Discourse Relation Configurations in Turkish and an Annotation Environment

Berfin Aktaş[*] and **Cem Bozsahin**[*†] and **Deniz Zeyrek**[*‡]
[*] Informatics Institute, [†] Computer Eng. and [‡] Foreign Language Education Dept.
Middle East Technical University, Ankara Turkey 06531
berfinaktas@gmail.com {bozsahin,dezeyrek}@metu.edu.tr

Abstract

In this paper, we describe an annotation environment developed for the marking of discourse structures in Turkish, and the kinds of discourse relation configurations that led to its design.

1 Introduction

The property that distinguishes a discourse from a set of arbitrary sentences is defined as coherence (Halliday and Hasan, 1976). Coherence is established by the relations between the units of discourse.

Systematic analysis of coherence requires an annotated corpus in which coherence relations are encoded. Turkish Discourse Bank Project (TDB) aims to produce a large-scale discourse level annotation resource for Turkish (Zeyrek and Weber, 2008). The TDB follows the annotation scheme of the PDTB (Miltsakaki et al, 2004). The lexicalized approach adopted in the TDB assumes that discourse relations are set up by lexical items called discourse connectives. Connectives are considered as discourse level predicates which take exactly two arguments. The arguments are abstract objects like propositions, facts, events, etc. (Asher, 1993). They can be linked either by explicitly realized connectives or by implicit ones recognized by an inferential process. We annotate explicit connectives; implicit connectives are future work. We use the naming convention of the PDTB. Conn stands for the connective, Arg1 and Arg2 for the first and the second argument, respectively. Conn, Arg1 and Arg2 are assumed to be required components of discourse relations. Supplementary materials which are relevant to but not necessary for the interpretation are also annotated.

Our main data is METU Turkish Corpus(MTC) (Say et al, 2002). MTC is a written source of Turkish with approximately 2 million words. The original MTC files include informative tags, such as the author of the text, the paragraph boundaries in the text, etc. We removed these tags to obtain raw text files and set the character encoding of the files to "UTF-8". These conversions are useful for programming purposes such as visualizing the data in different platforms and the use of third-party libraries.

We developed an annotation environment to mark up the discourse relations, which we call DATT (Discourse Annotation Tool for Turkish). DATT produces XML files as annotation data which are generated by the implementation of a stand-off annotation methodology. We present in §2 the data from Turkish discourse, which forced us to use stand-off annotation instead of in-line markup. The key aspect is potential crossing of the markup links. However, stand-off annotation is also advantageous for separate licensing. We present the design of data structure and the functionality of the tool in §3. We report some preliminary results in the conclusion.

2 Dependency analysis of discourse relations

The TDB has no a priori assumption on how the predicates and arguments are placed. We need to take into account potential cases to be able to handle overlappings and crossings among relations. We use the terminology proposed by Lee et al (2006), and follow their convention for naming the variations of structures we came across.

We looked at the connective tokens placed close to each other, and made an initial investigation to reveal how these predicates and their arguments are located in the text. Preliminary analysis of the data indicates that the components of two relations are placed in 7 different ways, two of which are special to Turkish (§2.5; §2.6). This section is devoted to the descriptions of observed patterns with

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 202–206,
Uppsala, Sweden, 15-16 July 2010.

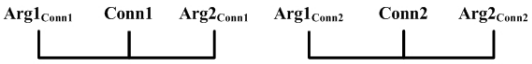
representative examples.[1]

In the examples the connective (Conn) is underlined, Arg1 is in *italics* and Arg2 is in **boldface**. A connective's relative order with respect to its own arguments is not shown in the graphical templates. It is made explicit in the subsequent examples.

2.1 Independent relations

The predicate-argument structure of the connectives are independent from each other (i.e., there is no overlap between the arguments of different connectives.) The template is (1). An example is provided in (2).

(1)



(2) *Akıntıya kapılıp umulmadık bir geceyi bölüştü benimle* ve **bu kadarla kalsın istedi belki.** *Eda açısından olayın yorumu bu kadar yalın olmalı.* Ama eğer böyleyse **benim için yorumlanması olanaksız bir düşten başka kalan yok geriye şimdi.**

She was drifted with a current and shared an unexpected night with me and **perhaps she wanted to keep it this much only.** *From the perspective of Eda, the interpretation of the incident should be that simple. But, if this is the case,* **now there is nothing left behind for me but a dream impossible to interpret.**

In (2), the relation set up by Ama is fully preceded by the relation set up by ve. There is no overlap between the argument spans of the connectives ve and Ama.

2.2 Full embedding

The text span of a relation constitutes an argument of another connective (3). An example is provided in (4).

(3)

(4)a. [..] madem **yanlış bir yerde olduğumuzu düşünüyoruz da doğru denen yere asla varamayacağımızı biliyoruz** , *senin gibi biri nasıl böyle bir soru sorar* ,[..]

b. [..] madem **yanlış bir yerde olduğumuzu düşünüyoruz** da *doğru denen yere asla varamayacağımızı biliyoruz* , senin gibi biri nasıl böyle bir soru sorar,[..]

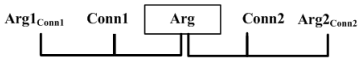
[..] if we think that we are in a wrong place, and we know that we will never never reach the right place; how come a person like you ask such a question? [..]

In (4), the span of the relation headed by da constitutes the Arg2 of the connective madem.

2.3 Shared argument

Two different connectives can share the same argument (5).

(5)



In some situations, different connectives can share both of their arguments as in the case of (6):

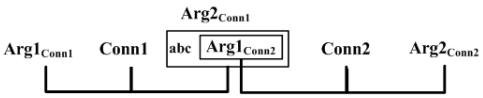
(6) Dedektif romanı içinden çıkılmaz gibi görünen esrarlı bir cinayetin çözümünü sunduğu için, *her şeyden önce mantığa güveni ve inancı dile getiren bir anlatı türüdür* ve bundan ötürü de **burjuva rasyonelliğinin edebiyattaki özü haline gelmiştir.**

Unraveling the solution to a seemingly intricate murder mystery, the detective novel is *a narrative genre which primarily gives voice to the faith and trust in reason* and being so, **it has become the epitome of bourgeois rationality in the literature.**

2.4 Properly contained argument

The argument span of one connective encapsulates the argument of another connective plus more text (7).

(7)



An example is provided in (8), where Arg2 of ve properly contains Arg1 of Tersine.

(8)a. *Kapıdan girdi* ve **söyler misin, hiç etkilenmedin mi yazdıklarından?, dedi.** Tersine, çok etkilendim.

b. Kapıdan girdi ve söyler misin, *hiç etkilenmedin mi yazdıklarından?,* dedi. Tersine, **çok etkilendim.**

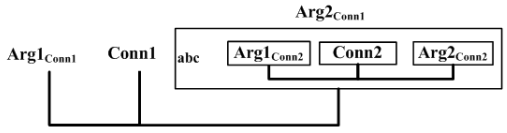
[1] All data in this paper are taken from MTC, unless stated otherwise. More examples can be found in Aktaş (2008).

S/he entered through the door and said “Tell me, are you not touched at all by what s/he wrote?”. On the contrary, I am very much affected.

2.5 Properly contained relation

The argument span of one connective covers the predicate-argument structure of another connective and more text (9), as exemplified in (10).

(9)



(10)a. *Burada bizce bir ifade bozukluğu veya çeviri yanlışı bahis konusu olabilir,* çünkü **elbiseler sanki giyildiği sürece ve yıpranmamışken yıkanamaz, fakat daha sonra yıkanabilirmiş gibi bir anlam taşımaktadır.**

b. Burada bizce bir ifade bozukluğu veya çeviri yanlışı bahis konusu olabilir, çünkü *elbiseler sanki giyildiği sürece ve yıpranmamışken yıkanamaz,* fakat **daha sonra yıkanabilirmiş** gibi bir anlam taşımaktadır.

Here a mistake of expression or mistranslation might be the case, because the meaning is as if the clothes cannot be washed as long as they are used and not worn out, but can be washed later.

In (10), the second argument of çünkü covers the whole relation headed by fakat and the text “gibi bir anlam taşımaktadır”, which is not part of it.

2.6 Nested relations

A relation is nested inside the span of another relation (11).

(11)

In (12), the relation headed by da is properly nested between the connective ve and its first argument.

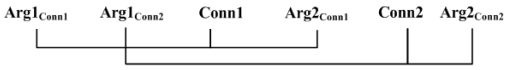
(12) *Büyük bir masada günlerce, gecelerce oturup konuşacağız - konuşmayı unuttum diyorum* da **gülüyorlar bana** - ve **biriniz kalkıp şiir okuyacak.**

We will sit and talk around a big table for days and nights - I say I have forgotten how to speak and they laugh at me - and one of you will stand up and recite poetry.

2.7 Pure crossing

The dependency structure of a relation interleaves with the arguments, or the connective of another relation (13), as exemplified in (14).

(13)



(14)a. (Constructed) *Kitabı okumaya başladım* : Okullar çoktan açılmıştı. Ardından **kapının çaldığını duydum ama yerimden kalkmadan okumaya devam ettim:** Ama bu okula henüz öğretmen atanmamıştı.

b. Kitabı okumaya başladım *Okullar çoktan açılmıştı.* Ardından kapının çaldığını duydum ama yerimden kalkmadan okumaya devam ettim: Ama **bu okula henüz öğretmen atanmamıştı.**

I started to read the book. The schools had long been opened. Then, I heard the door bell ring but I continued reading without getting up: But a teacher had not been appointed to this school yet.

3 The tool

DATT is an XML-based infrastructure for text annotation. It aims to produce searchable and trackable data. An initial investigation of connective and argument locations revealed that there is argument sharing of various sorts, and nested and crossing relations in Turkish discourse. The existence of such constructions lead us to use a stand-off annotation rather than an in-line method. These dependencies are violations of tree structure required by XML. Using the OCCURS feature of SGML for this purpose would lead to a less portable markup tool.

3.1 Data representation

In stand-off markup, annotations are stored separate from data. Since the base file is not modified during annotation, it is guaranteed that all the annotators are dealing with the same version of the data. The text spans of dependency constructions are represented in terms of character offsets from

the beginning of the text file. This is a highly error-prone way of storing annotation data. If there is a shift in the character indexes in the original text file, previously annotated data will be meaningless. To compensate for this, we keep the text spans of annotations for recovery purposes.

Annotation files are well-formed XML files. One can easily add new features to the annotations. XML facilities available as online sources such as the libraries for search and post-processing reduce the implementation effort of adding new features.

3.2 Search functionality

In the TDB, the annotation process is organized according to connective types and their tokens. The connective to be annotated is identified, and all the relations which are set up by the instances of that connective are marked. Therefore it is important to be able to find all the instances of a specific connective in the entire data source. DATT has a search functionality which walks through all resource files and shows the annotator which files have the token. We used "Apache Lucene Search Library" for this functionality.

Two distinguishing characteristics of Turkish, the vowel harmony and voicing, motivated us to enhance the search facilities by adding support for allomorphy. In Turkish, suffixes may have many different forms. The ability to search on these forms is crucial if connectives are attached to the inflected forms of words, which is very frequently the case. For instance, the "-dık"(the factive nominal) suffix has eight allomorphs (i.e. -dık, -dik, -duk, -dük, -tık, -tik, -tuk, -tük) depending on the phonological environment.

In Turkish discourse, the meaning of a connective may change according to the inflectional category of the word that precedes it. For example, the word just before the connective"için" can be inflected with "-dık" and "-mak"(the infinitive) suffixes. With "-dık" the connective bears the meaning of causal "since", while in the other case, the connective has the meaning of "so as to"(Zeyrek, Webber, 2008). Because of this semantic difference, it will be important for the annotator to cluster the instances of a connective token preceded by all the forms of a certain inflectional suffix in one search. DATT provides this opportunity with the allomorph search support.

In Turkish, connectives can be inflected. For example, the connective "dolayısıyla" (due to that) is the inflected form of "dolayı"(due to). The support for regular expression search is also added to DATT to retrieve the inflected forms of the same connective.

3.3 The user interface

The user interface of DATT is expected to allow the marking of dependency hierarchies mentioned in Section 2 in a user-friendly way. the TDB annotation requires at least three components, which are Arg1, Conn and Arg2. In DATT, in order to guide the annotator, we enforce the labeling of these mandatory components, while marking of the supplementary material is optional.

Another feature of DATT is the ability to mark discontiguous text spans as a unique relation, which is attested in Turkish discourse (15). Its connective-argument structure is shown below. The Arg1 of the connective -erek is interleaved with the second argument Arg2.

(15) *Yürü lan, dedi Katana,* **Ramiz'i kolundan** çekerek, *Miskoye korkuyo!*
"Hey you, move" said Katana, while dragging Ramiz by the arm, "Miskoye is freaked out."

Conn	Arg1	Arg2
-erek	Yürü ... Kat$ [5], Mis$ korkuyo	Ram$ ... çekerek

4 Conclusion

We adopt a lexical approach to discourse annotation. Connectives are words, and they take two text spans as arguments. An exploration of these structures shows that there is argument-sharing and overlap among relations. We are considering automatic detection of relation types for an appraisal of discourse relation distribution. For the time being, DATT has search support for allomorphy and regular expressions as an aid to finding the connectives.

Approximately 60 connective types and 100 tokens have been determined so far in the annotation process, using 3 annotators. 7,000 relation tokens headed by the connectives have been annotated using DATT, spanning approximately 300,000-word text. Work for agreement statistics is under way. We hope that machine learning techniques can discover more structure in the data once we have reasonable confidence with annotation.

[5]We use the notation "abc$" to refer to the word that begins with the string "abc".

References

Berfin Aktaş. 2008. Computational Aspects of Discourse Annotation. Informatics Institute, METU. *Unpublished master thesis*.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Alan Lee and Rashmi Prasad and Aravind Joshi and Nikhil Dinesh and Bonnie Webber. 2006. Complexity of dependencies in discourse. In *Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories*.

Eleni Miltsakaki and Rashmi Prasad and Aravind Joshi and Bonnie Webber. 2004. The Penn Discourse TreeBank. *LREC*, Lisbon, Portugal.

Bilge Say and Deniz Zeyrek and Kemal Oflazer and Umut Ozge. 2002. Development of a Corpus and a Treebank for Present-day Written Turkish. *11th International Conference on Turkish Linguistics*.

Deniz Zeyrek and Bonnie Webber. 2008. A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus. In *Proceedings of IJCNLP*.