Annotating Subordinators in the Turkish Discourse Bank

Deniz Zeyrek[a,0], **Ümit Turan**[b], **Cem Bozsahin**[a], **Ruket Çakıcı**[a],
Ayışığı Sevdik-Çallı[a], **Işın Demirşahin**[a], **Berfin Aktaş**[a], **İhsan Yalçınkaya**[a], **Hale Ögel**[a]
[a] Middle East Technical University, Ankara, Turkey
[b] Anadolu University, Eskişehir, Turkey

Abstract

In this paper we explain how we annotated subordinators in the Turkish Discourse Bank (TDB), an effort that started in 2007 and is still continuing. We introduce the project and describe some of the issues that were important in annotating three subordinators, namely *karşın*, *rağmen* and *halde*, all of which encode the coherence relation *Contrast-Concession*. We also describe the annotation tool.

1 Introduction

The Turkish Discourse Bank (TDB) is a project initiated by the joint effort of a group of researchers in Turkey. The project builds on an existing corpus, namely the METU Turkish Corpus (MTC) (Say et al., 2002), and extends it to a discourse level resource by following the principles of the PDTB (Prasad et al., 2007) in annotating discourse connectives and their arguments. The 2-million-word MTC contains 520 continuous texts from various genres written between 1991-2000.

From a semantic perspective, we take discourse connectives as predicates that take as their arguments tensed or untensed clauses with abstract object interpretations. Abstract objects are propositions, facts, events, situations, etc. (Asher, 1993). Connectives themselves may be realized explicitly or implicitly (Halliday, 1985; Prasad et al., 2007). Explicit connectives are simple or complex lexical items that encode a discourse relation, while implicit connectives can be inferred from related text spans that have coherence relations. The TDB project aims to annotate explicit connectives only.

In Turkish, discourse connectives are identified with three syntactic categories (Zeyrek and Webber, 2008): (a) Coordinating conjunctions (b) Subordinators (c) Discourse adverbials (or anaphoric connectives). All these discourse connectives have two and only two arguments, which are conveniently labeled as ARG1 and ARG2.[1] ARG2 is always the argument that syntactically hosts the connective.

The ARG1/ARG2 organization of discourse connectives is consistent with the following observations in discourse: Sentences in discourse are coherently related, and therefore when explicit discourse connectives are used, if they are really discourse connectives, they are bound to set up a relation between a consequent clause and its antecedent. (Note that the ARG2 designation does not imply that ARG2 is consequent or antecedent.) In certain cases presupposition needs a mediator, viz. the discourse connective. Nonconnectival discourse relations are certainly possible, but connective-engendered discourse relations are claimed to be more specific about their semantics, e.g. they bring about presuppositional meaning (van der Sandt, 1992; Webber et al., 1999).

In this regard, the ARG1/ARG2 classification is unlike syntactic subcategorization, which is a lexical property of functors (e.g. verbs) which are not necessarily presuppositional and hence they can differ in arbitrary ways (ditransitive, transitive, unergative, unaccusative etc.).

2 The Data

The MTC is preprocessed to obtain the raw texts keeping the title, author, publishing date and the text type information at the beginning of each file. Stand-off annotation is done on the textual rendering of the MTC.

To enable the data to be viewable universally without losing any character information, the file format (originally xcs) was converted to text, and the character encoding (originally Turkish-ISO-

[0]Corresponding author: dezeyrek@metu.edu.tr

[1]Whether or not discourse connectives in any language take more than two arguments is an open question that needs to be established in further research.

Text type	File Count	%	S1	%	S2	%	S3	%	S4	%
Novel	123	15.63%	31	15.74%	30	15.23%	31	15.82%	31	15.74%
Short story	114	14.49%	28	14.21%	29	14.72%	28	14.29%	29	14.72%
Research /Monograph	49	6.23%	13	6.60%	12	6.09%	12	6.12%	12	6.09%
Article	38	4.83%	9	4.57%	10	5.08%	9	4.59%	10	5.08%
Travel	19	2.41%	5	2.54%	5	2.54%	4	2.04%	5	2.54%
Interview	7	0.89%	2	1.02%	2	1.02%	2	1.02%	1	0.51%
Memoir	18	2.29%	4	2.03%	5	2.54%	5	2.55%	4	2.03%
News	419	53.24%	105	53.30%	104	52.79%	105	53.57%	105	53.30%
TOTAL	**787**		**197**		**197**		**196**		**197**	

Table 1: File count and percentage information according to text type for the preprocessed MTC and its subcorpora. (S:Subcorpus)

8859-9) was converted to the UTF-8. Finally, the processed MTC data were divided into four subcorpora by keeping the text type distribution, file count and word count as equal as possible in each subcorpus. The text type distribution, file count and word percentage information in each subcorpus are given in Table 1. In the project, we plan to annotate subcorpus 1.

3 Subordinating Conjunctions in Turkish: A Brief Overview

Subordinators have two subtypes. Converbs are suffixes attached directly to verb roots. For example, the suffix *-(y)ArAk* 'by (means of)' requires as its ARG2 a nominalized adverbial clause as in (1). Complex subordinators, e.g. *rağmen* 'despite, although', *karşın* 'although', *halde* 'despite, along with', *için* causal 'since', purposive 'so as to', etc. mostly take case-marked nominalized clauses as their ARG2.

(1) Hükümet ... **uyum paketini** onaylayarak ... *Erdoğan'n önündeki engellerden birini kaldırdı.*
By approving **the adaptation package ...**, the government *alleviated one of the obstacles for Erdoğan ...*

In this paper, we will not deal with converbs. We will also not deal with connectives taking as their ARG2 a finite clause because none of these subtypes have been annotated yet. We will focus on three postpositions taking a nominalized clause as ARG2, namely *rağmen*, *karşın* and *halde*, all of which encode the *Contrast-Concession* relation. In the PDTB, such clauses were not annotated as arguments. However, in Turkish, they are so common as arguments of subordinators that we would have missed an important property of Turkish discourse if we did not annotate them. In the rest of the paper, we provide examples taken from the MTC. We underline the connective, show ARG2 in bold letters and render ARG1 in italics.

3.1 The minimality principle

As in the PDTB, the minimality principle is invoked, according to which clauses, parts of clauses or sentences that are minimally necessary and sufficient for the discourse relation engendered by the connective are annotated as ARG1 or ARG2 (Prasad et al., 2007). Any other text span that is perceived to be important for the interpretation of the discourse relation can be selected as supplementary information in addition to ARG1 or ARG2.

3.2 Morphological properties of the arguments and their relative ordering

In Turkish, subordinate clauses are essentially nominalizations, which may be formed by *-DIK* or *-mA* suffixes (the factive nominal and the action nominal, respectively (Kornfilt, 1997)).

Two of the connectives, i.e. *rağmen* and *karşın* expect action nominals, the person agreement suffix, and the dative suffix *-(y)A* on their ARG2. On the other hand, the connective *halde* expects a factive nominal and the person agreement suffix. In the examples below, we show these suffixes with glosses on the English translations.

The arguments of subordinators are necessarily adjacent and mostly exhibit the ARG2-ARG1 order because Turkish is a left-branching language and subordinate clauses are on the left in canonical order. ARG2 can be postposed for backgrounding purposes or to express new and unexpected information, as in (2).

(2) ... aynı annesine oldu]ugu gibi ona da, *kimseye ba/uglanmayanlar kolayca bağlanıyordu*; üstelik **o öyle bir bağımlılık talep etmediği** halde.
... just as it happened to her mother, *people who can't easily commit themselves to anyone would*

easily commit themselves to her, although **she would not ask-FACTN-AGR for such a commitment**.

3.3 Issues in annotating the arguments

One of the challenges we have faced so far is the question of how to annotate connectives which are themselves a converb suffix (e.g. *-(y)ArAk*, as in (1)) or postpositions that choose a case-marked ARG2 as in (2). In both cases, we decided to annotate ARG2 by selecting the clause without separating the suffixes. In this way, we would not interfere with the annotators' intuitions since we would not be demanding them to have conscious knowledge of the morphological constraints on the arguments. This style of annotation was welcomed by the annotators. When all the annotations are completed, we plan to separate the suffixes with a morphological parser to provide a full view of the morphology of the arguments.

Another issue was how to annotate shared subjects in subordinate clauses. Turkish allows subject pro-drop and in complex sentences, the shared subject is shown by the person agreement suffix on the verb of the consequent clause. To capture this fact, we chose to exclude shared subjects from the annotation of the arguments. This style of annotation conforms to the minimality principle. As illustrated in (3), the subject, *Neriman*, which appears in its canonical clause-initial position in ARG2 is not selected because the verb of the subsequent clause carries the person agreement suffix.

(3) Neriman **yatak odasında sigara içilmesini istemediği** halde *şimdilik sigaraya ses çıkarmıyor*.
Although Neriman **does not want-FACTN-AGR people to smoke in her bedroom**, *(she) doesn't say-AGR anything for the moment*.[2]

If the subject is not shared, it is included in the annotation, even if it causes discontinuity. As it is illustrated in (4), ARG2 intervenes in ARG1 by separating it from its subject.

(4) *Rukiye*, **kendisinden üç yaş ufak olmasına** rağmen, *erkek kardeşini kendi oğlu sanıyordu*, ...
Rukiye, although **(he) is-ACTN-AGR-DAT three years younger than herself**, *thought-AGR that her brother was her son*...

[2]The pronoun is in parentheses to reflect pro-drop. The following abbreviations are used on the translations to show the morphological characteristics of the clauses: ACTN: Action nominal, FACTN: Factive nominal, AGR: Person agreement suffix, DAT: Dative case, ABL: Ablative case. NOM: Nominative case.

Example (5) shows that two nominalized clauses can be selected as the arguments of the subordinator *karşın* leaving out the shared subject. In this example, the subject is shown between square brackets for clarity. Note that, ARG1 is also a nominalized clause since it is embedded under the attribution verb *söyle* - 'say'.[3]

(5) ... [herkes yaratılan toplumsal değerden verdiği emek oranında pay alacak biçimindeki sosyalist iktisat ilkesinin] **aslında çok eşitlikçi gibi gözükmesine** karşın *eşitsizliği engelleyemeyeceğini*, ... söyler
... says that ... despite **(it) looks-ACTN-AGR-DAT quite egalitarian**, [the socialist principle, stating that everyone gets a share proportional to his labor] *will not prevent-ACTN-AGR inequality* ...

Finally, in annotating adjuncts, we follow the same principle we followed in annotating shared subjects. For instance in (6), the adjunct *yemekte* 'at dinner' is not annotated since it is shared by the arguments of the connective *rağmen*.

(6) *Gül* de yemekte **kilo aldırmasına** rağmen *Şam tatlılarından çok hoşlandığını* ifade etti.
At dinner, *Gül-NOM*, also said that although **(they$_i$) are-ACTN-AGR-DAT fattening**, *(he) likes Damascus deserts-ABL$_i$ very much*.

4 The Annotation Process

Before the annotation procedure started, a set of annotation guidelines were prepared. The guidelines include the basic principles, such as what discourse connectives are, where in the discourse one can find ARG1 and ARG2, how to annotate shared subjects, adjuncts, etc. Rather than being strict rules, the guidelines are aimed at being general principles guiding the annotators in their decision of selecting the text span that is minimally sufficient for interpreting the discourse relation encoded by the connective.

The annotation cycle consisted of 1) annotating a connective by at least three different people 2) measuring the agreement among them with the inter-anotator agreement tool 3) resolving the disagreements with an anonymous decision.

4.1 The annotation tool

We have an XML-based infrastructure for annotation. It aims to produce searchable and trackable data. Stand-off annotation has well-known advantages such as the ability to see layers separately, or

[3]In the PDTB, attribution is not taken as a discourse relation but it is annotated. Attribution is not annotated in the TDB.

	Overall			Annotator1			Annotator2			Annotator3		
Conn.	ARG1	ARG2	Overall	ARG1	ARG2	Overall	ARG1	ARG2	Overall	ARG1	ARG2	Overall
rağmen	0.37	0.343	0.444	0.476	0.493	0.538	0.810	0.889	0.83	0.591	0.550	0.660
karşın	0.394	0.546	0.364	0.771	0.781	0.724	0.677	0.833	0.71	0.677	0.62	0.676
halde	0.749	0.826	0.758	0.957	1	0.978	0.772	0.826	0.758	-	-	-

Table 2: Textspan inclusion agreement among three annotators for three subordinators with minimum success prob. >0.05. The first column shows the overall agreement among the three annotators. Other columns show the agreement of one annotator with the agreed/gold standard annotations. For *halde*, 2 annotators performed a common annotation (given as Annotator1) and a third annotator annotated it separately (given as Annotator2).

to distribute annotation without data due to licensing constraints. To this list we can add the empirical necessity that, the crossing links in a single layer of same kind of annotation might not be easy to do inline. They can be done inline using SGML OCCURS checks, but they are easier to annotate in stand-off mode.

The tool has a regular expression mode in which the annotator can use his/her knowledge of Turkish word structure to collect similarly inflected words without morphological analysis. For example, -ArAk$, in which the uppercase forms represent metaphonemes, will bring words ending with the allomorphs of the converb suffix due to vowel harmony: *erek, arak* etc.

5 Conclusion

The TDB project is a first attempt in creating an annotated resource for written Turkish discourse. The annotation process is still continuing. In this paper, the emphasis was on a small number of connectives, namely three postpositions, which form a subclass of subordinators. The paper described the role of certain morpho-syntactic facts in interpreting the coherence relation between two clauses, and how these facts were reflected in the annotations.

Three subjects separately annotated each of the subordinators on the annotation tool, and inter-rater reliability was calculated. The statistics were obtained from Cochran's Q test to the ARG1 and ARG2 spans. The annotation data were encoded with 1 if the character is in the span and 0 if it is not. The encoded data were put to the Q test. All the results were above the minimum success probability (>0.05), showing that the annotations were consistent (see Table 2). We will run another Cochran experiment in which we will test whether the annotators agree on ARG1/ARG2 boundaries, rather than just word inclusion in the text spans as above.

Given the distribution of agreements, Cochran provides the number of subjects who must agree so that a text span can be reliably considered an ARG1 or ARG2. This we believe is important to report with the final product (to be made public soon), so that its gold standard can be assessed by the community.

Acknowledgments

We thank TUBITAK for financial support. We also thank our two anonymous reviewers for their insightful comments.

References

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

Michael A. K. Halliday. 1985. *An Introduction to Functional Linguistics*. Edward Arnold Publishers Ltd.

Jaklin Kornfilt. 1997. *Turkish*. Routledge, London.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. 2007. The penn discourse treebank 2.0. annotation manual. Technical report, Institute for Research in Cognitive Science, University of Pennsylvania, March.

Bilge Say, Deniz Zeyrek, Kemal Oflazer, and Umut Özge. 2002. Development of a corpus and a treebank for present-day written turkish. In *Proceedings of the Eleventh International Conference of Turkish Linguistics*.

Rob van der Sandt. 1992. Presuppositon projection as anaphora resolution. *Journal of Semantics*, 9:333–377.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999. Discourse relations: A structural and presuppositional account using lexicalised tag. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 41–48, College Park, Maryland, USA.

Deniz Zeyrek and Bonnie Webber. 2008. A discourse resource for turkish: Annotating discourse connectives in the metu turkish corpus. In *The 6th Workshop on Asian Language Resources, The Third International Joint Conference on Natural Language Processing (IJNLP)*, Hyderabad,India, January.