Multilingual Extension of PDTB-Style Annotation: The Case of TED Multilingual Discourse Bank

Deniz Zeyrek[1], **Amália Mendes**[2], **Murathan Kurfalı**[1]
[1]Informatics Institute, Middle East Technical University, Ankara,
[2]Center of Linguistics, University of Lisbon, Lisbon
dezeyrek@metu.edu.tr, amaliamendes@letras.ulisboa.pt, kurfali@metu.edu.tr

Abstract

We introduce TED-Multilingual Discourse Bank, a corpus of TED talks transcripts in 6 languages (English, German, Polish, European Portuguese, Russian and Turkish), where the ultimate aim is to provide a clearly described level of discourse structure and semantics in multiple languages. The corpus is manually annotated following the goals and principles of PDTB, involving explicit and implicit discourse connectives, entity relations, alternative lexicalizations and no relations. In the corpus, we also aim to capture the characteristics of spoken language that exist in the transcripts and adapt the PDTB scheme according to our aims; for example, we introduce hypophora. We spot other aspects of spoken discourse such as the discourse marker use of connectives to keep them distinct from their discourse connective use. TED-MDB is, to the best of our knowledge, one of the few multilingual discourse treebanks and is hoped to be a source of parallel data for contrastive linguistic analysis as well as language technology applications. We describe the corpus, the annotation procedure and provide preliminary corpus statistics.

Keywords: discourse, parallel, multilingual corpus

1. Introduction

Over the last decade, there has been a growing interest in parallel corpora for linguistic research or language technology applications, such as Cettolo et al. (2012), Tiedemann and Nygaard (2004). While most of the existing resources are annotated at the word or syntactic level, as in Bentivogli and Pianta (2005), Haug et al. (2009), corpora enriched with discourse-level annotations are scarce but they exist. For example, Stede et al. (2016) introduce a corpus of parallel argumentative texts (German-English) annotated with respect to RST and SDRT; Popescu-Belis et al. (2012) describe pronoun and connective annotation over Europarl for English and French, and Samy and González-Ledesma (2008) report the development of an annotated corpus of UN documents in English, Spanish and Arabic by adopting a pragmatic perspective. Given the scarcity of discourse-annotated parallel corpora, we believe the community would benefit from new resources which involve various languages. Here we describe TED-Multilingual Discourse Bank (TED-MDB), a corpus of TED talks transcripts involving European languages as well as a non-European language, Turkish, annotated at the discourse level following the PDTB approach (Prasad et al., 2014). Our effort is based on years of discourse research and the principles that are being established in the science of annotation (Hovy and Lavid, 2010; Ide and Pustejovsky, 2017).[1]

1.1. Discourse Relations and the Role of Discourse Connectives in a Text

Discourse is a unit above the sentence level mainly structured in terms of lexical links, anaphoric relations and discourse relations. In this paper, our focus is on discourse relations (DRs), i.e. informational relations such as contrast, elaboration, causal, temporal, etc. that hold between two discourse units. DRs are low-level relations indicating discourse structure and they may be made salient by devices referred to as discourse connectives.[2] Discourse connectives are typically coordinating or subordinating conjunctions (*and, but, because*), prepositional phrases (*in sum*) or adverbs (*similarly*). Human readers easily infer a DR while reading a piece of text, for example, adjacent clauses or sentences often trigger a relation even when a discourse connective is absent.

DRs may be realized inter-sententially, as in example 1 or intra-sententially, as in 2. Examples 1 and 2 are referred to as explicit DRs due to the presence of an explicit connective relating two text spans consisting of clauses or clause complexes. These text spans are referred to as the arguments of a discourse connective (DC). In the examples throughout the paper, explicit DCs are underlined, implicit DCs are shown in parentheses. The arguments are rendered differentially, using italic fonts for the first argument and bold fonts for the second argument. The sense of the relation is shown in square parentheses where relevant. All the examples are from TED-MDB.

1. *About 80 percent of global CEOs see sustainability as the root to growth in innovation ..* <u>But</u> **93 percent see ESG as the future, or as important to the future of their business**.[Comparison:Contrast]
2. *I think it's reckless to ignore these things*, <u>because</u> **doing so can jeopardize future long-term return**. [Contingency:Cause:reason]

Example 3 below shows a text where a DC is lacking; these have been referred to as implicit relations. The inferred DR can be made explicit by inserting a discourse connective;

[1] The researchers and annotators involved in the creation of TED-MDB are members of the consortium initiated by the COST project Textlink (http://textlink.ii.metu.edu.tr). For information regarding the corpus, see http://www.textlink.ii.metu.edu.tr/ted-multilingual-discourse-bank.

[2] Low-level discourse structures differ from high-level discourse structures such as genre and topic Hobbs (1985).

thus, example 3 can be expressed with the discourse adverb *specifically* suggesting that the second clause specifies the meaning of the first clause. Here, it is both the adjacency of the clauses and their lexical content that guide a human reader to insert an appropriate adverb that would make the DR salient.

3. *We have a population that's both growing and aging*; (specifically) **we have seven billion souls today heading to 10 billion at the end of the century**. [Expansion: Level-of-detail-Arg2-as-detail]

1.2. PDTB Principles and the Goals of TED-MDB

Among the currently available frameworks for the investigation and annotation of discourse such as RST (Mann and Thompson, 1988) and SDRT (Asher and Lascarides, 2003), we have settled on PDTB. Our choice is motivated by several factors. For example, unlike RST and SDRT, PDTB aims to reveal discourse coherence and discourse structure at the local level and to the extent the relations are made explicit by DCs (Prasad et al., 2014). We believe the local, lexically based approach provides an easy starting point for the annotation process. Secondly, PDTB has a theory-natural approach, which would be appealing to linguists and NLP researchers working in different theoretical frameworks. Finally, the PDTB annotation guidelines have produced reliable results in corpora of written discourse developed for a range of languages, such as Hindi (Kolachina et al., 2012); Chinese (Zhou and Xue, 2015); Arabic (Al-Saif and Markert, 2010) and Turkish (Zeyrek et al., 2013). The PDTB principles have also been implemented in the creation of a corpus of conversational dialogues in Italian (Tonelli et al., 2010).

In the PDTB approach, DCs are discourse-level predicates with binary arguments. The arguments of a discourse connective (both explicit and implicit) are referred to as Arg1, Arg2 and always have an abstract object interpretation (Asher, 2012), i.e. in order for text spans to be selected as arguments to a DC, they need to be understood as eventualities, facts, propositions, etc. Arg2 is the text span that syntactically hosts the connective while Arg1 is conveniently the other argument.[3]

PDTB annotates texts for 5 major DR types to reveal relations between adjacent text spans: explicit relations, implicit relations, alternative lexicalizations (AltLex), entity relations (EntRel) and no relations (NoRel) (PDTB Group, 2008). By definition, explicit DCs are overtly expressed in texts and are easy to recognize (Pitler et al., 2008), as in examples 1, 2 above. By contrast, implicit relations lack an overt connective, where the relation between discourse units is inferred and shown by a potential explicit connective, as in example 3. AltLexs are any alternative means of lexicalizing a relation and can vary from fixed expressions such as *this is why* (example 4) to free expressions such as *that's the equivalent of* (example 5) (Prasad et al., 2010).

4. *.. long-term value creation requires the effective management of three forms of capital: financial, human, and physical.* This is why **we are concerned with ESG**. [AltLex: Contingency:Cause+Belief: Result+Belief]

5. *.. they yield savings of 23 million dollars in operating costs annually, and avoid the emissions of a 100,000 metric tons of carbon.* That's the equivalent of **taking 21,000 cars off the road** .. [Expansion:Equivalence]

EntRels are relations that hold between two entities rather than eventualities, as in example 6. Finally, NoRels are taken as those cases between adjacent sentences where no discernible DR exists (example 7).

6. *.. CalPERS is another example.* **CalPERS is the pension fund for public employees in California**. [EntRel]

7. Now *over almost eight years, they've outperformed by about two thirds.* So **yes, this is correlation**. [NoRel]

Adopting the PDTB principles, in the TED-MDB project, we annotate 5 DR types together with their binary arguments and sense (or senses), where relevant.[4] In all cases, PDTB's minimality principle is adopted; that is, the argument spans of a relation are annotated as minimally as possible as allowed by the sense of the relation.

While PDTB has inspired discourse annotation projects in various languages mentioned above, each team has developed their own corpora, made use of different annotation tools and used the PDTB annotation scheme in different ways. For example, Turkish Discourse Bank is a multi-genre corpus using a revised subset of the PDTB annotation categories. Arabic, Hindi and Chinese discourse banks involve news texts and again use revised subsets of the PDTB scheme. While such differences may be necessary for developing discourse-level corpora for different languages, they make cross-linguistic comparison difficult. With parallel corpora similarly annotated for discourse phenomena, (a) we may be able to avoid some of the differences arising from annotation efforts carried out separately for each language, (b) help discourse-level corpus annotation efforts to improve our understanding of discourse structure, and in turn (c) help to enhance language technology applications. Thus, our aim in developing TED-MDB is to investigate the phenomena surrounding discourse relations on the basis of texts belonging to the same genre by annotating them similarly across languages. The ultimate goal is to reveal a clearly defined level of discourse structure and semantics for multiple languages for linguistic and natural language processing research.

The structure of the rest of the paper is as follows. §2. introduces the corpus characteristics and its current coverage. §3. describes the annotation procedure, involving training of annotators (§3.1.), the steps in annotating the corpus

[3] Although arguments to a discourse connective can be adjacent or non-adjacent, for the sake of consistency, TED-MDB focuses on arguments that are adjacent to each other. This strategy also facilitates annotation without annotation projection, as explained in §3.2..

[4] Sense is not assigned to EntRels.

(§3.2.), an explanation of how sense annotation is carried out (3.3.), and an evaluation of the corpus covering English, Turkish and Portuguese (§3.4.). §4. describes how PDTB guidelines are extended to capture the interactive nature of TED talks reflected in the transcripts. Finally, §5. summarizes the paper and offers some future directions.

2. The Corpus

TED talks are examples of prepared (possibly scripted), formal monologues (cf. the structure of ICE, (Greenbaum, 1996)) delivered to a live audience. They are video-typed and stored in the TED website. The TED website also provides the transcripts of the talks together with the timestamps on the videos. The transcripts are prepared according to the norms of written language, e.g., they include punctuation and paragraph divisions matching the timestamps. The talks are translated to various languages by named volunteers who are required to follow the instructions provided in the TED website. The translated transcripts are checked by named experts.

We use the WIT3 website (Cettolo et al., 2012) to obtain the transcriptions in the original language, English and their existing translations in the target languages we focus on. We initially identified the common transcripts in English and two target languages (Turkish and European Portuguese). Out of this set of transcripts, we selected the texts for our corpus by reading them to make sure they had the expected level of translation quality and that they were easy to read and comprehend. Ambiguity was tolerated unless it hampered the natural flow of discourse or changed the meaning of the original text. We also ensured that the topics were varied and the texts that mostly relied on images and videos were not selected. Subsequently, the set of parallel texts was expanded to three more languages (German, Russian, and Polish). Table 1 shows the set of transcripts selected and annotated in 6 languages. This is the current coverage of the corpus.

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

Table 1: Annotated TED talks transcripts

3. Annotation Procedure

In this section we describe the annotation procedure we followed in creating the TED-MDB corpus.

3.1. Training the annotators

The teams involved in developing TED-MDB work with a primary annotator (often the leading researcher of the team) and an experienced secondary annotator or a researcher in discourse. The annotators and researchers are native speakers of the languages we focus on. For annotators already well-trained in discourse, a tutorial on PDTB guidelines explaining major issues such as the position and the span of the arguments, implicit relations and the method of sense assignment is often sufficient. Annotators who are less experienced in discourse-level annotation are trained differently, starting with PDTB annotation guidelines followed by a pre-annotation phase on .doc files. Then, they are asked to create sample annotations on TED talks transcripts proceeding sense by sense, annotating one top-level sense (and its lower levels) at a time. They discuss the annotations with the researcher of the team and proceed to the new top-level sense. The circle is completed when sample annotations are created for all the top-level senses. [5]

3.2. Annotating the Corpus

Unlike most annotated parallel corpora, TED-MDB does not start with annotation projection, where the annotation on one of the languages seeds the annotation on another language (Yarowsky et al., 2001; Ambati and Chen, 2007). Instead, trained or experienced annotators go through each transcribed text sentence by sentence independently and individually. Independent sentence-by-sentence annotation procedure ensures that each team annotates pairs of sentences incrementally and assigns a sense to each relation independent of the others in the team. In the usual annotation projection procedure, there is a risk for a specific language to influence the annotations of other languages. In our procedure, such potential influence is avoided.

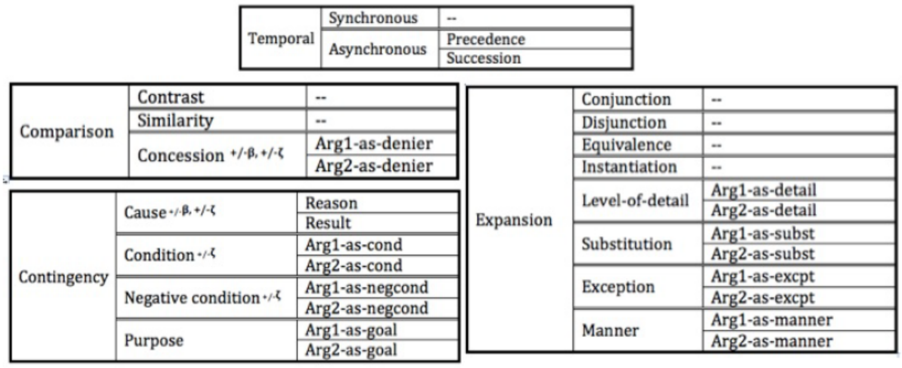
Once each team has annotated a text, the annotations are discussed in multilingual group meetings where all TED-MDB members are present. The aim in the discussions is to secure annotation consistency rather than to convince others on a specific type of annotation. The annotations are created using the PDTB annotation tool, which allows selecting discontinuous text spans (Lee et al., 2016) (see Figure 2.

3.3. Sense Annotation

For sense annotation, we use PDTB 3.0 relation hierarchy, which has 4 top-level senses (Expansion, Temporal, Contingency, Contrast) and their second- or in some cases third-level senses, as shown in Figure 1 (Webber et al., 2016).

PDTB 3.0 relation hierarchy is an enriched version capturing a larger number of cases missing in the PDTB 2.0 sense hierarchy. It also simplifies Level-3 senses either by moving them to Level-2 or eliminating them since they were

[5] English, Turkish, Portuguese and Polish annotations are provided by the primary annotator and checked by the researcher of the team. German and Russian annotations are provided by a primary annotator who is also a discourse researcher.



Temporal	Synchronous	--
	Asynchronous	Precedence
		Succession

Comparison	Contrast	--
	Similarity	--
	Concession $^{+/-\beta, +/-\zeta}$	Arg1-as-denier
		Arg2-as-denier

Contingency	Cause $^{+/-\beta, +/-\zeta}$	Reason
		Result
	Condition $^{+/-\zeta}$	Arg1-as-cond
		Arg2-as-cond
	Negative condition $^{+/-\zeta}$	Arg1-as-negcond
		Arg2-as-negcond
	Purpose	Arg1-as-goal
		Arg2-as-goal

Expansion	Conjunction	--
	Disjunction	--
	Equivalence	--
	Instantiation	--
	Level-of-detail	Arg1-as-detail
		Arg2-as-detail
	Substitution	Arg1-as-subst
		Arg2-as-subst
	Exception	Arg1-as-excpt
		Arg2-as-excpt
	Manner	Arg1-as-manner
		Arg2-as-manner

Figure 1: PDTB 3.0 Relation hierarchy. The superscripts indicate the implicit belief (-/+β) or speech-act (-/+ζ) features.

rare and difficult to find. With these improvements and the additional senses, PDTB 3.0 relation hierarchy has been an attractive option for the needs of a multilingual corpus; thus we chose to use it.

Following the PDTB guidelines, we also assign multiple senses for a DC (explicit or implicit). Example 8 provides an explicit DC token with multiple senses. An explicit DC may even trigger the inference of an implicit DC as in example 9 (Rohde et al., 2016). Such cases are annotated as two tokens, the first with the explicit connective (*and*), the second with an implicit adverbial connective (*consequently*). Each corresponding sense is also annotated.

8. .. when **they .. decide whether to invest**, *they look at financial data, metrics like sales growth, cash flow, market share, valuation ...* [Temporal:Synchronous], [Contingency:Condition:Arg2-as-cond]

9. .. *they are really complex* and (consequently) **they can seem really far off**.. [Expansion:Conjunction], [Contingency:Cause:result]

3.4. Intra-Annotator Agreement

To evaluate annotation stability, we measured intra-annotator agreement for three languages for now – English, Portuguese, Turkish. Each primary annotator re-created annotations for 20-23% of the total number of annotated tokens in the corpus. To avoid recall bias, we gave 8-10 months between the first annotations and the re-annotations. We measured intra-annotator agreement both on the discourse relation type, i.e. whether the annotator was consistent in annotating the type of a specific relation, and on the top-level sense, i.e. whether the annotator was consistent in annotating the sense of a specific relation at the top level. We regard the original annotations as gold-standard and calculate precision, recall and f-score following the equations given below. In equation (1), the denominator is the sum of reannotations. In equation (2), the denominator is the sum of gold standard annotations. In both equations, the numerator is the number of reannotated tokens.

$$Precision = \frac{\#\ of\ correct\ found\ annotations}{Total\ \#\ of\ found\ annotations} \quad (1)$$

$$Recall = \frac{\#\ of\ correct\ found\ annotations}{\#of\ correct\ expected\ annotations} \quad (2)$$

The intra-annotator agreement results are presented in Tables 2 and 3.

Language	Precision	Recall	F-Score
English	91.92%	91.92%	0.92%
Portuguese	75.97%	75.97%	0.76%
Turkish	71.8%	70.06%	0.71%

Table 2: Intra-annotator agreement results for discourse relation type in three languages

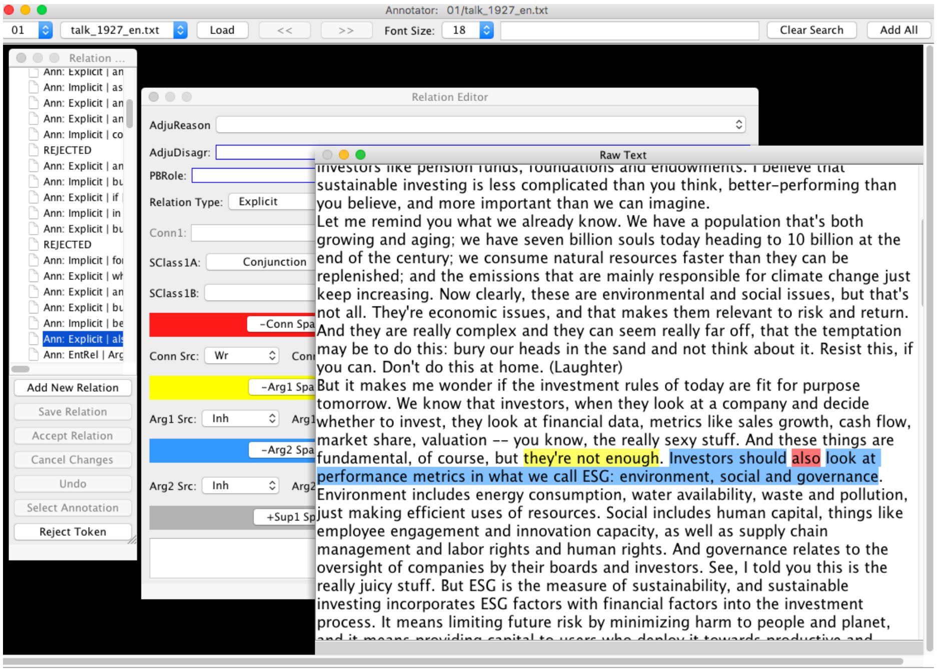
Language	Precision	Recall	F-Score
English	91.73%	93.27%	0.92%
Portuguese	73.28%	75.88%	0.74%
Turkish	72.6%	70.8%	0.71%

Table 3: Intra-annotator agreement results for top-level senses in three languages

We obtained high agreement results for English both for the discourse relation type and the top-level senses ($\geq$ 0.9). For Turkish and Portuguese, the coders achieved $\geq$ 0.7 in each case, which is acceptable for coherence phenomena (Spooren and Degand, 2010). This score is particularly satisfactory for our task, which presents the added difficulty of resolving ambiguities that arise due to the nature of translation. These results suggest that our annotation guidelines can be used consistently by the annotators.

4. Extending the PDTB Scheme

In the TED talks transcripts we looked at, we find that certain features of spoken discourse are maintained quite faithfully. For example, we observe that the connectives sometimes fulfill roles other than linking two text spans semantically. In particular, the connectives *but, so* exhibit a range of functions that could be defined as sequential linkage (speech management, topic structure, etc.) (Fischer, 2006; Redeker, 2000; Crible and Zufferey, 2015). We will



Relation Type: Explicit

SClass1A: Conjunction

Conn Src: Wr

Arg1 Src: Inh

Arg2 Src: Inh

investors like pension funds, foundations and endowments. I believe that sustainable investing is less complicated than you think, better-performing than you believe, and more important than we can imagine.
Let me remind you what we already know. We have a population that's both growing and aging; we have seven billion souls today heading to 10 billion at the end of the century; we consume natural resources faster than they can be replenished; and the emissions that are mainly responsible for climate change just keep increasing. Now clearly, these are environmental and social issues, but that's not all. They're economic issues, and that makes them relevant to risk and return.
And they are really complex and they can seem really far off, that the temptation may be to do this: bury our heads in the sand and not think about it. Resist this, if you can. Don't do this at home. (Laughter)
But it makes me wonder if the investment rules of today are fit for purpose tomorrow. We know that investors, when they look at a company and decide whether to invest, they look at financial data, metrics like sales growth, cash flow, market share, valuation -- you know, the really sexy stuff. And these things are fundamental, of course, but they're not enough. Investors should also look at performance metrics in what we call ESG: environment, social and governance.
Environment includes energy consumption, water availability, waste and pollution, just making efficient uses of resources. Social includes human capital, things like employee engagement and innovation capacity, as well as supply chain management and labor rights and human rights. And governance relates to the oversight of companies by their boards and investors. See, I told you this is the really juicy stuff. But ESG is the measure of sustainability, and sustainable investing incorporates ESG factors with financial factors into the investment process. It means limiting future risk by minimizing harm to people and planet, and it means providing capital to users who deploy it towards productive and

Figure 2: An explicit relation in the annotation environment

refer to these as discourse markers (Schiffrin, 1988). In addition, discourse particles (English *well*, Portuguese *bem*) that underlie the interactive nature of TED talks are also retained in the transcripts. Thus, TED talks transcripts represent a modality where certain features of spoken discourse is shared with written discourse. Our aim is to take this into consideration and annotate the properties of spoken discourse when they occur in the transcripts.

For example, we find that the transcripts involve hypophora, where the speaker asks a question and immediately provides a response himself. We have decided to add Hypophora to the annotation scheme as a feature impelled by the interactive nature of TED talks. We annotate Hypophora as a case of AltLex, where the question word is taken as the anchor, as in example 10 and its translated versions in Portuguese (11), Turkish (12) and Polish (13).

10. Why **is that hard**? Well *to see, let's imagine we take the Hubble Space Telescope and we turn it around ... We'll see something like that, a slightly blurry picture of the Earth.* .. [AltLex: Hypophora] [En]

11. Porque **é tão difícil**? Bem, *imaginemos que pegamos no Telescópio Espacial Hubble e o rodamos e o deslocamos.. Veríamos algo assim, Uma imagem algo difusa de a Terra.* [AltLex : Hypophora] [Por]

12. Neden **bu kadar zor**? *Bunu anlamak için, Hubble Uzay Teleskobu'nu tutup döndürdüğümüzü.. varsayalım. .. Görebildiğimiz tek şey, şuradaki .. yıldızın parıldayan büyük görüntüsü* .. [AltLex : Hypophora] [Tu]

13. Czemu **tego nie zrobiliśmy i czemu to takie trudne**? *Wyobraźmy sobie, że bierzemy Kosmiczny Teleskop Hubble'a, .. Zobaczymy coś takiego, nieco rozmazany obraz Ziemi* ..[AltLex: Hypophora] [Pol]

Hypophora cases may be annotated with a second sense, as shown in the German and the Russian equivalents of 10:

14. Warum **ist das schwer**? *Stellen Sie sich vor, wir nähmen das Hubble Weltraumteleskop , drehten es um .. Wir würden vermutlich ein leicht unscharfes Bild der Erde sehen..* [AltLex: Hypophora], [Contingency:Cause:Result] [Ger]

15. Почему это трудно? Чтоб понять, нужно представить, что мы берём космический телескоп Хаббл и поворачиваем его и перемещаем на орбиту Марса. Мы увидим что-то такое , слегка размытое фото Земли, ... [AltLex: Hypophora], [Contingency:Cause:Result] [Ru]

In addition to this, we spot the discourse marker use of connectives that otherwise signal a DR, such as *but, so*. In such cases, we annotate the arguments in the usual way and assign the label NoRel to the relation (see *so* in example 7 above). Thus, we tease apart discourse connectives from their discourse marker use in the speeches. In case of discourse particles such as *well*, we follow the same procedure. An example annotation from Portuguese *bem* is provided in 16.

16. *Seria o mesmo se erguesse o meu polegar e bloqueasse o ponto luminoso à frente dos meus olhos, poderia vê-los na última fila*. Bem, **o que está a acontecer**? [NoRel] [Por]
It would be the same thing if I put my thumb up and blocked that spotlight that's getting right in my eye, I can see you in the back row. Well, what's going on?

Finally, PDTB annotates attribution, i.e. whether the information in the arguments of discourse relations is categorized as fact or opinion. Here, PDTB's aim is to capture "the source and degree of factuality of abstract objects"

(Prasad et al., 2006). Given our scarce resources, we leave the annotation of attribution to a further stage.

Some preliminary corpus statistics regarding the current stage of TED-MDB are provided in Tables 4 and 5 below. Table 4 presents the total number of words as well as the total number of annotated tokens for 5 DR types per language. Table 5 presents the total number of annotations on 5 top-level senses per language.

5. Conclusion and Outlook

We introduced TED-MDB, a new multilingual corpus annotated at the discourse level following the approach and annotation principles of PDTB. We described the steps in corpus development and explained how we adapt the PDTB annotation categories. We introduced a new category, Hypophora, an aspect of spoken discourse kept in the transcripts. We also described how we teased apart the discourse connectives from their discourse marker use. A more precise characterization of such features of spoken discourse that exist in TED talks transcripts is further work.

We believe that in annotation tasks, annotation pace can be compromised for annotation quality; hence, we proceed carefully in a step-wise manner at all stages of our effort. Our initial intra-annotator results are promising. In an upcoming paper, we aim to report new statistics and intra-annotator agreement results for all the languages involved (Zeyrek et al., in preparation).

We believe that TED-MDB opens up various interesting research possibilities. Studying the specific strategies that each language (or language pair) encode discourse relations will be one line of research that would lead to understanding discourse structure and semantics across languages. Given that TED-MDB involves translated texts, this line of investigation would contribute to understanding native speakers' translation preferences in structuring the discourse of TED talks. Finally, a concrete output of TED-MDB would be its contribution to the discourse connective lexicon of individual languages and their translation to other languages, which can be used in various monolingual or multilingual natural language processing tasks.

6. Acknowledgements

We thank the members of TED-MDB: Maciej Ogrodniczuk, Yulia Grishina, Sam Gibbon. Thanks to Nuno Martins (for annotating the Portuguese transcripts), Celina Heliasz, Joanna Bilińska, Daniel Ziembicki (for annotating the Polish transcripts) and to Bonnie Webber, Manfred Stede, Alan Lee, Julia Lavid for helpful comments. We acknowledge the support of ISCH COST Action IS1312 Textlink, of TÜBİTAK, The Scientific Research and Technology Council of Turkey (project ID: 2219/1059B191500896), and of FCT, Fundação para a Ciência e a Tecnologia (project ID: PEst-OE/LIN/UI0214/2013). We are grateful to three anonymous reviewers, although any errors are ours.

7. Bibliographical References

Al-Saif, A. and Markert, K. (2010). The Leeds Arabic Discourse Treebank: Annotating discourse connectives for Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation - LREC 2010*.

Ambati, V. and Chen, W. (2007). Cross lingual syntax projection for resource-poor languages. CMU.

Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Asher, N. (2012). *Reference to abstract objects in discourse*, volume 50. Springer Science & Business Media.

Bentivogli, L. and Pianta, E. (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the multisemcor corpus. *Natural Language Engineering*, 11(3):247–261.

Cettolo, M., Girardi, C., and Federico, M. (2012). Wit3: Web inventory of transcribed and translated talks. In *Proc of the 16th Conference of the European Association for Machine Translation (EAMT)*, volume 261, page 268.

Crible, L. and Zufferey, S. (2015). Using a unified taxonomy to annotate discourse markers in speech and writing. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-11)*.

Fischer, K. (2006). Towards an understanding of the spectrum of approaches to discourse particles: introduction to the volume. *Approaches to discourse particles*, pages 1–20.

Greenbaum, S. (1996). *Comparing English world wide: The International Corpus of English*. Oxford University Press.

Haug, D. T., Jøhndal, M. L., Eckhoff, H. M., Welo, E., Hertzenberg, M. J., and Müth, A. (2009). Computational and linguistic issues in designing a syntactically annotated parallel corpus of Indo-European languages. *TAL*, 50(2):17–45.

Hobbs, Jerry, R. (1985). On the coherence and structure of discourse. *Technical report, CSLI* .

Hovy, E. and Lavid, J. (2010). Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, 22(1):13–36.

Ide, N. and Pustejovsky, J. (2017). *Handbook of Linguistic Annotation*. Springer.

Kolachina, S., Prasad, R., Sharma, D. M., and Joshi, A. K. (2012). Evaluation of discourse relation annotation in the Hindi Discourse Relation Bank. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation - LREC 2012*, pages 823–828.

Lee, A., Prasad, R., Webber, B. L., and Joshi, A. K. (2016). Annotating discourse relations with the PDTB Annotator. In *COLING (Demos)*, pages 121–125.

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

Language	Word count	Explicit	Implicit	AltLex	EntRel	NoRel	Total
English	7012	290	198	46	78	49	661
Russsian	5623	237	221	20	57	30	565
Polish	6520	218	195	11	104	52	580
Portuguese	7166	269	256	29	38	33	625
German	6366	240	214	17	59	30	560
Turkish	5164	276	202	59	70	51	658

Table 4: Number of words per language and distribution of discourse relation types across the corpus

Language	Temporal	Comparison	Expansion	Contingency	Hypophora	Total
English	46	71	281	132	11	541
Russsian	30	56	270	114	12	482
Polish	44	82	183	108	8	425
Portuguese	54	71	288	143	14	570
German	31	56	259	120	9	475
Turkish	41	74	307	146	14	582

Table 5: Distribution of top-level senses across the corpus

PDTB Group. (2008). The Penn Discourse Treebank 2.0 annotation manual. Technical report, Institute for Research in Cognitive Science, University of Philadelphia.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. K. (2008). Easily identifiable discourse relations. *Technical Reports (CIS)*, page 884.

Popescu-Belis, A., Meyer, T., Liyanapathirana, J., Cartoni, B., and Zufferey, S. (2012). Discourse-level annotation over Europarl for machine translation: Connectives and pronouns. In *Proc of the Eighth International Conference on Language Resources and Evaluation (LREC)*.

Prasad, R., Dinesh, N., Lee, A., Joshi, A. K., and Webber, B. L. (2006). Attribution and its annotation in the Penn Discourse Treebank. *TAL*, 47(2):43–64.

Prasad, R., Joshi, A., and Webber, B. (2010). Realization of discourse relations by other means: Alternative lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1023–1031. Association for Computational Linguistics.

Prasad, R., Webber, B., and Joshi, A. (2014). Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics*.

Redeker, G. (2000). Coherence and structure in text and discourse. *Abduction, Belief and Context in Dialogue*, pages 233–263.

Rohde, H., Dickinson, A., Schneider, N., Clark, C. N., Louis, A., and Webber, B. (2016). Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. *LAW X*, page 49.

Samy, D. and González-Ledesma, A. (2008). Pragmatic annotation of discourse markers in a multilingual parallel corpus (Arabic-Spanish-English). In *Proceedings of the sixth International Conference on Language Resources and Evaluation - LREC 2008*.

Schiffrin, D. (1988). *Discourse markers*. Number 5. Cambridge University Press.

Spooren, W. and Degand, L. (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, 6(2):241–266.

Stede, M., Afantenos, S. D., Peldszus, A., Asher, N., and Perret, J. (2016). Parallel discourse annotations on a corpus of short texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation - LREC 2016*.

Tiedemann, J. and Nygaard, L. (2004). The OPUS Corpus-parallel and free: http://logos. uio. no/opus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation - LREC 2004*.

Tonelli, S., Riccardi, G., Prasad, R., and Joshi, A. K. (2010). Annotation of discourse relations for conversational spoken dialogs. In *Proc. of the 8th International Conference on Language Resources and Evaluation-LREC 2010*.

Webber, B., Prasad, R., Lee, A., and Joshi, A. (2016). A discourse-annotated corpus of conjoined VPs. In *Proceedings of the LAW X*, pages 22–31.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proc. of the 1st international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.

Zeyrek, D., Demirşahin, I., Sevdik-Çallı, A., and Çakıcı, R. (2013). Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue and Discourse*, 4(2):174–184.

Zeyrek, D., Mendes, A., Grishina, Y., Gibbon, S., Kurfalı, M., and Ogrodniczuk, M. (in preparation). TED Multilingual Discourse Bank: A parallel corpus annotated in the PDTB style.

Zhou, Y. and Xue, N. (2015). The Chinese Discourse Treebank: a Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397–431.