **0.9** |
| | EN PT | 171 | 34 | 104 | 37 | 0.6 | 0.62 | 0.83 | **0.71** |
| | PT EN | 171 | 35 | 104 | 19 | 0.58 | 0.62 | 0.83 | **0.71** |
| | EN TR | 268 | 22 | 24 | 32 | 0.87 | 0.92 | 0.92 | **0.92** |
| | TR EN | 268 | 28 | 24 | 46 | 0.86 | 0.92 | 0.91 | **0.91** |
| Method II | EN LT | 288 | 2 | 15 | 41 | 0.95 | 0.95 | 0.99 | **0.97** |
| | LT EN | 288 | 1 | 15 | 66 | 0.96 | 0.95 | 1 | **0.97** |
| | EN PT | 273 | 5 | 37 | 31 | 0.88 | 0.88 | 0.98 | **0.93** |
| | PT EN | 273 | 2 | 37 | 17 | 0.88 | 0.88 | 0.99 | **0.93** |
| | EN TR | 279 | 10 | 37 | 20 | 0.86 | 0.88 | 0.97 | **0.92** |
| | TR EN | 279 | 12 | 37 | 38 | 0.87 | 0.88 | 0.96 | **0.92** |

(12) 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 argument span of the source relation are only partially selected as an argument in the target language*. In example 13, the English DR is a non-matching data. However, both methods fail in this instance and match the DR with a target language DR as it shares a part of the Arg1 span.

(13) *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 dos dois lados da questão da alteração climática . *Aposto que não adivinham de que lado estou* . Mas **o que eu gosto nisto é que me lembra uma coisa que Mark Twain disse :**

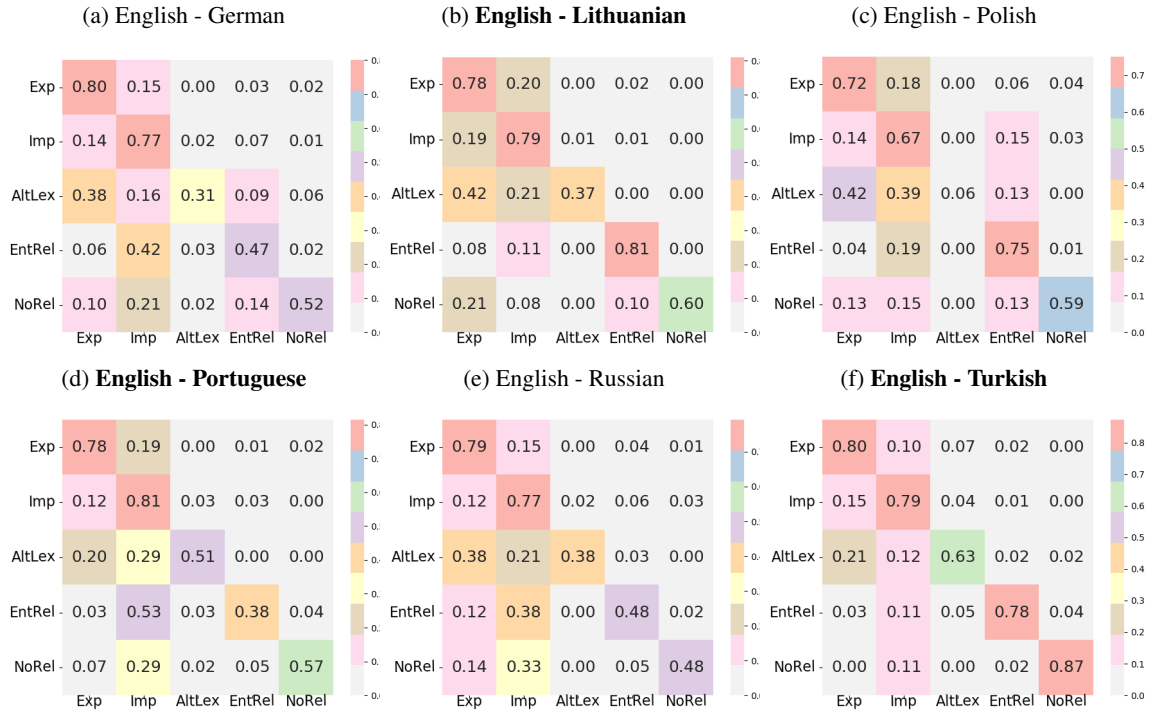



Fig. 3. Heatmap visualizations of the confusion matrices for relation type of the linked discourse relations. Rows correspond to the English relations and columns denote target languages. The matrices are normalized row-wise where each cell denotes the percentage of English relations converted to the respective label in the target language. Confusion matrices created from manually-corrected links are highlighted in bold.

**" Planeia para o futuro , " porque é onde vais passar o resto da tua vida.** [Explicit, Expansion:Conjunction] (European Portuguese, TED Talk no. 1927)

## 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. Yet, due to the scarcity of parallel corpora annotated for discourse relations on both sides, previous cross-lingual work has largely been confined to a specific aspect of discourse, e.g. omission of discourse markers [29, 30], mostly using parallel data with manual annotations on only one side. However, thanks to the availability of discourse information on both ends and the relation linking 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, we use the linked discourse relations. In order to maximize the reliability of our analysis, we used the manually-corrected links for three languages (LT, PT, TR) while using the automatically linked discourse relations in the second method for the remaining languages. Therefore, the observations should be approached cautiously due to possible incorrect links; yet, the high F-scores on capturing the semi-automatically obtained links (see Table 7) suggest that the reported results closely follow the distribution in semi-automatically linked data.

The following analysis is mainly confined with the descriptive analysis of the aforementioned points, 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 of each linked relation pair are compared with each other in a pair-wise manner. Figure 3 shows the heat-map visu-

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 | 19 | 8 | 1 | - | 9 | 10 | 1 | - |
| Lithuanian | 26 | 7 | 2 | 2 | 11 | 16 | 3 | 2 |
| Polish | 29 | 2 | 2 | 3 | 3 | 12 | 2 | - |
| Portuguese | 27 | 5 | 1 | 2 | 5 | 10 | 1 | - |
| Russian | 17 | 4 | - | 1 | 4 | 13 | - | 1 |
| Tukish | 16 | 3 | 2 | 1 | 7 | 14 | 2 | - |

alizations of the row-wise normalized confusion matrices for 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 3a reads "15% of English explicit discourse 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 3a to 3f, the target relations vary greatly with respect to English annotations in terms of their types. On average, 573.3 of the English relations are linked to each target language, and only 72% of them retained their type. Of the five relation types, the Explicit discourse relations (78.15%) and Implicit discourse relations (77%) are conserved most frequently, whereas 61.56% of the AltLexes are converted into other relation types. The language-specific breakdown of these variations can be read in Figures 3a to 3f.

When all language pairs are considered, the top three conversions (from English to the target languages) are as follows: 32.86% of AltLex relations become Explicit; 28.53% of EntRels become Implicit and 16.24% Explicit relations become Implicit.

Of these three, 78.76% of English EntRels 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 [31] as their distinction is very subtle. These two relations are semantically related to the extent that EntRels are exploited as implicit Expansion discourse relations to increase the available training data in implicit discourse relation recognition task, yielding increases in overall performances. [32, 33].

Finally, implicitation (the omission of a connective in the target text where there is a connective in the source text) 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 [34]; the results of the current work can be used safely in future crosslinguistic investigations of implicitation (or expliciation). In all language pairs in TED-MDB, at least 10% of the English discourse relations are found to be realized implicitly. These results raise a further question: are all explicit discourse relations equally likely to be realized implicitly in the target language? Interestingly, implicitation dominantly occurs with Expansion discourse relations (Table 8). The same is not true for explicitation, where Contingency discourse relations are relatively more frequently explicitated than others on average, but they are far from being as dominant as the implicitated expansion discourse relations.

***Cross-lingual Variation in Relation Sense:*** Unlike relation types, the discourse sense of the connectives are found to be more stable across languages. On average, 86.84% of English discourse relations retained their top-level sense in the target languages.[14] Comparison > Expansion seems to be the most frequent conversion (14.12%) followed by Contingency > Expansion (10.64%) cross-lingually.

When considered together with the higher level of variation in relation types, the cross-lingual consistency of the relation senses may suggest that translators take liberty in adapting the grammar of the source material into their languages; yet, naturally, these vari-

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

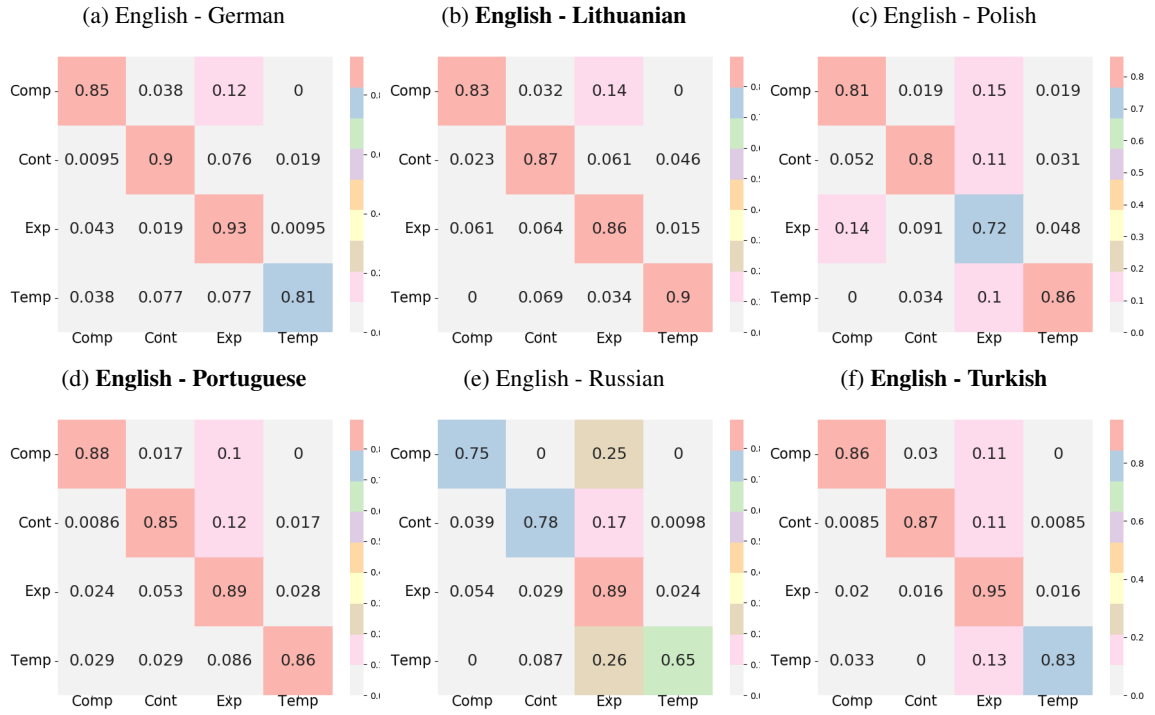



Fig. 4. Heatmap visualizations of the confusion matrices for the sense of linked discourse relations. Rows correspond to the English relations and columns denote target languages. The matrices are normalized row-wise, where each cell denotes the percentage of English relations converted to the respective label in the target language. Confusion matrices created from manually-corrected links are highlighted in bold.

ations in form did not affect the semantics as the senses of the relations are mostly preserved.

## 5. Building Bilingual Discourse Connective Lexicons

In addition to enabling linguistic investigations of cross-lingual discourse structures, a parallel corpus with linked relations has a number of practical use cases, where building bilingual discourse connective lexicons is one of them. Bilingual DC lexicons document the relationships between discourse connectives over two languages. Yet, the existing DC lexicons are overwhelmingly monolingual, where [15, 16, 35] are the only notable exceptions. In order to increase the breadth of the existing resources, the present paper exploits relation links to build such lexicons for each English-Language X pair. In the rest of the paper, the TED-MDB lexicons are introduced including our motivation to create them and their extraction procedure. Their coverage and limitations are also discussed.

### 5.1. *Motivation*

As mentioned before, discourse connective lexicons are important resources, facilitating the linking of connectives' various syntactic and semantic-pragmatic properties with their senses, which is a nontrivial task. They are also shown to be useful resources on a number of different fronts, including both human and machine translation [15, 36] and language learning and teaching [34, 37, 38].

Discourse connectives are shown to be challenging in multilingual settings such as machine translation [36] and second language learning [34, 38] due to their varying degrees of ambiguity across languages, which are not adequately accounted for in the standard resources. Standard dictionaries or similar lexical resources (e.g. word alignment databases such as Treq [39] or OPUS[15]) often fall short of providing an exhaustive list of translations for connectives, let alone grouping them according to their semantics [15, 35].

Moreover, monolingual DC lexicons have been utilized to facilitate the development of discourse-

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

Table 9
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 | 205 | 2 | 11 | 36 | 0.95 | 0.95 | 0.99 | 0.97 |
| LT EN | 205 | 3 | 9 | 60 | 0.96 | 0.96 | 0.99 | 0.97 |
| EN PT | 197 | 0 | 26 | 27 | 0.9 | 0.88 | 1 | 0.94 |
| PT EN | 197 | 1 | 21 | 24 | 0.91 | 0.9 | 0.99 | 0.95 |
| EN TR | 188 | 6 | 30 | 20 | 0.85 | 0.86 | 0.97 | 0.91 |
| TR EN | 188 | 6 | 28 | 40 | 0.87 | 0.87 | 0.97 | 0.92 |

annotated corpora [40] or the improvement of the shallow discourse parsing sub-tasks of connective identification and explicit discourse relation classification [41]. All these merits of monolingual DC lexicons can be straightforwardly expanded to a multilingual setting, given the suitable multilingual lexicons.

### 5.2. Procedure

One way of compiling a bilingual lexicon involves interlinking existing monolingual connective lexicons 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 [16]. Instead, in the current study, a more direct approach is adopted by exploiting the linked DRs. This alleviates the need for other resources. The procedure mimics the extraction of monolingual lexicons from an annotated corpus, closely following [35]. Using the relation links, connectives in different languages are mapped with one another, provided that they exist in a linked relation pair which conveys the same sense. The rationale behind our procedure is that 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.

The selection of discourse connectives and the languages solely rely on the TED-MDB annotations.[16] The extraction of bilingual connective lexicons from the linked 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 3). To remedy this situation and extend the coverage of the lexicons, implicit discourse connectives are also included, as in [35]. Specifically, the method consists of two steps, preceded by pre-processing:

0. In the pre-processing step, all linked relation pairs that include a non-Explicit or a non-Implicit discourse 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 discourse connective in the source language, translation equivalents are searched among the target language annotations using the relation links. Therefore, connective translations are provided (if any) separately for each sense. However, it is not uncommon for a matched discourse 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* 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 linked relation pairs obtained through the second method are used in the compilation of the lexicons.

### 5.3. Lexicons

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

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

Table 10

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) discourse senses per connective; (ii) translation equivalents available for each connective in the lexicons, respectively, e.g. an English connective is represented maximally by 6 German connectives.

| 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 | 2.50 |
| Turkish | 39 | 40 | 79 (67) | 1 | 5 | 1.43 | 1 | 4 | 1.84 |

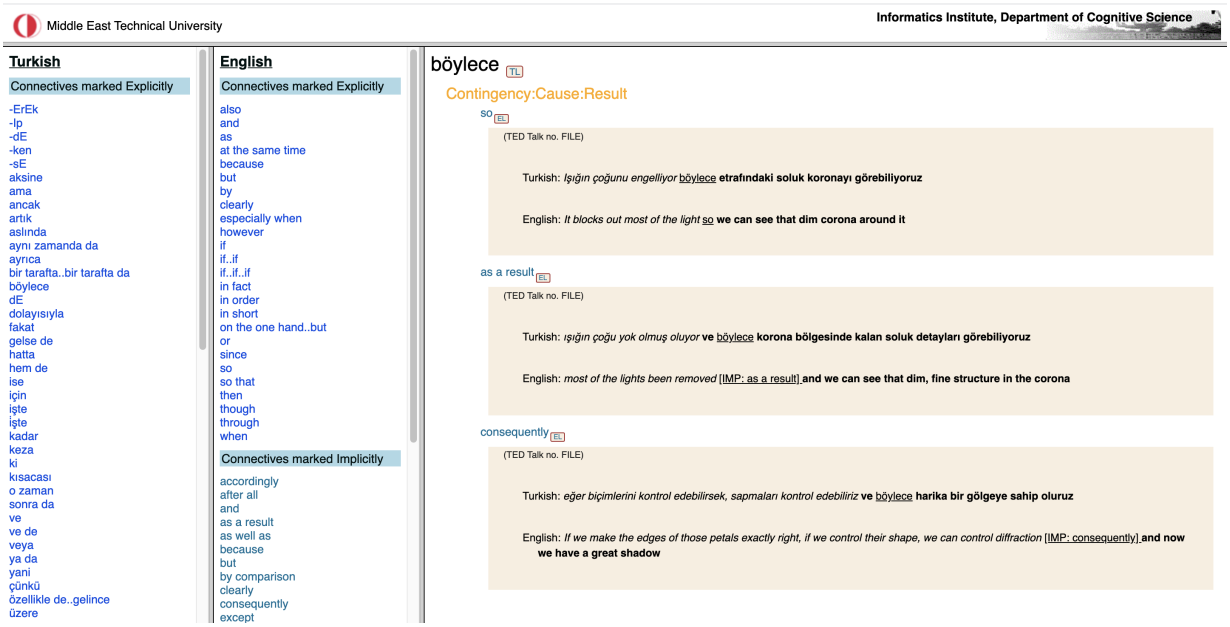
- **Connective**: Each lexicon entry is anchored to a discourse connective. The discourse connectives can be of any kind, single-word, multi-word or discontinuous (e.g. if...if). The discourse connectives are not processed in anyway, except being lower-cased.
- **Dimlex link**: The TED-MDB annotations, therefore the TED-MDB lexicons, do not include any syntactic/orthographic information regarding discourse connectives. In order to make that information available as well as creating a bridge between the bilingual and monolingual lexicons, each discourse connective and its translations are accompanied with a URL to their connective-lex[17] 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 discourse connectives appear, each translation candidate is accompanied with an example relation pair from TED-MDB.

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

A sample lexicon entry is illustrated in Figure 5. The statistics regarding each lexicon are provided in Table 10. As the entire lexicon induction phase is completely automatic, including the linking of the relations in the respective languages, the lexicons are bound to involve some errors. To evaluate the lexicons, firstly, the performance in linking Explicit discourse relations and Implicit discourse relations is checked, as those discourse relations constitute the basis of the lexicons (Table 9). In comparison to Table 7, these relation types turn out to be easier to link; in all directions, an average F-score of 0.94 is achieved. As a more direct evaluation, the lexicons generated from automatically linked pairs are compared against those from manually-corrected links that are available for three languages. On average, automatically generated lexicons capture 97.46% of the entries of the lexicons produced from the manually-corrected linkings, suggesting that the generated lexicons are of very high quality. Considering the typological variety in the evaluation languages (Lithuanian, Porgutuese, Turkish), it is safe to assume that the results are generalizable to other TED-MDB languages (German, Polish, Russion).

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 respec-

Fig. 5. A screenshot showing the entry for "böylece" in the Turkish-English lexicon.




tive languages, which would, otherwise, require a great deal of manual labor.

## 6. Conclusion

In the current work, two methods for linking the labels over the annotated 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 link relations across languages, is favored over the more traditional first method due to its higher performance. The second method is also preferable because it avoids the need for a large parallel corpus, 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 cross-lingual linking of relation annotations, which has its unique challenges. This leads to two promising results: First, a multilingual corpora with the cross-lingually linked relations 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.

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

[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] 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.

[5] J. Mírovskỳ, L. Mladová and Š. Zikánová, Connective-based measuring of the inter-annotator agreement in the annotation of discourse in PDT, in: *Coling 2010: Posters*, 2010, pp. 775–781.

[6] 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).

[7] A. Feltracco, E. Jezek, B. Magnini and M. Stede, LICO: A Lexicon of Italian Connectives., in: *CLiC-it/EVALITA*, 2016.

[8] 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.

[9] J. Mírovskỳ, P. Synková, M. Rysová and L. Poláková, CzeDLex-A Lexicon of Czech Discourse Connectives, *The Prague Bulletin of Mathematical Linguistics* **109**(1) (2017), 61.

[10] A. Mendes, I. del Rio, M. Stede and F. Dombek, A lexicon of discourse markers for portuguese–ldm-pt, in: *11th International Conference on Language Resources and Evaluation*, 2018, pp. 4379–4384.

[11] 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.

[12] R. Pasch, U. Braußе, 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.

[13] 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.

[14] W.C. Mann and S.A. Thompson, Rhetorical structure theory: Toward a functional theory of text organization, *Text-interdisciplinary Journal for the Study of Discourse* **8**(3) (1988), 243–281.

[15] 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.

[16] 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.

[17] C. Chiarcos and A. Pareja-Lora, 1 Open Data—Linked Data—Linked Open Data—Linguistic Linked Open Data (LLOD): A General Introduction, *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences* (2020), 1.

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

[19] 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.

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

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

[22] 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.

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

[24] 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.

[25] 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.

[26] J. Tiedemann, *Bitext alignment*, Morgan and Claypool Publishers, an Rafael, California, 2011.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] D. Meurers and M. Dickinson, Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics, *Language Learning* **67**(S1) (2017), 66–95.

[38] 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.

[39] 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.

[40] M. Stede and S. Heintze, Machine-assisted rhetorical structure annotation, in: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 2004, pp. 425–431.

[41] P. Bourgonje and M. Stede, Exploiting a lexical resource for discourse connective disambiguation in German, in: *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 5737–5748.

[42] T. Meyer and A. Popescu-Belis, Using sense-labeled discourse connectives for statistical machine translation, in: *Proceedings of the EACL2012 Workshop on Hybrid Approaches to Machine Translation (HyTra)*, 2012.

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

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.