

Features for an Internet Accessible Corpus of Spoken Turkish Discourse

Derya Çokal Karadaş

(Middle East Technical University)

Şükriye Ruhi

(Middle East Technical University)

Resumé:

In this paper we survey features of spoken corpora for a number of languages and focus on the design criteria, technological requirements and priorities in annotation for a spoken corpus for Turkish that aims to reflect its discursive and pragmatic features. The paper continues with a description of possible coding schemas for the annotation of the discursive and pragmatic features of Turkish. Taking as its basic premise that a time based data model with multiple tiers is best suited for constructing spoken corpora, the linguistic annotation on the visual and auditory recordings incorporated in METU Spoken Turkish Corpus will be accomplished with EXMARaLDA (*Extensible Markup Language for Discourse Annotation*) tools.

1. Introduction ¹

Constructed on the basis of linguistic principles and methodologies, spoken corpora are databases consisting of transcribed and tagged recorded samples of data, presented for use on computer-based or multimedia environments. There is a growing and immediate need for large-scale corpora of spoken Turkish (see, Tezcan Aksu 2006); however, there are no large-scale corpora of either Standard Turkish or Turkish dialects, consisting of richly annotated naturally occurring spoken data in Turkish.

In this article, we aim to present the design criteria, technological requirements and priorities in annotation for a spoken corpus for Turkish that aims to reflect its discursive and pragmatic features. We present how we handle these issues in METU Spoken Turkish Corpus Project, funded by TÜBİTAK. Before dwelling on the issues, we first provide below a very brief overview of METU Spoken Turkish Corpus Project.

2. METU Spoken Turkish Corpus Project

The METU Spoken Turkish Corpus Project (ODT-STD) aims for maximum reusability (i.e., “cross disciplinary acceptability” and “wide circulability” (Cattoni et al. 2002). We thus employ a stratified sampling of spoken discourse, taking into consideration text-analytic and sociolinguistic variables (e.g., location, age, gender, multi- vs. mono-party interaction, topic and genre variation, etc.). When completed, the project, aims to produce a corpus which will be accessible to researchers via internet and traditional media.

ODT-STD aims to construct a computer-based, searchable corpus of transcribed and tagged, naturally occurring samples of Turkish spoken in Turkey comprising at least one million words. Its annotation focus will be on certain aspects of the discursive features of Turkish spoken discourse. The project aims to produce the following products in the long and short run:

- 1- Audio and video recordings of daily speech (e.g., talk among the family and intimates, service encounters), focused conversations (e.g., classroom discourse and meetings), and mass media archives;
- 2- Transcription and annotation of linguistic and discursive features of spoken Turkish (e.g., morphological analysis; T/V use, speech formulae, repairs, and overlaps);
- 3- Metalanguage and gesture annotation (e.g., head and hand movements, laughing); and
- 4- Manuals for the transcription and tagging system utilized in the corpus and for annotation of metalanguage and gestures.

With the use of EXMARaLDA annotation tool, selected audio and video files and tagged files will be presented for use on computer-based or multimedia environments. Users will be able to search on the website and retrieve words, idioms, speech acts and morphological units (see, Schmidt 2004). In order to achieve these aims, ODT-STD converges both with the features of old generation corpora and with those of new generation corpora. To see why such convergence is important, it is necessary to discuss some of the major features of old generation corpora.

3. Some Properties of Old Generation Corpora

British National Corpus (BNC) and *American National Corpus* (ANC) can be considered as old generation corpora. Excluding BNC and ANC, the size of old generation spoken corpora is between 52.600 and one million words (e.g., *London-Lund Spoken English Corpus* (LLC); *Lancaster/IBM Spoken English Corpus* (SEC); *Santa Barbara Corpus of Spoken American English*). In this respect, ODT-STD, containing one million words, will be comparable in size to these smaller corpora.

Audio and video recordings of BNC were collected with two major criteria: contextual and demographic (Crowdy 1993). In order to present language use in different contexts, the following were recorded: lectures and conversations in educational and informational settings, a variety of

radio and television programs (e.g., news, discussion and talk-shows), business meetings (i.e. job interview, counseling), formal or informal talks (politic talk), and entertainment programs (e.g., sports programs and talk-show on sports activities).

All recordings were done systematically in twelve regions of England. The recordings of the second sample of BNC were done by 124 volunteers from four social groupings living in 38 different locations across the UK. The age range of the female and male volunteers was between 15 and over 60 years. Each volunteer carried personal audio-recorders and recorded his/her conversation over 2 or 15 days. After each conversation, those who participated in the recordings were asked to give permission for their speech to be included in the corpus. For the speech included, the following metadata are tagged in the corpus:

- (i) Location, date and time of recording
- (ii) Setting and talk features
- (iii) Topic of talk and surrounding activity
- (iv) Information about participants' gender, age, nationality, occupations, social status and dialects

Besides transcription of recordings, the following linguistic features are annotated in BNC:

- (i) Filled and unfilled pauses
- (ii) False starts
- (iii) Overlaps and repetitions
- (iv) Paralinguistic features

BNC is a frequently used corpus model (e.g., *Corpus of Spoken New Zealand English*). Nevertheless, it has certain weaknesses:

1. The written part is more comprehensive than the spoken part;
2. Researchers who plan to use BNC in their research state that BNC presents only the language use only at the end of the 20th century. Thus it does not present language change as does *London-Lund Spoken Corpus*;
3. Only a small part of the corpus represents prosodic features; and
4. Phonetic features are not annotated (Crowdy 1994).

In addition to the above, deep orthography is not used in BNC but some features related to spoken discourse (e.g., pauses and metalanguage) are annotated. However, BNC does not contain the annotation of pragmatic and discursive elements, which are seen as essential for new generation spoken corpora (McEnery et al. 2006; Garside et al. 1997; McEnery and Wilson 2001). DAMSL (Dialog Act Markup in Several Layers, Allen and Core 1997) and ADAM (Cattoni et al. 2002) can be given as sample corpora for the annotation of pragmatic and discursive features. In these new generation corpora and others, features such as requests, appreciation tokens, responses and

agreements are annotated in the light of speech act theory (Searle 1976) and discourse structure such as that developed by Coulthard (1977).

Similar to BNC, ANC and *Corpus of Spoken New Zealand English*, ODT-STD will be based on a variety of sociolinguistic criteria in the process of compilation and annotation. Hence, the corpus will play a role in the comprehensive representation of contemporary Turkish.

The ODT-STD project is modeled on new generation corpora, which contain conversational features (e.g., discursive and pragmatic features). It also aims to employ deep orthography and to annotate paralinguistic features.

4. Annotation Tool of ODT-STD: EXMARaLDA

In the ODT-STD Project, EXMARaLDA annotation tool will be used to transcribe recordings and annotate discursive and pragmatic features. EXMARaLDA annotation tool is composed of three parts: Partitur-Editor, Corpus Manager (Coma) and Exact (Search Engine)

Partitur Editor transcribes turns in a format similar to musical scores and links transcriptions with audio and video-recordings. This feature of Partitur Editor is essential for both annotators and users of the corpus since they can watch or listen text portions they select. The linguistic annotation of turns is tagged in Partitur Editor (e.g. utterance units, metalanguage, overlaps, false starts, word and utterance lists, and transcriber comments). In this respect, Partitur Editor supports the annotation of discursive and pragmatic features. The tool supports transcription systems like HIAT, GAT, DIDA and CHAT. ODT-STD will use HIAT transcription system in order to annotate filled and unfilled pauses, false starts, overlaps, repetitions and paralinguistic features.

The metadata of the recordings (i.e., age, gender, location, date and time of recording, topic of talk, the name of the transcriber) are encoded using Partitur Editor. In this respect, the metadata will be similar to those in BNC.

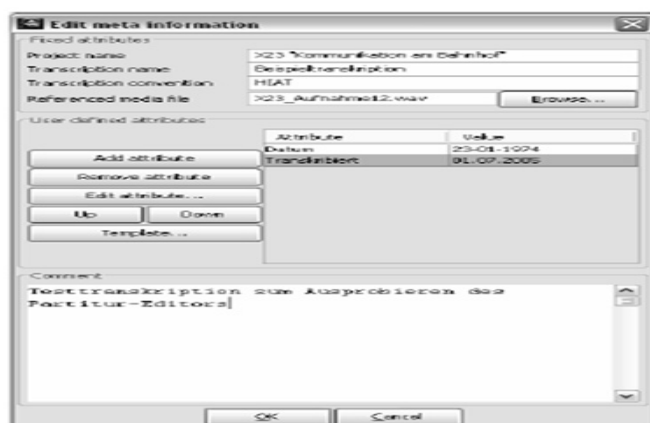


Image 1: A Snapshot from Partitur Editor

With the use of Partitur Editor, information about participants in recordings is also coded (see below, *A Snapshot from Partitur Editor for Participant Information*). In the ODT-STD project, demographic information (i.e., age, gender, nationality and occupation) will be coded in this part.

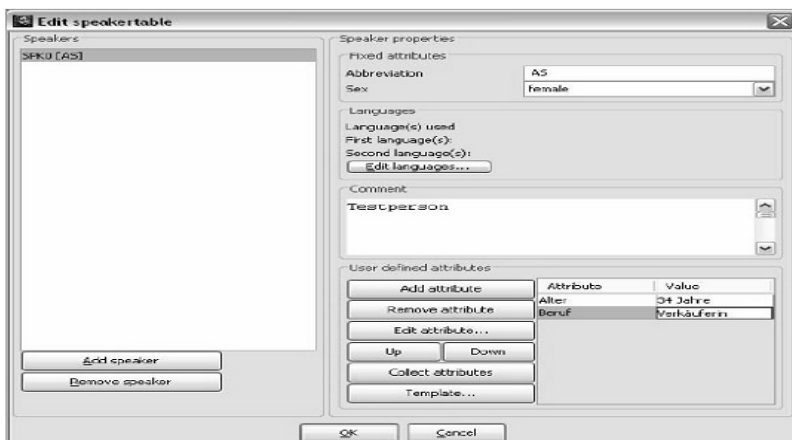


Image 2: A Snapshot from Partitur Editor for Participant Information

Another component of EXMARaLda is Corpus Manager. This component allows search for metadata attributes and lists attributes of transcripts and speakers. In addition, on the right side of the screen, it presents the number of speakers and their names in the recordings (see below, *A Snapshot from Coma*).

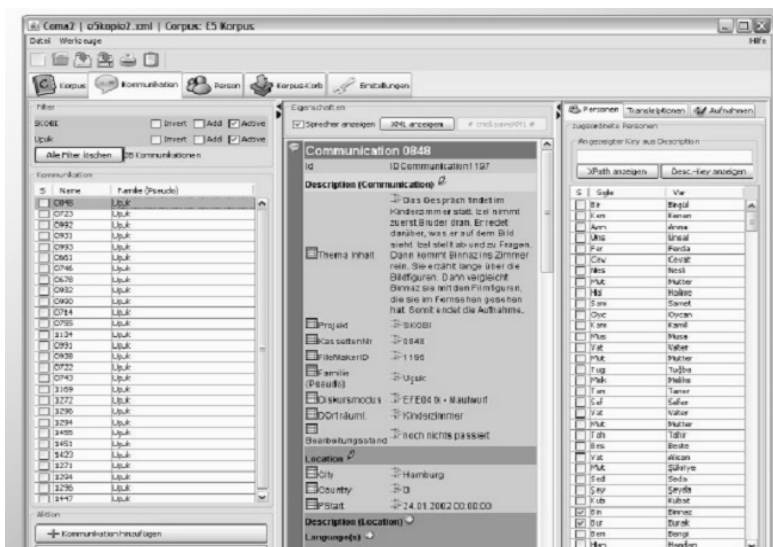


Image 3: A Snapshot from Coma

In EXMARaLda, another search and query instrument is Exact. It allows search for both linguistic features and metadata, and presents the search results within a larger textual span.

Comparing EXMARaLda with other annotation tools such as ELAN, ANVIL, TASX, the following advantages are observed:

- Transcripts are linked to audio and/or video files.
- Data can be transferred from other applications such as ELAN, TASX, and Praat.
- Transcription according to a number of systems can be implemented (e.g., HIAT, GAT, DIDA and CHAT).
- A few small-scale corpora have been compiled with this system.
- It was used in the compilation of small-scale Turkish corpora.

These features suggest that EXMARaLda can be used for compiling Turkish spoken corpora.

5. ODT-STD: Corpus Design

In the project audio and video recordings will be compiled with the methods below:

1. Recordings where the research team are co-participants
2. Recordings by volunteers, some of whom will also be involved in annotation
3. Telephone recordings
4. Video recordings by the research team

5.1 Criteria for Recordings

As mentioned above, ODT-STD will initially comprise one million words. This means that the corpus will not reflect all regional dialects with acceptable representative scope. The corpus will therefore give priority to register variation, a point which Biber (1993) emphasizes as being critical for general corpora. Considering the literature on Turkish corpora, there are no large-scale databases or resources for spoken Turkish. Hence, register variation is essential in the compilation of recordings in the project. The table below lists the registers that the corpus will comprise. In this manner ODT-STD aims to achieve representative validity.

Table 1

		PARTICIPATION FORMATS AND SETTINGS
Participation type:		TALK TYPE
Medium:	Topic of conversation:	Personal/impersonal
Face-to-face:	A. Chats	1) In the family; family with guests (e.g., at dinner) 2) Educational locations (e.g., chats during lunch or coffee) 3) Chats in business locations
	B. Institutional or semi-institutional	5) In hospitals/medical centers: (e.g.: doctor-patient encounters) 6) Rituals (e.g., engagements; festivities in business locations; condolences) 7) On public transportation (e.g. inter-city buses, taxi, on the <i>dolmuş</i> ¹) 8) Service encounters (e.g., making an appointment, malls, bazaar) 9) Business settings (e.g., meetings, talk in the secretary's office; job interviews) 10) Educational settings: meetings 11) Classroom discourse: Lectures; group activities
Telephone:	1) Institutional	2) Between family members and friends
Mass media:	1) TV and radio talk that is close to spontaneous talk (e.g., talk shows)	2) Scripted (e.g., excerpts from series) 3) Text reading (e.g., news)

Recordings of the regional variations listed above will be done as far as possible in the allocated time.

5.2 Linguistic Analysis and Annotation

An “agile” corpus design and annotation scheme is planned to be implemented in the ODT-STD Corpus. That is, compilation of the recordings and the annotation schemes will be revised cyclically. The following figure presents how this cyclical revision will take place.

¹ *dolmuş*: a minibus used for public transportation

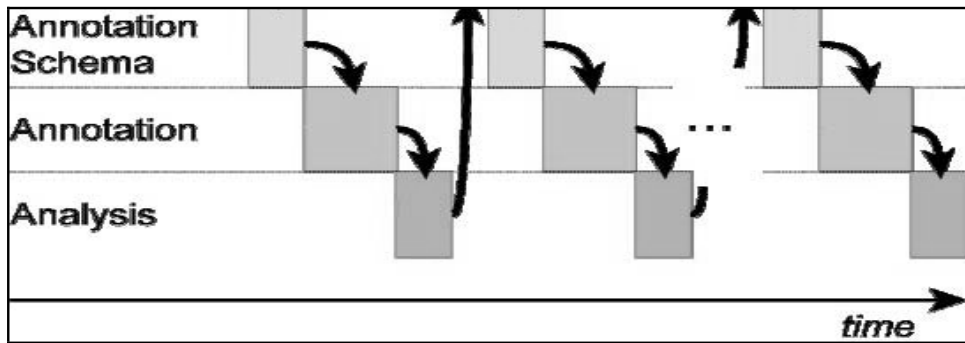


Image 4 (From Gut 2008; Voormann and Gut 2008)

In ODT-STD, the annotation will elaborate on the discursive and pragmatic features of spoken Turkish.

5.2.1 Transcription and Morphemic Analysis

Deep orthography will be applied (Cattoni et al. 2002). Dialectal variation and wrong enunciations will be kept as in the original and the standard forms will be indicated in transcriber tiers. HIAT will be used and improved for transcriptions. HIAT allows for tagging of the morphemic and discursive features of spoken language (i.e., overlapping, repairs and false start) (see Rehbein et al. 2004). Morphemic annotation will be done in some parts of the transcription. EXMARaLDA annotation tool automatically presents word list but a program will be developed in order to transfer the morphemic analysis to the annotation.

5.2.2 Pragmatic Annotation

Interactional sociolinguistics and the field of discourse analysis reveal the following features of spoken discourse as significant in interaction:

- Context and alignments (e.g., overlaps, repairs)
- Footing (e.g., address forms, agreements, paralinguistic features)
- Contextualization cues (e.g., register changes, code-switching)
- Interactional utterances (e.g., formulaic expressions)
- Other pragmatic markers: Discourse markers (e.g., *ancak* and *fakat*), discourse particles (e.g., *yani* and *işte*), and interjections (see, Goffman 1967, 1971)

ODT-STD aims to enable automatic search of pragmatic elements in Turkish. Therefore, priority will be given to the annotation of following pragmatic features to investigate the pragmatic features stated above:

- a. Pragmatic markers (e.g., primary and secondary interjections (Norrick 2008), discourse markers and discourse particles)
- b. Discourse deixis (e.g., pronominal *bu* (this) and *şu* (this/that))

- c. Overlaps, filled and unfilled pauses, repairs
- d. Discursive formulaic expressions (e.g., thanking formulae; (dis)agreement markers)
- e. (Im)politeness markers (address forms, T/V, tense/aspect)
- f. Metalanguage (laughing, puffing, etc.)

The annotation will be based on the principle of least interpretive work on the part of the transcriber. To illustrate, the overlaps will not be coded as interruption or collaboration. The macro-structure of the texts will not be annotated for the time being, as there is still much debate in the literature on how best to accomplish this (Carletta 1996; Allwood 2001). The available literature on Turkish discourse (Atabay et al. 1983) is being used to prepare the annotation scheme and pilot recordings are being examined to develop it.

The annotation of pragmatic markers follows a hierarchical coding system. That is, discourse markers are being annotated for morphology and semantic contribution. For instance, the following coding will be used for interjections:

- a. Onomatopoeic: *uff*, *vay*
- b. Lexical: *aman*
- c. Compound lexical: *aman yarabbim*
- d. Mixed (onomatopoeic and lexical): *yapma ya*

6. Conclusion

ODT-STD aims to integrate the features of old and new generation corpora. A compilation of a corpus for spoken Turkish is an endeavor that incorporates both research and analysis, as research on aspects of the (non-)linguistic characteristics of spoken Turkish is still a relatively new field, the findings of which are still not fully reflected in reference grammars of Turkish. Therefore, manuals for annotations will be developed during the transcription process.

[1] Note

A Turkish version of this paper was presented at the Mersin 2008 Symposium held at Mersin University, 19-21 November 2008.

References

- Allen, James & Core, Mark. (1997) "Dialog act markup in several layers", Retrieved from <ftp://ftp.cs.rochester.edu/pub/packages/dialog-annotation/manual.ps.gz>
- Allwood, Jens (ed.). (2001) "Dialog Coding — Function and Grammar Göteborg Coding Schemas", *Gothenburg Papers in Theoretical Linguistics* No. 85.
- Atabay, Neşe, Kutluk, İbrahim, Özel, Sevgi. (1983) *Sözcük Türleri*, TDK Yayınları, Ankara.
- Biber, Douglas. (1993) "Using register-diversified corpora for general language studies", *Computational Linguistics* 19 (2): 219-241.

BNC www.natcorp.ox.ac.uk

- Carletta, Jean, Isard, Stephen, Doherty-Sneddon Gwyneth, Isard, Amy, C. Kowtko, Jacqueline, H. Anderson, Anne. (1997) "The reliability of a dialogue structure coding scheme", *Computational Linguistics* 23 (1): 13-31.
- Cattoni, Roldona, Danieli, Morena, Sandrini, Vanessa, Soria, Claudia. (2002) "ADAM: The SI-TAL corpus of annotated dialogue", Retrieved from <https://nats-www.informatik.uni-hamburg.de/intern/proceedings/2002/LREC/pdf/237.pdf>
- Coulthard, Malcolm. (1977) *An Introduction to Discourse Analysis*, Longman, London.
- Crowdy, Steve. (1993) "Spoken corpus design", *Literary and Linguistic Computing* 8 (4): 259-265.
- Crowdy, Steve. (1994) "Spoken corpus transcription", *Literary and Linguistic Computing* 9 (1): 25-28.
- Garside, Roger, Leech, Geoffrey, McEnery, Tony. (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Longman, London.
- Goffman, Erving. (1967) *Interaction Ritual*, Anchor Books, New York.
- Goffman, Erving. (1971) *Frame Analysis*, Harper and Row, New York.
- Gut, Ulrike. (2008) "Corpus creation", *Summer School on Corpus Phonology*, University of Augsburg. Retrieved November 15, 2008 from www.corpho.eu/?u_act=download&dfile=Summerschool_corpus_creation.pdf
- Norrick, Neal. (in print) "Interjections as pragmatic markers", *Journal of Pragmatics*.
- McEnery, Tony, Xiao, Richard, Tono, Yukio. (2006) *Corpus-Based Language Studies*, Routledge, Oxon.
- McEnery, Tony, Wilson, Andrew. (2001) *Corpus Linguistics*, Edinburgh University Press, Edinburgh.
- Rehbein, Jochen, Schmidt, Thomas, Meyer, Bernd, Watzke, Franzisks, Herkenrath, Annette. (2004) "Handbuch für das computergestützte transkribieren nach HIAT", *Arbeiten zur Mehrsprachigkeit Folge B* (Nr. 56), Universität Hamburg: Sonderforschungsbereich Mehrsprachigkeit.
- Santa Barbara Corpus of Spoken American English*. <http://www.linguistics.ucsb.edu/research/sbcorpus.html>
- Searle, John. (1976) "A classification of illocutionary acts", *Language in Society* 5 (1): 1-23.
- Schmidt, Thomas. (2004) "Transcribing and annotating spoken language with EXMARaLDA", *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*, Paris: ELRA. http://www1.uni-hamburg.de/exmaralda/Daten/4D-Literatur/Paper_LREC.pdf
- Tezcan Aksu, Belgin. (ed.) (2006) *Bilgisayar Destekli Dil Bilimi Çalıştayı Bildirileri, 14 Mayıs 2005*. TDK Yayınları, Ankara.
- Voormann, Holger & Gut, Ulrike. (2008) "Agile corpus creation", *Corpus Linguistics and Linguistic Theory* 4 (2): 235-251.
- Wellington Corpus of Spoken New Zealand English*. khnt.hit.uib.no/icame/manuals/wsc/INDEX.HTM