THE ROLE OF ANNOTATION IN UNDERSTANDING DISCOURSE

Deniz Zeyrek*, Ümit Deniz Turan+, Cem Bozşahin*
*Middle East Technical University, +Anadolu University

1 Introduction

In this paper, we introduce our effort of building a resource by **e**xpanding an existing resource (METU Turkish Corpus) from a sentence-level resource to a discourse-level resource. The project shares the goals of the PDTB (www.seas.upenn.edu/~pdtb), a resource for English, which has been annotated for explicit and implicit connectives, their arguments and senses. The paper explains the linguistic goals of the project and describes how the annotated corpus can help us understand discourse structure. [1]

Our approach to discourse structure is corpus-based. As we briefly explain below, we have started our investigation of discourse connectives with a particular discourse theory, namely D-LTAG, but we are continuing largely in a theory-neutral way since our goal is to be able to empirically assess specific descriptions of discourse connectives with data from METU Turkish Corpus.

2 What Are Discourse Connectives?

2.1 Discourse Relations

Theories of discourse describe a text as coherent when it contains sentences that are related to each other in some way. For example, example (1a) is a coherent discourse because the sentences are closely related: *Ahmet* drank a glass of water because he was thirsty. In contrast, (1b) is incoherent as there is no conceivable relation between *Ahmet*'s being thirsty and the subsequent sentence about all countries being affected by the global economic crisis. The relations that hold between sentences are referred to as coherence relations, conjunctive relations, rhetorical relations or discourse relations. We will use the term discourse relations.

(1) a. Ahmet çok susamıştı. Kalkıp bir bardak su içti.
'Ahmet was very thirsty. He got up and drank a glass of water.'

b. # Ahmet çok susamıştı. Global ekonomik krizden tüm ülkeler etkileniyor.
'Ahmet was very thirsty. All countries are affected by the global economic crisis.'

Discourse relations may be semantic/informational or intentional. Example (1a) contains a

discourse whose sentences are related semantically/informationally. In contrast, example (2) is a discourse having an intentional discourse relation in which the intended effect of A's question is satisfied by B's answer.

(2) A. İndirime giren portakallar nerede?
'Where are the oranges with reduced price?'
B. Kaç kilo istiyordunuz?
'How many kilos did you want?'

We are not concerned with intentional relations in discourse; instead, we deal with the structural and semantic aspects of relations such as Concession, Contrast, Addition, Expansion, Result, Exemplification, Conditional, and Temporal etc. as elements of discourse coherence. We take connectives (e.g., *çünkü* 'since, because', *rağmen* 'although', *ama* 'but', *ve* 'and', etc.) as explicit markers signalling discourse relations, just like many discourse theories do (e.g., Halliday & Hasan 1976, Knott 1996, Kehler 2002, among others).

2.2 Discourse Connectives and Non-Discourse Connectives

Following D-LTAG introduced below briefly, we take a Discourse Connective (DC) as a predicate that takes as its arguments Abstract Objects (AOs) (Asher, 1993). AOs are eventualities (events and states), fact-like objects (situations, facts, possibilities), and proposition-like objects (questions, desires, commands, propositions).

Based on D-LTAG, we classify DCs in Turkish into two broad categories, as follows.

- Connectives which syntactically relate clauses: These include coordinating conjunctions (e.g. , pure connectives such as *ama 'but', fakat 'but', ve 'and'*), paired conjunctions (*hem … hem 'both and', ne … ne 'neither nor'*), and a range of complex (-A *rağmen* 'although', -dAn *dolayı* 'since') and simplex subordinators (*-(I)ken* 'while' (temporal), -ErEk 'by').
- Discourse adverbials, i.e., lexical items or phrases which signal relations between the semantic content of the clause they appear in and a second adjacent or nonadjacent clause (*aksine* 'on the contrary', *üstüne üstlük* 'and what's more') (Zeyrek & Webber 2008, Webber et al., 2003).

The basic linguistic unit in which AOs are realized is the clause, either tensed or untensed. Asher's definition of AOs, therefore, allows us to distinguish between DCs and non-Discourse Connectives (NDCs). For example, the connectives in (3) and (4) are NDCs since they link NPs rather than AOs. The connective in example (5) is taken to be a DC as both conjuncts of the connective *önce* 'before' have the AO interpretation of eventuality.

(3) Ali yemekten önce bir viski içmek istedi.
'Ali wanted to have whiskey before dinner.'

(4) Ali ve eşi Zeynep tatile çıktı.
'Ali and his wife Zeynep went on a holiday.'

(5) i. Yataktan kalkmadan önce
'before I woke up'
ii. kapı zilinin çaldığını duydum.
'I heard the bell ring.'
'I heard the bell ring before I woke up.'

Discourse connectives such as *aksine* 'on the contrary', *sonuçta* 'as a result', *yoksa* 'otherwise', etc. should be distinguished from all sentence modifying adverbials, such as sentential disjuncts *maalesef* 'unfortunately', *herhalde* 'perhaps', etc. and sentential adjuncts such as *dün* 'yesterday'. Naturally, all discourse connectives have two arguments, while sentential adverbials do not. There may, however, be cases in which one adverbial can be used both at clausal level and as a discourse connective. For example, *kısaca* 'shortly, briefly' can be used as a VP or sentential adjunct, or it can be used as a connective that links the previous discourse to its second argument in terms of a summarizing coherence relationship. Its function can be determined depending on its argument structure within its discourse context in a D-LTAG analysis. As a clausal modifier, it is of relevance to the syntactic level of analysis and its structure and meaning will not be dealt within the current project.

Cue phrases (*yani, işte)*, i.e., phrases used for management of discourse are excluded since we are mainly concerned with monologic and written discourse; and cue phrases are basically used in conversation, their analysis is within the scope of conversational analysis.

2.3 Arguments of discourse connectives

The starting point of our project is a lexicalized grammar which takes connectives as heads with binary arguments (see section 3.0). By the definition of the PDTB (Prasad et al., 2007), one argument is the clause that syntactically hosts the connective; the other argument comes from an adjacent or non-adjacent clause or clauses. In the absence of any criteria that can characterize the arguments of a connective in semantic terms (such as agent and patient), the PDTB gives convenient labels to the arguments, namely ARG2 (the host clause) and ARG1 (the other clause). This is the convention we also follow in our annotation effort. In the examples below, the connectives are underlined, ARG1s are shown in italics, ARG2s are typed in bold letters.

In examples (6) through (9), we show where DCs take their arguments from. In example (6), there are two DCs: the complex subordinator *–A rağmen* 'although, despite' (concession), and the simplex subordinator *–ErEk* [2]. In our annotation, the property of concession is lexicalized to this use of *rağmen*, i.e. it is part of the meaning of the connective; there may be other uses of *rağmen*. Clauses (ii) and (iii) together are ARG1 of *rağmen*, as shown in (6a). Clause (ii) is shared by the connectives: It constitutes (part of) *rağmen*'s ARG1 and*–ErEk's* (full) ARG2, as shown in (6b). Such cases of argument sharing are abundant in METU Turkish Corpus. We believe that our main emphasis on ARG1-CONN-ARG2 arrangement of discourse connectives

might bring out the intercalations and nested interactions of discourse structure, and we would be surprised to find out in discourse the kinds of constituencies one sees in syntax, such as unbounded crossing dependencies as observed in Dutch and Swiss-German. The argument overlaps and intercalations we have reported as in (6) are of bounded nature as far as we can tell.

(6a) i. **Gül Ahmet'e aşık olmasına** rağmen
'Although Gül was in love with Ahmet
ii. *onunla evlenmeyerek*
'by *not marrying him'*
iii. *herkesi şaşırttı.*
she surprised everyone.'
'Although Gül was in love with Ahmet, *she surprised everyone by not marrying him*.'

(6b) i. Gül Ahmet'e aşık olmasına rağmen
'Although Gül was in love with Ahmet
ii. **onunla evlenmey**erek
'by not marrying him'
iii. *herkesi şaşırttı.*
[she] surprised everyone.'
'Although Gül was in love with Ahmet, *she surprised everyone by* **not marrying him.'**

3 A Lexically Grounded Theory - Lexicalized Tree-Adjoining Grammar for Discourse (D-LTAG)

Tree Adjoining Grammar (TAG) is a formalism which uses trees as the fundamental building blocks of grammar (Joshi, 1985). A TAG consists of a finite set of elementary trees which can be combined with a substitution or an adjunction operation. Elementary trees are divided into two: initial trees and auxiliary trees. Initial trees represent the subcategorization relations, i.e. the obligatory predicate–argument structure; whereas auxiliary trees are used to illustrate recursion and optional constituents such as adjuncts. A derived tree is obtained from initial trees via either substitution or adjunction operations. Substitution simply replaces a substitution node in a tree with the same label. Adjunction is an operation through which an auxiliary tree is inserted into another tree. Taking into consideration the central role of lexicon in syntax, Shabes (1990) developed a lexicalized version of TAG, Lexicalized Tree Adjoining Grammar (LTAG). In LTAG each entry in the lexicon is associated with a finite set of tree structures which represent the related syntactic configurations. For example, the verb *yemek* 'eat' is associated with at least three diagrams shown in Figure 1, which represent *Ali pastasını yedi*, 'Ali ate his cake' and its permutations *Pastasını yedi Ali* and *Ali yedi pastasını*, respectively.

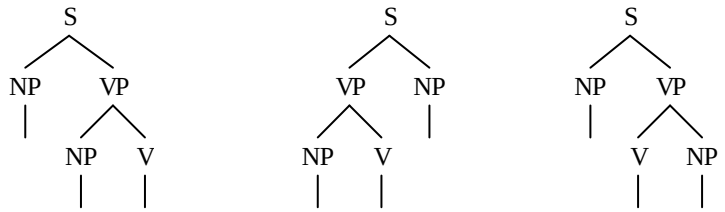

Figure1: Tree structures related with the verb *yemek* 'eat'

The leaves in these tree diagrams are substitution sites as they are reserved for constituents within the subcategorization frame of the verb. A manner adjunct PP such as *acele ile* 'with haste' can only be adjoined through an adjunction operation because it is not in the locality of the basic clause structure with its predicate and arguments.

LTAG is extended with the aim of analyzing the structure and semantics of low-level discourse (Webber and Joshi, 1998). This approach is known as Discourse Lexicalized Tree Adjoining Grammar (D-LTAG). It is by now a well-known fact that discourse is structure-dependent just like a clause is (Grosz and Sidner 1986, Mann and Thompson 1988, Polanyi 1995, among others). D-LTAG researchers (e.g. Prasad, et.al, 2007) have observed that the structure of discourse can be analyzed just like the structure of a clause. While the unit of syntactic analysis is limited to the clause-level, the unit of analysis of discourse grammar is beyond the clause. As has been already stated, in LTAG a verb — as a lexical item —is associated with various trees. The verb as the predicate selects its arguments to form a minimal clause. Likewise, D-LTAG conjectures discourse connectives as predicates that select their arguments which are clausal constituents denoting abstract objects in the sense of Asher (1993). The basic indivisible unit in D-LTAG is discourse clause (D_c). D-LTAG, like LTAG has a set of tree structures that are associated with –or anchored by— discourse connectives as predicates. For example, *çünkü* 'because' can have the following two tree structures:

Figure 2: Tree structures associated with the connective *çünkü* 'because'

Since connectives select their clausal arguments in D-LTAG, even subordinate conjunctions anchor initial trees through substitution operation.

D-LTAG views that syntax and compositional semantics, as well as inferences based on presupposition contribute to discourse coherence but in different ways as described above.

D-LTAG neither deals with global discourse structure nor with question-answer pairs, conversation or speaker intentions. We do not consider that these aspects of discourse are unimportant, but rather they are only out of the scope of D-LTAG. Thus, D-LTAG is not in competition with theories that deal with these issues, but it attempts to analyze low-level discourse structure and meaning to provide a substratum for theories that aim for a higher level discourse analysis. This approach provides a uniform way of analyzing clause and discourse structure and semantics as well as allowing a way of elaborating our understanding on the syntax–discourse relationship and interaction. It also provides a relatively straightforward and reliable way of annotating a large corpus in terms of local discourse structure.

4 The Steps in Annotation

The first step will be to annotate explicit DCs, which we see as a major corpus annotation effort in its own right. Implicit connectives, i.e., the conditions under which a discourse relation can be inferred from AO interpretations of the clauses will be integrated at a later stage.

Annotation will be done by trained human annotators. As in the PDTB, annotators will be given a list of connectives and will be asked to go through the corpus, annotating one connective at a time. In this way annotation will be quick as annotators will instantly take advantage of the skill they are gaining in annotating one connective (Webber et al., 2005). The alternative approach, i.e. asking annotators to proceed by annotating all the connectives in the list might easily lead to inconsistencies even in the annotations of a single annotator.

Annotators will use a simple annotation tool, which searches and highlights all examples of a certain connective. The tool will also include functions like CONN (connective), ARG1 (argument 1), ARG2 (argument 2), SUPP1 and SUPP1 (supplementary material for the first and the second arguments), and SENSE (of the connective). Annotators will proceed by accepting the highlighted connective as a DC and marking its arguments and sense/s, or they will reject it as a DC. These steps will go on until all the connectives in the list finish. Annotators will then proceed with the next list of connectives.

Once one phase of annotations is completed, inter-annotator agreement statistics will be run. The results of agreement statistics will inform us about the accuracy of the annotations. We will report the use metrics that can tell us the number of annotators required to reach a reliable conclusion, such as Cochran's Partial Q. We plan to run inter-agreement statistics at regular intervals to assure the reliability of annotations. We will make the corpus publicly available along with its reliability statistics. The first release of the annotated corpus is planned for December 2010.

5 Expected Gains

We hope the annotated data will reveal some facts about Turkish discourse and lead to further inquiry, both empirical and theoretical. For example, argument-taking characteristics of Turkish discourse connectives seem to rely quite heavily on morphology, as exemplified by *–A rağmen* and *–ErEk*: In the former example, the connective is a separate word and it subcategorizes for a factive clause with dative marking. In the latter example, the connective itself is morphologically bound.

Theoretical questions will undoubtedly arise. For example, our discussion so far might suggest that we take the presence of connectives in discourse as an empirical fact. We do not. We want to find out whether this is true or not, and any comparison with e.g., lexical cohesion theories of discourse such as Halliday and Hasan's (1976) and presuppositonal/logical models (e.g. Asher 1993) require that we establish an empirical base on which these theories can be tested. We see the annotated corpus as a resource toward that goal.

Adopting a certain view of discourse where the relations are established by connectives via their lexicalized categories is intended as a way of ensuring consistency of the annotation. It is already suggestive that all connectives have a binary argument structure in our annotation, which is quite unlike the constituent structure in syntax since some predicates may have one argument in a clause (e.g. intransitive verbs) or they may three (e.g. the verbs *give, put,* etc.). If we encounter unary and tertiary structures in discourse, arguments for a discourse grammar can be made more forcefully, otherwise it is a weak argument.

Comparisons with theories based on lexical cohesion will cross-fertilize further annotation work. As expected, such theories would expect semantic relations among words (hyponymy, meronymy etc.), rather than argument structure of connectives, to do the work of cohesion in discourse. They use resources such as WordNet. In a related study, Turhan-Yöndem and Bozşahin (2008) have found out that both grammatical (e.g., choice of referring expressions) and lexical-relational cues are good indicators of organization of discourse. Investigating whether these lexical relations concentrate between ARG1-ARG2 of our annotation requires both kind of annotation for the same data.

References

Andrew, K. (2002). *Coherence, Reference, and the Theory of Grammar***,** Stanford: CSLI Publications.

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

Dinesh, N., Lee, A., Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2005). Attribution and the (Non-) Alignment of Syntactic and Discourse Arguments of Connectives. *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky.* Ann Arbor, Michigan. June 2005.

Forbes-Riley, K., Webber, B., Joshi, A. (2006). Computing Discourse Semantics: The Predicate-Argument Semantics of Discourse Connectives in D-LTAG. *Journal of Semantics* 23 (1), 55—106.

Göksel, A., and Kerslake, C. (2005). *Turkish. A Comprehensive Grammar*. London and New York: Routledge.

Grosz, B. and C. Sidner. (1986). Attention, Intention and the Structure of Discourse. *Computational Linguistics, 12*(3), 175–204.

Halliday, M.A.K., Hasan, R. (1976). *Cohesion in English*. Oxford University Press.

Joshi, A. (1985). How Much Context-sensitivity is Necessary for Characterizing Structural Descriptions: Tree Adjoining Grammars. In Dowty, D. Karttunen, L., and Zwicky, A. (Eds.) *Natural Language Parsing,* Cambridge: Cambridge University Press, 206-250

Kerslake, C. (1996). The Role of Connectives in Discourse Construction in Turkish. Konrot, A. (Ed.). *Modern Studies in Turkish. Proceedings of the 6th International Conference on Turkish Linguistics. 12-14 August, 1992.* Eskişehir, Turkey,. 77-104.

Knott, A. (1996) *A Data-Driven Methodology for Motivating a Set of Coherence Relations.* PhD Dissertation. , Department of Artificial Intelligence, University of Edinburgh.

Mann, W. and Thompson, S. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text, 8*(3), 243–281.

Polanyi, L. (1995). The Linguistic Structure of Discourse. Retrieved from: http://cslipublications.stanford.edu/papers/ldmjuly.pdf

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A. (2007). The Penn Discourse Treebank Annotation Manual 2.0. The PDTB Research Group. (www.seas.upenn.edu/~pdtb).

Schabes, Y. (1990). *Mathematical and computational aspects of lexicalized grammars*. Ph.D. Dissertation, Department of Computer and Information Science, University of Pennsylvania.

Turhan-Yöndem, M., Bozşahin, C. (2008). Lexical and Grammatical Cohesion in Written Turkish Text. In submission.

Webber, B., Joshi, A., Miltsakaki, E. Prasad, R., Dinesh, N., Lee, A., Forbes, K. (2005) A Short Introduction to Penn Discourse TreeBank. *Copenhagen Working Papers in Language and Speech Processing.*

Webber, B. (2004) D-LTAG: Extending Lexicalized TAG to Discourse. *Cognitive Science* 28, 751–779.

Webber, B., Joshi, A., Stone, M., and Knott, A. (2003). Anaphora and Discourse Structure. *Computational Linguistics* 29 (4) 547-588.

Webber, B. and Joshi, A. (1998). Anchoring a Lexicalized Tree-Adjoining Grammar for discourse. In *COLING / ACL Workshop on Discourse Relations and Discourse Markers* (pp. 86–92).

Zeyrek, D. and Webber, B. (2008) A Discourse Resource for Turkish: Annotating Discourse Connectives in The METU Turkish Corpus. *Proc. of the 6th Workshop on Asian Language Resources. The Third International Joint Conference on Natural Language Processing (IJCNLP). January 11- 12 2008, Hyderabad, India.*

---

[1] The project is supported by TÜBİTAK:107E156 and has started in 2007.

[2] For practical reasons, the case suffix associated with the subordinator will not be shown in the rest of the paper.