Discourse Structure in Simultaneous Spoken Turkish

Işın Demirşahin
Middle East Technical University
Informatics Institute, Cognitive Science
ODTU, 06800, Ankara, TURKEY
disin@metu.edu.tr

Abstract

The current debate regarding the data structure necessary to represent discourse structure, specifically whether tree-structure is sufficient to represent discourse structure or not, is mainly focused on written text. This paper reviews some of the major claims about the structure in discourse and proposes an investigation of discourse structure for simultaneous spoken Turkish by focusing on tree-violations and exploring ways to explain them away by non-structural means.

1 Introduction

There is an ongoing debate about the nature of structure in discourse. Halliday and Hasan (1976) propose that although there is some structure in the text and structure implies *texture;* texture does not necessarily imply structure. Text is held together by a variety of non-structural cohesive ties: *reference, substitution, ellipsis, conjunction* and *lexical cohesion.* However, their notion of structure is strictly syntactic; and for other researchers, the elements that hold the text together, especially elements of conjunction, can be taken as indicators of structure in discourse.

If there is structure in discourse, the complexity of the said structure is of interest to linguistics, cognitive science and computer science alike. Is discourse structure more complex or more simple than that of sentence level syntax? How and to what degree is that structure constrained? In order to answer questions along these lines, researchers explore the possible data structures for discourse in natural language resources.

Section 2, reviews the current approaches to discourse structure. Section 3 introduces the current study, i.e., the search for deviations from tree structure in spontaneous spoken language. Section 4 presents a conclusive summary.

2 The Structure of Discourse

2.1 Tree Structure for Discourse

Hobbs (1985) takes it as a fact that discourse has structure. He argues that a set of coherence relations build a discourse structure that is composed of trees of successive and sometimes intertwining trees of various sizes connected at the peripheries.

Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) proposes that a text can be analyzed as a single tree structure by means of predefined rhetorical relations. Rhetorical relations hold between adjacent constituents either asymmetrically between a *nucleus* and a *satellite*, or symmetrically between two nuclei. The notion of *nuclearity* allows the units to connect to previous smaller units that are already embedded in a larger tree structure, because a relation is assumed to be shared by the nuclei of non-atomic constituents. In other words, a relation to a complex discourse unit can be interpreted as either between the adjacent unit and the whole of the complex unit, or between the adjacent unit and a nucleus of the complex unit.

One of the rhetorical structures in RST, *elaboration* is criticized by Knott et al. (2001) who propose an elaboration-less coherence structure, where the global focus defines linearly organized *entity chains,* which can contain multiple atomic or non-atomic RS trees, and which are linked via non-rhetorical resumptions.

Proceedings of the 2012 Student Research Workshop, pages 55–60,

Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

Discourse - Lexicalized Tree Adjoining Grammar (D-LTAG) (Webber, 2004) is an extension of the sentence-level Tree Adjoining Grammar (Joshi, 1987) to discourse level. Discourse connectives act as discourse level predicates that connect two spans of text with abstract object (Asher, 1993) interpretations. Coordinating and subordinating conjunctions such as *fakat* 'but' (1) and *rağmen* 'although' (2), take their host clauses by substitution and the other argument either by substitution or by adjoining; whereas discourse adverbials such as (3) take the host argument by adjoining, and the other argument anaphorically. In the examples below, the host argument is in boldface, the other argument is in italics and the connectives are underlined.

(1) *Araştırma Merkezi aşağı yukarı bitmiş durumda*, fakat **iç ve dış donanımı eksik.**

'The Research Center is more or less complete but **its internal and external equipments are missing.'**

(2) **Benim için çok utandırıcı bir durum olmasına** rağmen *oralı olmuyordum.*

'Although **it was a very embarrassing situation for me**, *I didn't pay much heed.'*

(3) İlgisizliğim seni şaşırtabilir. ama *üvey babamı görmek istemediğim için yıllardır o eve gitmiyorum.* ***Anneme çok bağlı olduğumu da söyleyemem*** ayrıca**.**

My indifference might surprise you, but *since I do not want to see my stepfather, I have not been to that house for years.* In addition**, I cannot say I am attached to my mom much.**

As in sentence level syntax, the anaphoric relations are not part of the structure; as a result, the discourse adverbials can access their first arguments anywhere in the text without violating non-crossing constraint of tree structure. When a structural connective such as *ve* 'and' and a discourse adverbial such as *bundan* ötürü 'therefore' are used together as in (4), an argument may have multiple parents violating one of the constraints of the tree structure; but since the discourse adverbial takes the other argument anaphorically, the non-crossing constraint is not violated.

(4) *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.**

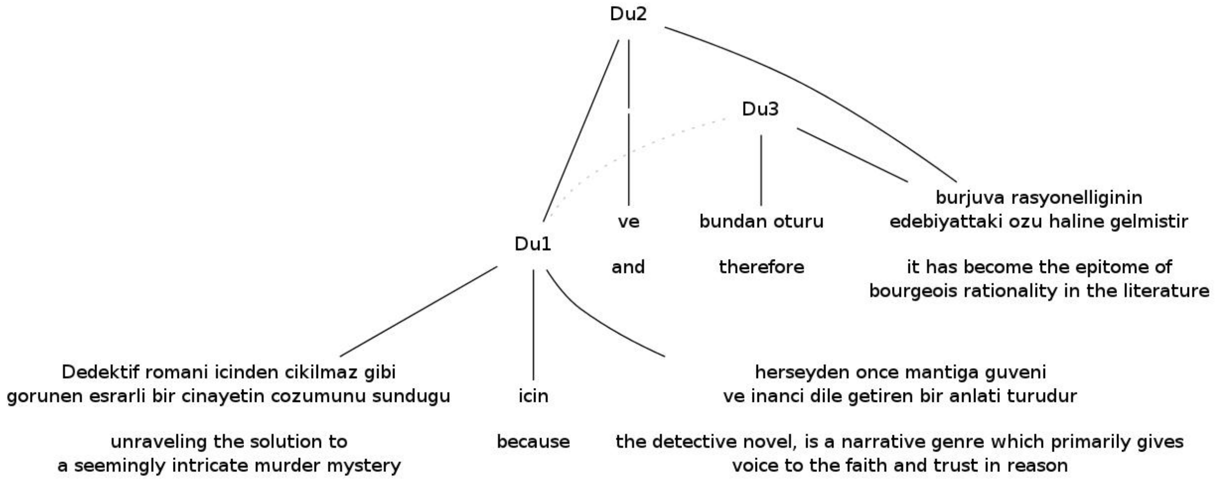
Du2
Du3
Du1
ve
and
bundan oturu
therefore
burjuva rasyonelliginin edebiyattaki ozu haline gelmistir
it has become the epitome of bourgeois rationality in the literature
Dedektif romani icinden cikilmaz gibi gorunen esrarli bir cinayetin cozumunu sundugu
unraveling the solution to a seemingly intricate murder mystery
icin
because
herseyden once mantiga guveni ve inanci dile getiren bir anlati turudur
the detective novel, is a narrative genre which primarily gives voice to the faith and trust in reason

Figure 1: Tree structure for (4). *Bundan ötürü* 'therefore' takes one argument anaphorically, shown as a dotted line in this representation. Since the anaphora is non structural, there is no crossing in (4). However, tree structure is still violated because Du2 and Du3 share an argument, resulting in multiple-parent structure.

Implicit connectives always link two adjacent spans structurally, the host span by substitution and the other by adjoining. Since after adjunction the initial immediate dominance configurations are not preserved, the semantic composition is defined on the derivation tree rather than the derived tree (Forbes et al., 2003; Forbes-Riley et al., 2005).

The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) is loosely based on D-LTAG, as the discourse connectives are annotated as discourse level predicates with two arguments; but the focus is no longer on the global structure of discourse but on individual relations, and the annotations are kept as theory-neutral as possible.

2.2 Deviations from Tree Structure

Wolf and Gibson (2005), judging from a corpus annotated for a set of relations that is based on Hobbs (1985), argue that the global discourse structure cannot be represented by a tree structure. They point out that the definition for the anaphoric connectives in D-LTAG seems to be circular, since they are defined by their anaphoric arguments which can be involved in crossing dependencies, and in turn they are defined as anaphoric and thus outside the structural constraints. They propose a chain graph-based annotations scheme, which they claim express the discourse relations more accurately than RST, because the relations can access embedded, non-nuclear constituents that would be inaccessible in an RST tree.

Since Wolf and Gibson use attribution and same relations, which are not considered discourse relations in D-LTAG or the PDTB, a direct comparison of chain graph annotations and the PDTB does not seem possible at this point; but violations of tree structure are also attested in the PDTB.

Lee et al. (2006, 2008) investigate the PDTB and identify dependencies that are compatible with tree structure, *independent relations* and *full embedding*; as well as incompatible dependencies, *shared argument, properly contained argument, partially overlapping arguments,* and *pure crossing*. They claim that only shared arguments (same text span taken as argument by two distinct discourse connectives) and properly contained arguments (a text span that is the argument of one connective properly contains a smaller text span that is the argument of another connective) should be considered as contributing to the complexity of discourse structure; the reason being that the instances of partially overlapping arguments and pure crossing can be explained away by anaphora and attribution, both of which are non-structural phenomena. The presence of shared arguments carries the discourse structure from tree to directed acyclic graphs (Webber et al., 2011).

Aktaş et al. (2010) have identified similar tree structure violations in the Turkish Discourse Bank (TDB) (Zeyrek et al., 2010). In addition to the dependencies in Lee et al. (2006), Aktaş have identified *properly contained relations* and *nested relations*. A full analysis of the TDB with respect to discourse structure is yet to be done.

Egg and Redeker (2008, 2010) argue that tree structure violations can be overcome by applying an underspecification formalism to discourse representation. They adopt a weak interpretation of *nuclearity*, where although the relation between an atomic constituent and a complex constituent is understood to hold between the atomic constituent ant the *nucleus* of the complex constituent, structurally the relation does not access the nucleus of the complex, and therefore does not result in multiple parenting. This approach is not directly applicable to PDTB-style relations, because of the *minimality principle*, which constrains the annotators to select the smallest text span possible that is necessary to interpret the discourse relation when annotating the arguments of a discourse connective.

Egg and Redeker also argue that most of the crossing dependencies in Wolf and Gibson (2005) involve anaphora, which is considered non-structural in discourse as well as in syntax. However, they admit that *multi-satellite constructions* (MSC) in RST, where one constituent can enter into multiple rhetorical relations as long as it is the nucleus of all relations, seems to violate tree structure. They state that only some of the MSCs can be expressed as atomic-to-complex relations, but they also state that those the MSCs that cannot be expressed so seems to be genre specific. The fact that both Egg and Redeker (2008) and Lee et al. (2006, 2008) cannot refute the presence of multiple parenting in discourse structure is striking.

2.3 Discourse Structure in Spoken Language

All studies in Section 2 investigates discourse structure in written texts. There are spoken corpora annotated for RST such as Stent (2000) and SDRT

(Baldridge & Lascarides, 2005), but the only PDTB-style spoken discourse structure annotation within the author's knowledge is part of the LUNA corpus in Italian (Tonelli, 2010).

The most striking change Tonelli et al. made in the PDTB annotation scheme when annotating spoken dialogues is to allow for implicit relations between non-adjacent text spans due to higher fragmentation in spoken language. They also added an *interruption* label for when a single argument of a speaker was interrupted. Some changes to the PDTB Sense Hierarchy was necessary including the addition of the GOAL type under CONTINGENCY class, fine tuning of PRAGMATIC subtypes, exclusion of LIST type from EXPANSION class and merging of syntactically distinguished REASON and RESULT subtypes into a semantically defined CAUSE type.

3 Proposed Study and Methodology

The aim of the current study is to determine whether tree structure is sufficient to represent discourse structure in simultaneous spoken Turkish. Unfortunately, due to time and budget constraints, continuous annotation of a large-scale corpus with multiple annotators is not possible for the short term. Therefore, the immediate goal is to extract excerpts of interest that include tree-violation candidates, annotate the violations along with their immediate context adopting a PDTB-style annotation with some adjustments for Turkish and spoken language; and explore means of explaining away these violations by non-structural cohesive ties defined by Halliday and Hasan (1976). Cohesive ties include the frequently discussed anaphora (*reference* in their terms), but also include other non-structural mechanisms such as *ellipsis* and *lexical cohesion*.

3.1 Extracting tree-violation candidates

The first step of the study is to examine the structural configurations in the TDB. Although the TDB is a written text source, it contains texts from multiple genres; and in some genres such as novels, stories and interviews, dialogues are annotated for discourse structure. We expect the TDB annotations to provide some insight that can be transferred to spoken language. For example, if a certain discourse connective, a particular attribution verb or some specific type of embedded clause seem to participate frequently in tree-violations in the TDB, searching for instances of that particular elements in spoken data may considerably hasten the search for tree-violation candidates.

The second step is the continuous annotation of small pieces of spoken data. The goal of this step is not to produce a fully annotated spoken corpus, but rather to gather some insight into the structures that are unique to spoken data. By annotating randomly selected small pieces of spoken data, we aim to discover structures that are unique to spoken data that cannot be extracted form the TDB. Like the first step, the goal is to identify elements that are likely to result in tree-violations that can be searched for in large amounts of unannotated data.

The last step is obviously to look for the identified elements in the first two phases in larger amounts of spoken data and annotate them. Currently considered spoken resources are the METU Spoken Turkish Corpus (Ruhi and Karadaş 2009) and freely available podcasts.

3.2 Anticipated adjustments to the PDTB annotation scheme

The TDB has already made some adjustments for Turkish on the PDTB style. One major adjustment is to annotate phrasal expressions that include deictic expressions (such as *bu sebeple* 'for this reason') as discourse connectives. Although the PDTB annotates some phrasal and multipart connectives, deictic and productive phrasal expressions such as *that's because* or *the reason is* were annotated as alternative lexicalizations rather than lexicalized discourse predicates. In the TDB, such expressions are annotated as discourse connectives because of the structural similarity between deictic phrasal expressions and subordinating discourse connectives. In addition, a *shared* span label was introduced to accommodate for text spans that belong to both arguments, such as sentential adverbials or subjects of subordinate clauses. Finally, in an ongoing attempt to add sense annotations to the TDB, some new sense labels such as OBJECTION and CORRECTION were added to the PDTB sense hierarchy.

In addition to Turkish-specific changes, we will consider adopting speech-specific changes such as the non-adjacent implicit connectives and the *repetition* label by Tonelli (2010) as needed.

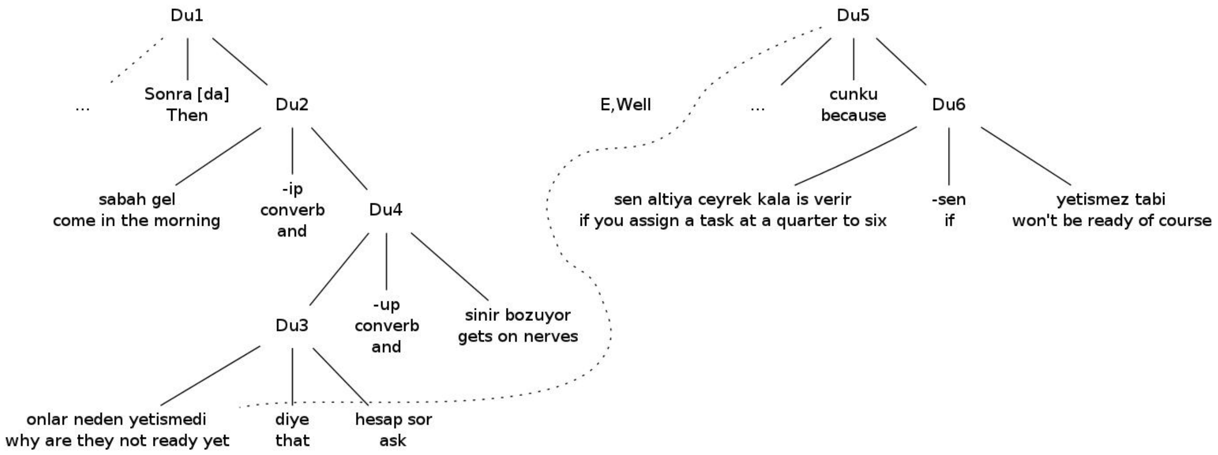

Figure2: An attempt at building a tree for (5). The first argument of *çünkü* in Du5 is either recovered from Du3 by non-structural means, or taken structurally form Du3, resulting in *pure crossing* and depending on the decision to annotate attribution as a discourse annotation or not, either *shared argument* or *properly contained argument.*

3.3 A sample tree-violation candidate

A sample excerpt of interest is (5). The context is that the speaker is complaining that the project manager assigns new tasks right before the end of working hours.

(5) Sonra da sabah gelip *onlar neden yetişmedi* diye hesap sorup sinir bozuyor. E, <u>çünkü</u> **sen altıya çeyrek kala iş verirsen yetişmez tabi**.

Then he comes in the morning and asks *why they are not ready yet* and (thus) he gets on my nerves. Well, <u>because</u> **if you assign the task at a quarter to six o'clock, they won't be ready of course**.

In (5), the first argument of the connective *çünkü* 'because' is the complement of asking, and is embedded in a sequence of events. Most importantly, it is neither the first nor the last event in the sequence, so structurally it should not be available to *çünkü*.

Once a tree-violation candidate such as (5) is identified, it will be analyzed to see if a plausible tree structure can be constructed, or the violation can be explained away by non-structural mechanisms or speech-specific features such as intonation. In this case, there doesn't seem to be an anaphoric explanation to get rid of the crossing dependency. However, left hand side argument of *çünkü* is embedded in a verb of attribution.

"Why are they not ready yet?" and the answer "Because if you give the task at a quarter to six o'clock, they won't be ready of course." make up a sub-discourse distinct from the structure of the main discourse. Another non-structural explanation is *ellipsis*, where the missing argument of *çünkü* is recovered from the preceding context. *Repetition* (an element of lexical cohesion) of *yetişmek* 'to catch up, be ready', may play a role in the recovery of the missing argument. At this point, we confine ourselves to identifying possible explanations, but refrain from committing ourselves to any one of the explanations. Further research should reveal whether this is a frequent dependency type a. for *çünkü* 'because', b. for lexically reinforced *ellipsis* and c. for arguments of attribution verbs d. for Turkish discourse, or e. for spontaneous speech. Each of this possibilities will have different ramifications, ranging from a discourse adverbial interpretation of *çünkü* 'because' to a graph structure for spoken discourse.

4 Conclusion

Whether tree structure is sufficient to represent discourse relations is an open question that will benefit from diverse studies in multiple languages and modalities. Here we have presented some of the arguments for and against tree structure in discourse. The current study aims to reveal the constraints in simultaneous spoken Turkish discourse structure. The proposed framework for discourse structure analysis is based on PDTB-style, with adjustments for Turkish and spoken language. The adjustments will be based on the existing PDTB-style studies in

Turkish and simultaneous speech, although they are likely to evolve further as research progresses. The methodology for the study is to search for possible tree-violations, and try to apply the explanations in the literature to explain them away. The violations that cannot be plausibly explained away by non-structural mechanisms should be accommodated by the final discourse model.

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) for the Turkish Discourse Bank.

References

Berfin Aktaş, Cem Bozşahin, Deniz Zeyrek. 2010. Discourse Relation Configurations in Turkish and an Annotation Environment. *Proc. LAW IV - The Fourth Linguistic Annotation Workshop*.

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

Jason Baldridge, Alex Lascarides. 2005. Annotating Discourse Structure for Robust Semantic Interpretation. *Proc. 6th International Workshop on Computational Semantics.*

Markus Egg, Gisela Redeker. 2008. Underspecified Discourse Representation. In A. Benz and P. Kuhnlein (eds) *Constraints in Discourse* (117-138). Benjamins: Amsterdam.

Markus Egg, Gisela Redeker. 2010. How Complex is Discourse Structure? *Proc. 7th International Conference on Language Resources and Evaluation (LREC 2010)* pp. 1619–23.

Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Anoop Sarkar, Aravind K. Joshi. 2003. D-LTAG System: Discourse Parsing with a Lexicalized Tree-Adjoining Grammar. *Journal of Logic, Language and Information*, 12(3), 261–279.

Katherine Forbes-Riley, Bonnie Webber, Aravind K. Joshi. 2005. Computing Discourse Semantics: The Predicate-Argument Semantics of Discourse Connectives in D-LTAG. *Journal of Semantics*, 23, 55-106.

Michael A. K. Halliday, Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Jerry R. Hobbs. 1985. On the Coherence and Structure of Discourse. *Report CSLI-85-37, Center for Study of Language and Information.*

Aravind K. Joshi. 1987. An Introduction to Tree Adjoining Grammar. In A. Manaster- Ramer (Ed.), *Mathematics of Language*. Amsterdam: John Benjamins.

Alistair Knott, Jon Oberlander, Michael O'Donnel, Chris Mellish. 2001. Beyond elaboration: The interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord & W. Spooren (Eds.), *Text Representation:Linguistic and psycholinguistic aspects* (181-196): John Benjamins Publishing.

Alan Lee, Rashmi Prasad, Aravind K. Joshi, Nikhil Dinesh, Bonnie Webber. 2006. Complexity of dependencies in discourse: are dependencies in discourse more complex than in syntax? *Proc. 5th Workshop on Treebanks and Linguistic Theory (TLT'06)*.

Alan Lee, Rashmi Prasad, Aravind K. Joshi, Bonnie Webber. 2008. Departures from tree structures in discourse. *Proc. Workshop on Constraints in Discourse III.*

William C. Mann, Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text,* 8(3), 243-281.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. *Proc. LREC'08 - The sixth international conference on Language Resources and Evaluation.*

Şükriye Ruhi, Derya Çokal Karadaş. 2009. Features for an internet accessible corpus of spoken Turkish discourse. *Working Papers in Corpus-based Linguistics and Language Education* 3, 311-320.

Amanda Stent. 2000. Rhetorical structure in dialog. *Proc. 2nd International Natural Language Generation Conference (INLG'2000). Student paper.*

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, Aravind Joshi. 2010. Annotation of Discourse Relations for Conversational Spoken Dialogues. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*.

Bonnie Webber, Matthew Stone, Aravind K. Joshi, Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*. 29 (4):545-587.

Bonnie Webber. 2004. D-LTAG: Extending Lexicalized TAG to Discourse. Cognitive Science, 28(5), 751-779.

Bonnie Webber, Markus Egg, Valia Kordoni. 2011. Discourse structure and language technology. *Natural Language Engineering,* doi: 10.1017/S1351324911000337, Published online by Cambridge University Press 08 December 2011.

Florian Wolf, Edward Gibson. 2005. Representing discourse coherence: a corpus-based study. *Computational Linguistics* 31: 249–87.

Deniz Zeyrek, Işın Demirşahin, Ayışığı Sevdik-Çallı, Hale Ögel Balaban, İhsan Yalçınkaya, Ümit Deniz Turan. 2010. The annotation scheme of Turkish discourse bank and an evaluation of inconsistent annotations. *Proc. 4th Linguistic Annotation Workshop (LAW IV).*