A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus

Deniz Zeyrek
Department of Foreign Language Education
Middle East Technical University
Ankara, Turkey
dezeyrek@metu.edu.tr

Bonnie Webber
School of Informatics
University of Edinburgh
Edinburgh, Scotland
bonnie@inf.ed.ac.uk

Abstract

This paper describes first steps towards extending the METU Turkish Corpus from a sentence-level language resource to a discourse-level resource by annotating its discourse connectives and their arguments. The project is based on the same principles as the Penn Discourse TreeBank (http://www.seas.upenn.edu/~pdtb) and is supported by TUBITAK, The Scientific and Technological Research Council of Turkey. We first present the goals of the project and the METU Turkish corpus. We then describe how we decided what to take as explicit discourse connectives and the range of syntactic classes they come from. With representative examples of each class, we examine explicit connectives, their linear ordering, and types of syntactic units that can serve as their arguments. We then touch upon connectives with respect to free word order in Turkish and punctuation, as well as the important issue of how much material is needed to specify an argument. We close with a brief discussion of current plans.

1 Introduction

The goal of the project is to extend the METU Turkish Corpus (Say et al, 2002) from a sentence-level language resource to a discourse-level resource by annotating its discourse connectives, and their arguments. The 2-million word METU Turkish Corpus (MTC) is an electronic resource of 520 samples of continuous text from 291 different sources written between 1990-2000. It includes multiple genres, such as novels, short stories, newspaper columns, biographies, memoirs, etc. annotated topographically, i.e., for paragraph boundaries, author, publication date, and the source of the text. A small part of the MTC, called the METU-Sabancı TreeBank (5600 sentences) has been annotated with morphological features and dependency relationships (e.g., modifier-of, subject-of, object-of, etc.). The result is a set of dependency trees. The MTC as a whole provides a large-scale resource on Turkish discourse and is being used in research on Turkish. To date, there have been 81 requests for permission to use the MTC and 31 requests to use the TreeBank sub-corpus. Most of the users are linguists, computer or cognitive scientists working on Turkish, or graduate students of similar disciplines. Some users have expressed a desire for the MTC to be extended by annotations at the discourse level, which provides further impetus for the present project.

The result of annotating discourse connectives will be a clearly defined level of discourse structure on the MTC. Annotation of text from the multiple genres present in the MTC will allow us to compare the distribution of connectives and their arguments across genres. The annotation will help researchers understand Turkish discourse by enabling them to give concise, clear descriptions of the issues concerning discourse structure and semantics, and support a rigorous empirical

characterization of where and how the free word-order in a language like Turkish is sensitive to features of the surrounding discourse. It can thus serve as a major resource for natural language processing, language technology and pedagogy.

2 Overview of Turkish Discourse Connectives

From a semantic perspective**,** a discourse connective is a predicate that takes as its arguments, abstract objects (propositions, facts, events, descriptions, situations, and eventualities). The primary linguistic unit in which abstract objects (AOs) are realized in Turkish is the clause, either tensed or untensed. Discourse connectives themselves may be realized explicitly or implicitly. An explicit connective is realized in the form of a lexical item or a group of lexical items, while an implicit connective can be inferred from adjacent text spans that realise AOs and whose AOs are taken to be related. To constrain the amount of text selected for arguments, a *minimality principle* can be imposed, limiting arguments to the minimum amount of information needed to complete the interpretation of the discourse relation. The project will initially focus on annotating explicit connectives, integrating implicit ones at a later stage.

One of the most challenging issues so far has been determining the set of explicit discourse connectives in Turkish (i.e., the various linguistic elements that can be interpreted as predicates on AO arguments) and the syntactic classes they are identified with. In the Penn Discourse TreeBank (PDTB), the explicit discourse connectives were taken to comprise (1) coordinating conjunctions, (2) subordinating conjunctions, and (3) discourse adverbials (Forbes-Riley et al, 2006). But coordinating and subordinating conjunctions are not classes in Turkish *per se*. Moreover, most of the existing grammars of Turkish describe clausal adjuncts and adverbs in semantic (e.g., temporal, additive, resultative, etc.) rather than syntactic terms. We therefore made a rough classification first and determined the broad syntactic classes by considering the morpho-syntactic properties shared by elements of the initial classification.

As a result of this process, we have come to identify explicit discourse connectives in Turkish with three grammatical types, forming five classes:

(a) Coordinating conjunctions such as single lexical items *çünkü* 'because', *ama* 'but,' *ve* 'and', and the particle *dA*. (N.B., *dA* can also function as a subordinator.)

(b) Paired coordinating conjunctions such as *hem .. hem* 'both and,' *ne .. ne* 'neither nor' which link two clauses, with one element of the pair associated with each clause in the discourse relation.

(c) Simplex subordinators (also termed as converbs), i.e., suffixes forming non-finite adverbial clauses, e.g. *–(y)kAn*, 'while', *-(y)ArAk* 'by means of'.

(d) Complex subordinators, i.e., connectives which have two parts, usually a postposition (*rağmen* 'despite', *için* 'for', *gibi* 'as well as') and an accompanying suffix on the (non-finite) verb of the subordinate clause.[1]

(e) Anaphoric connectives such as *ne var ki* 'however', *üstelik* 'what is more', *ayrıca* 'apart from this', *ilk olarak* 'firstly', etc.

In the PDTB, non-finite clauses have not been annotated as arguments. However, since all non-finite clauses are marked with a suffix in Turkish (see sections 4.1 and 4.2 below) and encode a relation between AOs, we would have missed an important property of the language if we had not identified them as discourse connectives (cf. Prasad et al., 2008).

All the discourse connectives above have exactly two arguments. So as in English, while verbs in Turkish can vary in the number of arguments they take, Turkish discourse connectives take two and only two arguments. These can conveniently be called ARG1 and ARG2. It remains an open question whether there is any language in which discourse connectives take more than two arguments.

In the following, we give representative examples of each of the above five classes of discourse connectives and discuss the assignment of the argument labels, linear order of arguments and types of arguments. By convention, we label

[1] Postpositions correspond to prepositions in English, though there are many fewer of them. They form a subordinate clause by nominalizing their complements and marking them with the dative, ablative, or the possessive case. In the examples given in this paper, suffixes are shown in upper-case letters. Case suffixes are underlined in addition to being presented in upper-case letters.

the argument containing (or with an affinity for) the connective as ARG2 (presented in boldface) and the other argument as ARG1 (presented in italics). Discourse connectives are underlined. This annotation convention is used in the English translations as well. Except for examples (12), (13), (19), (20), all examples have been taken from the MTC.

3 Coordinating conjunctions

3.1 Simple coordinating conjunctions

Coordinating conjunctions are like English and combine two clauses of the same syntactic type, e.g., two main clauses. They are typically sentence-medial and show an affinity with the second clause (evidenced in part through punctuation and their ability to move to the end of the second clause). Whether a coordinating conjunction links clauses within a single sentence or clauses across adjacent sentences (cf. Section 6), it shows an affinity with the second clause. Thus ARG2 of these conjunctions is the second clause and ARG1 is the first clause.

(1) *Yapılarını kerpiçten yapıyorlar, ama sonra taşı kullanmayı öğreniyorlar. Mimarlık açısından çok önemli*, çünkü **bu yapı malzemesini başka bir malzemeyle beraber kullanmayı, ilk defa burada görüyoruz.**
'*They constructed their buildings first from mud-bricks but then they learnt to use the stone. Architecturally, this is very important* because **we see the use of this construction material with another one at this site for the first time**.'

The particle *dA* can serve a discourse connective function with an additive (Example 2) or adversative sense (Example 3). In contrast with coordinating conjunctions, the order of arguments to *dA* is normally ARG2-ARG1, thus exhibiting a similarity with subordinators (see below). However, since *dA* combines two clauses of the same syntactic type, we take it to be a simple coordinating conjunction.

(2) **Konuşmayı unuttum diyorum** da *gülüyorlar bana*.
'**I said I've forgotton to talk** and *they laughed at me.'*

(3) **Belki bir çocuğumuz olsa onunla oyalanırdım** da *Allah kısmet etmedi.*
'**If we had a child I would keep myself busy with her/him** but *God did not predestine it*.'

3.2 Paired coordinating conjunctions

Paired coordinating conjunctions are composed of two lexical items, with the second often a duplicate of the first element. These lexical items express a single discourse relation, such as disjunction as in example (4). The order of arguments is ARG1-ARG2 and the position of the conjunctions is clause-initial.

(4) *Birilerinin* ya *işi vardır, aceleyle yürürler*, ya **koşarlar**.
'*Some people are* either *busy and walk hurriedly*, or **they run**.'

4 Subordinators

4.1 Simplex subordinators

When a subordinate clause is reduced in Turkish, it loses its tense, aspect and mood properties. In this way, it becomes a nominal or adverbial clause associated with the matrix verb. The relationship of an adverbial clause with the AO expressed by the matrix verb and its arguments is conveyed by a small set of suffixes corresponding to English 'while', 'when', 'by means of', 'as if', or temporal 'since', added to the non-finite verb of the reduced clause. This pair of non-finite verb and suffix, we call a 'converb'. The normal order of the arguments of a converb is ARG2-ARG1, where the converb appears as the last element of ARG2. The following example illustrates *–(y)ArAk* 'by means of' and its arguments:

(5) Kafiye Hanım beni kucakladı, **yanağını yanağıma sürt**erek *iyi yolculuklar diledi.*
'Kafiye hugged me and **by rubbing her cheek against mine**, *she wished me a good trip*.'

4.2 Complex subordinators

Complex subordinators constitute a larger set than the set of simplex subordinators. Here, a lexical item, usually a postposition, must appear with a nominalizing suffix and, if required, a case suffix as well. If the verb of the clause does not have a subject, it is nominalized with *–mAk* (the infinitive suffix). If it has a subject, it is nominalized with *-DIK* (past) or *–mA* (non-past) and carries the possessive marker agreeing with the subject of the

verb. The normal order of the arguments of a complex subordinator is the same as with converbs, i.e., ARG2-ARG1. The nominalizer, the possessive and the case suffix (if any) appear attached to the non-finite verb of ARG2 in that order. The connective appears as the last element of ARG2.

Some postpositions have multiple senses, depending on the type of nominalizer attached to the non-finite verb. For example, the postposition *için* means causal 'since' with *–DIK* (Example 6), and 'so as to' with *–mA* or *–mAk* (Example 7). In these examples, the lexical part of the complex subordinator is underlined, and the suffixes on the non-finite verb of ARG2 rendered in small caps.

(6) **Herkes çoktan pazara çıkTIĞI** için *kentin o dar, eğri büğrü arka sokaklarını boşalmış ve sessiz bulurduk.*
'Since **everyone has gone to the bazaar long time ago**, *we would find the narrow and curved back streets of the town empty and quiet*.'

(7) **[Turhan Baytop] Paris Eczacılık Fakültesi Farmakognozi kürsüsünde görgü ve bilgisini arttırMAK** için *çalışmıştır.*
'**Turhan Baytop** *worked* **at Paris Pharmacology Faculty** so as to **increase his experience and knowledge**,'

Since postpositions also have a non-discourse role in which they signal a verb's arguments and/or adjuncts, we will only annotate postpositions as discourse connectives when they have clausal elements as arguments. Given that a clausal element always has a nominalizing suffix, the distinction will be straightforward. For example, in (8) *için* takes an NP complement (marked with the possessive case) and will not be annotated, while in (9) *rağmen* 'despite' comes with a nominalizer and the dative suffix, and it will be annotated:

(8) Bunun için paraya ihtiyacımız var.
'We need money for this.'

(9) **Çok iyi bir biçimde yayılmış olMASINA** rağmen *Celtis (çitlenbik) poleninin yokluğu* dikkate değerdir.
'Despite **not dispersing well**, *the absence of the Celtis [tree] polen* is worthy of attention.'

In general, both parts of a complex subordinator must be realized in the discourse. An exception is 'if' *eğer* and its accompanying suffix *–sE* (and the marker agreeing with the subject of the subordinate clause where necessary). The suffix suffices to introduce a discourse relation on its own, even without the postposition *eğer*:

(10) **Salman Rushdi öldürülürSE** *İslam dini bundan bir onur mu kazanacak*?
'If **Salman Rushdi was to be killed**, *would the Islam religion be honoured*?'

(11) Eğer **sigarayı bırakmak için mükemmel zamanı bekliyorSAnız** *asla sigarayı bırakamazsınız.*
'If **you are waiting for the best time to stop smoking**, *you can never stop smoking'*

5 Anaphoric connectives

The fifth type of explicit discourse connectives are anaphoric connectives. Anaphoric connectives are distinguished from clausal adverbs like *çoğunlukla* 'usually', *mutlaka* 'definitely, *maalesef* 'regrettably', which are interpreted only with respect to their matrix sentence. In contrast, anaphoric connectives also require an AO from a sentence or group of sentences adjacent (Example 12) or non-adjacent (Example 13) to the sentence containing the connective. Another important property of anaphoric connectives is that they can access the inferences in the prior discourse (Webber et al 2003). This material is neither accessible by other types of discourse connectives nor clausal connectives. For example, in example (14), the anaphoric connective *yoksa* 'or else, otherwise' accesses the inference that the organizations have not united and hence did not introduce political strategies unique to Turkey.

(12) *Ali hiç spor yapmaz.* Sonuç olarak çok istediği halde **kilo veremiyor**.
'*Ali never exercises*. Consequently, **he can't lose weight** although he wants to very much.'

(13) Zeynep önceleri Bodrum'da oturdu. *Krediyle deniz kenarında bir ev aldı***.** Evi dayadı, döşedi, bahçeye yasemin ekti. Ne var ki **banka kredisini ödeyemediğinden evi satmak zorunda kaldı.**
'Zeynep first lived in Bodrum. *She bought a house by the sea on credit*. She furnished it fully and planted jasmine in the garden. However, **she had to sell the house because she couldn't pay back the credit.'**

(14) *Bu örgütlerin birleşerek Türkiye'yi etkilemesi ve Türkiye'ye özgü politikaları gündeme getirmesi lazım.* Yoksa **Tony Blair şöyle yaptı şimdi biz de şimdi böyle yapacağızla olmaz.**
'*These organizations must unite, have an impact on Turkey and introduce political strategies unique to Turkey*. Or else **talking about what Tony Blair did and hoping to do what he did is outright wrong**.'

6 Ordering flexibility of explicit discourse connectives and their arguments

In Turkish, the linear ordering of coordinating conjunctions and subordinators and the clauses in which they occur shows some flexibility as to where in the clause they appear or as to the ordering of the clauses. For example, coordinating conjunctions may appear at the beginning of their ARG2, i.e. S-initially. This was shown earlier in Example (1). The sentences below illustrate *ama* 'but' and *çünkü* 'because' used at this position.

(15) *Hatem Ağa'nın malına kimse yanaşamaz, dokunamazdı*. Ama **Osman gitmiş, Hatem Ağa'nın çiftliğini yakmıştı**.
'*No one could approach and touch Agha Hatem's property*. But **Osman had burnt Agha Hatem's ranch.'**

(16) *Söz özgürlüğünün belli yasalar, belli ilkeler çerçevesinde kalmak zorunda* olduğunu biliyoruz. Çünkü, **bütün özgürlükler gibi, belli sınırlar aşılınca, başkalarına zarar vermek, başkalarının özgürlüklerini zedelemek söz konusu oluyor.**
'We know that *freedom of speech should remain within the limits of certain laws and principles*. Because, **like all the other freedoms, when certain constraints are violated, one may harm others' freedom.'**

But coordinating conjunctions may also appear at the end of their ARG2 and so will appear S-finally in sentences with ARG1-ARG2 order. Below, we illustrate two cases of *ama* 'but' and *çünkü* 'because'.

(17) *Kazıyabildiğini sildi, biriktirdi mendilinin içine*. **Çaba isteyen zor bir işti bu yaptığı** ama.
'*He wiped the area he had scraped and saved all he could scrape in his rag*. But **what he was doing was a difficult job, requiring effort**.'

(18) *Kimi müşteriler dore rengi kumaşlarla, sarı taftalarla gelirdi de, elim dolu yapamam, diye geri çevirirdi, pek anlam veremezdim*. **Parayı severdi** çünkü.
'*Some customers would come with gold coloured fabrics and yellow taffeta weaves but he would reject them saying his hands were full, which I could not give any meaning to*. Because **he loved money.**'

In contrast, the position of a subordinator (both simplex and complex) in its ARG2 clause is fixed: it must appear at the end of the clause, as shown in example (19). However, the clause is free in the sentence and may be moved to the right of the sentence, as in example (20). It is a matter of empirical research to find out whether different genres vary more in how clauses are ordered and what motivates preposing of ARG1.

(19) **Ayşe konuşur**ken *ben dinlemiyordum*.
'*I was not listening* while **Ayşe was talking**.'

(20) *Ben dinlemiyordum* **Ayşe konuşur**ken.
'*I was not listening* while **Ayşe was talking**.'

7 Issues and plans

As mentioned above, we also plan to annotate implicit connectives between adjacent sentences or clauses whose relation is not explicitly marked with a discourse connective. This we will do at a later stage, after explicit connectives have been annotated, following the procedure used in annotating implicit connectives in the PDTB (PDTB-Group, 2006). Preliminary analysis has shown that punctuation serves as a useful hint in inserting a coordinating conjunctions such as 'and' or an anaphoric connective such as 'then' or 'consequently' between the multiple adjacent main clauses that can occur in a Turkish sentence separated by a comma. Example (21) illustrates these cases.

(21) Yürüyor, Imp = THEN oturuyor, resim yapmaya çalışıyor ama yapamıyor, tabela yazmaya çalışıyor ama yazamıyor, Imp= CONSEQUENTLY sıkılıp sokağa çıkıyor, Imp=AND bisikletine atladığı gibi pedallara basıyor.
'He walks around, then sits down and tries to draw, but he can't. He tries to inscribe words on the wooden plaque, but again he can't. Consequently he gets bored, goes out, and hops on his bike and pedals.'

A second important issue that will have to be tackled in the project is determining how much material is needed to specify the argument of a discourse connective. Annotation will be on text spans, rather than on syntactic structure. This reflects two facts: First, there is only a small amount of syntactically treebanked data in the MTC, and secondly, as has been discovered for English, one can not assume that discourse units map directly to syntactic units (Dinesh et al, 2005). Preliminary analysis also shows that discourse units may not coincide with a clause in its entirety. For example, in examples (9) and (16), one can take ARG1 to cover only the nominal complement of the matrix verb: The rest of the clause is not necessary to the discourse relation. The ways in which the arguments of a discourse connective may diverge from syntactic units must be characterized for Turkish as is being done for English (Dinesh et al, 2005).

A third issue we will investigate is whether different senses of a subordinator may be identified simply from the type of nominalizing suffix required on the subordinate verb. For example, we have noted in examples (6) and (7) that the two senses of the postposition için (namely, 'since (causal)' and 'in order to') are disambiguated by the nominalizing suffixes. The extent to which morphology aids sense disambiguation is an empirical issue that will be further addressed in the project.

Acknowledgement

We would like to thank Sumru Özsoy, Aslı Göksel and Cem Bozşahin for their comments on an earlier version of this paper. The first author also thanks the Caledonian Research Foundation and the Royal Society of Edinburgh for awarding her with the European Visiting Research Fellowship, which made this research possible. All remaining errors are ours.

References

Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi and Bonnie Webber (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.

Katherine Forbes-Riley, Bonnie Webber and Aravind Joshi (2006). Computing Discourse Semantics: The Predicate-Argument Semantics of Discourse Connectives in D-LTAG. *Journal of Semantics* 23, pp. 55—106.

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

‚Jacklin Kornfilt (1997). *Turkish.* London and New York: Routledge.

PDTB-Group (2006). The Penn Discourse TreeBank 1.0 Annotation Manual. *Technical Report IRCS 06-01*, University of Pennsylvania.

Rashmi Prasad, Samar Husain, Dipti Sharma and Aravind Joshi (2008). Towards an Annotated Corpus of Discourse Relations in Hindi. *The Third International Joint Conference on Natural Language Processing*, January 7-12, 2008.

Bilge Say, Deniz Zeyrek, Kemal Oflazer and Umut Özge (2002). Development of a Corpus and a TreeBank for Present-day Written Turkish. *Proceedings of the Eleventh International Conference of Turkish Linguistics,* Eastern Mediterranean University, Cyprus, August 2002.

Bonnie Webber, Aravind Joshi, Matthew Stone and Alistair Knott (2003). Anaphora and Discourse Structure. *Computational Linguistics* 29 (4) 547-588.

Appendix: A preliminary list of explicit discourse connectives found in the MTC belonging to five syntactic classes and their English equivalents

Simple coordinating conjunctions	**English equivalent**
ama	but
fakat	but
çünkü	because
dA	and, but
halbuki	despite
oysa	despite
önce	before
sonra	after
ve	and
veya	or
ya da	or
veyahut	or

Paired coordinating conjunctions	**English equivalent**
hem .. hem	both and
ya .. ya	either or
gerek .. gerek(se)	either or

Simplex subordinators (Converbs)	**English equivalent**
-ArAk	by means of
-Ip	and
-(y)kEn	while, whereas
-(y)AlI	since
-(I)ncA	when

Complex subordinators	**English equivalent**
-Ir gibi	as if, as though
-eğer (y)sE	if
-dIğI zaman	when
-dIğI kadar	as much as
-dIğI gibi	as well as
-dAn sonra	after
-dAn önce	before
-dAn dolayı	due to
-(y)sE dA	even though
-(y)Incaya kadar/dek	until
-(y)AlI beri	since (temporal)
-(n)A rağmen/karşılık	despite, although
-(n)A göre	since (causal)

Anaphoric connectives	**English equivalent**
aksi halde	if not, otherwise
aksine	on the contrary
bu nedenle	for this reason
buna rağmen/karşılık	despite this
bundan başka	besides this
bunun yerine	instead of this
dahası	moreover, in addition
ilk olarak	firstly, first of all
örneğin	for example
mesela	for example
sonuç olarak	consequently
üstelik	what is more
yoksa	otherwise
ardından	afterwards