TED-MDB Lexicons: Tr-EnConnLex, Pt-EnConnLex

Murathan Kurfalı[†*], **Sibel Özer**[‡*], **Deniz Zeyrek**[‡*], **Amália Mendes**[§]
[†]Linguistics Department, Stockholm University, Stockholm, Sweden
[‡]Graduate School of Informatics, Middle East Technical University, Ankara, Turkey
[§]Center of Linguistics, University of Lisbon, Lisbon, Portugal
murathan.kurfali@ling.su.se e159606,dezeyrek@metu.edu.tr
amaliamendes@letras.ulisboa.pt

Abstract

In this work, we present two new bilingual discourse connective lexicons, namely, for Turkish-English and European Portuguese-English created automatically using the existing discourse relation-aligned TED-MDB corpus. In their current form, the Pt-En lexicon includes 95 entries, whereas the Tr-En lexicon contains 133 entries. The lexicons constitute the first step of a larger project of developing a multilingual discourse connective lexicon.

1 Introduction

During the past decade or so the interest in discourse studies have dramatically increased following the release of the PDTB 2.0 corpus (Prasad et al., 2008) and, later, with the TextLink initiative[1]. In parallel to this interest, available resources annotated for various discourse-level phenomena have expanded, where discourse relational devices (DRDs) have received a special interest leading to the Connective-Lex database (Stede et al., 2019). ConnLex is a joint online database project, which is the first attempt to bring together connective lexicons of different languages. It currently hosts the connective lexicons of nine different languages providing a web-based interface together with a cross-linguistically applicable XML schema. The entries in the lexicons provide fundamental information about discourse connectives, such as orthography, syntactic category, and their senses. The ConnLex project pursues the aim of expanding the database both in coverage (by adding new languages) and depth of the information. However, except for a few resources, most of the previous effort on devising discourse connective lexicons has relied on monolingual resources and any multilingual links that were provided have not gone beyond offering English equivalents. Few exceptions involve the bilingual Italian–German contrastive/concessive connective lexicon based on the cross-lingual projection of monolingual lexicons for Italian and German (Bourgonje et al., 2017), and the very recent GeCzLex, Anaphoric Connective Lexicon for Czech and German (Poláková et al., 2020).

The main contributions of the present study are (1) proposing an alternative way of producing bilingual lexicons, potentially applicable to building multilingual lexicons, (2) providing new bilingual discourse connective lexicons for European Portuguese-English and Turkish-English by (3) considering not only the explicit discourse relations but also the implicit relations in a recent multilingual discourse bank, namely TED-Multilingual Discourse Bank (TED-MDB) annotated in the PDTB style (Zeyrek et al., 2019). The lexicon entries are extracted from TED-MDB, where each relation in the source language is aligned to its semantic equivalent in the target languages (Turkish and European Portuguese) (see §2.1). In their current form, the Pt-En lexicon includes 95 entries covering 51 connectives in Portuguese and 57 connectives in English, while the Tr-En lexicon contains 133 entries with 72 connectives in Turkish and 56 in English.

The rest of the study is structured as follows: We firstly summarize the main data source, TED-MDB followed by the discourse relation alignment procedure (§2), the output of which is used as inputs to construct bilingual lexicons. §3 describes the construction of the bilingual lexicons in detail. In §4, we discuss issues concerning our lexicon construction procedure. §5 concludes the paper presenting some future directions.

2 TED-MDB

TED-MDB is a resource of TED talk transcripts comprising 7 languages manually annotated for discourse relations. It includes English, the source language (SL) along with transcribed texts in Ger-

[*]Authors contributed equally.
[1]http://textlink.ii.metu.edu.tr/

Proceedings of the First Workshop on Computational Approaches to Discourse, pages 148–153
Online, November 20, 2020.
https://doi.org/10.18653/v1/P17

man, Lithuanian, European Portuguese, Russian, Turkish and Polish (target languages, or TLs). [2] Following the rules and principles of the PDTB, it annotates five discourse relations types (henceforth, DRs) with respect to the PDTB-3 sense hierarchy (Webber et al., 2016) and ultimately aims to provide a clearly described level of discourse structure and semantics in multiple languages, thus engendering discourse parsing studies in multiple languages. TED-MDB currently involves 6 TED talk transcripts annotated with 5 DR types (Explicit, Implicit, AltLex, EntRel, NoRel), their senses and binary arguments, amounting to a total of 3649 tokens. The annotations have been carried out by native speaker annotators of the languages involved using the PDTB annotation tool. (Lee et al., 2016)[3] This tool stores the DR annotations in separate pipe-delimited files.

2.1 Alignment Procedure

To create TED-MDB, each monolingual team annotated the texts independently of the original texts to avoid the risk of the original language influencing the annotations. Yet, due to cross-lingual variation in rendering DRs, this design criterion led to tokens not existing in the original language (Zeyrek et al., 2019). As the extraction of bilingual DC lexicons requires aligned relations, in the present study, our pipeline starts with the alignment of DRs following Özer and Zeyrek (2019). Firstly, the DR annotations originally kept in pipe-delimited files were transferred onto the base text files of both TLs generating an ID for each. Then, word- and punctuation-tokenization as well as sentence alignment procedures were performed, followed by manual corrections of the latter. For DR alignment, all DRs in each bi-text unit were paired constructing DR matrices. The text pieces constituting discourse relations were translated into the SL using the Google Translate API and stop words were removed. Next, semantic similarity, taken in terms of cosine distance, was calculated between the source and target text segments using Word2Vec (Mikolov et al., 2013) within the range of 0 ("no similarity") to 1 ("perfect similarity"). DR pairs with a similarity over 0.7 were further evaluated for alignment.

For DR pairs with acceptable scores, the similarity of the DR sense and type was evaluated using a ranking algorithm which depends on the sense tags on the DRs. A score that reflected the SL-TL match was added to the semantic similarity score, where the DR type and the

SL sense were both considered. The DR pair with the maximum score was marked as an aligned pair, and the same procedure was repeatedly applied until no DR pair was left in the matrices.

All the aligned pairs were manually checked by the authors.

The alignment algorithm has an F-score of 0.78 for Turkish-English and 0.81 for European Portuguese-English distributed over six documents accepting English annotations as the gold standard.

3 TED-MDB Lexicons

As shown in Poláková et al. (2020) and Bourgonje et al. (2017), preparing a bilingual lexicon of discourse connectives is not a straightforward task requiring a variety of resources to compute a translation candidate table including monolingual DC lexicons of the TLs and a large parallel corpus (with at least 2M parallel sentences). A monolingual discourse connective lexicon exists for Portuguese (Mendes and Lejeune, 2016) and one is being developed for Turkish (Zeyrek and Başıbüyük, 2019) but parallel corpora of the required size are absent for the language pairs under investigation. Thus, the current study is built on the observation that just as monolingual lexicons can be compiled from annotated resources, bilingual dictionaries of discourse connectives can be constructed from a similar though low scaled parallel corpus such as TED-MDB. This corpus includes 375 bi-sentence units for English-Turkish and 364 for English-European Portuguese. The rest of the section describes the method employed to create two such bilingual DC lexicons of English-Turkish and English-European Portuguese.

3.1 Populating lexicon entries automatically

Given the availability of TED-MDB, we propose an alternative way of building bilingual DC lexicons, which can be seen as the multilingual extension of extracting DC lexicons from annotated resources as in Mendes and del Río (2018); Das et al. (2018).

The method accepts a set of aligned DRs as input. For pre-processing, we firstly filter out all aligned pairs which contain a non-Explicit or non-Implicit relation followed by the removal of the pairs which are not annotated with exactly the same sense. This step helps us to eliminate the translation-based

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

[3] https://www.cis.upenn.edu/ pdtb/annotator.html

Language	Explicit	Implicit	AltLex	EntRel	NoRel	Total
English	290 (44%)	198 (30%)	46 (7%)	78 (12%)	49 (7%)	661
Russian	237 (42%)	221 (39%)	20 (4%)	57 (10%)	30 (5%)	565
Polish	218 (37,5%)	195 (33,5%)	11 (2%)	104 (18%)	52 (9%)	580
Portuguese	269 (43%)	256 (41%)	29 (5%)	38 (6%)	33 (5%)	625
German	240 (43%)	214 (38%)	17 (3%)	59 (11%)	30 (5%)	560
Turkish	276 (42%)	202 (30,5%)	59 (9%)	70 (10,5%)	51 (8%)	658
Total	1530	1286	182	406	245	3649

Table 1: Distribution of discourse relation types in TED-MDB (Zeyrek et al., 2019)

noise in the corpus as it is not uncommon for the senses of DRs to be lost or modified during translation.

After the pre-processing step, the bilingual lexicons are constructed in the following way:

- For each connective in the SL, the list of senses in the input is computed.
- The translation equivalents of the given connective are found in the TL using the aligned DRs. The translations are grouped under the senses found in the first step. Hence, we create different entries for each sense conveyed by the connective in SL. For example, in Tr-En, the "but/ama" pair appears both under the Comparison:Concession:Arg2-as-denier sense and the Comparison:Contrast sense.

Due to the limited number of explicit DRs in TED-MDB (Table 1), we also include in our lexicon *implicit connectives* which are the connectives inserted to implicit DRs by the annotators (Prasad et al., 2008). An inserted *implicit connective* can be regarded as the most suitable overt marker for a given implicit relation; hence, the pair of implicit connectives extracted from an aligned implicit DR is as valid an entry for our lexicon just as a pair of explicit connectives extracted from an aligned explicit DR. However, in order to keep things separated and facilitate further research, we create different entries for explicit and implicit usages of connectives in our lexicon. The detailed statistics about the lexicons are provided in Table 2.

3.2 Post-process

The inspection of the automatically extracted connective pairs reveals several issues, which can mostly be attributed to translation strategies. In certain cases, translators use a completely different linguistic construction in the TL; yet, they manage to preserve the sense of the SL text (Example 1). Since both relations are annotated with the same sense, our method erroneously assumes these different connectives form a valid pair.

(1) by **investing sustainably,** *we're doing two things ..*
Quando **investimos na sustentabilidade** *estamos a fazer duas coisas*

'When **we invest in sustainability**, *we are doing two things..*'

In order to fix such cases, we firstly adopted a fully automatic approach where we tried to eliminate the unacceptable pairs by checking them against comprehensive bilingual dictionaries similar to Poláková et al. (2020). To this end, we used Treq (Škrabal and Vavřín, 2017) and the OPUS word alignment database.[4] However, both resources turned out to be unsuitable for our purposes. The translation candidate tables created from these resources eliminate a nontrivial amount of acceptable pairs as most of the time, valid translations are either absent in the databases or are assigned a very low probability, making it virtually impossible to determine an appropriate threshold between unacceptable and acceptable translations. That some of the Turkish connectives are suffixal connectives further render the use of dictionaries impractical. Therefore, we manually went through each entry in the lexicons in order to reach gold pairs. As the lexicons are not large and the task of deciding whether two words are translation equivalents is not too challenging, the manual control was completed within hours. The decision was made unanimously, which resulted in the removal of 9 pairs from Portuguese and only 2 from Turkish. It is also worth noting that the eliminated pairs overwhelmingly had the label Expansion:Level-of-detail:Arg2-as-detail, which "is used when Arg2 describes in more detail, the situation in Arg1"(Webber et al., 2019).

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

Language	# of Connectives				# of Sense	# of Translations		
	Exp	Imp	Total (Unique)	Monolingual		Min	Max	Avg
English	26	31	57 (48)	142	1.23	1	6	1.36
Portuguese	26	25	51 (42)	-	1.49	1	4	1.36
English	24	32	56 (47)	142	1.29	1	7	1.83
Turkish	34	38	72 (62)	226	1.44	1	4	1.26

Table 2: Statistics regarding the constructed lexicons. "Exp" and "Imp" refers to the number of Explicit and Implicit connectives, respectively. The "Total" column represents the number of connectives when implicit and explicit connectives are counted as separate entries and when their type is disregarded (within parenthesis). The "Monolingual" column represents the number of connectives in the the respective language's monolingual lexicon (retrieved from (Stede et al., 2019)) . The last column presents the minimum, maximum and the average number of translation equivalents in the target language.

A close examination showed that this subsense was not conveyed by the annotated DC tokens in the SL but rather inferred from the arguments, leading the translator to render the DR almost freely with a mismatching token in the TL. The removed pairs are as follows:

- **Pt-En:** e - *rather*, e - *for that matter*, enquanto - *and*, assim - *that is*, de facto - *specifically*, e - *as well as*, e - *lastly*, isto é - *clearly*, assim - *specifically*
- **Tr-En:** özetle - *clearly*, yani - *clearly*, işte - *clearly*

3.3 The Structure of the Lexicons

Each entry in the TED-MDB lexicons corresponds to a specific connective in the TL and a list of its possible translations in the TL grouped under the sense the connective conveys. Specifically, an entry consists of the following components (illustrated in Figure 1):

- **Connective**: The head of each entry is a DC represented in its lemmatized form.
- **Dimlex link**: Each DC and its translations are accompanied with an URL to their respective connective-lex entry,[5] which serves as a bridge between the bilingual and monolingual lexicons.
- **Sense list:** The list of the senses that the head connective conveys in TED-MDB is displayed in the main screen of the interface sorted by the corpus frequencies of the senses.
- **List of translation candidates**: For each sense in the list, the translation candidates specified in TL texts are provided. The translation candidates also have their own entries and are accessible just by clicking.
- **Example sentence**: Each connective pair is accompanied with a randomly selected sentence pair from TED-MDB.

4 Discussion

To the best of our knowledge, the TED-MDB lexicons presented here constitute the first attempt to construct a bilingual connective lexicon directly from an annotated parallel resource. Compilation of bilingual lexicons in this way has a number of practical benefits, where the main advantage is being not dependent on external resources. It alleviates the need for parallel corpora required to extract the translation candidates to map the connectives in different languages onto each other and does not necessitate monolingual DC lexicons, a challenging and time consuming effort especially when started from scratch (Roze et al., 2012). Also, as all entries are populated from an annotated corpus, the lexicons are guaranteed to be symmetrical, and the bilingual examples provide an opportunity to observe the usage of connectives in context in two languages. It must also be noted that despite being compiled from a set of merely 300+ relations in each language set, our bilingual lexicons roughly account for 30% of the documented connectives of these languages; hence, their coverage is more impressive than it looks (Table 2).

As explained in Section 3.2, there are certain cases where the connectives from an aligned DR pair do not form valid lexicon entries. This issue revealed the larger problem that translational candidate tables, even those from a large parallel corpus like InterCorp (Škrabal and Vavřín, 2017) cannot

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

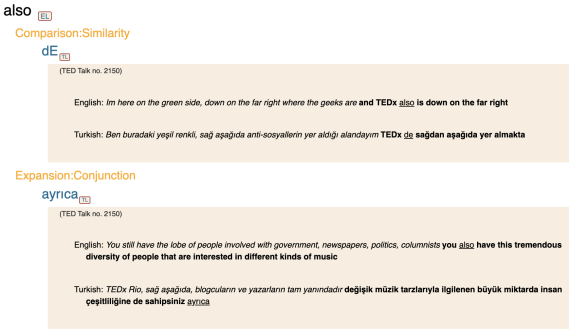

Figure 1: The entry for 'also'in Tr-En lexicon

adequately capture the translation equivalents of connectives. We believe this finding further highlights the need for such bilingual lexicons.

Finally, as the manual control of the DR alignments constitutes one of the two non-automatic steps of our pipeline, we investigated its effect on the final lexicons to guide future research. To our surprise, the automatic alignment procedure turns out to be more than satisfactory: we were able to fetch more than 96% of all entries in the gold lexicons, suggesting that even a multilingual lexicon involving all languages in TED-MDB can be automatically constructed. This is left for future work.

5 Conclusion

In translation, the choice of a DC that best conveys the sense of a relation and renders the relation in a natural way is not a trivial task. At a minimum, it requires careful consideration of the multiple senses of the connectives and their parts-of-speech. Even a bilingual dictionary is not always helpful for a translator. Bilingual lexicons built on the basis of naturalistic data is important to aid both machine and human translation as well as second language learners. In this study, we described a method of building two bilingual lexicons using aligned DR annotations. Both lexicons are available online as HTML web pages. [6] In contrast to previous bilingual lexicon studies, we did not use monolingual connective lexicons or dictionaries as the former was absent (at least for Turkish), and the latter caused loss of useful data. Although the alignment and the lexicon extraction procedures have been applied to two languages so far, this approach has the potential to be extended to other language pairs covered in the TED-MDB corpus, and this is what we plan to do as a future study.

[6] http://metu-db.info/mdb/ted/resources.jsf

References

Peter Bourgonje, Yulia Grishina, and Manfred Stede. 2017. 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*, pages 53–58.

Debopam Das, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede. 2018. Constructing a lexicon of english discourse connectives. In *Proceedings of the 19th Annual SIGDIAL Meeting on Discourse and Dialogue*, pages 360–365.

Alan Lee, Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2016. Annotating discourse relations with the pdtb annotator. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 121–125.

Amália Mendes and Pierre Lejeune. 2016. Ldm-pt-a portuguese lexicon of discourse markers. In *Conference Handbook of TextLink–Structuring Discourse in Multilingual Europe Second Action Conference*, pages 89–92. Debrecen University Press.

Amália Mendes and Iria del Río. 2018. Using a discourse bank and a lexicon for the automatic identification of discourse connectives. In *International Conference on Computational Processing of the Portuguese Language*, pages 211–221. Springer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Sibel Özer and Deniz Zeyrek. 2019. An automatic discourse relation alignment experiment on TED-MDB. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 31–34, Florence, Italy. Association for Computational Linguistics.

Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, and Jiří Mírovskỳ. 2020. Geczlex: Lexicon of czech and german anaphoric connectives. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1089–1096.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. Lexconn: a french lexicon of discourse connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (10).

Michal Škrabal and Martin Vavřín. 2017. The translation equivalents database (treq) as a lexicographer's aid. In *Electronic lexicography in the 21st century.*

Proceedings of eLex 2017 conference. Leiden: Lexical Computing.

Manfred Stede, Tatjana Scheffler, and Amália Mendes. 2019. 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*, (24).

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A discourse-annotated corpus of conjoined vps. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 22–31.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual.

Deniz Zeyrek and Kezban Başıbüyük. 2019. Tcl-a lexicon of turkish discourse connectives. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 73–81.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. Ted multilingual discourse bank (tedmdb): a parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–27.