Annotating Discourse Connectives in Spoken Turkish

Işın Demirşahin
Middle East Technical University
Informatics Institute
Department of Cognitive Science
e128500@metu.edu.tr

Deniz Zeyrek
Middle East Technical University
Informatics Institute
Department of Cognitive Science
dezeyrek@metu.edu.tr

Abstract

In an attempt to extend Penn Discourse Tree Bank (PDTB) / Turkish Discourse Bank (TDB) style annotations to spoken Turkish, this paper presents the first attempt at annotating the explicit discourse connectives in the Spoken Turkish Corpus (STC) demo version. We present the data and the method for the annotation. Then we reflect on the issues and challenges of transitioning from written to spoken language. We present the preliminary findings suggesting that the distribution of the search tokens and their use as discourse connectives are similar in the TDB and the STC demo.

1 Introduction

Turkish Discourse Bank (TDB) is the first discourse-annotated corpus of Turkish, which follows the principles of Penn Discourse Tree Bank (PDTB) (Prasad *et al*., 2008) and includes annotations for discourse connectives, their arguments, modifiers and supplements of the arguments. The TDB is built on a ~ 400,000-word sub-corpus of METU Turkish Corpus (MTC) (Say *et al*., 2002), a 2 million-word multi-genre corpus of post-1990 written Turkish[1].

In both PDTB and TDB, the discourse connectives link two text spans that can be interpreted as, or can be anaphorically resolved to abstract objects (Asher, 2003). The PDTB includes annotations for both explicit and implicit connectives, whereas TDB has covered only explicit connectives so far.

The explicit discourse connectives annotated in TDB come from a variety of syntactic classes, namely coordinating conjunctions (*ve* 'and'), subordinating conjunctions (*için* 'for/since') and discourse adverbials (*ancak* 'however'). It also annotates phrasal expressions (Zeyrek *et al,* 2013).

The coordinating and subordinating conjunctions are 'structural' discourse connectives that take their arguments syntactically, whereas discourse adverbials only take one argument syntactically, and the other one anaphorically (Forbes-Riley *et al*. 2006). For all syntactic types, the argument that syntactically accommodates the discourse connective is called the second argument (Arg2). The other argument is called the first argument (Arg1). In TDB, phrasal expressions consist of an anaphoric element and a subordinating conjunction. In PTDB, similar expressions are annotated as AltLex, a subtype of implicit connectives (Prasad *et al.* 2008). For example, *onun için* 'because of that' in (1) is annotated as a discourse connective with its two argument spans. In the rest of the paper, the connective is underlined, the Arg2 is in bold face, and Arg1 is shown in italics. Supplementary materials and modifiers are shown in square brackets labelled with subscripts when necessary.

[1] The MTC and the TDB are freely available to researches at http://ii.metu.edu.tr/corpus and http://medid.ii.metu.edu.tr, respectively.

LAW VIII - The 8th Linguistic Annotation Workshop, pages 105–109,
Dublin, Ireland, August 23-24 2014.

(1) O ses dinleme cihazı. *Ödev var da*. Onun için **sizi dinliyorum şu anda.**
"That is a recording device. *I have homework.* Because of that **I'm recording you right now."**

The TDB has chosen to annotate these expressions as discourse connectives because they are highly frequent and have limited compositional productivity. Furthermore, some phrasal expressions such as *bunun aksine* 'contrary to this', *aksi takdirde* 'however' are so frequent that the native speakers perceive them as single lexical entries.

In an attempt to extend the PDTB/TDB style discourse annotation to spoken Turkish, we have annotated the search tokens in TDB on the Spoken Turkish Corpus (STC) (Ruhi *et al.*2009; 2010) demo version[2]. Following the TDB conventions, we annotated the phrasal expressions such as *onun için* 'because of that' and *ondan sonra* 'after that'. The annotation of 77 search tokens identified in TDB yielded a total of 416 relations in the STC demo.

In this paper we first present the data from the STC demo release, the method we used for annotations, and issues and challenges we have met. Then we present our preliminary findings. Finally, we discuss the methods and the findings of the study and draw a road map for future work.

2 Annotating Spoken Turkish

2.1 The Data

The Spoken Turkish Corpus demo version is a ~20,000-word resource of spoken Turkish. The demo version contains 23 recordings amounting to 2 hours 27 minutes. Twenty of the recordings include casual conversations and encounters, comprising 2 hours 1 minutes of the total, the 3 remaining recordings are broadcasts lasting a total of 26 minutes. The casual conversations include a variety of situations such as conversations among families, relatives and friends, and service encounters. The broadcasts are news commentaries. The topics of conversation range from daily activities such as infant care and naming babies to biology e.g. the endocrine system, to politics such as European Union membership process or the clearing of the mine fields on Syrian border. Such wide range of topics provide for a wide coverage of possible uses of discourse connectives even in such a relatively small corpus.

2.2 Annotation Method

Since our main aim was to follow the PDTB/TDB style, we chose to use the Discourse Annotation Tool for Turkish (DATT) (Aktaş et al., 2010). We used the transcription texts included in the STC demo version as the DATT input and provided the annotators with separate audio files.

This approach was a trade-off: the annotators could not make use of the rich features of the time-aligned annotation of the STC; but by importing text transcripts directly into an existing specialized annotation tool we did not have to go through any software development and/or integration stage. The annotators reported only slight discomfort in matching the text and the audio file during annotation, but stated that it was manageable as none of the files are long enough to get lost between the two environments.

2.3 Issues and Challenges

Some of the challenges of annotating discourse connectives we have already observed in written language transfer to the spoken modality. For example, in written discourse it is possible for an expression to be ambiguous between a discourse and non-discourse use, as the anaphoric elements can refer to both abstract objects and non-abstract entities. This applies to spoken language as well.

(2) SER000062: Şey Glomerulus o yuvarlak topun adı mıydı (bu)? Ordan şey oluyor…
AFI000061: hı-hı hı-hı
AFI000061: Süzülme ondan sonra oluyor ama. Şu Henle kulpu falan var ya. Şöyle geri.

[2] The STC demo version is freely available to researches at http://std.metu.edu.tr

"SER000062: Um Glomerulus was (this) the name of that round ball? Stuff happens there …
AFI000061: Yes, yes.
AFI000061: Filtration occurs after that, though. That Loop of Henle and such. Reverse like this."

In (2) *ondan sonra* 'after that' could be interpreted as resolving to the clause 'Stuff happens there', which is an abstract object although a vague one. The pronoun can also refer to the glomerulus, which is an NP. This was exactly the case during the annotation of this specific example: one annotator interpreted it as a temporal discourse connective that indicates the order of two sub-processes of kidney function, whereas the other annotator interpreted that *o* 'that' refers to the NP and did not annotate this instance of *ondan sonra*. As a TDB principle, if an expression has at least one discourse connective meaning, it is annotated. As a result, this example was annotated as per the first annotator's annotation.

In spoken language, particularly spontaneous casual dialogues, phrasal expressions can take their first arguments from anywhere in the previous discourse. This is very much like discourse adverbials. For example, *için* in (3) displays an unattested use in TDB, as it appears distant from *both* its arguments, allowing the participant to question the discourse relation between two previous text spans. Given the supplemental material "thyroxin increases the metabolism" in line (a) by speaker AFI, speaker SER provides two propositions, "thyroxin is secreted by the thyroid gland" in line (b) and "people with overactive thyroids tend to be hyperactive" in line (e). In line (h), AFI offers a discourse connective "because" in order to show her understanding of the preceding discourse, i.e., something like '(so they tend to be very active) because of that?', where the material in parentheses are elided. One can argue that this connective builds a new discourse relation with one anaphoric and one elliptic argument. Nevertheless, we kept the annotations as shown in the example, because (a) it was the most intuitive annotation according to the annotators and (b) the DATT does not allow annotation of ellipsis as arguments for now.

(3) (a) AFI000061: [$_{SUPP1}$Tiroksin. Ha bak. Metabolizma hızını arttırıyor.]
[…]
(b) SER000062: *Tiroit bezinden tiroksin salgılanıyor*.
(c) AFI000061: Hmm salgılanıyor dedin sen. Tamam. Doğru.
(d) SER000062: Tamam.
(e) SER000062: Hatta tiroit şey olan… Emm **tiroidinde sorun olanlar çok ee şey olur ya aktif olur ya.**
(f) AFI000061: Hmm?
(g) SER000062: Çok hareketli olurlar. Evet.
(h) AFI000061: **Onun için** [$_{MOD}$mi]?

"(a) AFI000061: [$_{SUPP1}$Thyroxin. Oh look. It speeds up the metabolism.]
[…]
(b) SER000062: *Thyroxin is secreted by the thyroid gland.*
(c) AFI000061: Hmm you said secreted. Ok. Right.
(d) SER000062: Ok.
(e) SER000062: Actually thyroid is the one that… Emm **you know, those who have problems with thyroid are ee they tend to be very active.**
(f) AFI000061: Hmm?
(g) SER000062: They tend to be very energetic. Yes.
(h) AFI000061: [$_{MOD}$ Is (it)] **because of that**?"

Another problem with spoken corpus is that some elements may be missing. There are many examples that could not be annotated as discourse connectives, because the speakers were interrupted before they could complete, or at times even start, the latter argument of a possible discourse relation. In other examples, the argument may be there but not recorded clearly, or may be completely inaudible even though they were uttered because of background noise or overlapping arguments.

3 Preliminary Findings

In this section we present some of our preliminary findings and compare them to the TDB to the extent possible. Because of the large difference in size between the two corpora, we converted the raw numbers to frequencies. We used number/1000 words as the frequency unit in Table 1.

The top five most frequent connectives in the TDB in descending order are *ve* 'and', *için* 'for', *ama* 'but', *sonra* 'later' and *ancak* 'however' and the top five most frequent connectives in the STC are *ama* 'but', *ve* 'and', *mesela* 'for example', *sonra* 'later' and *için* 'for'. Here we compare the four most frequent connectives, namely, *ve*, *için, ama* and *sonra*, which make up 4951 (58.3%) of the total 8484 annotations in TDB and 217 (52.2%) of the total 416 relations annotated in the STC.

	TDB						**STC demo**					
	Discourse connectives			**Total instances**			**Discourse connectives**			**Total instances**		
Conn	**#**	***f***	**%**	**#**	***f***	**%**	**#**	***f***	**%**	**#**	***f***	**%**
ve 'and'	2112	5.31	28.2	7501	18.86	100	50	2.40	48.1	104	5.00	100
için 'because'	1102	2.77	50.9	2165	5.44	100	32	1.54	61.5	52	2.50	100
ama 'but'	1024	2.57	90.6	1130	2.84	100	96	4.61	80.7	119	5.72	100
sonra 'later'	713	1.79	56.7	1257	3.16	100	39	1.87	72.2	54	2.60	100

Table 1 - Written and spoken uses of *ve, için, ama,* and *sonra.*

Although both the frequency of the total occurrences of the connectives and their discourse uses seem to be lower in the spoken corpus, chi square tests show that the differences are not statically significant ($p>0.5$). The percentage of the use of tokens as discourse connectives across modalities is not significant either ($p>0.5$). The preliminary results indicate that the distribution of these five connectives and their uses as discourse connective are similar in written and spoken language.

The similarity is expected, as the MTC and the subcorpus that the TDB is built on are multi-genre corpora. Specifically, the TDB includes novels and stories, which in turn include dialogues. Also, there are interviews in news excerpts, which are basically transcriptions of spoken language. As a result, the TDB texts reflect some aspects of spoken language. In addition, 3 of the 23 files of the STC demo are news broadcasts and interviews, which are probably scripted and/or prepared. Thus they may not necessarily reflect all aspects of spontaneous spoken language.

4 Discussion and Conclusion

In this paper we presented a preliminary attempt at annotating Turkish Spoken Language in PDTB/TDB style. We used the transcripts and audio files of STC demo as our source, and used DATT of TDB to annotate the discourse relations. As future work, we intend to integrate the discourse annotation to the time-aligned annotation of the STC, thus allowing the users to benefit from the features of both annotation schemes.

During the annotation process, we encountered the use of discourse connectives unattested in TDB, specifically *için* 'since/for' in a predicative/interrogative position, where the connective occurs with its deictic Arg1. We assume that the question in which this connective is used has a rhetorical role, possibly expressing the speaker's understanding of the discourse relation in the previous discourse. Apart from this newly attested use, the distribution of the search tokens and their use as discourse connectives remain largely similar to that of the TDB. We conclude that this similarity results from the fact that the TDB includes some features of the spoken language just as the STC demo may include scripted recording. Yet, we suspect that the occurrence of discourse connectives with a deictic Arg1 is quite frequent in spoken language. We leave the investigation of such occurrences, and other issues such as the genre breakdown of the frequency of discourse connectives in STC for further study.

Our goal for the near future is to complete at least a second set of double-blind annotations and the agreement statics on the STC, so that the discourse-level annotation of spoken Turkish can be compared to those of the TDB.

Reference

Aktaş, B., Bozsahin, C., & Zeyrek, D. 2010. Discourse relation configurations in Turkish and an annotation environment. *Proceedings of the Fourth Linguistic Annotation Workshop*, pp. 202-206.

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

Demirşahin, I., Sevdik-Çallı, A., Ögel Balaban, H., Çakıcı, R. & Zeyrek, D. 2012. Turkish Discourse Bank: Ongoing Developments. *Proceedings of LREC 2012 The First Turkic Languages Workshop.*

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.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. (LREC'08).

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

Ruhi, Şükriye, Eröz Tuğa, Betil, Hatipoğlu, Çiler, Işık-Güler, Hale, Can, Hümeyra, Karakaş, Özlem, Acar, Güneş, Eryılmaz, Kerem (2010). Türkçe için genel amaçlı sözlü derlem oluşturmada veribilgisi, çeviriyazı ölçünleştirmesi ve derlem yönetimi. *XXIV. Dilbilim Kurultayı*, 17-18 Mayıs, 2010, Yuvarlak Masa Toplantısı.

Say, B., Zeyrek, D., Oflazer, K., and Özge, U. 2002. Development of a Corpus and a Treebank for Present-day Written Turkish. In *Proceedings of the Eleventh International Conference on Turkish Linguistics* (ICTL 2002).

Zeyrek, D., and Webber, B. (2008). A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Turkish Corpus. In *Proceedings of the 6thWorkshop on Asian Language Resources, The Third International Joint Conference on Natural Language Processing (IJNLP)*.

Zeyrek, D., Demirşahin, I., Sevdik-Çallı, A. B., Çakıcı, Ruket. 2013. Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language. *Discourse and Dialogue* 4 (3), 174-184.