Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

Lang Resources & Evaluation
https://doi.org/10.1007/s10579-019-09445-9

PROJECT NOTES

TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style

Deniz Zeyrek[1] · **Amália Mendes**[2] · **Yulia Grishina**[3] · **Murathan Kurfalı**[1,4] · **Samuel Gibbon**[5] · **Maciej Ogrodniczuk**[6]

© Springer Nature B.V. 2019

Abstract TED-Multilingual Discourse Bank, or TED-MDB, is a multilingual resource where TED-talks are annotated at the discourse level in 6 languages (English, Polish, German, Russian, European Portuguese, and Turkish) following the aims and principles of PDTB. We explain the corpus design criteria, which has three main features: the linguistic characteristics of the languages involved, the interactive nature of TED talks—which led us to annotate Hypophora, and the decision to avoid projection. We report our annotation consistency, and post-annotation alignment experiments, and provide a cross-lingual comparison based on corpus statistics.

Keywords Discourse · Discourse relations · Corpus creation · Annotation · Multilingual corpus

1 Introduction

Manual and automatic annotation efforts started with what was seen as "low-hanging fruit" (Joshi 2012): PoS tagging, morphological and syntactic parsers, referential links, named entities, etc. More recently however, attention has shifted to

✉ Deniz Zeyrek
dezeyrek@metu.edu.tr

[1] Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

[2] Centre of Linguistics, University of Lisbon, Lisbon, Portugal

[3] University of Potsdam, Potsdam, Germany

[4] Department of Linguistics, Stockholm University, Stockholm, Sweden

[5] Centre for Neuroscience in Education, University of Cambridge, Cambridge, UK

[6] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

higher levels of language, namely semantics and discourse, resulting in various semantically-annotated corpora, such as FrameNet (Baker et al. 1998), PropBank (Palmer et al. 2005), Groningen Meaning Bank (Basile et al. 2012), and the Penn Discourse TreeBank, or PDTB (Prasad et al. 2014). Despite the growing number of discourse-annotated corpora being developed for individual languages, discourse-annotated corpora for multiple languages are still rare. They are, however, very much needed as they would contribute to the empirical investigations of discourse cross-linguistically, enhance the science of annotation (Hovy and Lavid 2010; Ide and Pustejovsky 2017), and simulate language technology applications that need discourse parsing, such as question-answering and summarization. TED Multilingual Discourse Bank, or TED-MDB, is a corpus of transcribed TED talks involving multiple European languages (English, German, Russian, European Portuguese, Polish) as well as one non-European language, Turkish, also annotated at the discourse level following the PDTB approach (Zeyrek et al. 2018).[1] The corpus aims to serve three purposes. The first is to provide an empirical basis for a cross-lingual comparison of discourse relations and discourse structure. Second, it aims to induce discourse parsers, particularly for languages other than English. Two important steps for discourse parsing are discourse connective identification, and sense disambiguation. For English, Pitler and Nenkova (2009) extracted explicit discourse connectives in the PDTB and disambiguated their senses. Other work in sense identification includes Marcu (2000) and Lin et al. (2014), as well as the CoNLL Shared Task (http://www.cs.brandeis.edu/clp/conll15st/). But for most languages involved in TED-MDB other than English, work on discourse parsing is either scarce or non-existent. For example, for Brazilian Portuguese, tools for manual and automatic discourse annotation in the RST and CST frameworks (RST Toolkit, DiZer, CSTParser) have been developed (Aleixo and Pardo 2008; Maziero and Pardo 2012) based on corpora annotated with discourse information (CSTNews, CorpusTCC, Rhetalho, Summ-it), but no such resources exist for the European variety of Portuguese. Hence, the second goal is to contribute to the development of state-of-the-art discourse parsers for new languages. This in turn will help identify whether discourse relations are conveyed similarly across languages. Thus, the third aim of TED-MDB is to identify similarities and differences in discourse structure across languages.

The rest of the paper is structured as follows: we first summarize the data, providing our decisions concerning design, as well as what we leave out of scope (Sect. 2). Section 3 introduces how discourse connectives in different languages are specified and annotated with the major categories of the PDTB. In Sect. 4 we define Hypophora, question/response pairs with a rhetorical function that reflects the interactive nature of TED talks—a novelty of the corpus that differs from the PDTB 2.0. Section 5 introduces one of our design criterion, namely the avoidance of projection, describes our annotation cycle, and presents an evaluation of the corpus. It then describes a post-annotation alignment experiment on two annotated talks and

[1] The TED-MDB initiative is taken by a group of researchers involved in a consortium brought together by the ISCH COST Action (IS1312), *Textlink: Structuring discourse in Multilingual Europe*, http://textlink.ii.metu.edu.tr/.

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

discusses potential reasons for non-aligned tokens. Section 6 starts with corpus statistics on TED-MDB and compares them with other PDTB-inspired corpora. It also presents a cross-lingual comparison of the languages involved in TED-MDB and argues that valuable cross-linguistic facts can be revealed by analyzing the aligned as well as the non-aligned annotation tokens. Section 7 brings the paper to an end and presents some future directions.

2 Data, assumptions, and what we annotate

This section summarizes the data, presents our linguistic assumptions as well as the annotation decisions based on these assumptions, and explains what is left out of scope.

2.1 The data

The data comprise a collection of TED talk transcriptions, selected from the WIT3 corpus (Cettolo et al. 2012).[2] By settling on TED talks, we take advantage of the availability of parallel texts covering numerous languages. TED-MDB has 6 talks annotated uniformly, in 6 languages (Table 1), comprising a total of 3649 relations (Table 2).[3]

Our starting point is that adjacency matters for incremental interpretation of texts, and that adjacent clauses or sentences are likely to trigger a discourse relation. We reflect this notion in our annotation style by asking annotators to search for a discourse relation between each adjacent clause or groups of clauses. Discourse relations can also be sought among non-adjacent text segments; we leave the relations between non-adjacent text units for further research.

2.2 Assumptions, how and what we annotate

As in the PDTB, we assume that discourse connectives are predicates with binary arguments, referred to as Arg1, Arg2, where the criterion for argumenthood is Asher's abstract objects (Asher 1993)—eventualities and other abstract objects. Adopting the lexicalized approach of the PDTB, we ask annotators to mark discourse relations anchored to a connective, whether explicit (example 1) or implicit (example 2). Because explicit connectives are easy to recognize, we annotate discourse relations conveyed by explicit connectives used inter-sententially as well as intra-sententially. We annotate implicit relations that only hold inter-sententially, leaving intra-sentential implicit relations for further work. Implicit relations are annotated by inserting a connective that would make the inferred relation explicit. Other categories of the PDTB, i.e. alternative lexicalizations, entity

[2] https://wit3.fbk.eu/.

[3] TED-MDB is freely available to researchers and can be accessed at: https://github.com/MurathanKurfali/Ted-MDB-Annotations. The corpus now includes annotations on the transcripts of the same TED talks in a new language—Lithuanian—introduced in Oleskeviciene et al. (2018).

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

Table 1 TED talks annotated in TED-MDB

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

Table 2 Distribution of discourse relation types across the corpus

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

REVISED PROOF

relations and no relations are also annotated. We provide examples from as many languages as possible, but for reasons of space we sometimes limit the examples to a few representative languages. Where multiple languages are introduced as examples for the issues under discussion, they are presented in alphabetical order of the language name.

Throughout the paper, we show annotated tokens by underlining the connective; Arg1 is rendered in italics, Arg2 in bold type. The labels Arg1, Arg2 do not imply any kind of ordering, such as cause-consequence. Arg2 is the text segment that is syntactically related to the discourse connective, Arg1 is the other text segment. This approach is useful for a multilingual relation bank because it gives the monolingual teams freedom to determine how the arguments are ordered in a sentence, and where the discourse connective is positioned in the respective language. Unless otherwise noted, the English transcriptions of non-English examples are provided in parentheses.

1. *Ich bin in Sierra Leone geboren und aufgewachsen, einem kleinen und sehr schönen Land in Westafrika, einem Land reich sowohl an Bodenschätzen als auch an kreativen Talenten.* Allerdings **ist Sierra Leone berüchtigt für einen jahrzehntelangen Rebellenkrieg in den 90ern, in dem ganze Dörfer niedergebrannt wurden.**
   [Comparison:Concession:Arg2-as-denier] (German, TED Talk no. 1971) (I was born and raised in Sierra Leone, a small and very beautiful country in West Africa, a country rich both in physical resources and creative talent. However,

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

Table 3 PDTB 3.0 relation hierarchy (Webber et al. 2016)

Temporal	Synchronous		Comparison	Contrast	
				Similarity	
	Asynchronous	Precedence		Concession	Arg1 as denier
		Succession			
Contingency	Cause	Reason			Arg2 as denier
		Result		Concession+SpeechAct	Arg2 as denier+SpeechAct
	Cause+Belief	Reason	Expansion	Conjunction	
		Result		Disjunction	
	Cause+SpeechAct	Reason		Specification	Arg2 as denier
		Result			Arg1 as denier
	Purpose	Arg1 as goal		Equivalence	
		Arg2 as goal		Instantiation	
	Condition	Arg1 as condition		Exception	Arg1 as exception
		Arg2 as condition			Arg2 as exception
	Condition+SpeechAct			Substitution	Arg1 as subst
	Negative Condition	Arg1 as negcond			Arg2 as subst
		Arg2 as negcond		Manner	Arg1 as manner
	Negative Condition+SpeechAct				Arg2 as manner

Sierra Leone is infamous for a decade-long rebel war in the '90s when entire villages were burnt down.)

2. *Мне очень повезло начать карьеру в Музее Современного Искусства на ретроспективе работ Элизабет Мюррей.* (Implicit = поскольку) **Я столькому научилась у неё.** [Contingency:Cause:Reason] (Russian, TED Talk no. 1978)
   (I feel so fortunate that my first job was working at the Museum of Modern Art on a retrospective of painter Elizabeth Murray. I learned so much from her.)

In determining argument spans, we follow the minimality principle of the PDTB, which states that the smallest text spans that correspond to the sense of the relation are to be selected as arguments to a discourse connective (Prasad et al. 2014), e.g. see example 3.

3. We have a population that is both *growing* and **aging**. [Expansion:Conjunction] (English, TED Talk no. 1927)

For marking the sense of discourse relations, we use the PDTB 3.0 sense hierarchy, which is an enriched and revised form of the PDTB 2.0 (Table 3). We show the sense(s) of the relations in square brackets after each example where relevant.

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

2.3 What is not annotated

There are several levels of information that we do not include at this stage in our annotation scheme.

Attribution: We have left attribution out of scope. In our annotation scheme, we leave the attributive phrase unmarked except when it is an essential part of either argument, and necessary to complete the meaning of the relation. For example, *think* in Arg2 of token 4 could not be omitted.

4. *That's why I got into doing this*, because **I think that will change the world**. [Contingency:Cause:Reason] (TED Talk no. 1976)

Pragmatic markers: Since TED Talks are transcribed public speeches, they include many pragmatic markers frequently found in spoken registers. In TED-MDB, we focus for now on discourse connectives and do not annotate pragmatic markers that signal hesitations, filled pauses, turn beginning and closing, attitudinal meaning, etc.

Modified connectives: These indicate cases where the discourse connective is modified by an adverb. Annotating the modifying adverb is necessary as the adverb might constrain the sense of the relation. In our annotation scheme, we do not assign a separate tag for the modifier but annotate it together with the discourse connective, as in examples 5, 6, and 7, leaving the analysis of the modifier for post-processing.

5. 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**. (English, TED Talk no. 1927)
6. Und ich befürchte, *dass Investoren einigen der größten Veränderungen nicht genügend Aufmerksamkeit schenken*. Insbesondere wenn **es um Nachhaltigkeit geht**. (German, TED Talk no. 1927)
7. ... endişem o ki *yatırımcılar değişimin en büyük faktörlerinden bazılarına yeterince dikkat etmiyorlar*, özellikle de **iş sürdürülebilirliğe gelin**ce. (Turkish, TED Talk no. 1927)

We have observed variations in the use of adverb modifiers. In Russian for example, the equivalent of *especially* in example 5 is separated from the connective by a comma and not annotated (8); by contrast, in Polish the relation is rendered within a conjoined nominal phrase and no discourse relation is annotated (9).

8. Мир изменяется основательным образом, и я беспокоюсь, *что инвесторы не уделяют достаточного внимания некоторым крупнейшим двигателям перемен*, особенно, когда **речь идёт об устойчивости развития**. (Russian, TED Talk no. 1927)

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

9. Świat ulega głębokim zmianom, a mnie martwi to, że inwestorzy zwracają zbyt mało uwagi na główne motory tych zmian, a zwłaszcza na zrównoważony rozwój. [not annotated] (Polish, TED Talk no. 1927)
(The world is undergoing profound changes, and it worries me that investors pay too little attention to the main drivers of these changes, and especially to sustainable development.)

3 Determining discourse connective types

This section describes how we specified and annotated discourse connectives in different languages with the major annotation categories of the PDTB, and how we extended the NoRel tag to suit our purposes (Sect. 3.5).

3.1 Explicit and implicit connectives across languages

The TED-MDB team conveniently gleans discourse connective types from three well-known syntactic classes: (a) coordinating conjunctions (*and, but, so*), (b) subordinating conjunctions (*because, although, when*), (c) discourse adverbials (*however, nevertheless, therefore*). Prepositions and prepositional phrases form yet another class of potential discourse connective types (*for example, in summary, in sum*).

We take it as a fact that discourse connectives are a closed set of items; thus, the syntactic classes above are merely a starting point to determine the set of explicit discourse connectives in each language. We allow and encourage each monolingual team to specify discourse connectives that go beyond the syntactic classes above. To illustrate, in Turkish there are numerous suffixal subordinators that largely correspond to the senses conveyed by conjunctions in English. These are referred to as converbs in the literature (e.g. *-da* 'when', *-arak* 'by means of'/'and', *-se* 'if'). Converbs typically have Arg2-Arg1 ordering, where Arg2 is a non-finite nominalized clause linked to the finite Arg1 clause, as in example 10 and the original English transcript in example 11.

10. **Teleskobun içinde saçıl**-arak, *gezegeni görülemeyecek hale getiren* bu aşırı parlak görüntüyü ... [Expansion:Manner:Arg2-as-manner; Contingency:Cause: Result] (Turkish, TED Talk no. 1976).
11. It's scattering inside the telescope, creating that very bright image that washes out the planet. (no annotation) (English, TED Talk no. 1976)

In other languages, token 11 is rendered either as an inter-sentential implicit relation as in Polish and Russian, or as an explicit relation encoded by a coordinating conjunction, as in German.

12. *Das Licht vom Stern wird gebeugt, im Inneren des Teleskops gestreut,* und **erzeugt das sehr helle Bild, das den Planeten verblassen lässt.** [Expansion: Conjunction] (German, TED Talk no. 1976)

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

13. *Rozprasza się wewnątrz teleskopu*, (Implicit = i w efekcie 'and as a result') **tworząc ten jasny obraz, który zamazuje planetę**. [Contingency:Cause: Result:Arg2-as-result] (Polish, TED Talk no. 1976)
14. *Свет от звезды преломляется.* (Implicit = затем) **Он рассеивается внутри телескопа**, **создавая очень яркое изображение**, **которое затмевает планету.** [Temporal:Asynchronous:Precedence] (Russian, TED Talk no. 1976)

German also has discourse connectives that do not fit the well-known syntactic classes mentioned above. Specifically, a large number of connectives exhibit an anaphoric morpheme and therefore form a special class of the so-called 'anaphoric' connectives (as opposed to 'structural' connectives; Webber et al. 2003). They are event anaphors that additionally signal a coherence relation, as illustrated in *Dadurch* 'thereby' in example 15 below. The English version of this token is provided in example 16.

15. *Diese Initiativen schaffen einen mobileren Arbeitsplatz und reduzieren unseren Immobilien-Bedarf.* Dadurch **werden jährlich 23 Mio. Dollar an Betriebskosten gespart und die Emission von 100,000 Tonnen Kohlenstoff vermieden.** [Expansion:Manner:Arg1-as-manner] (German, TED Talk no. 1927)

While these types of connectives are common for German, they are not typical for other languages in TED-MDB, and as a result the corresponding relation might be expressed by other means in other languages. For example, in English (16) and Turkish (19) two clauses are connected with the intra-sentential explicit conjunction *and*; in Portuguese (17) two independent sentences are linked with an implicit inter-sentential relation, while the Russian equivalent (18) is expressed via an implicit intra-sentential relation (which is not marked according to our current guidelines).

16. 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. [Expansion:Conjunction] (English, TED Talk no. 1927)
17. Estas iniciativas criam um ambiente de trabalho mais móvel e *reduzem a nossa pegada imobiliária*. (Implicit = e 'and') **Permitem uma economia em custos operacionais na ordem de 23 milhões de dólares anuais e evitam emissões de 100 mil toneladas métricas de carbono**. [Expansion:Conjunction] (Portuguese, TED Talk no. 1927)
18. Эти действия создают большее количество мобильных рабочих мест, сокращают рабочие площади, позволяют сохранить 23 миллиарда долларов в эксплуатационных расходах ежегодно и избежать выброса 100 000 тонн углерода. [no relation marked] (Russian, TED Talk no. 1927)
19. ... *işletme maliyetlerinde yıllık olarak 23 milyon dolar tasarruf sağlıyor* ve **100.00 metrik ton karbon emisyonunu önlüyor**. [Expansion:Conjunction] (Turkish, TED Talk no. 1927)

UNCORRECTED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

3.2 Co-occurring connectives

For all languages in TED-MDB, we observed cases of multiple connectives, i.e. connectives that co-occur (*and then, so finally*), as pointed out for English by Webber et al. (2001), and for Catalan and Spanish by Cuenca and Marín (2009). These connective pairs often contain a conjunction and a discourse adverb. We create multiple tokens for such connective pairs in an attempt to reveal their senses and to understand which discourse pieces they relate.[4] Below is a German example *und deshalb* 'and hence' based on the single connective token in English.

20. (a) *Es sind auch Wirtschaftsthemen*. Und **deshalb sind sie für die Risiko und Renditebewertung sehr wichtig**. [Expansion:Conjunction] (German, TED Talk no. 1927)
(b) *Es sind auch Wirtschaftsthemen*. **Und** deshalb **sind sie für die Risiko und Renditebewertung sehr wichtig**. [Contingency:Cause:Result] (German, TED Talk no. 1927)
(*They're economic issues*, and **that makes them relevant to risk and return**.)

In addition to co-occurring explicit multiple connectives, we annotate cases where there is a single explicit connective in the discourse but one can infer an additional, implicit relation from the linguistic context (Rohde et al. 2016). For example, particularly in the case of the conjunction *and*, our annotators often infer an additional implicit sense. In these cases, we annotate the explicit connective with its relevant sense and create an implicit relation token in that context, as illustrated in Portuguese (example 21). This example also shows that, in our annotation, although we find cases of implicit intra-sentential relations, they are always associated with an explicit connective in the linguistic context.

21. (a) ... *venderam o seu principal negócio de ferramentas elétricas* e **reinvestiram o que apuraram no negócio da água**. [Expansion:Conjunction] (Portuguese, TED Talk no. 1927)
(b) ... *venderam o seu principal negócio de ferramentas elétricas* **e** (Implicit = a seguir 'then') **reinvestiram o que apuraram no negócio da água**. [Temporal: Asynchronous:Precedence] (Portuguese, TED Talk no. 1927) (... *they sold their core power tools business* and **reinvested those proceeds in a water business**).

Finally, we also annotate multiple senses for implicit relations, where necessary. For example, Portuguese has an implicit relation token with two senses (example 22).

22. *Está em querer permanentemente preencher o fosso entre onde estamos e onde queremos estar*. (Implicit = ademais 'in addition') **A mestria é sacrificarmo – nos pela nossa arte e não pelo amor de traçar a nossa carreira**. [Expansion. Conjunction], [Expansion:Level-of-detail:Arg2-as-detail] (Portuguese, TED Talk no. 1978)

[4] Our annotation procedure for capturing co-occurring multiple connectives has been to annotate each connective separately as a different token, and assign a meaning to each respective token, following the annotation principles of the PDTB. Multiple connectives could also be selected as a single token, as it has been the procedure of Cuenca and Marín (2009) and Crible (2007), among others.

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

(*It's in constantly wanting to close that gap between where you are and where you want to be*. **Mastery is about sacrificing for your craft and not for the sake of crafting your career**.)

3.3 Alternative lexicalizations (AltLex)

In the PDTB, AltLexes are alternative ways of lexicalizing discourse relations that lie beyond the closed set of discourse connectives (Prasad et al. 2010), and are indicators of a discourse relation. They include multi-word expressions that exhibit a range of syntactic constructions. An English example is presented below (example 23) together with its equivalents in other languages.

23. The moon has moved in front of the sun. *It blocks out most of the light so we can see that dim corona around it.* 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. [Expansion:Equivalence] (English, TED Talk no. 1976)
24. Der Mond hat sich vor die Sonne geschoben. *Er deckt den Großteils des Lichts ab und wir sehen um ihn herum eine matte Korona.* Es ist wie ('it is as') **wenn ich den Daumen hochhalte und den Strahler abblocke, der mich blendet:** Ich kann Sie in der hinteren Reihe sehen. [Expansion:Equivalence] (German, TED Talk no. 1976)
25. *Zasłonił większość światła tak, że widać wokół niego przyćmioną koronę*. To tak, jakbym ('It's just like') **palcem zasłonił światło wpadające do oka**, **widzę was w tylnym rzędzie**. [Comparison:Similarity] (Polish, TED Talk no. 1976)
26. A Lua colocou-se à frente do Sol. *Bloqueou a maior parte da sua luz por isso podemos ver a coroa ténue à sua volta.* Seria o mesmo ('It would be the same') **se erguesse o meu polegar e bloqueasse o ponto luminoso à frente dos meus olhos**, poderia vê-los na última fila. [Expansion:Equivalence] (Portuguese, TED Talk no. 1976)
27. Луна встала перед солнцем. *И заблокировала большинство света, поэтому видим тусклую корону вокруг.* То же самое ('The same if'), **если я наведу палец и заблокирую тот прожектор**, **который светит мне в глаз**, я могу увидеть вас на последнем ряду. [Expansion: Equivalence] (Russian, TED Talk no. 1976)
28. Işığın çoğunu engelliyor, böylece *etrafındaki soluk koronayı görebiliyoruz*. **Eğer başparmağımı kaldırıp, tam gözüme gelen şu spot ışığını engellersem de** aynı şey olacaktı ('would be the same thing') [Expansion:Equivalence] (Turkish, TED Talk no. 1976).

As these set of examples suggest, an AltLex in the original language tends to be captured as a translated version of that AltLex in the other languages. The opposite of the pattern also holds, for example there are cases where an explicit connective in the original language is captured by an AltLex in another language. This is commonly observed in Turkish, which has frequently occurring phrasal expressions based on postpositions conveying causal, resultative or concessive senses, e.g. *bunun için* 'for this reason', *bunun sonucunda* 'as a result of this', *buna rağmen*

UNCORRECTED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

'despite this'. In Turkish Discourse Bank, these expressions are easily identified by the deictic element and grouped as a subclass of AltLex (Demirşahin and Zeyrek 2017).

3.4 Entity relations (EntRel)

Entity relations represent identity relations between persons or objects mentioned in text segments. In this sense, they are different from the semantic relations that hold between text segments. Teasing apart a semantic relation from an entity relation can sometimes be difficult. To alleviate some of the difficulties, we limit entity relations to adjacent sentences and use the EntRel label as the last-resort strategy. That is, we annotate a pair of adjacent sentences as EntRel when the relation between the text segments is based on an attribute of an entity, rather than a relation that holds between eventualities. An example from English is provided in 29, followed by its multilingual versions in examples 30–33.

29. The reason, I would come to find out, was *their prosthetic sockets were painful because they did not fit well.* **The prosthetic socket is the part in which the amputee inserts their residual limb, and which connects to the prosthetic ankle.** [EntRel] (English, Ted Talk no. 1971)
30. *Der Grund*, wie ich später herausfand, *waren die Prothesenschäfte, die Schmerzen verursachten, weil sie nicht gut passten.* **Der Prothesenschaft ist der Teil, in welchen der Amputierte seinen Stumpf steckt, der mit der eigentlichen Prothese verbunden ist.** [EntRel] (German, Ted Talk no. 1971)
31. A razão, como vim a saber mais tarde, era que *o encaixe das próteses era doloroso por não ser um encaixe perfeito*. **O encaixe de uma prótese é a parte em que o amputado insere o coto do membro, e que liga com a articulação prostética**. [EntRel] (Portuguese, Ted Talk no. 1971)
32. Я выяснил, что причина была в том, что *их культеприемые гильзы вызывали боль*, *потому что не подходили по размеру.* **Культеприемые гильзы это часть**, **куда инвалид вставляет свою культю и которая соединяется с протезом.** [EntRel] (Russian, Ted Talk no. 1971)
33. Sebebi, sonradan öğrendiğim üzere *protez soketlerinin düzgün oturmadığı için canlarını yakmasıymış*. **Protez soketi, uzvu kesilmiş kişinin kesik uzvuna taktığı ve böylece uzvu protez ayağa bağladığı parçadır**. [EntRel] (Turkish, Ted Talk no. 1971)

3.5 No relation (NoRel)

For the sake of completeness, and to distinguish between discourse relations and non-discourse relations in the corpus, we use the NoRel tag to annotate pairs of adjacent sentences that are neither related by a discourse relation nor by an entity relation. For example, adjacent pairs of sentences involving a topic shift as in example 34 are annotated as NoRel.

34. *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,**

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

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 1971)

The second sentence of the cases annotated as NoRel might sometimes be related to a non-adjacent sentence in the text. For example, the last sentence of 35 relates to a listing of examples that answer a question raised higher up in the text. But since we limit NoRels to adjacent sentences, we mark token 35 and the corresponding instances as NoRel (cf. 36 and 37).

35. That's the equivalent of taking 21,000 cars off the road. *So awesome, right*? **Another example is Pentair**. [NoRel] (English, Ted Talk no. 1927)
36. Das sind 21,000 Autos weniger auf den Straßen. *Genial, oder*? **Ein weiteres Beispiel ist Pentair.** [NoRel] (German, Ted Talk no. 1927)
37. Isto equivale a retirar das ruas 21 mil carros. *É muito bom, não é*? **Outro exemplo é a Pentair.** [NoRel] (Portuguese, Ted Talk no. 1927)

Finally, in many cases, the connectives seem to have a rhetorical role or discourse organizing function rather than instantiating a semantic relation. For example, the connective *but* in token 38 does not convey a contrast relation; rather, it marks a topic shift. We annotate these cases as NoRel, as also shown in the Turkish version (example 39).

38. *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*. **But it makes me wonder if the investment rules of today are fit for purpose tomorrow**. [NoRel] (English, TED talk no. 1927)
39. *Gerçekten de karmaşık ve uzak görünebilirler, ki bu da şunu yapmamızı cazip kılabilir: Kafamızı kuma gömüp, bunun hakkında düşünmemek. Yapabilirseniz, buna karşı koyun. Bunu evde denemeyin*. **Ama bu beni bugünkü yatırım kurallarının yarınki amaca uygun olup olmadığı konusunda meraklandırıyor**. [NoRel] (Turkish, TED talk no. 1927)

In Russian, the equivalent of example 38 does not contain any connectives and is also marked as NoRel (example 40). This supports our use of the NoRel tag for instances where a connective is used for rhetorical or other purpose.

40. ...*Не повторяйте этого дома*. (Смех) **Это заставляет меня сомневаться, соответствуют ли правила инвестирования сегодняшнего дня делам завтрашнего**. [NoRel] (Russian, TED talk no. 1927)

4 Rhetorical level: Q/R pairs conveying the hypophora function

TED talks represent a specific genre where the aim of the speaker is to convince the audience that their story is true and worth listening to. The transcripts contain question-response pairs, where the question is both asked and answered by the speaker. Such Q/R pairs reflect the interactive nature of TED talks and are usually

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

meant to motivate the listener, attract their attention, or convince them to think in a specific way; thus they have a rhetorical function. Such Q/R pairs present a figure of speech called hypophora, defined as a pragmatic figure with an appealing function (Lanham 1991; Mayoral 1994) and also as a figure oriented towards the audience (see *subiectio* in Lausberg 1998).

In the PDTB 2.0, question and answer pairs are not treated differently, rather they are tagged either as an explicit relation, as in example 41, or as an implicit relation, as in examples 42, 43. In both cases, they are tagged with the appropriate sense:

41. *Why constructive*? Because **despite all the media prattle about comedy and politics not mixing, they are similar in one respect: Both can serve as mechanisms for easing tensions and facilitating the co-existence of groups in conflict**. [Contingency:Cause:Reason] [wsj-2369]
42. *How does a nice new tax, say 5% on any financial transaction sound*? **That ought to make sure we're all thinking for the long term**. (Implicit = indeed) [Expansion] [wsj-0118]
43. *Are you kidding*? **Looking for a job was one of the most anxious periods of my life—and is for most people**. (Implicit = because; so) [Contingency: Pragmatic cause:justification] [wsj-2373]

We extend the PDTB sense hierarchy with the new, top-level sense Hypophora, to mark such Q/R pairs; when applicable, we create an additional discourse relation sense. We use hypophora both to annotate Q/R pairs with an explicit question word or particle and to annotate Q/R pairs where the question is only intonationally marked (and shown with a question mark in the text).

4.1 Hypophora as an AltLex relation

In Q/R pairs that convey the hypophora function, we take the relation between the question and the response as one of alternative lexicalization. Thus, in wh-questions, we take the wh-word itself as the evidence for alternative lexicalization (as shown in example 44 and the equivalent tokens in Portuguese and Turkish).

44. What **gets us to convert success into mastery**? *This is a question I've long asked myself.* (English, TED Talk no. 1978)
45. O que é que **nos leva a transformar o êxito em mestria**? *Há muito que faço a mim mesma esta pergunta.* (Portuguese, TED Talk no. 1978)
46. **Başarıyı ustalığa dönüştürmemizi sağlayan şey** ne? *Uzun zamandır kendime sorduğum soru bu.* (Turkish, TED Talk no. 1978)

In polar questions, we search for other kinds of evidence that lexicalizes the hypophora function between the question and the response. Thus, the AltLex would be an auxiliary, as in English (example 47), or the question particle, as in Turkish ('mu', example 48).

47. Do **companies that take sustainability into account really do well financially**? *The answer that may surprise you is yes.* [Expansion:Level-of-detail:Arg1-as-detail; Hypophora]

UNCORRECTED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

48. **Özel sektör bu konuya dikkat ediyor** mu? Evet, *gerçekten güzel olan şey çoğu genel müdürün dikkat etmesi*. [Expansion:Level-of-detail:Arg1-as-detail; Hypophora] (Turkish, TED Talk no. 1927)

4.2 Hypophora as an implicit relation

In spoken registers of Romance languages, polar questions can be expressed by intonational structure without resorting to subject-verb inversion or the use of a question particle. By comparison, in written registers the only way to differentiate declarative clauses from such polar questions is through the use of a question mark. Therefore, when we come across intonationally expressed questions (and their responses) in TED-MDB that we identify as hypophora, due to the presence of a question mark, we take them as implicitly conveyed hypophora. As a result, the Portuguese equivalent of example 47 is marked as implicit hypophora (example 49). We have not observed such implicit relations in the other languages annotated in TED-MDB because they do not allow for non-explicitly marked questions. When more languages are added to the corpus, we are likely to observe more cases of implicit hypophora.

49. (Implicit = será que 'is it the case that') Estes casos são casos isolados? ... **As companhias que praticam a sustentabilidade estão mesmo bem financeiramente?** *A resposta pode surpreender-vos, mas é: "Estão, sim"* [Hypophora] (Portuguese, TED Talk no. 1927)
(... are these just isolated cases? ... Do **companies that take sustainability into account really do well financially**? *The answer that may surprise you is yes.*)

5 Annotation procedure and evaluation

For multilingual annotation efforts, annotation projection is an important step (Padó and Lapata 2009; Laali and Kosseim 2017). However, for discourse annotation efforts, this has the potential risk for the original language to seed the annotations in the other languages. Thus, we settled on starting the project without annotation projection. Based on this design criterion, this section describes our annotation cycle and presents our experiments on annotation consistency. Then, it presents a post-annotation alignment experiment followed by a discussion on the non-aligned tokens.

5.1 Annotation cycle

Each mono-lingual team minimally consisted of a primary annotator, who was typically an experienced researcher, or the lead researcher of the team, and a secondary annotator. The primary annotator annotated the entire corpus, going through each text sentence by sentence and marking all the relevant discourse types together with their binary arguments and senses. Where appropriate, supplementary

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

information supporting the meaning of the arguments was captured using the tags Supp1 and Supp2, as in the PDTB. We used the PDTB annotation tool (Lee et al. 2016).

The annotation cycle consisted of the following steps.

- Preparing the annotation guidelines: Prior to annotating the corpus, each annotator read through the guidelines—a summary of the main points of the PDTB principles, including our own examples and style (inexperienced annotators were trained differently, as explained in (Zeyrek et al. 2018)).
- Annotating the texts: The annotation flow involved going through each file, and annotating discourse relations as they appeared in the text. In this way, the annotators were able to pay attention to the incremental flow of discourse, just as in real life reading.
- Holding cross-lingual team meetings: After each text had been annotated cross-lingual meetings were held. In these meetings the teams went over each annotated token and examined their own and others' annotations token by token. In addition to this, the lead researcher of the team performed further checks where needed.[5] This helped identify mistakes or impossibilities (with regard to the annotation guidelines). Although the pace of annotation in following this procedure can be rather slow, we feel that the resulting cross-lingual consistency is well worth the time.
- Revising guidelines: Cross-lingual team meetings may lead to new or sharper annotation guidelines. These are added to the annotation guidelines where necessary.
- Repeating the cycle: After the addition of new guidelines, the cycle is repeated.

5.2 Experiments on annotation consistency

There are various methods being used to measure annotation (or annotator) reliability, e.g. (Artstein and Poesio 2008; Hovy and Lavid 2010). The most commonly used methods are inter-annotator agreement (calculating the reproducibility of a task performed by different annotators) and/or intra-annotator agreement (calculating the consistency of annotators on a specific task over time). Here we present inter-annotator agreement results for TED-MDB, where a new, independent annotator annotated approximately 25% of the data (corresponding to 2 transcripts per language) following the annotation cycle described in Sect. 5.1, but skipping the cross-lingual meeting step.

We adopted a different method than the one described in Zeyrek et al. (2018) to measure agreement and proceeded in two phases; firstly we calculated agreement on discourse relation spotting, i.e. whether or not the annotators identified a relation between the same discourse units. In the second phase, we measured agreement among the common annotations on the discourse relation type (whether or not the discourse relation identified in two sets of annotations is of the same type, e.g. Explicit, AltLex, etc.) and on the sense of the discourse relation (whether or not the

[5] The German and Russian annotations were carried out and checked by a single, bilingual researcher.

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

discourse relation identified in two sets of annotations is of the same top level sense of PDTB's relation hierarchy). In this procedure, we do not adopt a strict approach in terms of argument spans. E.g. we wanted to rule out tokens such as 50 and 51 as disagreement as the only difference in the second annotation is the inclusion of the adjunct *with this kind of relaxed focus* in Arg2.

50. I stood and watched as the coach drove up these women in this gray van and they exited.
51. I stood and watched as the coach drove up these women in this gray van and they exited with this kind of relaxed focus.

We only require a match between the selected connectives (for the Explicits and AltLexes), and a match of the end point of the first text span and the beginning of the second span point.[6] We measured precision, recall, and F1-score using formulae (1)–(3), where the "correct" tokens refer to the tokens in the first annotations. The results are presented in Table 4.

$$Precision = \frac{\#\ of\ correct\ found\ tokens}{Total\ \#\ of\ found\ tokens} \tag{1}$$

$$Recall = \frac{\#\ of\ correct\ found\ tokens}{\#\ of\ correct\ expected\ tokens} \tag{2}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

In the second phase, we measured type and sense agreement using simple ratio agreement (i.e. the ratio of all tokens with the same sense over all tokens shared by the annotation sets per language), as well as Cohen's κ. The results are provided in Tables 5 and 6.

Annotating discourse relations presents a number of difficulties. For example, discourse relations can be ambiguous (multiple readings are assigned to a single relation), or vague (the sense of the relation is nonspecific). There are also hard cases—rare instances that are difficult to categorize using existing annotation guidelines. In addition, different genres and modalities present different annotation challenges. For example, translators of TED talks have to obey certain rules, an important one being the need to translate texts in bits, i.e. the translators need to translate the text pieces between time stamps on the videos. This might lead translators to concentrate on one text piece at a time, disregarding the global coherence of the text; the resulting translation could influence the way discourse relations are conveyed. Given such added challenges, we consider Cohen's $\kappa \geq 0.70$ a good standard (Spooren and Degand 2010).

Tables 5 and 6 indicate that this minimal level of inter-annotator agreement is reached on type and sense assignment in all sections of the corpus, which suggests

[6] For convenience, here we refer to the linear ordering of the selected text spans Mírovskỳ et al. 2010, cf. Sect. 3.3.

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

Table 4 Inter-annotator agreement results on discourse relation spotting

Language	Precision	Recall	F-score
English	0.71	0.75	0.73
German	0.85	0.83	0.84
Polish	0.86	0.89	0.88
Portuguese	0.83	0.75	0.79
Russian	0.75	0.65	0.70
Turkish	0.86	0.84	0.85

Table 5 Inter-annotator agreement results on discourse relation type

Language	Simple ratio agreement	Cohen's κ
English	0.90	0.80
German	0.85	0.78
Polish	0.95	0.92
Portuguese	0.84	0.74
Russian	0.81	0.70
Turkish	0.86	0.80

Table 6 Inter-annotator agreement results on top-level senses

Language	Simple ratio agreement	Cohen's κ
English	0.91	0.86
German	0.80	0.71
Polish	0.84	0.77
Portuguese	0.89	0.84
Russian	0.83	0.75
Turkish	0.82	0.73

REVISED PROOF

574 that the PDTB guidelines can be used quite reliably for multilingual annotation
575 efforts.

576 **5.3 Post-annotation alignment experiment**

577 Before moving on to the next set of annotations in the project, we present a proof-
578 of-concept experiment, where we reveal to what extent annotated relations in other
579 languages are aligned with those annotated for English.
580 For this task, 20–23% of all the annotated relations, amounting to two TED talk
581 transcripts per language (TED talk no. 2009 and 2150), were aligned with respect to
582 English. Alignment was achieved through semi-automatic means:

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

Table 7 Distribution of discourse relation types in two TED talks

	Explicit	Implicit	AltLex	EntRel	NoRel	Total
English	75	39	11	8	9	142
German	45	43	4	15	4	111
Polish	42	52	2	16	6	118
Portuguese	47	54	5	9	7	122
Russian	50	36	4	11	9	110
Turkish	64	37	8	13	11	133

Table 8 Number of aligned relations and the number of annotated relations in two texts per language

English	Talk no. 2009		Talk no. 2150	
	Aligned –	Total 47	Aligned –	Total 95
German	32 (0.68%)	38	65 (0.68%)	73
Polish	33 (0.70%)	46	60 (0.63%)	72
Portuguese	46 (0.98%)	47	74 (0.78%)	75
Russian	40 (0.85%)	43	63 (0.66%)	67
Turkish	42 (0.89%)	51	73 (0.77%)	82

Table 9 Alignment performance in terms of f-score

	Talk no. 2009	Talk no. 2150
German	0.75	0.77
Polish	0.71	0.72
Portuguese	0.98	0.87
Russian	0.89	0.78
Turkish	0.86	0.82

F-scores are computed by regarding English annotations as gold annotation

– Firstly, discourse relations were extracted from the annotations via a simple script. Then, these relations were aligned using the LFAligner.[7]
– The performance of LFAligner was checked by the teams and wrong alignments were manually corrected.

Table 7 displays the distribution of discourse relation types in two talks on which the post-annotation alignment experiment was performed.

Table 8 presents the number of aligned relations with respect to the number of annotated relations and Table 9 reveals the alignment performance, with an f-score of ≥ 0.70 in all the language sets (see the corresponding confusion matrices in the Appendices).

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

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

Given the expected cross-lingual variation in rendering discourse relations and the fact that total alignment is linguistically unlikely, we consider the alignment performance satisfactory.

5.4 An assessment of the non-aligned tokens

An examination of the non-aligned tokens suggests that the mismatches are mostly due to the nature of the data, i.e. the translators' preferences, and an interaction of their preferences with our design choices. For example, our decision to leave out intra-sentential implicits results in unsupported annotations if there exists an intra-sentential explicit connective in the target language sentence corresponding to the implicit intra-sentential relation of the English sentence (or vice versa). Such cross-linguistic differences have already been mentioned in Sect. 3 (see examples 10–14 and 16–19) and it is no surprise that they compromise the alignment performance.

An extension of this issue is frequently observed in Polish texts, where implicit intra-sentential relations of the original text tend to be rendered as entity-related sentences. For example, while the English sentence 52 has no annotation, the Polish equivalent (example 53) has two sentences linked with an entity-based relation:

52. In 1988, she won the gold in the heptathlon and set a record of 7,291 points, a score that no athlete has come very close to since. [no annotation] (English, TED talk no. 1978)
53. *W 1988 roku wygrała złoty medal w siedmioboju i ustanowiła rekord na 7291 punktów*. **Rekord**, **do którego dotąd żaden sportowiec się nie zbliżył**. (Polish, TED talk no. 1978) [EntRel]

Moreover, because we annotate an additional implicit relation when the context of an explicitly conveyed relation enables it (see example 21), unsupported annotations may appear when one token has an explicit connective triggering an additional implicit relation, as opposed to only one implicit relation in the corresponding relation of the other language. This is illustrated in the explicit and implicit tokens created for English (examples 54–55), and the translation into Portuguese (example 56). In this case, although tokens 55 and 56 are aligned, token 54 does not have an aligned equivalent in Portuguese.

54. *There was a deep restlessness in me, a primal fear that I would fall prey to a life of routine and boredom*. And **many of my early memories involved intricate daydreams ...** [Expansion:Conjunction] (TED Talk no. 2009)
55. *There was a deep restlessness in me, a primal fear that I would fall prey to a life of routine and boredom*. **And** (Implicit = so) **many of my early memories involved intricate daydreams ...** [Contingency:Cause:Reason] (TED Talk no. 2009)
56. *Sentia uma profunda inquietação, um medo primordial de que seria vítima de uma vida de rotina e aborrecimento*. (Implicit = por isso 'so') **Muitas das minhas primeiras memórias envolviam sonhar acordada ...** [Contingency: Cause:Reason] (Portuguese, TED Talk no. 2009)

UNCORRECTED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

So far, we have discussed some constraints in the data that arise from our annotation choices coupled with the translators' tendencies. Some of these problems can be alleviated when TED-MDB is more richly annotated. But there are also other mismatches, which will remain as a challenge to any alignment or projection task involving discourse relations. For example, restrictive relative pronouns (*who, which, that*), which are not annotated according to our guidelines, may be translated to the target language with an explicit connective ('and') and get annotated. Example 57 and its translation into Turkish (example 58) illustrate this situation.

57. Now, on the other side of the network, you tend to have primarily African-American and Latino folks who are really concerned about somewhat different things than the geeks are ... [no annotation] (TED Talk no. 2150)
58. *Ağın diğer tarafında başlıca Afro-Amerikalılar ve Latin toplumu yer almakta* ve **bunlar anti-sosyallerden kısmen daha farklı şeylerle ilgilenirler.** [Expansion: Conjunction] (TED Talk no. 2150)

Secondly, across all language sets, clauses with an abstract object interpretation in English may be translated to the other language as nominal phrases (NPs) with no abstract object interpretation. In the aligned data, we find numerous examples of this phenomenon, as in 59–60: the clause 'mapping cities' is translated as the non eventive NP *mapas de cidades* 'city maps'. As a result, and following our guidelines, the English sentence (59) is annotated as a case of explicit intra-sentential conjunction, while 60 is not annotated.

59. ... there's other ways to think about *mapping cities* and **how they got to be made** [Expansion:Conjunction] (TED Talk no. 2150)
60. ... há outras formas de pensar em mapas de cidades e na forma como devem ser feitos ... [no annotation] (Portuguese, TED Talk no. 2150)

Finally, in each language set, we found some annotation errors; in particular, explicit intra-sentential connectives and implicit relations (i.e, only those that hold across sentences) appear to be easily missed. Though these errors are not frequent, in cases where they occur in one file but the corresponding file of the other language is correctly annotated for the same tokens, non-aligned relations are inevitable.

6 Cross-lingual explorations

In this section, we first compare TED-MDB with other PDTB-inspired corpora through corpus statistics. Then, we present a cross-lingual comparison of the languages involved in TED-MDB on the basis of the results of the alignment experiment and TED-MDB corpus statistics.

6.1 TED-MDB and other PDTB-inspired corpora

Table 10 is an extension of the comparisons provided in Prasad et al. (2014) with TED-MDB in terms of the distribution of explicit vs. non-explicit relations. The table shows that in all these corpora, there exists a difference in explicit vs. non-

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

Table 10 The percentage of explicit relations versus other types of relations in PDTB-based corpora and TED-MDB

	# of all tokens	# of explicit(%)	# of other relations (%)
Chinese discourse TB	5534	1223 (22%)	4311 (78%)
Hindi discourse RB	602	189 (31%)	413 (69%)
PDTB	40600	18459 (46%)	22141 (54%)
Turkish DB 1.1	1924	868 (45%)	1056 (55%)
TED-MDB			
English	661	290 (44%)	371 (56%)
German	560	240 (43%)	320 (57%)
Polish	580	218 (38%)	362 (62%)
Portuguese	625	269 (43%)	356 (57%)
Russian	565	237 (42%)	328 (58%)
Turkish	658	276 (42%)	382 (58%)
TED-MDB—Total	3649	1530 (41%)	2119 (59%)

explicit relations, with larger differences displayed by Chinese and Hindi Discourse TreeBanks, possibly because intra-sentential implicits are also annotated in these corpora. It will suffice to say that the current explicit-non-explicit difference in TED-MDB will change when intra-sentential implicit relations are added to the corpus.

The top-level senses in PDTB 2.0 presents an order of Expansion (0.42%) > Comparison (0.23%) > Contingency (0.22%) > Temporal (0.13%). This is preserved in TED-MDB to a great extent: Expansion (0.52%) > Contingency (0.25%)> Comparison (0.13%) > Temporal (0.08%) with Contingency relations being more frequently expressed than the Comparison relations. The distribution of the top-level senses in all sections of TED-MDB are very similar to each other, as shown in Zeyrek et al. (2018) (cf. Table 5 therein), which is expected as we are dealing with translations that aim to remain loyal to the meaning of the source texts. Among top-level senses, Expansion relations are the most frequent, while Temporal relations are the least frequent, which might be due to the topic of the TED talks chosen. Finally, the frequency of Hypophora is about 0.02% per language—although this frequency is quite low, we believe it enables an understanding of the types of Hypophora and provides a starting point for examining the role of question/answering in TED talks.

6.2 Discourse relations across languages: the view from TED-MDB

Despite the current small size of TED-MDB, we are able to reach some conclusions based on our study. The quantitative data in Table 2 and the data obtained from the aligned talks point to some conclusions. As in the previous sections of the paper, the term implicit refers only to inter-sentential implicit relations.

Explicit relations: according to Table 2, the percentage of explicit relations is quite stable across languages and falls between 42 and 44%, though Polish is an

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

exception (37%). This shows that conveying a discourse relation by explicit means is the preferred mode in TED-MDB. Any other differences are related to the distribution of the non-explicit relation types across languages, as we explain below.

Implicit relations: the percentage of implicit relations among the language sets ranges between 30 and 41%, placing English and Turkish at one end of the spectrum, and Portuguese at the other end. Portuguese has the highest percentage of implicit relations in TED-MDB; in fact the percentage of implicit relations is almost the same as the explicit relations (41% vs. 43%). This raises the hypothesis that there is a high frequency of contexts where the explicit connective is omitted in the translations from English to Portuguese. Table 13 supports this conclusion and shows that in the talks we experimented with, there are 62 English explicit contexts aligned with Portuguese, out of which 41 contexts are kept as explicit, while 19 cases are rendered as implicit. According to the table, there are in fact more implicit tokens than explicit ones in the two talks (54 vs. 46). This confirms the data found in Table 2, but should be compared with original Portuguese texts to understand if implicitation is indeed more frequent in Portuguese.

Russian has the second highest percentage of implicit contexts in TED-MDB, and the percentages of explicit and implicit relations are close to those found for Portuguese. On the other hand, Table 14 shows that only 8 contexts eliminate the connective found in the English talk, rendering them as implicit relations in 7 cases, and as an EntRel in 1 case; in addition, the total number of implicit tokens in the two aligned talks is not as high as that in Portuguese (32 vs. 54). Thus, the two aligned talks may not be enough to observe the implicitation tendencies in Russian and better conclusions would be reached after the alignment of the entire set of talks with English. The Turkish annotation closely follows the distribution of the English annotations in terms of the split between explicit and implicit relations. This is interesting, as Turkish and English are furthest apart in terms of typology when considering all languages in TED-MDB. However, in many cases the type of connective might differ, as mentioned in Sect. 3.1, and might explain the typological difference of Turkish with English, and the other languages.

In the Polish set, the percentage of implicit relations is lower than the explicit relations (cf. Table 2), but the picture changes when we consider the distribution of explicitly conveyed relations vs. the relations that lack a clear signal (implicit relations, EntRels, and AltLexes). Then, the split is 218 vs. 310 (37.5% vs. 53.5%). Table 12 also confirms this and shows that in the two aligned talks, the combined frequency of implicit and EntRel tokens where an explicit connective is omitted is 24, slightly higher than the 22 cases that are kept as explicit. This behaviour seems specific to Polish transcripts in TED-MDB.

Entity relations: In TED-MDB, the frequency of the EntRel category ranges between 6 to 18%. Portuguese exhibits the lowest number of contexts labeled as EntRel and Polish displays the highest number of contexts (Table 2), which may be due to the way English sentences are split into two sentences and linked with entity-based relations (cf. Example 53). The confusion matrices show that in Polish and Portuguese (Tables 12, 13) the aligned EntRel tokens of the English texts are captured as EntRels only in half of the cases, the other half is rendered as implicit tokens. In German and Russian (Tables 11, 14), the 7 EntRel relations in English are

UNCORRECTED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

labelled as EntRels in 4 cases, and as implicit tokens in 3 cases. This suggests that the different translations of English EntRel contexts lead to some hesitation in some languages; we attribute this to the fact that implicit and EntRel contexts are both cases of a relation that are not lexically marked by a discourse connective.

Alternative lexicalizations: In general, the AltLex category occurs at low percentages in TED-MDB. Turkish exhibits the highest percentage (9%) (Table 2), while Polish shows the lowest percentage (2%). Table 15 also shows that in Turkish, the frequency of AltLexes in the two aligned talks is the highest of the six languages in the corpus, and confirms the observation related to the prevalence of the AltLex type in Turkish.

No relations: According to Table 2, the percentage of contexts marked as having no relation is quite stable across languages. Given our annotation guidelines regarding NoRels, these numbers indicate that topic shifts, listing relations, and the rhetorical links between adjacent clauses are captured fairly closely to their originals.

To sum up, our analysis suggests that the languages in TED-MDB converge on the distribution of the explicit relation type but diverge on certain matters such as the tendency for implicitation across sentences (Portuguese and Russian), the frequent use of a subtype of AltLexes based on postpositions (Turkish), and the high number of EntRels (Polish). Furthermore, by proceeding without annotation-projection, we were able to reveal some cross-linguistic issues surrounding discourse relations. An examination of the aligned relations generally supported our conclusions from TED-MDB's overall corpus frequencies, and the non-aligned data gave us valuable information about translation tendencies and cross-linguistic facts, which could have been disregarded in an approach that uses projection. Our analysis suggests that the annotation without projection approach lends itself well to contrastive linguistic analysis as it is free of bias, though it suffers from difficulties of synchronization of multilingual teams.

7 Conclusion

The main contributions in this paper have been

- to highlight the major design criteria of TED-MDB, including a consideration of the linguistic differences in conveying discourse relations across languages, and an approach that allows annotators to use their intuition during the annotation process, and subsequently mitigating projection;
- to compare ways in which discourse relations are conveyed in different language sections of the corpus and in other PDTB-inspired resources;
- to present the variations of hypophora in the corpus (a new top-level sense category) that illuminates the interactive nature of TED talks;
- to describe a post-annotation alignment exercise.

There are numerous ways this study can be extended. First, an annotation projection framework can be adapted or developed to identify discourse connectives and their arguments on parallel texts; the results could then be compared with those obtained from the current TED-MDB-style annotation. Second, TED-MDB can be extended

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

with more annotations on more texts to enable language technology applications; it can also attempt to better capture the interactive nature of TED talks by developing new annotation categories. Finally, future work can extend the cross-linguistic issues revealed in our study, and can explore deeper whether they are an effect of translation or due to linguistic characteristics of each language.

Acknowledgements We thank our annotators (Robin Goodfellow Malamud, Robin Schäfer, Olha Zolotarenko, Nuno Martins, Aida Cardoso, Celina Heliasz, Joanna Bilińska, Daniel Ziembicki, İpek Süsoy). The research has been partially supported by Textlink, by the Scientific and Technological Research Council of Turkey—BIDEB-2219 Postdoctoral Research program, by the Polish National Science Centre (Contract Number 2014/15/B/HS2/03435) and by the FCT—Fundação para a Ciência e a Tecnologia (project ID: PEst-OE/LIN/UI0214/2013). The support of Bonnie Webber and Manfred Stede is greatly acknowledged though all errors are our own.

Appendix

Here we present confusion matrices of the aligned relations in two talks. Rows show the English tokens aligned to language X, and columns show language X aligned to English. For example, in Table 11, the sum of the first row (47) is the sum of explicit relations (in English) aligned with a discourse relation in German. Of those relations, 31 are also conveyed explicitly in German, while 13 are realized as implicits and 3 as EntRels. The total number of explicit relations in the two English talks is 75 (also see Table 7 above), with 28 non-aligned explicit relations. Bold fonts indicates that the number of tokens in language X matches the number of tokens in English.

Table 11 German

	Exp.	Imp.	AltLex	EntRel	NoRel	Total aligned	Total Eng. tokens	Non-aligned
Exp.	**31**	13	0	3	0	47	75	28
Imp.	1	**23**	0	3	0	27	39	12
AltLex	3	0	**3**	0	0	6	11	5
EntRel	0	3	0	**4**	0	7	8	1
NoRel	0	4	0	1	**3**	8	9	1
Total	35	33	3	11	3			

Table 12 Polish

	Exp.	Imp.	AltLex	EntRel	NoRel	Total aligned	Total Eng. tokens	Non-aligned
Exp.	**22**	19	0	5	2	48	75	27
Impl.	5	**15**	0	6	0	26	39	13
AltLex	0	2	**2**	0	0	4	11	7
EntRel	0	3	0	**3**	0	6	8	2
NoRel	4	2	0	1	**2**	9	9	0
Total	31	41	2	15	4			

REVISED PROOF

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

Table 13 Portuguese

	Exp.	Imp.	AltLex	EntRel	NoRel	Total aligned	Total Eng. tokens	Non-aligned
Expl.	**41**	19	0	2	0	62	75	13
Impl.	2	**27**	0	3	2	34	39	5
AltLex	2	2	**4**	0	0	8	11	3
EntRel	0	4	0	**4**	0	8	8	0
NoRel	1	2	0	0	**5**	8	9	1
Total	46	54	4	9	7			

Table 14 Russian

	Exp.	Imp.	AltLex	EntRel	NoRel	Total aligned	Total Eng. tokens	Non-aligned
Exp.	**44**	7	0	1	0	52	75	23
Imp.	0	**20**	0	6	4	30	39	9
AltLex	3	0	**1**	0	0	4	11	7
EntRel	0	3	0	**4**	0	7	8	1
NoRel	0	2	0	0	**6**	8	9	1
Total	47	32	1	11	10			

Table 15 Turkish

	Exp.	Imp.	AltLex	EntRel	NoRel	Total aligned	Total Eng. tokens	Non-aligned
Exp.	**45**	6	3	4	0	58	75	17
Impl.	2	**27**	1	1	1	32	39	7
AltLex	3	1	**3**	2	0	9	11	2
EntRel	0	2	0	**5**	1	8	8	0
NoRel	0	0	0	1	**7**	8	9	1
Total	50	36	7	13	9			

REVISED PROOF

References

Aleixo, P., & Pardo, T. A. (2008). CSTTool: um parser multidocumento automático para o Português do Brasil. In *Proceedings of the IV workshop on M.Sc dissertation and Ph.D thesis in artificial intelligence (WTDIA)* (pp. 140–145). Salvador, Bahia.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Asher, N. (1993). *Reference to abstract objects in discourse*. Dordrecht: Kluwer.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics (COLING-ACL '98)* (Vol. 1, pp. 86–90). Montreal: Association for Computational Linguistics.

Basile, V., Bos, J., Evang, K., & Venhuizen, N. (2012). Developing a large semantically annotated corpus. In *Proceedings of the eighth international conference on language resources and evaluation (LREC 2012)* (pp. 3196–3200). Istanbul: European Language Resources Association (ELRA).

Journal : 10579 Dispatch : 30-3-2019 Pages : 27
Article No. : 9445 LE TYPESET
MS Code : LREV-D-18-00038 CP DISK

Cettolo, M., Girardi, C., & Federico, M. (2012). WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th conference of the European association for machine translation (EAMT)* (Vol. 261, p. 268). Trento.

Crible, L. (2007). Discourse markers and (dis)fluency across registers: A contrastive usage-based study in English and French. Ph.D thesis, Louvain.

Cuenca, M. J., & Marín, M. J. (2009). Co-occurrence of discourse markers in Catalan and Spanish oral narrative. *Journal of Pragmatics*, *41*, 899–914.

Demirşahin, I., & Zeyrek, D. (2017). Pair annotation as a novel annotation procedure: The case of Turkish Discourse Bank. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 1219–1240). Berlin: Springer.

Hovy, E., & 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., & Pustejovsky, J. (Eds.). (2017). *Handbook of linguistic annotation*. Berlin: Springer.

Joshi, A. (2012). Rememberance of ACLs past. Keynote speech, ACL 50th anniversary lectures. Jeju Island: The Association for Computational Linguistics. https://www.aclweb.org/mirror/acl2012/program/sub01.asp.html. Accessed 25 Feb 2018.

Laali, M., & Kosseim, L. (2017). Improving discourse relation projection to build discourse annotated corpora. *Recent advances in natural language processing meet deep learning (RANLP)* (pp. 407–416). Varna.

Lanham, R. (1991). *A handlist of rhetorical terms*. Berkeley: University of California Press.

Lausberg, H. (1998). *Handbook of literary rhetoric: A foundation for literary study*. Leiden: Brill.

Lee, A., Prasad, R., Webber, B. L., & Joshi, A. K. (2016). Annotating discourse relations with the PDTB Annotator. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Demos* (pp. 121–125). Osaka.

Lin, Z., Ng, H. T., & Kan, M.-Y. (2014). A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, *20*(02), 151–184.

Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. Cambridge: MIT Press.

Mayoral, J. A. (1994). *Figuras retóricas*. Madrid: Editorial Sintesis.

Maziero, E. & Pardo, T. A. (2012). CSTParser: A multi-document discourse parser. In *Proceedings of the international conference, PROPOR 2012: Demonstration*. Coimbra. http://conteudo.icmc.usp.br/pessoas/taspardo/PROPOR2012Demo-MazieroPardo.pdf. Accessed 25 Feb 2018.

Mírovskỳ, J., Mladová, L., & Zikánová, Š. (2010). Connective-based measuring of the inter-annotator agreement in the annotation of discourse in PDT. In *Proceedings of the 23rd international conference on computational linguistics: Posters Volume* (pp. 775–781). Beijing: Association for Computational Linguistics.

Oleskeviciene, G. V., Zeyrek, D., Mazeikiene, V., & Kurfalı, M. (2018). Observations on the annotation of discourse relational devices in TED talk transcripts in Lithuanian. In S. Kübler & H. Zinsmeister (Eds.), *Proceedings of the workshop on annotation in digital humanities co-located with ESSLLI 2018* (Vol. 2155, pp. 53–58). Sofia. CEUR-WS.org.

Padó, S., & Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, *36*, 307–340.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*(1), 71–106.

Pitler, E. & Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 conference: Short papers* (pp. 13–16). Singapore: Suntec, Association for Computational Linguistics.

Prasad, R., Joshi, A., & Webber, B. (2010). Realization of discourse relations by other means: Alternative lexicalizations. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp. 1023–1031). Uppsala: Association for Computational Linguistics.

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

Rohde, H., Dickinson, A., Schneider, N., Clark, C. N., Louis, A., & Webber, B. (2016). Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In *Proceedings of the 10th linguistic annotation workshop held in conjunction with ACL 2016* (pp. 49–58). Berlin: Association for Computational Linguistics.

UNCORRECTED PROOF

Journal : 10579 | Dispatch : 30-3-2019 | Pages : 27
Article No. : 9445 | LE | TYPESET
MS Code : LREV-D-18-00038 | CP | DISK

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

Webber, B., Knott, A., & Joshi, A. (2001). Multiple discourse connectives in a lexicalized grammar for discourse. In H. Bunt & R. E. Muskens Thijsse (Eds.), *Computing meaning, Studies in Linguistics and Philosophy* (Vol. 77, pp. 229–245). Berlin: Springer.

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2016). A discourse-annotated corpus of conjoined VPs. In *Proceedings of the 10th Linguistics Annotation Workshop* (pp. 22–31). Berlin: Association for Computational Linguistics.

Webber, B., Stone, M., Joshi, A., & Knott, A. (2003). Anaphora and discourse structure. *Computational Linguistics*, *29*(4), 545–587.

Zeyrek, D., Mendes, A., & Kurfalı, M. (2018). Multilingual extension of PDTB-style annotation: The case of TED Multilingual Discourse Bank. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)* (pp. 1913–1919). Miyazaki: European Language Resources Association (ELRA).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REVISED PROOF