Linking Discourse-level Information and the 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, School of Arts and Humanities, 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

Editors: Julia Bosque-Gil, University of Zaragoza, Spain; Milan Dojchinovski, Czech Technical University in Prague, Czech Republic; Philipp Cimiano, Bielefeld University, Germany
Solicited review: Seven Anonymous Reviewers

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 a 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 English with high accuracy, as verified against the relations of three diverse languages semi-automatically linked with relations over English texts. The resulting corpus has a great potential to reveal the divergences in local discourse relations, as well as leading to new resources, as exemplified by the induction of bilingual discourse connective lexicons.

Keywords: discourse relations, discourse connectives, discourse connective lexicons, 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, wordnets, and dictionaries, to name a few. Following the TextLink project[1], there has been an effort to synchronise the various lexicons, one of the most prominent of these being the online Connective-Lex database [1]. The entries in the Connective-Lex database provide information on discourse connectives (*but, once, although*) such as their orthography, syntactic category (coor-

[†] Equal contribution.

[1] http://textlink.ii.metu.edu.tr/

dinating conjunction, adverb, subordinating conjunction), and the senses they convey (contrast, temporal, concession). The web-based interface allows users to filter a selected lexicon based on the available information. But it supports partial cross-linguistic investigation as only two lexicons are linked to other lexicons in the database: using the feature 'synonyms', the Italian lexicon is linked to the German DiMLex, and the Portuguese lexicon is linked to the English DiMLex.

A collaborative effort that evolved during the lifespan of the TextLink project, 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). Since most of the languages involved in the project did not have a discourse connective lexicon when this resource was created, no discourse connective lexicons were utilized during its development. 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 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 for carrying out cross-lingual discourse analysis and inducing monolingual and bilingual discourse connective lexicons for a new set of languages. But this resource presents a challenge for both aims because discourse relations in one language are annotated blind to the annotations performed on other languages in order to accurately account for the differences exhibited across languages. Such natural cross-lingual discrepancies could hinder any efforts of cross-linguistic comparisons or induction of new resources such as bilingual lexicons among the languages included in the corpus. A sentence-to-sentence alignment of the texts would not suffice for the induction of lexicons, neither would it enable cross-linguistic analysis, as one would not make sense of which connectives are kept or omitted, and how the meaning of a connective varies. A data linking task, more specifically, a relation linking task[2] must be performed on TED-MDB, which involves the alignment of annotated relations and linking of the labels over annotated relations.

The main contributions of the paper are: (1) to introduce two alternative methods for relation linking 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 an overview of the discourse structures across TED-MDB languages facilitated by the relation linking task, and (4) to automatically induce new bilingual discourse connective lexicons for each TED-MDB language (target languages) and English (source language), 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. In §3, a short review on existing bilingual and multilingual discourse connective lexicons is provided. §4 describes the data linking task, highlighting its challenges in §4.1, followed by the details of the two proposed methods in §4.2 and §4.3. §5 provides an evaluation of the linked data and an error analysis, also introducing the structure of the resulting data in XML format (§5.3). In §6, an overview of the discourse structures observed in TED-MDB is presented together with the statistics obtained from the relation linking task. In §7, the bilingual lexicons that link the connectives induced from the linked data are described. The paper ends with a conclusion and some future directions for further research (§8).

2. 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 [2]. The wide coverage of TED talks in terms of topics and translated languages

[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 between source and the target language.

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

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 are 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

Annotation Scheme of TED-MDB: The texts included in TED-MDB 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 are referred to as Explicit relations when marked by a discourse connective, Implicit when no overt connective is present.

In an Explicit discourse relation, the relation that holds between the two arguments are made salient by the connective, 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)

In the absence of a discourse connective, 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 dos pacientes, deixando-os com escaras e feridas.**
[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 blisters.

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]

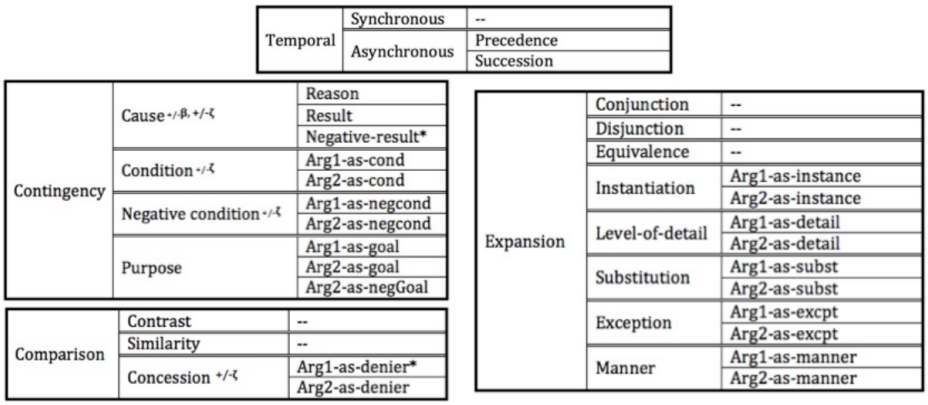
(3) *... many of my early memories involved intricate daydreams where I would walk across borders, forage for berries, and meet all kinds of strange people living unconventional lives on the road*. Years have passed, but **many of the adventures I fantasized about as a child – traveling and weaving my way between worlds other than my own — have become realities through my work as a documentary photographer.**
[AltLex, Temporal:Asynchronous: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 [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) and shows the highest semantic category in the hierarchically organized semantic categories. The second level sense, or Level2 sense (Cause) is a subsense of the top category, itself containing a subsense one level down, referred to as the

[4]The examples are taken from TED-MDB. In all the examples, the discourse connective or AltLex 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 tag. As in the PDTB, Arg2 is the discourse segment hosted by the discourse connective or AltLex, Arg1 is the other discourse unit.

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

Fig. 1. PDTB 3.0 sense hierarchy [5]



third level sense, or Level3 sense (Result). In this way, the sense tags provide 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 the entity introduced in the other argument. These cases are annotated as an Entity Relation (EntRel), as illustrated in Example 4. Finally, when no relation holds between the two adjacent sentences, the relation is of the type NoRel (Example 5). The annotation scheme is summarized in Table 2. [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 in TED-MDB, which applies in cases where the speaker asks a question and immediately answers it. In monologues, not all questions are asked to be answered: in TED Talk transcripts, Hypophora has 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, the texts in each language were 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 led to different sets of relations annotated for each language. Table 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, ~20% of the whole corpus (i.e. two TED talks per language) are annotated by an independent annotator, using the annotation scheme and following the annotation principles summarized above. The inter-annotator agreement (IAA) is performed on two levels following [6]: (i) whether or not annotators spotted a relation between the same discourse units, (ii) whether or not the 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.

[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.

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 relation	Arg1, Arg2	Yes
Entity Relation (EntRel)	None	Arg1, Arg2	No
No Relation (NoRel)	None	Arg1, Arg2	No

3. Discourse Connective Lexicons

The last decade has seen an upsurge in the development of monolingual discourse connective lexicons, such as LexConn [7], LiCo[8], DiMLex [9], CzeDLex [10], LDM-PT [11], and TCL [12], all included in the Connective-Lex database, designed to act as a hub for publicly available monolingual connective lexicons. In the back-end, the lexicons are stored in XML format with a simple structure; hence, the integration of a new lexicon in the database is a straightforward process. Currently, the database hosts lexicons of 11 different languages. Researchers have envisioned linking the existing lexicons but the linking task poses certain challenges as they may be modeled on different relation taxonomies [1], or the existing connective inventories may not be machine-readable as is the case for many of the discourse connective inventories not integrated into Connective-Lex . The connective inventories may also vary in depth and detail making the mapping of linguistic descriptions difficult. For example, the *Spanish Diccionario de partículas discursivas del español* (DPDE – [13]) includes explicit information on discourse particles in Spanish but it excludes conjunctions and prepositions. The German resource *Handbuch der Konnektoren* [14, 15] contains detailed linguistic descriptions of discourse connectives, including their possible positions in a sentence, also the register and possible modifiers.

Nevertheless, there exist a few multilingual discourse connective lexicons [16, 17], and a recent work [18] achieves the first comprehensive edition of machine-readable discourse marker lexicons developed in accordance with web standards and Linguistic Linked Open Data (LLOD) principles. This work links the sense definitions and annotation schemes in Connective-Lex along with other existing discourse marker inventories such as TED-MDB, under a common standard.

4. Data Linking in TED-MDB

The main task of the present work, relation linking, involves linking the components of a relation in different languages, i.e. the labels for relation type, arguments, connectives, and their discourse senses (if available), as explained in §2. The objective, therefore, is not aligning the words and sentences of different languages, but linking the connectives, sentences or parts of sentences annotated as a relation together with the linguistic labels over those relations. The outcome of this work will enable access to the discourse labels over texts in different languages on the level of format, as well as permitting easy access to the discourse structures of different languages by means of the reference to existing labels [18]. Furthermore, the main task of the current paper will not only support the induction of a multilingual discourse connective lexicon, but it will also allow immediate access to different datasets within TED-MDB. In the sections that follow, we present the steps of a novel approach of linking texts and the labels over those texts, and we describe the current format of the resulting resource. However, it must be noted that Linked Open Data Principles [19] are yet to be applied to this resource.

The relation linking 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 [20–22]. 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

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

is not a straightforward task. In the rest of this section, the challenges surrounding the task undertaken and the methods to address these challenges are provided.

4.1. 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 as described below. Typically, the argument spans of the relations tend to vary across languages as illustrated in Example 7.

(7) *... we take the Hubble Space Telescope* and **we turn it around ...** [Explicit, Expansion: Conjunction] (English, TED talk 1976)

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

Here, none of the arguments match completely. Both arguments of the English connective consist of full clauses unlike the Turkish arguments. The Arg2 of the Turkish relation only consists of the verb (*tut-* which refers to 'take' in this context) with no object or subject. The subject information is conveyed in the other argument; however, Turkish Arg2 still lacks the object *it* that refers to *The Hubble Space Telescope*, because objects can be dropped in Turkish as in this case. The relation could have been translated to Turkish mimicking the same syntactic structure of English; yet, the translator did not opt to that, probably because it would sound less natural.

Secondly, the PDTB 3.0 annotation manual allows multiple relations over similar text spans [5]. In the English sentence of Example 8, for example, two relations are annotated over the same text spans, an Explicit relation (*and*), and an Implicit one (*as a result*). In the equivalent Portuguese sentence, an Explicit relation (*por* 'due to') (which has no counterpart in the English sentence) is annotated in addition to an Explicit relation (*e* 'and'), which should be linked to the Explicit relation conveyed by *and* in English. As the Arg2 of the Portuguese relation (*por*) contains the Arg2 of the Implicit English relation (*as a result*), the linking task becomes challenging.

(8) Thomas Gilovich and his team from Cornell studied this difference and found that the frustration silver medalists feel compared to bronze, who are typically a bit more happy *to have just not received fourth place* and (implicit = as a result) **not medaled at all, gives silver medalists a focus on follow-up competition**.
and: [Explicit, Expansion:Conjunction]
as a result: [Implicit,Contingency:Cause:Result]
(English, TED Talk no. 1978)

Thomas Gilovich e a sua equipa da Universidade de Cornell estudaram esta diferença e descobriram que a frustração que os 'medalhas de prata' sentem, comparada com a dos 'de bronze', que *normalmente se sentem mais felizes* por ***não terem ficado em quarto lugar*** **e não terem recebido nenhuma medalha**, dá aos 'medalhas de prata' uma concentração na competição seguinte.
e: [Explicit,Expansion:Conjunction]
por: [Explicit,Contingency:Cause:Reason]
(Portuguese, TED Talk no. 1978)

4.2. Method I: Relation Linking through Word Alignments

The first method attempts to link the relations annotated over the texts of different languages through

word alignments, adapted from the conventional annotation projection practice. The details are presented below.

4.2.1. 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 the 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], which is based on the hunalign algorithm [23], 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 target text, as listed in Table 4. Since any error in this step would be propagated through the pipeline, we settled on aligning each target 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

4.2.2. 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 learned from a large parallel data by using the model 3 of EFLOMAL[8] [24]. The parallel corpora are created for each language pair by concatenating the largest corpus of each language pair available in the OPUS database [25]. All the corpora are obtained and processed using OpusTools[9] [26]. The data sizes of each corpus are listed in Table 5.

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

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 where the source language is set as the non-English language. Yet, using alignments directly from either direction is reported to underperform [20, 22]; 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 strictest heuristic and leads to fewer but precise alignments.
- **Grow-diag**: Grow-diag expands on the intersection set by adding the diagonally neighbouring data points.
- **Grow-diag-final**: Adds another step on grow-diag 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.

4.2.3. Relation Linking

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 §4.1, 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 textual components of that relation, namely Arg1, Arg2, and the discourse connec-

[7] https://sourceforge.net/projects/aligner/
[8] https://github.com/robertostling/eflomal
[9] https://github.com/Helsinki-NLP/OpusTools

tive (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 it is added to the score (also see §4.3.2).

4.3. Method II: Relation Linking through Cross-lingual Sentence Embeddings

The second method uses the modern, language agnostic sentence encoders which are capable of assigning similar representations to semantically similar linguistic units across languages. The method is a continuation of a previous study [27] which 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, the 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 (§4.3.1), 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). Semantic similarity is calculated as the cosine similarity between the LASER embeddings [29] of each relation's text segments.

[10] A bitext unit is a pair of source and target sentences which have either partial or full translation equivalence [28].

4.3.1. 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, training is performed for language pairs involving the source language and three of the target languages in TED-MDB, namely, Turkish, Portuguese, and Lithuanian (EN-TR, EN-PT, EN-LT). First, the relation labels over English texts are automatically matched with those on the texts of these three languages. Then, the performance of the automatic process is manually checked by the authors and wrong matches are corrected. In the training phase, further performance evaluation is done using this manually checked data. Throughout the paper, we refer to this data as manually-corrected or semi-automatically linked data.

For training, the six English files are split into the train and test data sets 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 leaving the other two to be used as the test set.[12] Using semantic threshold values starting from 0 to 0.95 and increasing 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 the 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 often results in linking English relations with wrong target relations. This can be seen in Example 9, where the relation anchored by *and* in the English sentence is falsely linked to that of

[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.

ve 'and' in Turkish, as their senses and relation types match. However, different from the English relation, the Turkish relation holds between thinking about *how to create maps* and *how to draw them.*

(9) 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)

4.3.2. Relation Linking

In the relation linking stage, a scoring algorithm is developed, where the links over the relations are determined on the basis of the total score obtained from the semantic similarity between the relations, the degree of correspondence for sense levels, and the relation types. Each step is described below.

1. The similarity score between the relation pairs are calculated considering their text segments. The pairs that do not exceed the similarity threshold learned in the previous step (described in §4.3.1) are discarded.
2. The semantic similarity score is combined with another score that reflects the *semantic* match between the relation pairs in each source-target language pair. That is, in a ranked manner, a match on Level1 sense is given a score of 1000, a match on Level2 sense 100, a match of Level3 sense 10, and 1 is assigned for relation type match (Explicit, Implicit etc.). If there is no match on the Level1 sense, scoring is not done for other sense levels and relation types, but especially for NoRels and EntRels, relation type match plays a key role as they are not assigned a sense label.
3. For each source relation, the target relation that yields the highest 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 illustrated on a sample text provided in Example 10. The text involves three Explicit relations in two languages (EN, TR) signaled by (*but, as, and*) and (*ama 'but', kadar 'as', ve 'and'*) as summarized in Example 11. As the first step, all pairwise combinations of these relations are calculated, resulting in a (3x3) matrix as shown in Table 6. Then, following the scoring procedure, each pair is assigned a score. On the Turkish side, the connective label *Ama* (the first column of Table 6) matches the English connective label *But* in relation type and relation sense in all levels; similarly, the connective label *kadar* in Turkish matches the English connective label *as* in relation type and relation sense in all levels. The same situation holds for the third Turkish-English connective label pair *ve* and *and*. For a non-linking case, consider the third English relation conveyed by *and*, which has no match with the first Turkish connective *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

(10) 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)

(11) *English* :

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

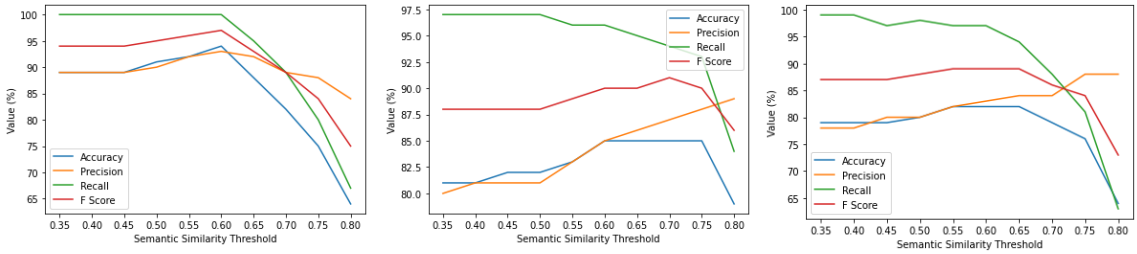


Fig. 2. The change of evaluation metrics (Accuracy, Precision, Recall and F-Score) at different levels of semantic threshold values. Although the threshold is searched between 0 and 0.95 at increments of 0.05, to achieve better visualization, only the values between 0.3 and 0.85 are provided.

- DR[13]-Explicit-Comparison.Concession.Arg2-as-denier-DC[14]-**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

The Relation Scoring Matrix for Example 10. The numbers refer to the scores based on sense/type agreement + semantic similarity of the 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 the results from our relation linking procedure revealed the need for certain revisions. As mentioned before, it is common for more than one discourse relation to hold between similar arguments ([30]), 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 bitext 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.

[13]DR stands for Discourse Relation.
[14]DC is used for Discourse Connective.

5. Evaluation and an Error Analysis

In the literature, data linking quality is evaluated by using the standard precision, recall and 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 [6, 30].

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 relation linking, accuracy should also be taken into consideration; providing information on the non-matching data pairs is as important as providing matching data. In a data linking task such as ours, non-matching data offers valuable insights into linguistics, machine translation and in particular, into the assessment of the annotation quality.

5.1. Evaluation of Method I and Method II

The methods proposed in the current work are evaluated against the manually-corrected data that exists

for English texts and the corresponding Lithuanian, European Portuguese and Turkish texts.[15] The linking performance of the proposed methods is 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.

5.2. Error Analysis

Regardless of the language pair, the relation linking task is challenged if argument span lengths in the source and target relation differ, if the text is translated freely, or if the argument spans of the source and target relation have partial overlap. In such cases, both methods fail and performance decreases due to an increase either in False Negatives (see Example 12) or False Positives (see Examples 13 and 14). An increase in those numbers affects all the performance metrics. In the following, we report some instances that led to performance drop under three headings. 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.

[15] Unfortunately, relation links for the remaining language pairs did not go through a manual correction procedure.

Different argument span lengths: As discussed in §4.1, the variation in argument spans across languages lead to mismatches. Example 12 further illustrates this problem, showing that neither method could link the Lithuanian relation to its English counterpart due to longer Arg2 annotation in Lithuanian.

(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)

Different realization of discourse relations: Translation may lead to different realizations of discourse relations. In Example 13, the relation conveyed by *and* in the English sentence does not exist in Turkish but the English relation is contained in Arg2 of the Turkish relation in a freely translated form. Our data linking models erroneously link English *and* with Turkish *sanki* 'as if'.

(13) Lord, grant that I desire more than I can accomplish, Michelangelo implored, as if to that Old Testament God on the Sistine Chapel, and he himself was that Adam *with his finger outstretched* and **not quite touching that God's hand**. [Explicit, Expansion:Conjunction] (English, TED Talk no. 1978)

Tanrım, bana başarabileceğimden daha fazlasını istemeyi bahşet, diye yakarmıştı Michelangelo, sanki **Sistina Şapeli'ndeki Eski Ahit Tanrısı'na ve kendisi de uzattığı parmağı Tanrı'nın eline tam değmeyen Âdem'di.** [Explicit, Comparison:Similarity] (Turkish, TED Talk no. 1978)

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 as explained in §4.3.1. The first three parts refer to the results of the first method grouped by the symmetrization heuristics, ranked from the most restrictive to the least restrictive, as explained in §4.2.2.

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**

Partially overlapping argument spans: In certain cases, even though source and target relations do not match in a bitext unit, they have partially overlapping argument spans. In Example 14, while the English Implicit discourse relation signaled by *as a result* has no counterpart in the Portuguese sentence, there is an Explicit relation conveyed by *sem* 'without' in Portuguese, partially overlapping with the Arg2 of the English discourse relation; for this reason, both methods erroneously link the two relations.

(14) And *I saw that gave her more tenacity*, (implicit = as a result) **and she went after it again and again**. [Implicit, Contingency:Cause:Result] (English, TED Talk no. 1978)

E vi que isso deu-lhe mais persistência, e *continuou, continuou*, sem **parar**. [Explicit, Expansion:Conjunction] (Portuguese, TED Talk no. 1978)

5.3. Dissemination of the Linked Relations

The resulting set of linked relations are made publicly available as a set of XML files to facilitate further research.[16] For each language pair (English-Language X), the resulting set of both linked as well as non-linked relations are stored in a separate XML file. A sample English-Turkish file is provided in Figure 3 in order to illustrate the underlying structure. The relations are stored under the <*relation*> nodes with children of Arg1, Arg2 and Conn (if there is any), and five attributes that encode the sense, type information of the relation as well as the meta-level information regarding its language, its source TED talk ID, and the unique relation ID. The linked relations are combined under the <*relation_pair*> elements which further encode the *alignment score*, showing the confidence of Method II in linking the relations. The set of linked re-

[16] https://github.com/MurathanKurfali/Ted-MDB-Annotations

Fig. 3. A sample file showing the structure of the adopted XML schema in the published linked relations.

```
<doc>
  <linked_relations>
    <relation_pair alignment_score="1111.85">
      <relation RelID="DR64" TalkID="talk_1978" Language="en" Type="Explicit" Sense="Contingency:Cause:Reason">
        <Arg1>I stayed</Arg1>
        <Arg2>I realized I was witnessing whats so rare to glimpse, that difference between success and mastery</Arg2>
        <Conn>because</Conn>
      </relation>
      <relation RelID="DR58" TalkID="talk_1978" Language="tr" Type="Explicit" Sense="Contingency:Cause:Reason">
        <Arg1>orada kaldım</Arg1>
        <Arg2>çok nadir olarak görülecek bir şeye tanıklık ettiğimi anladım, başarı ve ustalık arasındaki o farka</Arg2>
        <Conn>çünkü</Conn>
      </relation>
    </relation_pair>
  </linked_relations>
  <non_paired_relations>
    <relation RelID="DR16" TalkID="talk_1978" Language="en" Type="Implicit" Sense="Expansion:Level-of-detail:Arg2-as-detail">
      <Arg1>this is the thing</Arg1>
      <Arg2>What gets us to convert success into mastery</Arg2>
      <Conn>actually</Conn>
    </relation>
  </non_paired_relations>
</doc>
```

lations are grouped under the *<linked_relations>* element. Additionally, the set of non-linked relations, which do not have any counterpart in the other language, are listed under the *<non_paired_relations>*. We believe that these relations are equally informative regarding the discourse structures of the languages involved. It must also be noted that these links are created automatically so there may be some false negatives in this set which can be corrected via manual effort.

6. 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 [31, 32], 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 target languages 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 relations. In order to maximize the reliability of our analysis, we used the manually-corrected data which exists for three target languages (Lithuanian, Turkish, Portuguese) and the automatically linked data obtained through Method II for the remaining languages. Therefore, the observations should be approached cautiously due to possible incorrect links; yet, high F-scores are obtained in capturing the semi-automatically prepared links (see Table 7). This suggests 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 for future work.

Cross-lingual Variation in Relation Types: In order to answer the first question, the relation types of each linked relation are compared with each other in a pairwise manner. Figure 4 shows the heat-map visualizations 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 4a 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 4a to 4f, the target relations vary greatly with respect to English relations in terms

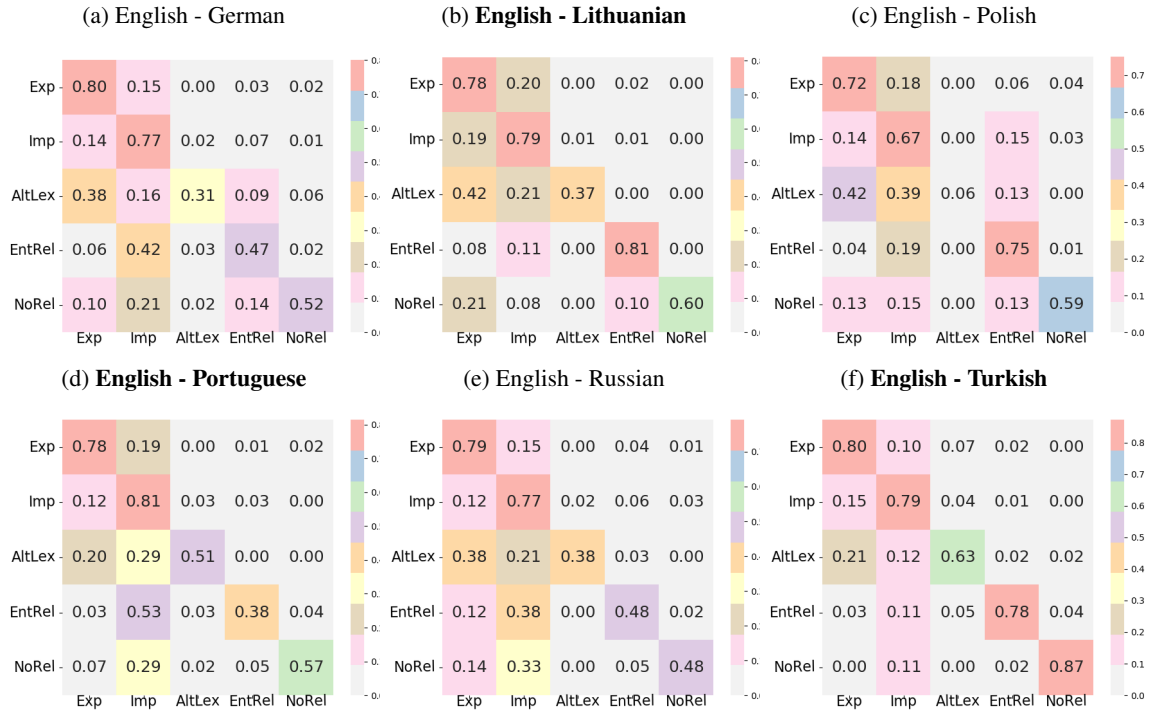


Fig. 4. Heatmap visualizations of the confusion matrices for relation type of the linked discourse relations. Rows correspond to the English relations and columns denote target languages. The matrices are normalized row-wise where each cell denotes the percentage of English relations converted to the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

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	23	10	1	4	12	19	2	-
Lithuanian	39	9	4	3	17	28	6	5
Polish	36	3	5	5	10	23	6	-
Portuguese	36	10	3	4	7	21	1	3
Russian	22	11	4	2	5	21	1	2
Turkish	20	3	2	5	12	24	3	1

of their types. On average, 573.3 of the English relations are linked to the relations in 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 are presented in Figures 4a to 4f.

When all language pairs are considered, the top three conversions (from English to the target languages) are as follows: 32.86% of AltLexes become Explicit; 28.53% of EntRels become Implicit and 16.24% Explicit relations become Implicit. Explicit relations becoming Implicit and AltLexes becoming Explicit are linguistically motivated and arise due to language-specific properties, the translator's choice, or both. The EntRel cases are often methodologically motivated: the annotator in one language decides to label a relation as EntRel while another annotator as Implicit.

Moreover, EntRels and Implicit relations have been reported to be the most easily confused pairs even within the same language [33] 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 tasks, yielding increases in overall performance [34, 35]. In fact, we see the same tendency in our corpus, where 78.76% of English EntRels that become Implicit are annotated as Implicit Expansion relations in the non-English language, in accordance with the previous observations.

Finally, implicitation of a discourse connective (the omission of a connective in the target text where there is one 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 of discourse connectives is an actively studied topic in discourse [36], the results of the current work can be used safely in future crosslinguistic investigations of implicitation (or its reverse, expliciation) of discourse connectives. 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 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 often explicitated than others on average, but they are far from being as dominant as the implicitated Expansion discourse relations.

Cross-lingual Variation in Discourse Sense: Unlike relation types, the discourse sense of the connectives are found to be quite stable across languages. On average, 86.84% of English discourse relations retained their top-level sense in the target languages.[17] Level1 sense annotation differences involve Comparison relations, an average of 14.12% of which are annotated as Expansion relations in the target languages.

7. 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 discourse connective lexicons document the relationships between discourse connectives over two languages. Yet, the existing discourse connective lexicons are overwhelmingly monolingual, where [16–18, 37] 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.

7.1. Motivation

As mentioned before, discourse connective lexicons are important resources that facilitate the linking of syntactic and semantic-pragmatic properties of connectives as well as 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 [16] and language learning and teaching [38].

Discourse connectives are also known to be challenging in multilingual settings such as machine translation [39] and second language learning [36, 40] 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 [41] or OPUS[18]) often fall short of providing an exhaustive list of translations for connectives, let alone grouping them according to their semantics [16, 37].

Moreover, monolingual discourse connective lexicons have been utilized to facilitate the development of discourse-annotated corpora [42] or the improvement of the shallow discourse parsing sub-tasks of connective identification and explicit discourse relation classification [43]. All these merits of monolingual discourse connective lexicons can be straightforwardly expanded to a multilingual setting, given the suitable multilingual lexicons.

7.2. Procedure

One way of compiling a bilingual lexicon involves interlinking existing monolingual discourse connective lexicons by exploiting translation candidate tables

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

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

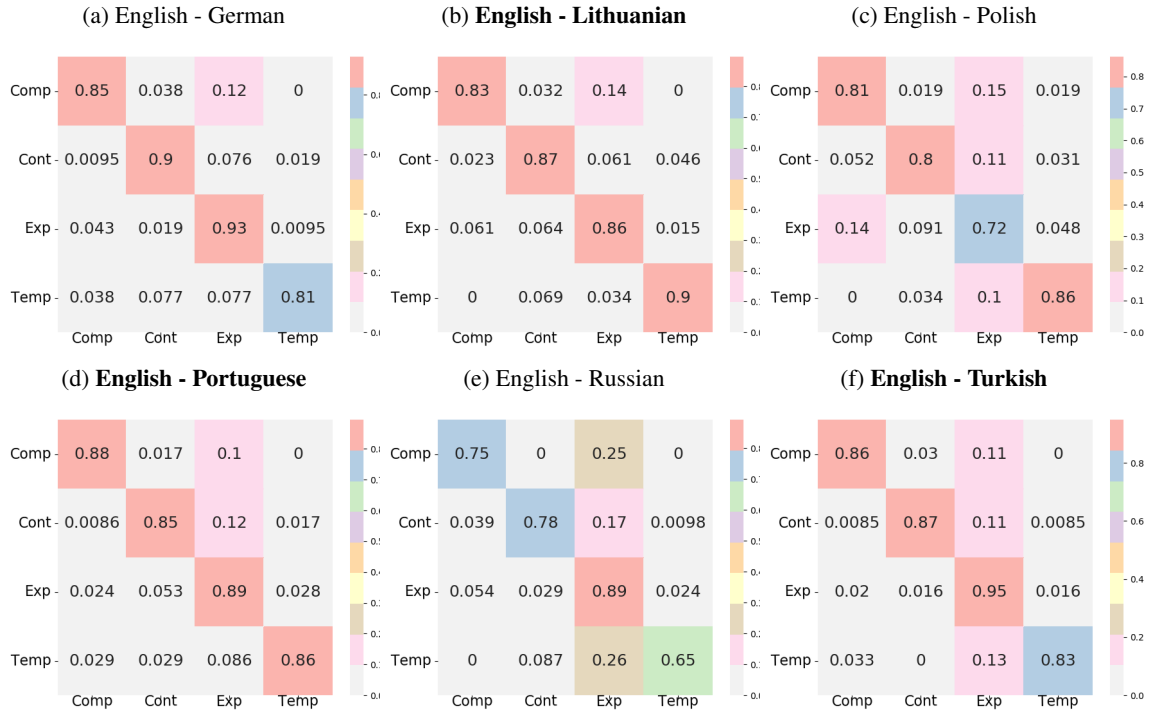


Fig. 5. Heatmap visualizations of the confusion matrices for the sense of linked discourse relations. Rows correspond to the English relations and columns denote target languages. The matrices are normalized row-wise, where each cell denotes the percentage of English relations converted to the respective label in the target language. Confusion matrices created from manually-corrected links are highlighted in bold.

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) 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

calculated from large parallel corpora. To arrive at the bilingual discourse connective lexicon, the translation candidates are filtered in a way that for each possible sense of the source language connective, only those translations that can signal the same sense (determined by the discourse connective lexicons of those partic-

Table 10
The performance of 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

ular languages) are kept [17]. Instead, in the current study, a more direct approach is adopted by exploiting the linked discourse connectives. This alleviates the need for other resources. The procedure mimics the extraction of monolingual lexicons from an annotated corpus, closely following [37]. Using the relation links, discourse connectives in different languages are mapped with one another, provided that they exist in a linked relation which conveys the same sense. The rationale behind our procedure is that bilingual discourse connective 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.[19] The extraction of bilingual discourse 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 [37]. Specifically, the method consists of two steps, preceded by pre-processing:

0. In the pre-processing step, all linked relations that include a non-Explicit or a non-Implicit discourse relation on 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 linked relations. 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 relations obtained through Method II are used in the compilation of the lexicons.

7.3. Lexicons

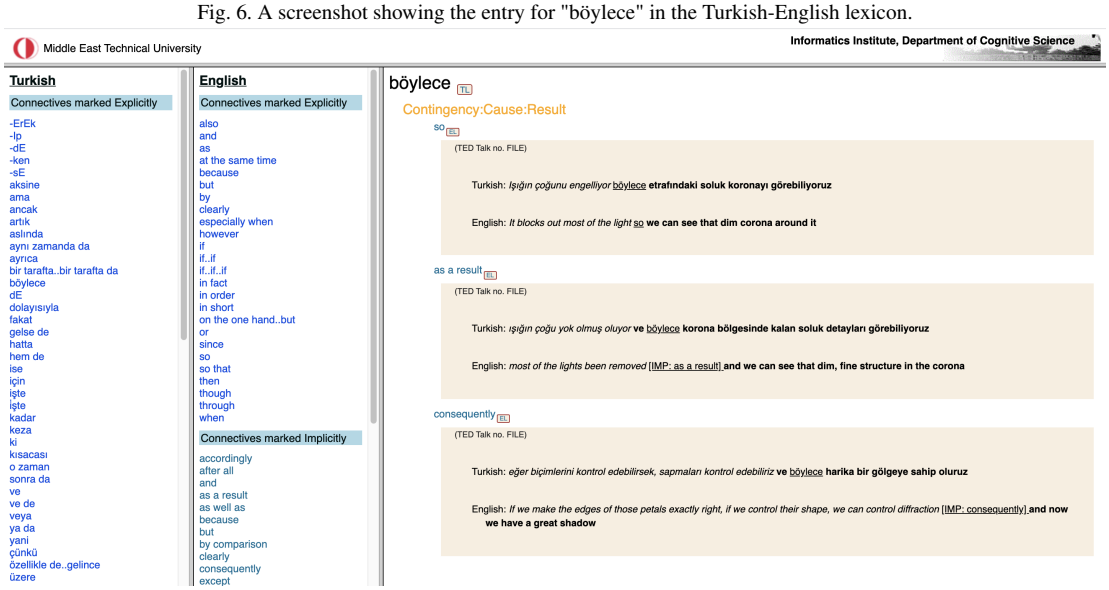
The generated TED-MDB lexicons adopt a common structure. To repeat:

- **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 connectives are represented in lower-case letters.
- **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[20] entry.

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

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

Fig. 6. A screenshot showing the entry for "böylece" in the Turkish-English lexicon.



- **Sense list:** The list of observed senses (according to the PDTB 3.0 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.

A sample lexicon entry is illustrated in Figure 6. The statistics regarding each lexicon are provided in Table 9. As the entire lexicon induction phase is completely automatic, including relation linking 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 10). 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 relation links are compared against manually-corrected links that are available for English and the translations in three languages. On average, lexicons generated from relation links capture 97.46% of the entries of the lexicons produced from the manually-corrected links, suggesting that the generated lexicons are of very high quality. Considering the typological variety in the evaluation languages (Lithuanian, Portuguese, Turkish), it is safe to assume that the results are generalizable to other TED-MDB languages (German, Polish, Russian).

Overall, through adopting a fully automatic pipeline, a number of high quality bilingual discourse connective 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 researchers of the respective languages, which would, otherwise, require a great deal of manual labor.

8. Conclusion

In the current work, in order to improve the data structure of TED-MDB and facilitate further research, two methods for data linking (more specifically, relation linking) are proposed, one of them using word alignments and the other relying on distributional semantics. The data linking task attempted in the current work resulted in a resource with a better and more informative data structure than the original TED-MDB and enabled the induction of bilingual lexicons of six language pairs, which are presented in a user-friendly format, where discourse connectives, both Explicit and Implicit, are linked. The new resource will facilitate

cross-lingual investigations of the discourse structures of the languages included in the corpus, as a preliminary examination already illustrates. Due to the challenges specific to the current task, each relation linking method was tailored to the current context through a set of heuristics. Overall, the second method, which employs multilingual embeddings to relation linking, 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 applied the data linking concept 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 corpus with the linked relations would enable many cross-linguistic studies to be performed, including machine translation, shallow discourse parsing, etc. Secondly, the bilingual discourse connective lexicons were extracted purely contextually. These lexicons can be useful in many domains of information technology.

In the future, we plan to extend the bilingual lexicons to the multilingual level to enable a better perspective on the use of discourse connectives across multiple languages. We also intend to integrate the dictionaries into the linked discourse connective lexicon of [18], offering new functionalities such as cross-lingual querying over the linked relations. In this way, a resource in machine-readable format compatible with the Linked Open Data Principles will be obtained.

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). doi:10.4000/discours.10098.

[2] A. Mendes and D. Zeyrek, The discourse markers *well* and *so* and their equivalents in the Portuguese and Turkish subparts of the TED-MDB corpus, in: *Corpora in Translation and Contrastive Research in the Digital Age. Recent advances and explorations*, J. Lavid-López, C. Maíz-Arévalo and J.R. Zamorano-Mansilla, eds, John Benjamins, 2021. doi:10.1075/btl.158.08men.

[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* **54** (2) (2020), 587–613. doi:10.1007/s10579-019-09445-9.

[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. doi:10.1162/COLI_a_00204.

[5] B. Webber, R. Prasad, A. Lee and A. Joshi, The Penn Discourse Treebank 3.0 Annotation Manual, *Philadelphia, University of Pennsylvania* (2019).

[6] J. Mírovský, L. Mladová and Š. Zikánová, Connective-Based Measuring of the Inter-Annotator Agreement in the Annotation of Discourse in PDT, in: *Coling 2010: Posters*, Coling 2010 Organizing Committee, Beijing, China, 2010, pp. 775–781. https://aclanthology.org/C10-2089.

[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). doi:10.4000/discours.8645.

[8] A. Feltracco, E. Jezek, B. Magnini and M. Stede, LICO: A Lexicon of Italian Connectives., in: *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it*, Torino: Accademia University Press, 2016. doi:10.4000/books.aaccademia.1770.

[9] 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)*, European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 1008–1013. https://aclanthology.org/L16-1160.

[10] 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. doi:10.1515/pralin-2017-0039.

[11] A. Mendes, I. del Rio, M. Stede and F. Dombek, A Lexicon of Discourse Markers for Portuguese – LDM-PT, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, 2018, pp. 4379–4384. https://aclanthology.org/L18-1693.

[12] D. Zeyrek and K. Başıbüyük, TCL - a Lexicon of Turkish Discourse Connectives, in: *Proceedings of the First International Workshop on Designing Meaning Representations*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 73–81. doi:10.18653/v1/W19-3308.

[13] 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.

[14] 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.

[15] 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.

[16] 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, Torino: Accademia University Press, 2017, pp. 53–58. doi:10.4000/books.aaccademia.2360.

[17] L. Poláková, K. Rysová, M. Rysová and J. Mírovský, GeCzLex: Lexicon of Czech and German Anaphoric Connectives, in: *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2020, pp. 1089–1096. https://aclanthology.org/2020.lrec-1.137.

[18] C. Chiarcos and A. Pareja-Lora, *1: Open Data—Linked Data—Linked Open Data—Linguistic Linked Open Data (LLOD): A General Introduction*, in: *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*, 2019, pp. 1–17.

[19] C. Chiarcos, J. McCrae, P. Cimiano and C. Fellbaum, Towards open data for linguistics: Linguistic linked data, in: *New Trends of Research in Ontologies and Lexical Resources*, Springer, 2013, pp. 7–25.

[20] Y. Versley, Discovery of ambiguous and unambiguous discourse connectives via annotation projection, in: *Workshop on the Annotation and Exploitation of Parallel Corpora (AEPC)*, 2010, pp. 83–82. http://hdl.handle.net/10062/15953.

[21] 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*, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 577–587. https://aclanthology.org/C14-1055.

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

[23] 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. doi:10.1075/cilt.292.32var.

[24] R. Östling and J. Tiedemann, Efficient Word Alignment with Markov Chain Monte Carlo, *The Prague Bulletin of Mathematical Linguistics* **106** (1) (2016), 125–146. doi:10.1515/pralin-2016-0013.

[25] J. Tiedemann, Parallel Data, Tools and Interfaces in OPUS, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 2214–2218. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

[26] 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.

[27] 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.

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

[29] M. Artetxe and H. Schwenk, Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond, *Transactions of the Association for Computational Linguistics* **7** (2019), 597–610. doi:10.1162/tacl_a_00288.

[30] 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*, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 33–42. https://aclanthology.org/E17-4004.

[31] 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. doi:10.1016/j.pragma.2017.10.010.

[32] S. Zufferey, Discourse connectives across languages: Factors influencing their explicit or implicit translation, *Languages in Contrast* **16** (2) (2016), 264–279. doi:10.1075/lic.16.2.05zuf.

[33] D. Zeyrek and M. Kurfalı, TDB 1.1: Extensions on Turkish Discourse Bank, in: *Proceedings of the 11th Linguistic Annotation Workshop*, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 76–81. doi:10.18653/v1/W17-0809.

[34] 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*, Association for Computational Linguistics, Seoul, South Korea, 2012, pp. 108–112. https://aclanthology.org/W12-1614.

[35] 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. doi:10.1162/tacl_a_00142.

[36] 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. doi:10.1177/0267658315573349.

[37] M. Kurfalı, S. Ozer, D. Zeyrek and A. Mendes, TED-MDB Lexicons: Tr-EnConnLex, Pt-EnConnLex, in: *Proceedings of the First Workshop on Computational Approaches to Discourse*, Association for Computational Linguistics, Online, 2020, pp. 148–153. doi:10.18653/v1/2020.codi-1.15.

[38] D. Meurers and M. Dickinson, Evidence and Interpretation in Language Learning Research: Opportunities for Collaboration With Computational Linguistics, *Language Learning* **67** (2017), 66–95. doi:10.1111/lang.12233.

[39] T. Meyer, A. Popescu-Belis, N. Hajlaoui and A. Gesmundo, Machine Translation of Labeled Discourse Connectives, in: *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*, Association for Machine Translation in the Americas, San Diego, California, USA, 2012. https://aclanthology.org/2012.amta-papers.20.

[40] 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. doi:10.3390/languages5030035.

[41] 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.

[42] M. Stede and S. Heintze, Machine-Assisted Rhetorical Structure Annotation, in: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*,

COLING, Geneva, Switzerland, 2004, pp. 425–431. https://aclanthology.org/C04-1061.

[43] 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*, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 5737–5748. doi:10.18653/v1/2020.coling-main.505.

[44] T. Meyer and A. Popescu-Belis, Using Sense-labeled Discourse Connectives for Statistical Machine Translation, in: *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, Association for Computational Linguistics, Avignon, France, 2012, pp. 129–138. https://aclanthology.org/W12-0117.

[45] F.J. Och and H. Ney, A Systematic Comparison of Various Statistical Alignment Models, *Computational linguistics* **29** (1) (2003), 19–51. doi:10.1162/089120103321337421.

[46] 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)*, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 228–234. doi:10.18653/v1/P18-2037.

[47] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, Distributed Representations of Words and Phrases and Their Compositionality, in: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, Curran Associates Inc., Red Hook, NY, USA, 2013, pp. 3111–3119.

[48] 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)*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 277–281. doi:10.18653/v1/W19-5438.

[49] P. Christen and K. Goiser, *Quality and Complexity Measures for Data Linkage and Deduplication*, in: *Quality Measures in Data Mining*, F.J. Guillet and H.J. Hamilton, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 127–151. ISBN 978-3-540-44918-8. doi:10.1007/978-3-540-44918-8_6.

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

[51] 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. doi:10.1075/ijcl.22.2.05dup.

[52] C. Chiarcos and M. Ionov, Linking Discourse Marker Inventories, in: *3rd Conference on Language, Data and Knowledge (LDK 2021)*, D. Gromann, G. Sérasset, T. Declerck, J.P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo and B. Heinisch, eds, Open Access Series in Informatics (OASIcs), Vol. 93, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2021, pp. 40:1–40:15. ISSN 2190-6807. ISBN 978-3-95977-199-3. doi:10.4230/OASIcs.LDK.2021.40.