

Building monolingual and multilingual discourse banks and implications for understanding discourse structure

Deniz Zeyrek

Middle East Technical University & Bosphorous University, Turkey
dezeyrek@metu.edu.tr

May 25, 2025, Polish Academy of Sciences

Introduction and the outline of presentation

- Turkish Discourse Bank (TDB) and TED-Multilingual Discourse Bank (TED-MDB), are resources where texts are annotated at the discourse level following the aims and principles of the Penn Discourse TreeBank (PDTB).

In this talk, I will discuss the corpus design criteria of these resources remaining within the PDTB framework and considering three main features:

- the linguistic characteristics of the language(s)
- consideration of the mode of texts (written/spoken/signed, etc.), both of which may led us to give new decisions
- approaches to evaluation.

Outline of presentation

- **PART I (slides 1-32)**
- Aims and motivations in building discourse corpora
- Two PDTB-based corpora: Turkish Discourse Bank, TED-Multilingual Discourse Bank
- Corpus design principles
- Discourse annotation styles
- Annotation workflow for consistency
- What can make annotation inconsistent?
- **PART II (slides 33-65)**
- Attention to language-specific characteristics in building discourse corpora
- What we annotate
- How we annotate: two cases of connective-led annotation in TDB
- Attention to language mode characteristics
- Evaluation
- Conclusion: Implications for local discourse structure

Aims and motivations in building discourse corpora

- Semantics does not only concern the meaning of clauses and sentences but also the senses associated with their relations to each other in text, known as discourse relations (cause, contrast, expansion, etc.)
- Discourse relations can be conveyed explicitly via discourse connectives (then, moreover, in contrast, etc.) or can be inferred, which are known as implicit relations.

The PDTB annotates discourse relations (DRel), both explicit and implicit, hence revealing

- relation semantics,
- how coherence is achieved at the local level (coherence among the clauses of texts).

Aims and motivations

Both the TDB and TED-MDB adopt these goals.

We created these corpora to support linguistic research and NLP applications that need coherence languages in multiple languages.

Two PDTB-based corpora: Turkish Discourse Bank (TDB) and TED-MDB

- The TDB is a corpus of written Turkish texts based on one-fourths of the 2 million-word-METU Turkish corpus.
- TÜBİTAK (Turkish Scientific and Technological Research Council) (2008-2011) and Middle East Technical University research funds.

to provide an empirical basis for discourse relations

to induce discourse parsers for Turkish

Three versions have been released:

- TDB 1.0 (inter- and intra-sentential explicit single- and multi-word connectives and their argument spans) - 8300 annotations over 400.000 words
- TDB 1.1 (smaller version extended with new connectives with senses) - 1856 annotations over 40.000 words
- TDB 1.2 (final, corrected version with all PDTB DRel realization types and senses) - 4000 annotations over 40.000 words
- TDB versions are available upon request and via DISRPT webpages (see below).

TED-MDB is a corpus of transcribed TED talks in English and their translations into multiple languages (German, Russian, European Portuguese, Polish, Turkish) later extended to Lithuanian.

It is built within the framework of the Cost Action, Textlink (2014-2018).



<https://github.com/MurathanKurfali/Ted-MDB-Annotations>.

Shared Task

Discourse Relation Parsing and Treebanking (DISRPT)

Shared Task on Discourse Segmentation, Connective and Relation Identification across Formalisms

In conjunction with [ACL 2023](#) and [CODI 2023](#) workshop

July, 2023

News 04/05/23: data have been updated (fixed mistakes in .rels of surprise datasets TEDm and CRPC), don't forget to pull the new data files.

News: deadline extension, systems and papers are due on **May, 14**

News: Test and surprise data are now available in our repository: <https://github.com/disrpt/sharedtask2023>

News 17/04/23: data have been updated, don't forget to pull the new data files.

Please check our [FAQ](#) page for more information about participation, evaluation etc.!

Study of coherence relations in frameworks such as RST (Mann & Thompson 1988), SDRT (Asher & Lascarides 2003) and PDTB (Miltsakaki et al. 2004), has experienced a revival in the last few years, in English and many other languages (Matthiessen & Teruya 2015; Maziero et al. 2015; da Cunha 2016; Iruskieta et al. 2016; Zeldes 2016, 2017). Multiple sites are now actively engaged in the development of discourse parsers (Lin et al. 2014, Feng and Hirst 2014; Ji and Eisenstein 2014; Joty et al. 2015; Surdeanu et al. 2015; Xue et al. 2016; Braud et al. 2017; Guz & Carenini 2020; Kobayashi et al. 2020; Nguyen et al. 2021; Kobayashi et al. 2021; Zhao et al. 2021; Yu et al. 2022; Atwell et al. 2022; Kurfali 2022; Nishida and Matsumoto 2022; Huber et al. 2022), as a goal in itself, but also for applications such as sentiment analysis, argumentation mining, summarization, question answering, or machine translation evaluation (Benamara et al., 2017; Gerani et al. 2019; Durrett et al. 2016; Peldszus & Stede 2016; Scarton et al. 2016; Schouten & Frasincar 2016; Xu et al. 2020; among many others). At the same time, evaluation of results in discourse parsing has proven complicated (see Morey et al. 2017), and progress in

- In the PDTB, discourse connectives are predicates with **binary arguments** (Arg1, Arg2), where the criterion for argumenthood is Asher's abstract objects¹ – eventualities and other abstract objects.
- **Adjacency** matters for the incremental interpretation of texts; adjacent clauses or sentences are likely to trigger a DRel.
- We reflect this notion in our annotation style by asking annotators to search for a DRel between each adjacent clause.
- We ask annotators to mark DRels anchored to a connective, whether explicit or implicit.

¹Nicholas Asher. *Reference to Abstract Objects in Discourse*. Dordrecht: Kluwer, 1993.

Corpus design principles

- The PDTB focuses on local and relational coherence.
- Unlike the RST, the PDTB does not follow the nuclearity principle.
- It does not assume a hierarchical discourse structure, so the annotators do not have to keep the global discourse structure in their memory.
- Instead, the notions of Arg1, Arg2 are evoked. They are relatively easy to apply to texts (more on this below!)
- **Arg2** - is based on a syntactic concept and it is the text span that hosts the connective (bold)
- *Arg1* - the other text span (italics)

Nuclearity in RST (from B. Webber & M. Stede's slides in TextLink)

- In RST, the first instruction for the annotators is: “Is the text composed of recognizable topical units? If so, mark the boundaries.” (These will be boundaries between larger units of analysis.)
- When assigning a coherence relation to a pair of text spans, one of them may be more “central” to the author's purposes: nucleus versus satellite
- Whichever approach we take, there is a single shallow discourse structure which is constructed incrementally.

- In the PDTB, relation senses are organized hierarchically, with **4 top-level** senses at the top of the hierarchy, specified by second-level and in certain cases, third-level senses, e.g.
CONTINGENCY:Cause:reason
- The hierarchical organization of the senses reflect the idea that there is “a small core set of relations that can hold between the situations described in the arguments of the connectives.”²
- For **symmetric** relations (e.g. Conjunction, Disjunction, Instantiation), the hierarchy stops at level 2.
- For **asymmetric** relations (e.g. Asynchronous, Cause), there are level 3 senses.

²Rashmi Prasad et al. “The Penn Discourse TreeBank 2.0”. In: *LREC*. 2008. 

PDTB 3.0 Sense hierarchy

Temporal	Synchronous		Comparison	Contrast	
	Asynchronous	Precedence		Similarity	
		Succession		Concession	Arg1 as denier Arg2 as denier
Contingency	Cause	Reason		Concession+ <u>SpeechAct</u>	Arg2 as <u>denier+Speech Act</u>
		Result			
	<u>Cause+Belief</u>	Reason Result	Expansion	Conjunction	
	<u>Cause+SpeechAct</u>	Reason Result		Disjunction	
	Purpose	Arg1 as goal Arg2 as goal		Specification	Arg2 as detail Arg1 as detail
				Equivalence	
	Condition	Arg1 as condition Arg2 as condition		Instantiation	
				Exception	Arg 1 as exception Arg2 as exception
	<u>Condition+SpeechAct</u>			Substitution	Arg1 as <u>subst</u> Arg2 as <u>subst</u>
	Negative Condition	Arg1 as <u>negcond</u> Arg2 as <u>negcond</u>		Manner	Arg1 as manner Arg2 as manner
	Negative Condition+ Speech act				

The PDTB annotation tool

The screenshot displays the TDB (Text Development Browser) interface. At the top, the 'Annotator: TDB/00003121.txt' is shown. Below the toolbar, the 'Relation Editor' window is open, showing a list of relations on the left and a table of relations in the center. The table has columns for 'Relation', 'PBVerb', 'Provenance', 'Conn2', 'Arg2-as-manner', 'SClass2A', and 'SClass2B'. The 'Relation' column lists various relations like 'ANN: NoRel | Ar', 'ANN: Implicit | w', 'ANN: Explicit | ip', 'ANN: Explicit | ai', 'ANN: Explicit | o', 'ANN: Explicit | v', 'ANN: Implicit | s', 'ANN: Implicit | w', 'ANN: Implicit | v', 'ANN: Explicit | e', 'ANN: Explicit | ü', 'ANN: Implicit | v', and 'ANN: Fvnlir | c'. The 'PBVerb' column is empty. The 'Provenance' column is 'DEFAULT'. The 'Conn2' column is empty. The 'Arg2-as-manner' column is empty. The 'SClass2A' column is empty. The 'SClass2B' column is empty. The 'Relation Editor' window also has buttons for 'Add New Relation', 'Save Relation', 'Accept Relation', 'Cancel Changes', 'Undo', 'Select Annotation', 'Reject Token', 'Expand All', and 'Delete'. The 'Raw Text' window is open, showing the text of the document. The text is in Turkish and describes a scene where a person is sitting on a bench, looking at a person who is walking away. The text is: 'Burada, bu ilçede... Bir kişi biliyor artık onu. Büyük ve haydut gibi görünmek isteyen, aralarına girmek isteyip giremeyen, alaylı bakış ve sözlü satışlara hedef olan o acılı ve gülünc çocukluğunu, varoluşunun en zor günlerini bilen biri. Arada karşılaşılıyorlar. **Kimi başını eğerek selamlıyor onu** ilhamı, kimi de sağ elini büküp parmaklarının ucunu hafifçe şapkasının siperine götürerek. Kahvede, pencere önündeki masalardan birinde oturuyorsa -geç kız ve kadınları yaralayan bıçak gibi bakışlar fırlattırdı ordan- sesleniyor; "Gel de bir çay iç," diyor. Çoğunlukla gitmiyor yanına; "İşim var," diyor. "Yorgunum," diyor. Hantal hantal yürüyor sokaklarda. Oyun oynayan çocuklara rastlarsa, gizli gizli izliyor onları. Daha çok evde, üst kattaki odanın penceresinin başındayken; perdenin ucunu aralayıp gözlüyor. Şimdi ona güç veren tek şey çocuklar, yitmiş, bir uygarlığın kalıntılarıymış gibi görünen oyunları ve çocukluğunu anımsatan her şey. Arka bahçenin kapısında dikilip bir zamanlar bu evin en büyük çocuğu olan, bu evin en büyük çocuğu olan, bu evin en büyük çocuğu olan...

Corpus design principles

In determining argument spans, the PDTB suggests to follow the **minimality principle**.³

Two examples from the PDTB:

Workers described "clouds of blue dust" *that hung over parts of the factory, even though* **exhaust fans ventilated the area.**
(COMPARISON:Concession)

The theory was *that the Voice is a propaganda agency* and **this government shouldn't propagandize its own people.**
(EXPANSION:Conjunction)

See how they are annotated :

³Rashmi Prasad, Bonnie Webber, and Aravind Joshi. "Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation". In: *Computational Linguistics* (2014).

PDTB Browser (18469 Results)

New Query Prev Next 00 93 Load Close Tab Split Font 15

0003

Conn: once

Conn: Although

Conn: even though

connHead	sClassA	Source	Type	Polarity	Det	r
though	Comparison.Concession.Expectation	Wr	Comm	Null	Null	ev

Conn: But

In Western industrialized countries, he said.
The plant, which is owned by Hollingsworth & Vose Co., was under contract with Lorillard to make the cigarette filters.

The finding probably will support those who argue that the U.S. should regulate the class of asbestos including crocidolite more stringently than the common kind of asbestos, chrysotile, found in most schools and other buildings, Dr. Talcott said.

The U.S. is one of the few industrialized nations that doesn't have a higher standard of regulation for the smooth, needle-like fibers such as crocidolite that are classified as amphiboles, according to Brooke T. Mossman, a professor of pathology at the University of Vermont College of Medicine.
More common chrysotile fibers are curly and are more easily rejected by the body, Dr. Mossman explained.

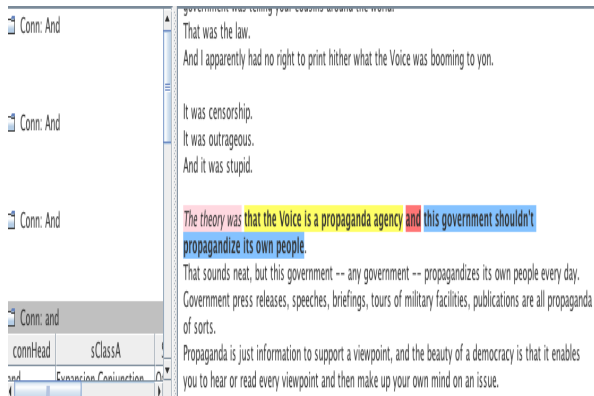
In July, the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos.
By 1997, almost all remaining uses of cancer-causing asbestos will be outlawed.

About 160 workers at a factory that made paper for the Kent filters were exposed to asbestos in the 1950s.
Areas of the factory were particularly dusty where the crocidolite was used.
Workers dumped large burlap sacks of the imported material into a huge bin, poured in cotton and acetate fibers and mechanically mixed the dry fibers in a process used to make filters.
Workers described "clouds of blue dust" that hung over parts of the factory. Even though exhaust fans ventilated the area.

"There's no question that some of those workers and managers contracted asbestos-related diseases," said Durrell Phillips, vice president of human resources for Hollingsworth & Vose. "But you have to recognize that these events took place 35 years ago.
It has no bearing on our work force today."

Conn/ATLAS Conn/ATLex Attr Arg1 Arg1 Attr Arg2 Arg2 Attr Sup1 Sup2

Minimality principle



The screenshot shows a text editor window with a transcript. The left pane shows a list of connections, and the right pane shows the corresponding text. The text is as follows:

Conn: And government was setting your standards around the world.

That was the law.

And I apparently had no right to print hither what the Voice was booming to yon.

Conn: And It was censorship.

It was outrageous.

And it was stupid.

Conn: And The theory was that the Voice is a propaganda agency and this government shouldn't propagandize its own people.

That sounds neat, but this government -- any government -- propagandizes its own people every day.

Government press releases, speeches, briefings, tours of military facilities, publications are all propaganda of sorts.

Propaganda is just information to support a viewpoint, and the beauty of a democracy is that it enables you to hear or read every viewpoint and then make up your own mind on an issue.

Conn: and	
connHead	sClassA
end	Expansion Connection

Annotating nominalizations

- Annotated by the PDTB in two strictly restricted contexts (quoted):⁴
- when they allow for an existential interpretation, as in the example below where the Arg1 selection can be interpreted existentially as that there will be major new liberalizations:
- Economic analysts call his trail-blazing liberalization of the Indian economy incomplete, and many are hoping *for major new liberalizations*, if **he is returned firmly to power..** (2041)

⁴Rashmi Prasad et al. “The penn discourse treebank 2.0 annotation manual”. In: (2007).

- when they involve a clearly observable case of a derived nominalization, as in example below, where the Arg1 selection can be assumed to be transformationally derived from such laws to be resurrected:
- *But in 1976, the court permitted resurrection of such laws, if **they meet certain procedural requirements.*** (0426)

Discourse annotation styles

- Segment-based annotation (RST)
- Punctuation-based annotation (PDTB, Chinese Discourse Bank)
TDB and TED-MDB applied this style to a limited extent!
- Connective-based annotation (PDTB)

Segment-based annotation (from Webber & Stede's TextLink slides)

1. Divide the text into minimal segments
2. Link (adjacent) segments via a coherence relation
3. If there is a complete spanning tree, stop; otherwise go to (2)

This requires:

- Clear rules on what defines a “minimal segment”
- Definitions of relations
- Some procedure specifying the order in which segments are linked

Punctuation-based annotation (from Webber & Stede's TextLink slides)

An explicit discourse connective that doesn't appear with punctuation will not be annotated.

Connective-led annotation: annotation workflow

Potential discourse connectives were identified (coordinating subordinating conjunctions, discourse adverbials).

Each text is automatically pre-annotated, highlighting the presence of potential connectives. Each occurrence was examined in turn (PDTB style)

If it expressed an independent relation between Abstract Objects, it was annotated.

Connective-led annotation

The notion of abstract objects (events, states, propositions, and so on) usually led to the annotation of **clauses**, finite or nonfinite.

Connective **modifiers** (adverbs that constrain the sense of the connective) are annotated (e.g. *only then*, *largely because*) in the TED-MDB.

Guidelines - essential regardless of style of annotation (from Webber & Stede's TextLink slides)

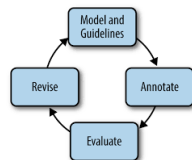
Guidelines tell annotators:

- What should and should not be annotated;
- How to annotate a token;
- How not to annotate a token;
- What else they might (or must) annotate, in addition to the basic elements of a discourse relation.

Comparison

	No. of words	Exp-Imp	Intra- /inter-S exp.	Intra- /inter-S imp.	Conn. Mod.	Sense	Attr.
PDTB	2 million	Yes-Yes	Yes	Yes	Yes	Yes	Yes
TED-MDB	5K-7K	Yes-Yes	Yes	Partly	Yes	Yes	No
TDB 1.0	400K	Yes-No	Yes	No	Yes	No	No
TDB 1.2	40K	Yes-Yes	Yes	Partly	Yes	Yes	No

Annotation workflow for consistency



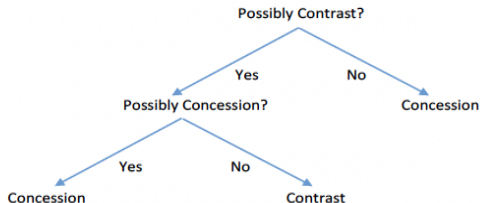
- ① Write guidelines.
- ② Annotate the texts going through the entire text sentence by sentence.
- ③ While annotations are going on, hold meetings with the annotators for a thorough discussion and control (including error analysis).
- ④ Check annotation consistency (IAA)
- ⑤ ↑ If IAA is low, go to (1) and revise guidelines.
- ⑥ Repeat the cycle.

What can make annotation inconsistent? (from Webber & Stede's TextLink slides)

- 1 Ambiguity: Different readings have been assigned different labels.
- 2 Annotation errors: Annotators have made the wrong selection
- 3 Guideline changes: Not propagated to earlier annotation
- 4 Linguistically “hard” cases: Guidelines may be inconclusive, e.g. contrast and concession in the PDTB 2.0.

Examples of hard cases from the PDTB (from Webber & Stede's TextLink slides)

Consistency and “hard cases”



Created an explicit guide to deciding which sense to use.

Examples of hard cases from the PDTB (from Webber & Stede's TextLink slides)

- Relations involving “yield” may be **intentional** (Purpose) or **unintentional** (Result).
- There is \$81.8 million of 7.20 term bonds due 2009 priced at 99 $\frac{1}{4}$ to yield 7.272 (Purpose)
- The offering was priced with an 8.95% coupon rate at 99.1875 to yield 9.19 (Purpose)
- The 12% notes due 1995 fell $\frac{9}{32}$ to 103 $\frac{3}{8}$ to yield 11.10 (Result)
- Britain's benchmark 11 $\frac{3}{4}$ bond due 2003/2007 rose $\frac{2}{32}$ to 111 $\frac{1}{2}$ to yield 10.14 (Result)

Examples of hard cases from the PDTB (from Webber & Stede's TextLink slides)

- In the PDTB 2.0, Arg2 of contingency relation was always taken as the antecedent:
- If **Mr. Krenz sticks to rigid policies** *the pressure from the Soviet Union could intensify.*
- Should **USX be left with only Marathon,** *Mr. Corry might well feel pushed to scout out other energy companies.*

Examples of hard cases from the PDTB (from Webber & Stede's TextLink slides)

- However, while correcting the annotations, the PDTB team noticed that the Arg1 of 'and' expresses the antecedent of a condition relation (and updated the annotation guidelines accordingly):
- Call Jim Wright's office in downtown Fort Worth, Texas, these days and the receptionist still answers the phone, "Speaker Wright's office."
- Add it all up and it means "that the Fed has a little leeway to ease its credit policy stance without the risk of rekindling inflation,"
- Guidelines were updated accordingly.

End of Part I



Any questions, comments so far?

- Attention to language-specific characteristics in building discourse corpora
- What we annotate
- How we annotate: two cases of connective-led annotation in TDB
- Attention to language-mode characteristics
- Evaluation
- Implications for local discourse structure: the case of TED-MDB

Language-specific characteristics

- Turkish uses suffixes (converbs) as discourse connectives as well as single-words and “phrasal expressions” (a form of AltLexes)
- 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)
- **Suffixal connective with a modifier:** özellikle de ... -ce ‘especially when’
- ...endişem o ki *yatırımcılar değişimin en büyük faktörlerinden bazılarına yeterince dikkat etmiyorlar, özellikle de iş sürdürülebilirliğe gelince.* (Turkish, TED Talk no. 1927)

In TDB, postpositions are also annotated if they combine discourse units that have independent abstract object interpretations:

Ali'nin göster-diği gibi *resim yaptım*. 'I drew as Ali showed'.

What we annotate

- A detailed set of annotation **guidelines** was prepared on the basis of the PDTB manual, taking into consideration language-specific characteristics of Turkish.
- **Explicit** relations were quite easy to detect.
- **Implicit** relations (both inter- and intra-sentential relations) were annotated exactly in the order they appeared in the text.
- **Punctuation**: Annotators tagged each implicit discourse relations that holds between adjacent sentences demilited by: a full-stop, a colon, a semicolon, a queson mark.
- For each implicit DRel, annotators inserted an **explicit** connective that best expresses the discourse relation.

Entity Relations (EntRel)

- A type of implicit relations; a sense category is not assigned to them.
- The second sentence provides more information about one or more entities in the previous sentence:
 - EntRels may be conveyed directly (via pronouns, overt NPs or prodrop in Turkish)
e.g. 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)
 - Indirectly (via bridging inference)
The house was painted white; the doors were green.

Alternative Lexicalization of a DRel (AltLex)

- AltLexes are a large group ranging from relatively fixed forms ('in response') to syntactically and lexically free forms ('that compares with').⁵
- In Turkish, we found that most AltLexes involved a deictic element, e.g. *buna rağmen* 'despite this.'
- Zastonił większość światła tak, że widać wokół niego przyćmiona korone. To tak, jakbym ('It's just like') palcem zastonił światło wpadające do oka, widze was w tylnym rzędzie. [Comparison:Similarity] (Polish, TED Talk no. 1976)

⁵Rashmi Prasad, Aravind Joshi, and Bonnie Webber. "Realization of discourse relations by other means: Alternative lexicalizations". In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics. 2010, pp. 1023–1031.

No Relations (NoRel)

- In the TED-MDB, NoRels generally indicate listing relations or topic shifts.
- In the TDB, NoRel types are more varied, involving many cases where no relation could be inferred, cases of weak coherence relations or absence of coherence relations. .

How we annotate - two cases of connective-led annotation

In constructing TDB, we used a number of variants of the connective-led annotation approach.

In **TDB 1.0**, only single- and multi-word explicits and their argument spans were annotated.

One connective was searched anywhere in the text by an in-house-built annotation tool.

Once an explicit connective was found, it was annotated together with its binary arguments and modifiers.

In **TDB 1.1**, we shifted to the PDTB annotator.

We changed our annotation style and annotated implicits, EntRels and AltLexes, their binary arguments and senses by reading each text sentence-by-sentence.

In **TDB 1.2** we followed the same style of TDB 1.1: checked the entire data, added NoRels and intra-sentential implicits as well as their argument spans and senses.

NoRels and EntRels were searched only at the inter-sentential level.

How we annotate - two cases of connective-led annotation

In TED-MDB, individual teams read each text sentence-by-sentence annotating all DRel types, their binary arguments and senses as they go along.

ExplicitS were annotated both at the intra- and inter-sentential level.

ImplicitS, EntRels and NoRels were searched at the inter-sentential level.

For Turkish, Portuguese and Lithuanian, intra-sentential implicitS were added at a later stage (only those conveyed by “and”).

Annotating inter- and intra-sentential relations

The snapshots in the following slides show how an inter-sentential (implicit) and an intra-sentential explicit connective are annotated over the same sentence. Both examples are from the PDTB.

Relation List

gold: Explicit | once | Ter
 gold: Explicit | with | Con
 gold: EntRel | Arg1(202...
 gold: Explicit | although |
 gold: Implicit | in fact | Ex
 gold: Implicit | in compar
 gold: Implicit | accordingl
 gold: EntRel | Arg1(1046
 gold: (1) Explicit | and |
 gold: (1) Implicit | then |
 gold: EntRel | Arg1(1315
 gold: Implicit | in addition
 gold: Implicit | and | Exp
 gold: EntRel | Arg1(2066
 gold: EntRel | Arg1(2639
 gold: (2) Explicit | and |
 gold: (2) Implicit | and |

Add New Relation

Save Relation

Accept Relation

Cancel Changes

Undo

Select Annotation

Reject Token

Expand All

Delete

Relation Editor

Raw Text

uReason

uDisagr:

ole:

ation Type:

nn1:

ass1A:

ass1B:

n Src:

Arb

1 Src:

Arb

2 Src:

Arb

A Lorillard spokeswoman said, "This is an old story.

We're talking about years ago before anyone heard of asbestos having any questionable properties.

There is no asbestos in our products now."

Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes. "We have no useful information on whether users are at risk," said James A. Talcott of Boston's Dana-Farber Cancer Institute.

Dr. Talcott led a team of researchers from the National Cancer Institute and the medical schools of Harvard University and Boston University.

The Lorillard spokeswoman said asbestos was used in "very modest amounts" in making paper for the filters in the early 1950s and replaced with a different type of filter in 1956.

From 1953 to 1955, 9.8 billion Kent cigarettes with the filters were sold, the company said.

Among 33 men who worked closely with the substance, 28 have died -- more than three times the expected number. Four of the five surviving workers have asbestos-related diseases, including three with recently diagnosed cancer. The total of 18 deaths from malignant mesothelioma, lung cancer and asbestosis was far higher than expected, the researchers said.

"The morbidity rate is a striking finding among those of us who study

Annotator: 00/wsj_0003

00 wsj_0003 Load << >> Font Size: 18 Clear Search Add All

Relation List

- gold: Explicit | once | Ter
- gold: Explicit | with | Con
- gold: EntRel | Arg1(202..
- gold: Explicit | although |
- gold: Implicit | in fact | Ex
- gold: Implicit | in compar
- gold: Implicit | accordingl
- gold: EntRel | Arg1(1046
- gold: (1) Explicit | and |
- gold: (1) Implicit | then |
- gold: EntRel | Arg1(1315
- gold: Implicit | in addition
- gold: Implicit | and | Exp
- gold: EntRel | Arg1(2066
- gold: EntRel | Arg1(2639
- gold: (2) Explicit | and |
- gold: (2) Implicit | and |

Add New Relation

Save Relation

Accept Relation

Cancel Changes

Undo

Select Annotation

Reject Token

Expand All

Delete

Relation Editor

Raw Text

uReason: results appear in today's New England Journal of Medicine, a forum likely to bring new attention to the problem.

uDisagr: []

role: []

ation Type: A Lorillard spokeswoman said, "This is an old story. We're talking about years ago before anyone heard of asbestos having any questionable properties. There is no asbestos in our products now."

nn1: and

ass1A: Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes. "We have no useful information on whether users are at risk," said James A. Talcott of Boston's Dana-Farber Cancer Institute.

ass1B: Dr. Talcott led a team of researchers from the National Cancer Institute and the medical schools of Harvard University and Boston University.

1 Src: Arb The Lorillard spokeswoman said asbestos was used in "very modest amounts" in making paper for the filters in the early 1950s and replaced with a different type of filter in 1956.

2 Src: Arb From 1953 to 1955, 9.8 billion Kent cigarettes with the filters were sold, the company said.

Among 33 men who worked closely with the substance, 28 have died -- more than three times the expected number. Four of the five surviving workers have asbestos-related diseases, including three with recently diagnosed cancer. The total of 18 deaths from malignant mesothelioma, lung cancer and asbestosis was far higher than expected, the researchers said.

Attention to language mode-specific characteristics

Hypophora

- TED talks - a specific genre, a mix of spoken and written registers
- Speakers aim 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 - usually meant to motivate the listener, attract their attention, or convince them to think in a specific way; thus they have a rhetorical function.
- Hypophora - a top-level sense in the TED-MDB, also annotated as an Expansion category in the PDTB 3.0⁶

⁶Rashmi Prasad, Bonnie Webber, and Alan Lee. "Discourse Annotation in the PDTB: The Next Generation". In: *Proceedings 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*. 2018, pp. 87–97.

- Hypophora involves questions and meaningful answers given to them.
- It doesn't include, e.g. rhetorical questions, check questions, or other questions that do not require (or do not have) an answer in the local discourse context.
- Needs detailed guidelines (see PDTB 3.0 guidelines)

Hypophora?

- ① What **gets us to convert success into mastery?** *This is a question I've long asked myself.* (English, TED Talk no. 1978)
- ② 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)
- ③ **Başarıyı ustalığa dönüştürmemizi sağlayan şey** ne? *Uzun zamandır kendime sorduğum soru bu.* (Turkish, TED Talk no. 1978)

- For the TED-MDB, two types of inter-annotator agreement are calculated to assess the reliability of the annotations.
- Discourse relation spotting (whether or not the annotators identified a relation between the same discourse units.) – Precision, Recall, F1 measure
- Whether or not the discourse relation identified in two sets of annotations is of the same type, e.g. Explicit, Implicit, AltLex, etc.
- DRel type (Explicit, Implicit, AltLex, etc.) – Cohen's Kappa

Evaluation of DRel spotting

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

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

$$\textit{F1} = \frac{2 * \textit{Precision} * \textit{Recall}}{\textit{Precision} + \textit{Recall}} \quad (3)$$

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 4: Inter-annotator agreement results on discourse relation spotting

- In TED-MDB, most teams had a primary annotator and secondary annotator
- " # of correct found tokens" refers to the relations that are annotated by both annotators.
- " # of correct expected tokens", on the other hand, refers to the relations that are annotated by the primary annotator.

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 5: Inter-annotator agreement results on discourse relation type

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

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

Evaluation of senses - the case of TDB

In earlier stages of TDB 1.1 (IPRA 2015), we evaluated the agreement between annotators over the senses using the exact match criterion.

LEVEL	IMPLICIT
CLASS	0.52
Type	0.43
subtype	0.34

ENTREL
0.795

LEVEL	EXPLICIT
CLASS	0.842
Type	0.711
subtype	0.29

Evaluation of senses - the case of TDB

In later stages, we calculated agreement among common annotations using the exact match.⁷

Sense	Explicit	Implicit	AltLex
Level-1	88.4%	85.7%	93.9%
Level-2	79.8%	78.8%	79.5%
Level-3	75.9%	73.1%	73.4%

Table 3: IAA results for sense agreement in TDB 1.1

⁷Deniz Zeyrek and Murathan Kurfalı. “TDB 1.1: Extensions on Turkish Discourse Bank”. In: *Proceedings of the 11th Linguistic Annotation Workshop*. Ed. by Nathan Schneider and Nianwen Xue. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 76–81. DOI: 10.18653/v1/W17-0809. URL: <https://aclanthology.org/W17-0809/>.

Explicit relations:

- The percentage of explicit relations is quite stable across languages and falls between 42% and 44%, though Polish is an exception (37%).
- This shows that conveying a discourse relation by explicit means is the preferred mode in TED-MDB.

Conclusion: Implications for local discourse structure: TED-MDB

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%) – Implication?

EntRels: The frequency of the EntRel category ranges between 6% to 18%. Polish displays the highest number of contexts, which may be due to the way English sentences are split into two sentences and linked with entity-based relations:

- ① 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)
- ② *W 1988 roku wygrała złoty medal w siedmioboju i ustanowiła rekord na 7291 punktów. **Rekord, do którego dotąd żaden sportowiec się nie zbliżył.*** [EntRel] (Polish, TED talk no. 1978)

AltLexes (non-connective expressions):

- The AltLex category occurs at low percentages in TED-MDB.
- Turkish exhibits the highest percentage (9%), while Polish shows the lowest percentage (2%).
- 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.

NoRels:

- The percentage of contexts marked as having no relation is quite stable across languages.

- In the English section and translations into Turkish, Portuguese and Lithuanian, implicit VP conjunctions ('and's) are annotated.⁸
- DRel annotations over English are (automatically) aligned with the annotations over other languages⁹ and made publicly available.

⁸Deniz Zeyrek, Giedrė Valūnaitė Oleškevičienė, and Amalia Mendes. “Multiple Discourse Relations in English TED Talks and Their Translation into Lithuanian, Portuguese and Turkish”. In: *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*. Ed. by Pierre Zweigenbaum, Reinhard Rapp, and Serge Sharoff. Torino, Italia: ELRA and ICCL, May 2024, pp. 125–134. URL: <https://aclanthology.org/2024.bucc-1.14/>.

⁹Sibel Özer et al. “Linking discourse-level information and the induction of bilingual discourse connective lexicons”. In: *Semantic Web 13.6 (2022)*, pp. 1081–1102.

Implications for local discourse structure: TDB

Table 4

Number of different realizations of discourse relations and their Level-1 sense tags in TDB 1.2

	Expansion	Temporal	Comparison	Contingency	DRs with no sense tag	Total
Implicit	1090	158	162	333	0	1743
Explicit	540	400	259	268	0	1467
AltLex	33	32	14	67	0	146
EntRel	0	0	0	0	233	233
Hypophora	0	0	0	0	78	78
NoRel	0	0	0	0	203	203
Total	1663	590	435	668	514	3870

Implications for local discourse structure: TDB

- We investigated discourse dependencies in TDB 1.2, following a research line explored by the PDTB team, as well as the Czech group.
- We examined discourse dependencies among three linearly ordered discourse units (DUs), where DU means any text span selected as an argument by one or both of the discourse connectives.
- The object of our investigations can be represented as: DU1 - DC1 - DU2 - DC2 - DU3.
- Examining TDB 1.2 with a Python script, we investigated the dependencies among three discourse units belonging to two consecutive DRels related by explicit or implicit discourse connectives (other discourse relations were out of scope of our analysis).

Implications for local discourse structure: TDB

An implicit-implicit dependency structure: “shared argument”¹⁰

- (3) Bu ben değildim, (çünkü) ben yere bakmazdım, (bilakis) gözüne gözüne bakardım insanların.
This was not me (because) I would not look down, (rather) I would look into people's eyes.

11

¹⁰Alan Lee et al. “Complexity of dependencies in discourse: are dependencies in discourse more complex than in syntax?” In: *5th International Workshop on Treebanks and Linguistic Theories*. 2006.

¹¹Deniz Zeyrek and Mustafa Erolcan Er. “A description of Turkish Discourse Bank 1.2 and an examination of common dependencies in Turkish Discourse”. In: *The International Conference on Agglutinative Language Technologies as a challenge of Natural Language Processing, ALTNLP'22, June 6, Koper, Slovenia (2022)*.

Implications for local discourse structure: TDB

A full-embedding dependency structure - 'after' and its binary arguments are fully embedded as an argument to a suffixal connective on the left side, -arak once'.

- (5)

Hukuk Fakültesini yarım bırak	-arak	<table border="1" style="display: inline-table; vertical-align: middle;"><tr><td>anneannesinin yanına gel</td><td>-ip</td><td>Ankara'ya yerleşmesinin</td></tr></table>	anneannesinin yanına gel	-ip	Ankara'ya yerleşmesinin
anneannesinin yanına gel	-ip	Ankara'ya yerleşmesinin			

nedeni ...
'the reason why

after	moving to her grandmother's	she settled in Ankara
-------	------------------------------------	-----------------------

once

she quitted the Law School

' ...

Implications for local discourse structure: TDB

A properly contained dependency structure - similar to fully embedded dependency, except that some material is left out in the embedded argument.

The subordinate clause (DU2) and its matrix clause (DU3) are selected entirely as the second argument to DC1.

- (6)
- | | | | | | |
|--|---------|-----|---------------|-------------------|---------------|
| çarşafarla geceden giderek terasa saklandı | (sonra) | ... | çarşafını giy | -erek | terastan indi |
| he hid at the terrace with the hijab | (then) | ... | after | wearing the hijab | he came down |

Ideas for further research

- There is room for more research on discourse dependencies.
- Our findings show that the implicit discourse relation recognition task can be improved by considering shared arguments, because, e.g. three adjacent implicit discourse relations is a highly likely sequence in Turkish discourse.
- An automatic argument span detection system can be induced by considering the availability of an entire discourse relation anchored by postpositions or suffixal connectives as an argument, as fully embedded and properly contained dependency patterns reveal.

Thank you! Any questions?