Turkish Discourse Bank: Ongoing Developments

Işın Demirşahin*, Ayışığı Sevdik-Çallı*, Hale Ögel Balaban†, Ruket Çakıcı*, and Deniz Zeyrek*

*Middle East Technical University
Ankara, Turkey
†İstanbul Bilgi University
İstanbul, Turkey
demirsahin@ii.metu.edu.tr, ayisigi@ii.metu.edu.tr, hogel@bilgi.edu.tr, ruken@ceng.metu.edu,tr, dezeyrek@metu.edu.tr

Abstract

This paper describes the first release of the Turkish Discourse Bank (the TDB), the first large-scale, publicly available language resource with discourse-level annotations for Turkish. We describe the features of the source corpus and the sub-corpus annotated for discourse connectives. We provide information about the annotations and other contents of the first release of the TDB. Finally, we describe the ongoing developments including annotating the sense and the class of the connectives, and the morphological features of the nominalized arguments of subordinating conjunctives.

Keywords: Turkish, discourse bank, discourse connectives

1. Introduction

Turkish Discourse Bank (the TDB) is the first large-scale publicly available language resource with discourse level annotations for Turkish. Following the style of Penn Discourse Tree Bank (PDTB) (Prasad et al., 2008), annotations include discourse connectives, modifiers and arguments of connectives, and supplementary materials for the arguments. In (1), a sample annotation is given. The connective is underlined; the first argument is in italics and the second argument in bold face.

(1) *İnsanlar tabiattan eşit doğarlar.* Dolayısıyla **özgür ve köle ayrılığı olmamalıdır.**
People are born equal by nature. As a result, **there should be no such distinction as the freeman and the slave.**

The annotations were carried out using the tool designed specifically for the TDB (Aktaş, et al., 2010). The annotations were performed by either three independent annotators, or by a pair of annotators and an independent individual annotator (Zeyrek et al., 2010; Demirsahin et al, ms).

2. Contents of the First Release

The TDB can be requested from www.tdb.ii.metu.edu.tr. The first release of the TDB includes the raw text files, annotation files, annotation guidelines, and a browser.

2.1. Text Files

The TDB is built on a ~ 400,000-word sub-corpus of METU Turkish Corpus (the MTC) (Say et al., 2002). the MTC is a 2 million-word resource of post-1990 written Turkish from multiple genres. A total of 159 files, 83 columns and 76 essays were excluded from the TDB, because these genres lack the conventional paragraph structure and make extensive use of boldface. These characteristics were not transferred to the MTC, which might have interfered with the reliable interpretation of the discourse relations and the specification of the extent of the arguments.

For the rest of the genres, the TDB preserves the genre distribution of the MTC, as shown in Table 1.

	the MTC		**the TDB**	
Genre	**#**	**%**	**#**	**%**
Novel	123	15.63	31	15.74
Story	114	14.49	28	14.21
Research/Survey	49	6.23	13	6.60
Article	38	4.83	9	4.57
Travel	19	2.41	5	2.54
Interview	7	0.89	2	1.02
Memoir	18	2.29	4	2.03
News	419	53.24	105	53.30
Total	787	100	197	100

Table 1: Distribution of the genres in the MTC and the TDB

2.2. Annotations

For each annotated text span, the text and the offsets for the beginning and the end of the span are kept in a standoff XML file. All tags except NOTE denote text spans. The annotation files include the content text and the beginning and end offsets for text spans. A sample XML tree for the connective span of (1) is provided in (2).

(2)
```
<Conn>
   <Span>
      <Text>dolayisiyla</Text>
      <BeginOffset>15624</BeginOffset>
      <EndOffset>15635</EndOffset>
   </Span>
</Conn>
```

The following subsections provide details for tree nodes and the note attribute.

2.2.1. CONN (Connective)

The discourse connective is regarded as an immediate discourse-level predicate (Webber and Joshi, 1998; Webber, 2004) with two abstract object arguments (Asher, 1993). Connectives that link non-abstract objects or sentential adverbs are not annotated. Table 2 shows five most frequent discourse connectives, compared to their total instances in the TDB.

	Discourse connectives		Other uses		Total instances	
Conn	**#**	**%**	**#**	**%**	**#**	**%**
ve 'and'	2112	28.2	5389	71.8	7501	100.0
için 'because'	1102	50.9	1063	49.1	2165	100.0
ama 'but'	1024	90.6	106	9.4	1130	100.0
sonra 'later'	713	56.7	544	43.3	1257	100.0
ancak 'however'	419	79.1	111	20.9	530	100.0

Table 2: Percent of discourse connectives and other uses

In the first release of the TDB, only explicit connectives are annotated. The discourse connectives are gleaned from coordinating conjunctions, subordinating conjunctions and discourse adverbials (Zeyrek & Webber, 2008). In addition to these, phrasal expressions are also annotated. These are subordinating conjunctions that take a deictic argument, which resolves to an abstract object. For instance, the postposition *rağmen* 'despite, although' can either take a nominalized subordinate clause or a deictic element such as *bu* 'this', resulting in the phrasal expression *buna rağmen* 'despite this'. Although syntactically the argument of the postposition is the deictic element, the TDB annotations select the whole phrasal expression as the connective, and annotate the abstract object the anaphora resolves to as the argument, in order to more explicitly reflect the discourse relations between the abstract objects.

A total of 8483 relations are annotated in the TDB. The annotators searched for 77 tokens. This number includes various forms of one root, such as *amaçla* 'goal+INS' and *amacıyla* 'goal+POS+INS'. 143 distinct text spans were annotated as discourse connectives, including phrasal expressions and constructions based on a token. For instance, *buna rağmen* 'despite this', *bunlara rağmen* 'despite these', *herşeye rağmen* 'despite everything', are annotated as distinct connectives. Likewise, the token *yandan* 'side+ABL' returns *bir yandan* 'on one hand' and a a variety of phrases as its second part, such as *bir yandan da, diğer yandan, öbür yandan,* and *öte yandan*, all of which come to mean 'on the other hand'. Most variations of connectives can be collapsed to few common roots as exemplified in Table 3.

Root	**Variations**
amaç- 'goal'	bu amaçla, amacıyla, amacı ile
dolayı- 'because'	dolayı, dolayısıyla, dolayısı ile, bundan dolayı, bu sebepten dolayı
neden- 'reason'	bu nedenle, o nedenle, bu nedenlerle, yukarıdaki nedenlerle, nedeniyle, nedeni ile
sonuç- 'result'	sonuçta, sonucunda, sonuç olarak, bunun sonucunda, bunların sonucunda
zaman- 'time'	zaman, bir zamanda, aynı zamanda, o zaman, ne zaman…o zaman

Table 3: Some of the common roots for morphological varieties of connectives

2.2.2. MOD (Modifier)

The modifiers are spans that specify or intensify the meaning of the connective, or signify the modality of the relationship. For example, the discourse adverbial *sonra* 'later' can be modified for duration by *iki gün* 'two days' or the relation indicated by the subordinator *için* 'because/ for' can be modified for modality by *belki* 'perhaps'.

2.2.3. ARG1, ARG2 (First and Second Argument)

Similar to the PDTB, the argument that syntactically hosts the connective is called the second argument (ARG2) and the other argument is called the first argument (ARG1). Arguments of the discourse connectives can be single or multiple verb phrases, clauses or sentences, i. e., any text span with an abstract object interpretation.

2.2.4. SHARED (Shared Material)

The SHARED span was introduced to the TDB for the spans that belong to both Arg1 and Arg2 of a connective. A shared material may be the common subject, object or adjunct.

2.2.5. SUPP (Supplementary Material)

Supplementary materials are selected for the arguments or shared spans: SUPP1 for ARG1, SUPP2 for ARG2 and SUPP_SHARED for SHARED. These tags specify the spans of text necessary to fully interpret the arguments. In the TDB, the supplementary materials are extensively used to include the resolutions of discourse-level anaphora in the arguments.

2.2.6. NOTE

NOTE is an attribute of the relation tag, as in (3)[1].

(3) `<Relation note="" sense="" type="EXPLICIT">`

The annotators can enter free text in the notes field. This filed is used for entering the rationale of the annotation, the problems annotators encountered during the annotation, or alternative annotations to the current one.

[1] The first release of the TDB does not include sense annotation. The sense attribute of the relation tag is included to easily implement sense annotation in future releases and to ensure the compatibility of the sense tag with the current release of the browser.

2.3. Annotation Guidelines

The annotation guidelines provide the definitions of key terms and general criteria for the annotations. The guidelines are supported with rich examples of both the annotated and unannotated cases.

2.4. The Browser

A browser specifically created for the TDB (Şirin, et al., 2012) is included in the first release. The browser enables the users to view all annotations on each file. The quick search feature enables the user to filter the files for connectives and genre. The advanced search feature offers the means to perform text and regular expression searches. A user manual is included in the distribution of the first release.

3. Ongoing Developments

Most discourse connectives have multiple uses. In the TDB, we have encountered connectives that can belong to multiple syntactic classes, such as subordinator and discourse adverbial. Also, most discourse connectives are polysemous to various degrees. In order to disambiguate such ambiguities, we introduce connective class and Arg2 feature annotations, as well as a PDTB-style sense annotation (Miltsakaki et al., 2005; Prasad et al, 2008).

3.1. CLASS (Connective Class)

The roots like *amaç-* 'goal', *neden-* 'reason', *netice-* 'result', *saye-* 'thanks to', and *yüz-* 'due to' may form subordinators and phrasal expressions. The subordinators are in the form root+POS+INS whereas their corresponding phrasal expressions have the form root+INS. However, the syntactic class of all such connectives cannot be figured out directly from the morphology of the connective. Some roots such as *sonuç-* 'result', form the subordinator *sonucunda* 'result+POS+LOC', as well as phrasal expressions, e.g. *bunun sonucunda* 'as a result of this'. Since phrasal expressions are annotated with the anaphoric expression in the text span, the connective class of *sonucunda* can be disambiguated from the CONN span. Still, there are connectives that are completely ambiguous in terms of subordinator and discourse adverbial uses, such as *sonra* in (5) and (4), respectively.

(4) **Sana aşık olduktan** sonra *karısından boşandı.*
He divorced his wife after **falling in love with you.**

(5) *Adam öldüğünü sandı,* **öldürüldüğünü** sonra.
The man thought he was dead; then (he thought) **that he was murdered.**

CLASS is a relation attribute like sense and notes. It has a limited set of values: CON for coordinating conjunctions, SUB for subordinating conjunctions, ADV for discourse adverbials and PHR for phrasal expressions. In addition, parallel constructions are marked with PAR together with the connective class of the compulsory item in the construction. For example, PAR CON for the parallel construction of the coordinating conjunctive *ya…ya* 'either…or', or PAR PHR for *ne zaman…o zaman* "when…then".

The preliminary connective class annotations have provided the connective class breakdown for the following ambiguous spans, given in Table 4[2]:

Span	Subordinating Conjunctive	Discourse Adverbial	Total
ardından 'following'	32	37	69
dolayısıyla 'as a result of'	2	64	66
önce 'first, before'	76	45	121
sonra 'than, later'	273	376	649

Table 4: Connective class disambiguation for ambiguous spans

3.2. ARG2FEAT (Feature Annotation for Second Arguments of Subordinators)

Most of the subordinating conjunctives in Turkish take nominalized clauses as their second arguments. These nominalizations can have a variety of morphological features, which makes the TDB a valuable source for studying nominalized abstract objects.

The morphological properties of the nominalized arguments also allow a further degree of disambiguation in case of *için* 'because, for'. *İçin* can express goal or cause driven relations. The sense of the relation can be disambiguated between goal and cause by simply looking at the morphology of the second argument. In (6), the *–mek için* marks a goal driven relation by taking an infinitival clause as argument, and in (7) *-dığım için* marks a cause driven relation by taking a factive clause (see also Table 5 below).

(6) **Onu görmek** için *tüm zamanınızı o parkta geçirmeye başlarsınız.*
In order to **see her** *you start to spend all your time in that park.*

(7) **Üvey babamı görmek istemediğim** için *yıllardır o eve gitmiyorum.*
Since **I don't want to see my step father,** *I haven't been to that house for years.*

Like CLASS, the ARG2FEAT is a relation attribute, which will be left blank for classes other than subordinating connectives.

A preliminary morphological annotation for (6) is INF which stands for infinitive, and for (7) FAC + AGR which stands for factive clause with person agreement. Other examples would be NOM MA + POS AGR + ABL CASE (nominalized with –mA, with person agreement on possessive case, attributed ablative case by the postposition) for … *olmalarından dolayı* 'although they are …', and CNV CA + DAT CASE (converb –cA, attributed dative case by the postposition) for *duyuncaya kadar* 'until hearing'.

[2] This table does not include parallel constructions and phrasal expressions including these spans, because their CONN spans already disambiguate their connective class; for instance the span *bunun ardından* 'following this' is unambiguously a phrasal expression as *ilk olarak...ardından* 'first…then' is a parallel construction.

Table 5 shows the disambiguation of *için* annotations in the TDB with respect to goal and cause driven relations.

Goal driven	
inf (-mAk) için	510
-mA + pos agr için	239
-mA için	6
-Iş + pos agr için	6
-Iş için	2
-Im + pos agr için	7
Goal Total	770
Cause driven	
-dIğI + agr için	276
- (A)cAğI + agr için	12
Cause total	288
İçin total	1058

Table 5: Goal - cause disambiguation for subordinator *için*

3.3. SENSE

Some connectives such as the subordinator *gibi* 'like, as/just as" cannot be disambiguated by morphology. *–dIğI gibi* marks an expansive relation in (8), a similarity relation in (9), and a temporal immediate succession relation in (10), with no morphological distinction on its argument.

(8) **Kahve değirmeninin nerede olduğunu bilmediği** gibi*, bulacağını da sanmıyordu*
In addition to **not knowing where the coffee mill is**, *he didn't think that he would be able to find it*.

(9) **Sizin yaptığınız** gibi *açık konuşacağım.*
I will speak frankly just like **you do.**

(10) **Bisikletine atladığı** gibi *pedallara basıyor.*
As soon as **he jumps on the bicycle**, *he hits the pedals.*

In addition to connectives like *gibi* that mark distinct sense classes such as EXPANSION and TEMPORAL relations, most connectives signal several types and subtypes of senses. For example, *ama* 'but' can signal CONTRAST, CONCESSION, EXCEPTION as well as PRAGMATIC variants of these senses.

For sense annotation we have taken the PDTB sense hierarchy (Prasad, 2007) as a starting point. Similar to Tonelli (2012), who discovered that the PDTB sense tags need to be expanded for spoken corpus annotations because of the extensive pragmatic uses, we have discovered that the rich variation of genres in the TDB calls for expansion of the sense hierarchy. In preliminary sense annotations, we have encountered a wide variety of pragmatic uses of *ama* 'but' including OBJECTION (11) and CORRECTION (12).

(11) - *Sana kahve yapacağım.* - Ama **çok içmedim.**
- *I will make you some coffee* - But **I haven't drunk much.**

(12) *Öyle bir kadın var!* Ama **o başkası!**
There is such a woman! But **she is someone else!**

The sense annotations are at a very early stage and the sense hierarchy is likely to be modified more as annotations progress.

4. Conclusion

In this paper we have introduced the features of the first release of the TDB. We also presented the ongoing developments for further enrichments, namely connective class annotation, Arg2 feature annotation and sense annotation.

The first two of these developments are well underway, and have already revealed detailed descriptives, such as the total connective class breakdown of disambiguated connectives in the TDB (Table 6). The number of distinct connectives increased from 143 to 150 (cf. § 2.2.1); because after the disambiguation processes, spans such as *ardından*-sub and *ardından*-adv or *için*-goal and *için*-cause are counted as distinct connectives.

		Single	**Parallel**	**Total**
Coordinating Conjunctive	Spans	15	12	27
	Relations	4348	129	4477
Subordinating Conjunctive	Spans	31	1	32
	Relations	2285	2	2287
Discourse Adverbial	Spans	32	18	50
	Relations	1152	73	1225
Phrasal Expression	Spans	40	1	41
	Relations	490	4	494
Total	Spans	118	32	150
	Relations	8275	208	8483

Table 6: Connective class breakdown of disambiguated connectives in the TDB

We believe that connective class, Arg2 feature, and sense annotations will contribute to the further study of Turkish in particular and provide a unique perspective to the studies in discourse in general.

5. Acknowledgements

We gratefully acknowledge the support of Turkish Scientific and Technological Research Council of Turkey (TUBITAK) and METU Scientific Research Fund (no. BAP-07-04-2011-005).

6. References

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.

Aktaş, B., Bozşahin, C., and Zeyrek, D. (2010). Discourse Relation Configurations in Turkish and an Annotation Environment. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV)*.

Demirşahin, I., Yalçınkaya, İ., and Zeyrek, D. (ms). Pair Annotation: Adaption of Pair Programming to Corpus Annotation.

Miltsakaki, E., Dinesh, N., Prasad, R., Joshi, A., and Webber, B. (2005). Experiments on sense annotation

and sense disambiguation of discourse connectives. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B. (2007). The Penn Discourse Treebank 2.0 Annotation Manual. Technical Report 203, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, Pennsylvania.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation. (LREC'08).*

Say, B., Zeyrek, D., Oflazer, K., and Özge, U. (2002). Development of a Corpus and a Treebank for Present-day Written Turkish. In *Proceedings of the Eleventh International Conference on Turkish Linguistics (ICTL 2002).*

Şirin, U., Çakıcı, R., and Zeyrek, D. (2012). METU Turkish Discourse Bank Browser. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

Tonelli, S., Riccardi, G., Prasad, R., and Joshi. A. (2010). Annotation of Discourse Relations for Conversational Spoken Dialogues. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC).*

Webber, B. (2004). D-LTAG: Extending lexicalized TAG to discourse. *Cognitive Science*, *28*(5), 751-779

Webber, B., and Joshi, A. (1998). Anchoring a lexicalized tree-adjoing grammar for discourse. In Stede, M., Wanner, L., Hovy, E. (Eds) *Discourse Relations and Discourse Markers: Proceedings of the Conference*, pp. 86–92. Association for Computational Linguistics.

Zeyrek, D., and Webber, B. (2008). A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Turkish Corpus. In *Proceedings of the 6th Workshop on Asian Language Resources, The Third International Joint Conference on Natural Language Processing (IJNLP)*.

Zeyrek, D., Demirşahin, I., Sevdik-Çallı, A., Balaban, H. Ö., Yalçınkaya, İ., and Turan. Ü. D. (2010). The Annotation Scheme of the Turkish Discourse Bank and an Evaluation of Inconsistent Annotations. In *Proceedings of the Fourth Linguistic Annotation Workshop* (*LAW IV).*