FEATURES FOR AN INTERNET ACCESSIBLE CORPUS OF SPOKEN TURKISH DISCOURSE

Şükriye RUHİ
sukruh@metu.edu.tr
Derya ÇOKAL KARADAŞ
cokal@metu.edu.tr
Middle East Technical University

THE METU SPOKEN TURKISH DISCOURSE PROJECT (ODT-STD)

- October 2008-October 2010
- A TUBİTAK EVRENA project
- Research Team
  - Şükriye Ruhi
  - Betil Eröz
  - Çiler Hatipoğlu
  - Derya Çokal Karadaş
  - Hale Işık-Güler
  - Güneş Acar
  - Kerem Eryılmaz
  - Hümeyra Can

PROJECT PRODUCTS

- Using EXMARaLDA for annotation and transcription, ODT-STD will consist of the following products in the long run:
  - Audio and video-recorded everyday talk (e.g., family talk and talk among intimates, service encounters), talk for specific purposes (e.g., meetings, classroom discourse), mass media archives;
  - Transcription and annotation of linguistic and discursive features of spoken Turkish (e.g., mophological analysis; T/V use, speech formulae, repairs, overlaps)
  - Metalanguage and gesture annotation (e.g., head and hand movements, laughing)
- Manuals for transcription and annotation
  - Manual for transcription
  - Manual for annotation of pragmatic elements
  - Manual for annotation of metalanguage and gestures

SOME PROPERTIES OF SPOKEN CORPORA

- Most spoken corpora, excluding BNC ve ANC, are between 52,600 - 1 million words. E.g.,
  - *London-Lund Corpus* **(**LLC**)**
  - *Lancaster/IBM Spoken English Corpus* (SEC)
  - *Santa Barbara Corpus of Spoken American English*

FEATURES OF THE SPOKEN COMPONENT OF BNC

- Demographic: 38 locations, 4 different socio-economic groups; age range: 15 - 60 and above, 124 men and women – all volunteers; 2000 hour recording.
- Method of recording: Volunteers recorded talk during their daily activities over 2-15 days.
- Audience was informed of recording and allowed to erase recording.

ANNOTATION IN BNC - 2

- **Metadata coding**:

    (i) Location, date and time of recording

    (ii) Setting and talk features

    (iii) Topic of talk and surrounding activity

    (iv) Gender, age, race, occupation, education, social class, relationship, dialect

- **Annotations in transcription**:

    (i) filled and unfilled pauses

    (ii) False starts

    (iii) Overlaps and repetitions

    (iv) Paralinguistic features

NEW GENERATION SPOKEN CORPORA

- Deep orthography (e.g., *bir* vs. *bi* in Turkish)
- Metalanguage features
- Dialogue annotation
  - Speech acts (e.g., requests, appreciation tokens)
  - Discourse moves (e.g., responses; agreements)

SPOKEN CORPORA IN TURKISH

- *OrienTel Turkish Database*
- *Turkish Speecon Database*
- *Turkish Continuous and Isolated Word Speech Database*
- *Multilingual Turkish Corpus*
- *Interpreting in Hospitals*
- *Linguistic Connectivity in Bilingual Turkish-German Children*

ANNOTATION TOOL of ODT-STD: EXMARalDA

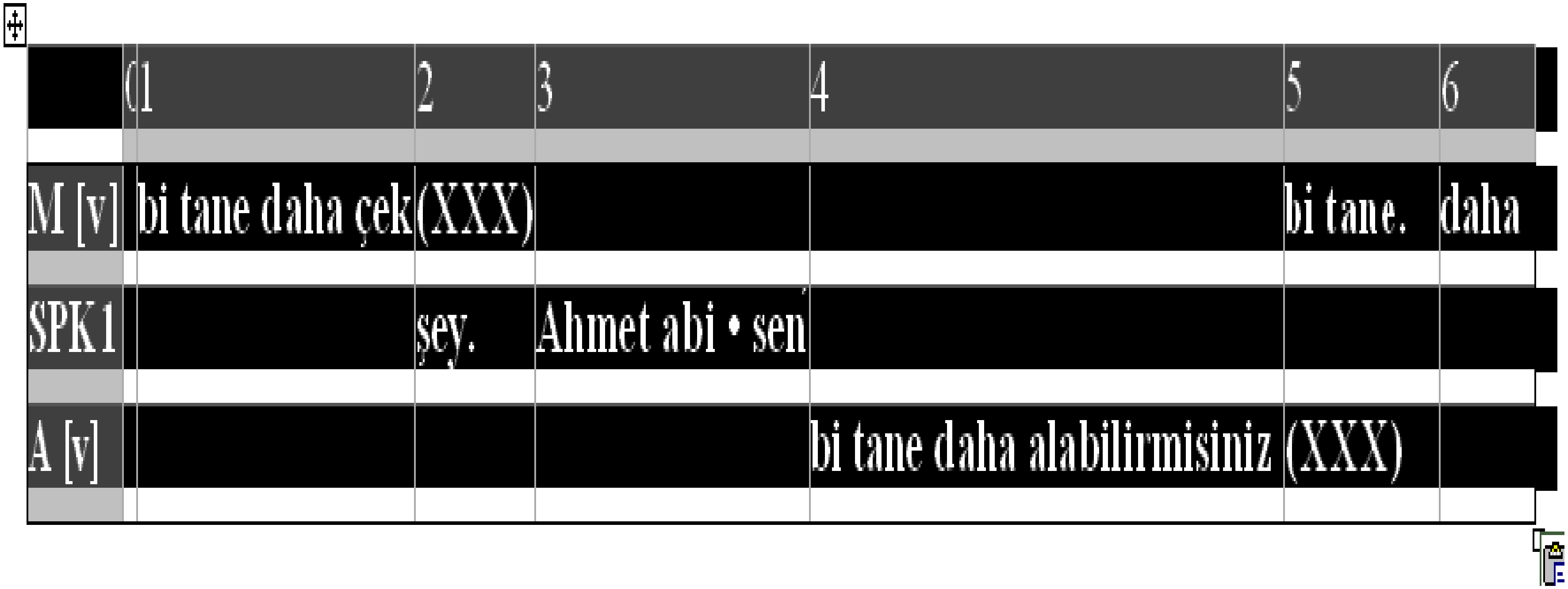
- **EXMARaLDA's components**:
  - Partitur-Editor
  - Corpus Manager (CoMA)
  - Exact (search engine)
- **Partitur-Editor**:

  1. Transcribing turns in a format similar to musical scores

  2. Linking transcriptions with audio and video-recordings

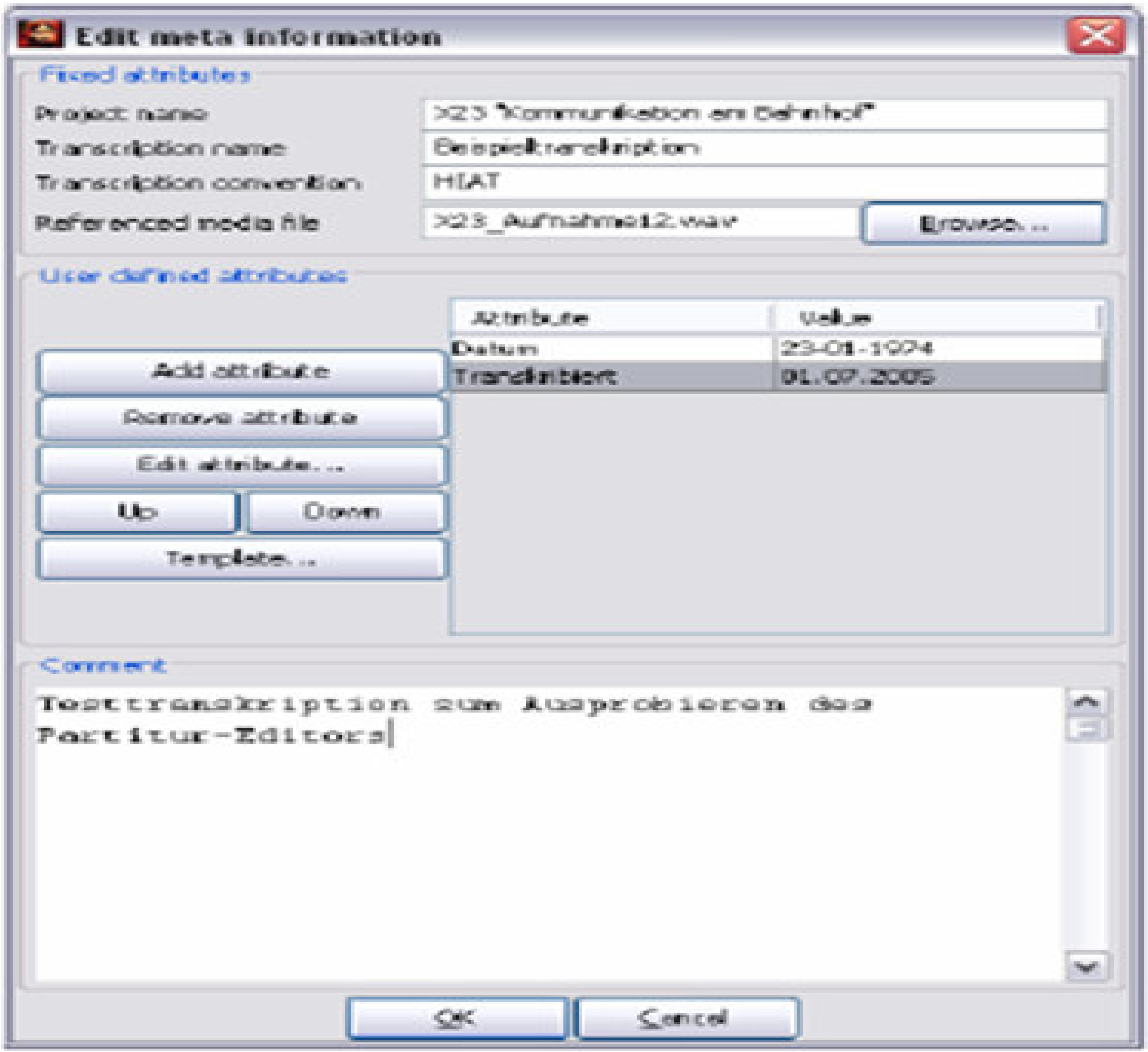
  3. Linguistic annotation of turns (e.g., utterance units, metalanguage, overlaps, false starts, word and utterance lists, transcriber comments)

SAMPLE TRANSCRIPTION WITH PARTITUR



	0	1	2	3	4	5	6
M [v]		bi tane daha çek	(XXX)			bi tane.	daha
SPK1			şey.	Ahmet abi • sen´			
A [v]					bi tane daha alabilirmisiniz	(XXX)	

ENTERING METADATA WITH PARTITUR



Project name: X23 "Kommunikation am Bahnhof"

Transcription name: Beispieltranskription

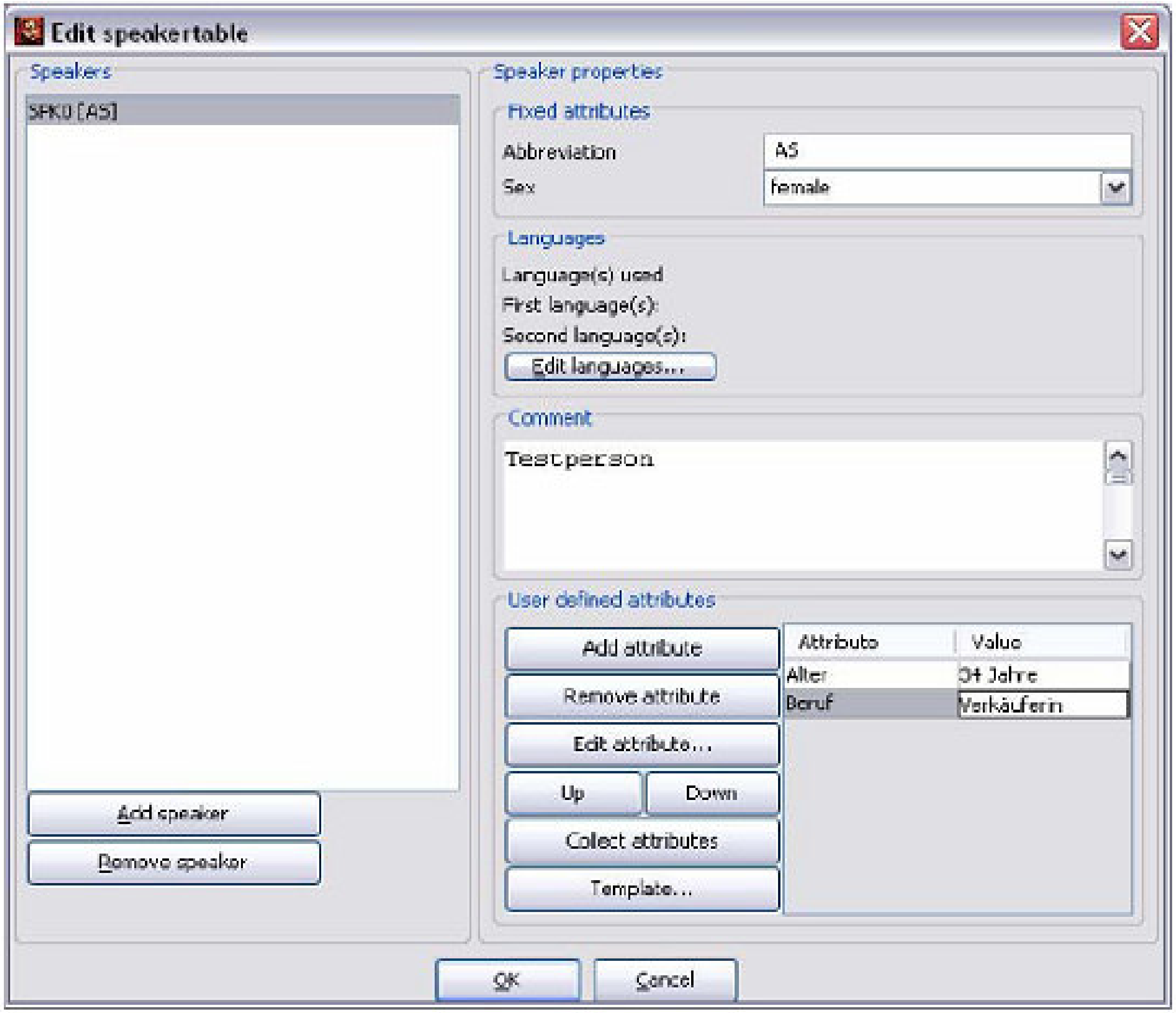
Transcription convention: HIAT

Referenced media file: X23_Aufnahme12.wav

Attribute	Value
Datum	23-01-1974
Transkribiert	01.07.2005

Testtranskription zum Ausprobieren des Partitur-Editors

ENTERING SPEAKER METADATA WITH PARTITUR



Edit speakertable

Speakers

SPK0 [AS]

Speaker properties

Fixed attributes

Abbreviation: AS

Sex: female

Languages

Language(s) used

First language(s):

Second language(s):

Comment

Testperson

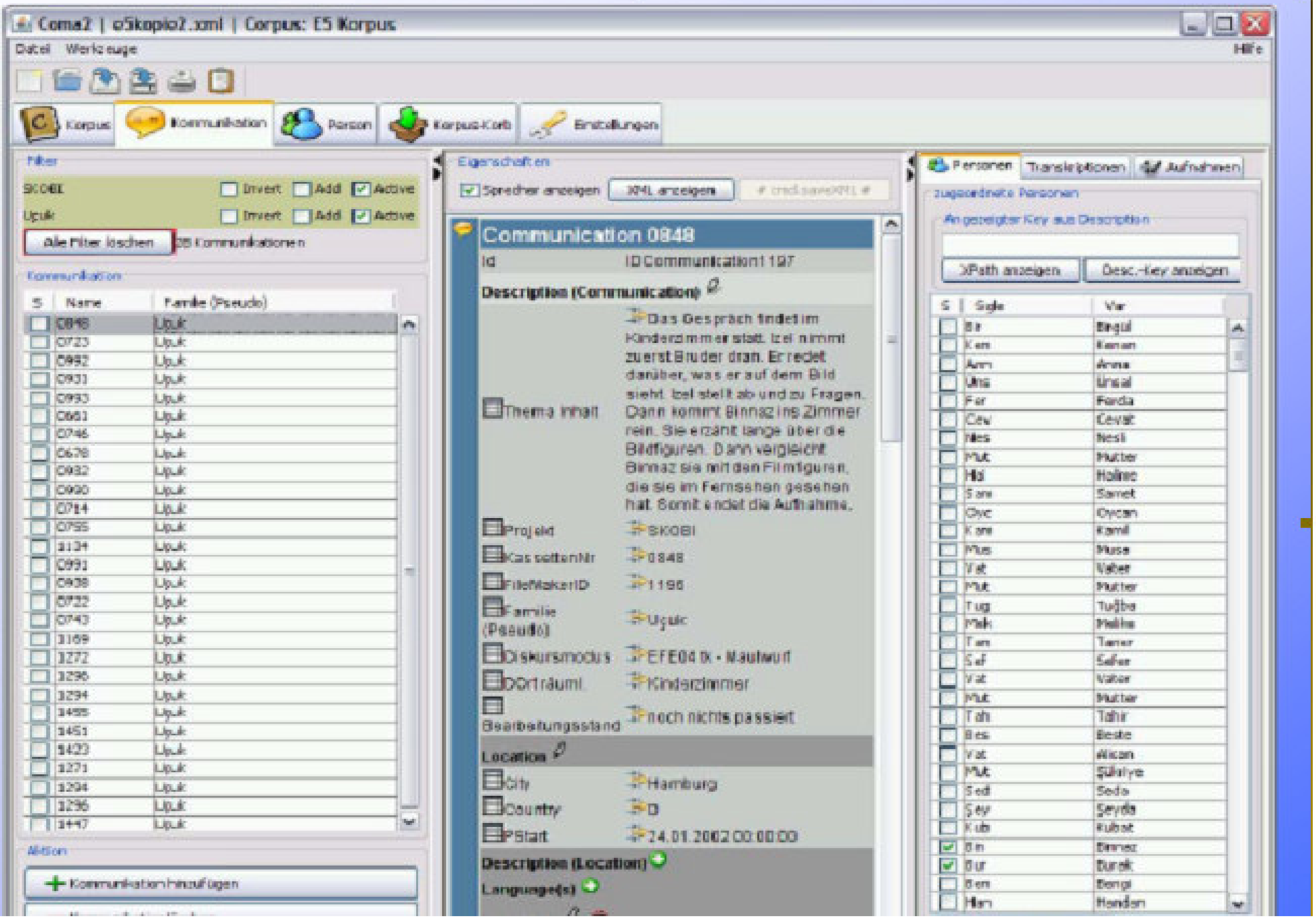
User defined attributes

Attribute	Value
Alter	34 Jahre
Beruf	Verkäuferin

EXMARaLDA - CoMA

- Corpus Manager:
  - Corpus metadata (identity of transcribers, method used in annotation; speaker attributes, location of talk, etc.)
  - Allows search for metadata attributes;
  - Lists attributes of transcripts and speakers.

VIEW FROM CoMA



Coma2 | e5kopie2.xml | Corpus: E5 Korpus

Filter

SKOBI	□ Invert	□ Add	☑ Active
Uçuk	□ Invert	□ Add	☑ Active

Alle Filter löschen 28 Kommunikationen

Kommunikation

S	Name	Familie (Pseudo)
□	0848	Uçuk
□	0723	Uçuk
□	0992	Uçuk
□	0931	Uçuk
□	0993	Uçuk
□	0661	Uçuk
□	0746	Uçuk
□	0678	Uçuk
□	0932	Uçuk
□	0930	Uçuk
□	0714	Uçuk
□	0755	Uçuk
□	1134	Uçuk
□	0991	Uçuk
□	0938	Uçuk
□	0722	Uçuk
□	0743	Uçuk
□	1169	Uçuk
□	1272	Uçuk
□	1296	Uçuk
□	1294	Uçuk
□	1455	Uçuk
□	1451	Uçuk
□	1423	Uçuk
□	1271	Uçuk
□	1294	Uçuk
□	1296	Uçuk
□	1447	Uçuk

Aktion

Kommunikation hinzufügen

Eigenschaften

☑ Sprecher anzeigen XML anzeigen # cmdi:saveXML #

Communication 0848

Id	IDCommunication1197
Description (Communication)	
Thema Inhalt	Das Gespräch findet im Kinderzimmer statt. Izel nimmt zuerst Bruder dran. Er redet darüber, was er auf dem Bild sieht. Izel stellt ab und zu Fragen. Dann kommt Binnaz ins Zimmer rein. Sie erzählt lange über die Bildfiguren. Dann vergleicht Binnaz sie mit den Filmfiguren, die sie im Fernsehen gesehen hat. Somit endet die Aufnahme.
Projekt	SKOBI
KassettenNr	0848
FileMakerID	1196
Familie (Pseudo)	Uçuk
Diskursmodus	EFE04 tk - Maulwurf
DOrt räuml	Kinderzimmer
Bearbeitungsstand	noch nichts passiert
Location	
City	Hamburg
Country	D
PStart	24.01.2002 00:00:00
Description (Location)	
Language(s)	

Zugeordnete Personen

Angezeigter Key aus Description

XPath anzeigen Desc.-Key anzeigen

S	Sigle	Vor
□	Bir	Birgül
□	Ken	Kenan
□	Ann	Anna
□	Uns	Ünsal
□	Fer	Ferda
□	Cev	Cevat
□	Nes	Nesli
□	Mut	Mutter
□	Hal	Halime
□	Sam	Samet
□	Oyc	Oycan
□	Kam	Kamil
□	Mus	Musa
□	Vat	Vater
□	Mut	Mutter
□	Tug	Tuğba
□	Mel	Meltem
□	Tan	Taner
□	Saf	Safer
□	Vat	Vater
□	Mut	Mutter
□	Tah	Tahir
□	Bes	Beste
□	Vat	Akcan
□	Mut	Şükriye
□	Sed	Seda
□	Şey	Şeyda
□	Kub	Kubat
☑	Bin	Binnaz
☑	Bur	Burak
□	Ben	Bengi
□	Han	Handan

SOME FEATURES OF EXMARaLDA

- Transcripts are linked to audio and/or video files.
- Allows for data transfer from other applications such as ELAN, TASX, and Praat.
- Allows for transcription according to a number of systems (e.g., HIAT, GAT, DIDA and CHAT).
- A few small-scale corpora have been compiled with this system.

ODT-STD: CORPUS DESIGN

- Audio and video recordings will be compiled with four methods (see, below).
- Where there is no permission for links to the audio files, only transcriptions will be made and identifiers will be changed (e.g., names).

  1.Recordings where the research team is a co-participant

  2. Recordings by volunteers, some of whom will also do transcriptions

  3. Telephone recordings

  4. Video recordings by the research team

CRITERIA FOR RECORDINGS

- The initial size of the corpus will be 1 million words. Given this small size, the corpus, initially, will **not** be able to reflect all regional dialects.
- The corpus will thus give **priority to register variation** (Biber 1993).
- Table 1 lists the registers that the corpus will comprise. In this respect, ODT-STD will achieve **representative** validity.
- (We expect to achieve a 10 million word corpus by 2013.)

Yakınlık :	**Talk Type**	**PARTICIPATION FORMATS AND SETTINGS**
Topic of conversation :	Personal/impersonal	
Participation type	1) Monologue	2) Dialogue a. 2 -5 persons b. 6 -10 persons c. More than 10
Medium	1) face-to-face	2) Telephone 3) Mediated (e.g. broadcasts)
Face-to-face	**A**. *Sohbet* (chats)	1) In the family; family with guests (eg., at dinner; family get togethers) 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.: Kız isteme; engagements; festivities in business locations; condolences 6) On public transportation: E.g. inter-city buses, taxi, on the *dolmuş*) 7) Service encounters: E.g., making an appointment, malls, bazaar 8) Business settings: E.g. meetings; talk in the secretary's office; job interviews 9) Educational settings: meetings 10) 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) Scrpited (e.g., excerpts from series) 3) Text reading (e.g., news)

DEMOGRAPHIC DATA

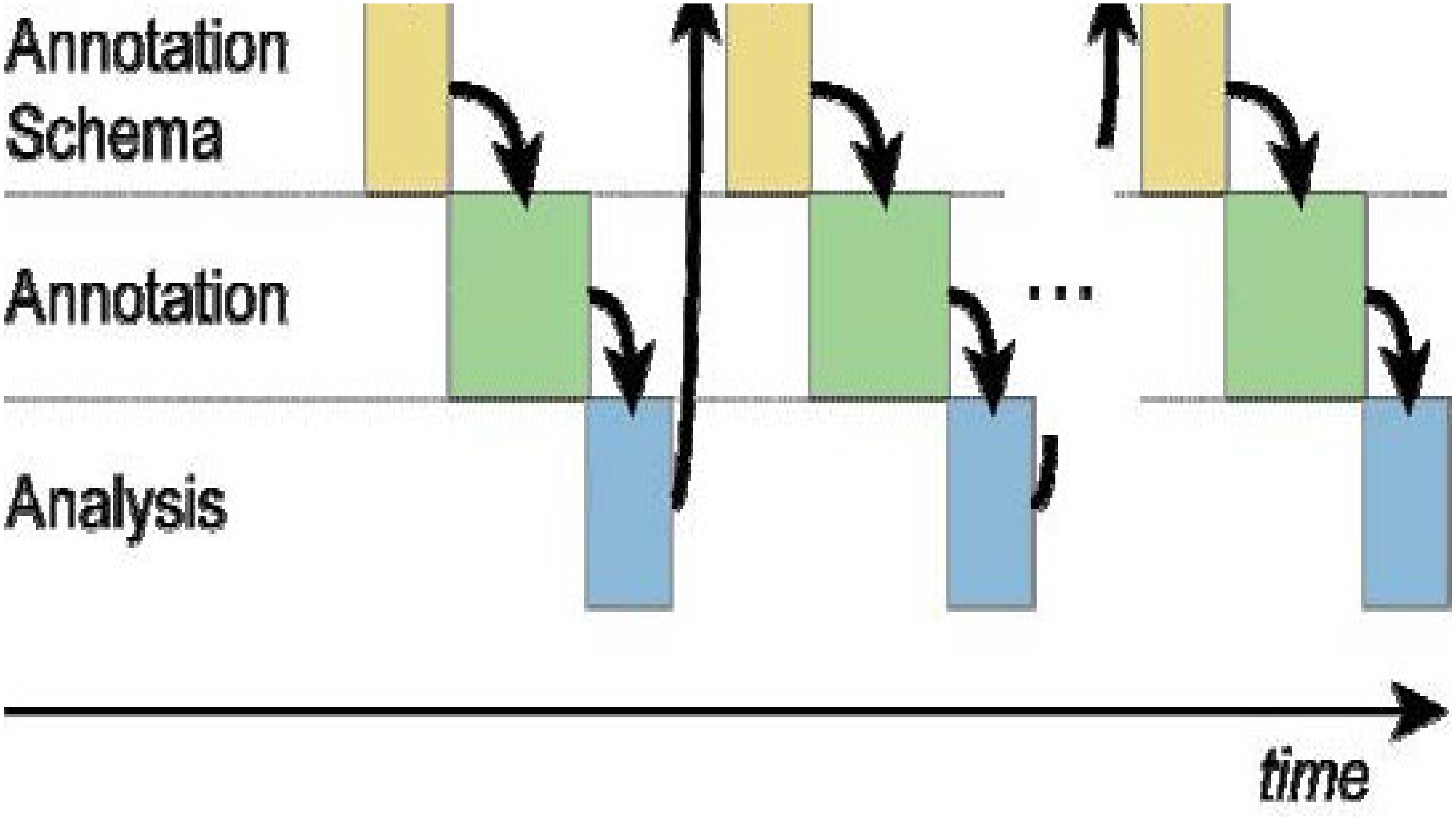
1 Citizenship	8 Geographical location
2 Age	9 Place of birth
3 Gender	10 Residency
4 Marital status	11 Languages spoken
5 Education	12 Residency outside TR
6 Occupation	13 Duration of residency outside TR
7 Number of children	.....

LINGUISTIC ANALYSIS AND ANNOTATION - 1

- 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 for transcriptions.

LINGUISTIC ANALYSIS AND ANNOTATION - 2

From Gut 2008 and Voormann and Gut (2008). : An “agile” corpus design and annotation scheme is implemented. That is, both the compilation of the recordings and the annotation schemes will be revised cyclically.



PRAGMATIC ELEMENTS NOTED IN THE LITERATURE

- Interactional sociolinguistics and the field of discourse analysis reveal the following as significant in interaction:
  - *Context and alignments* (e.g., overlaps, repairs)
  - *Footing* (e.g., address forms, agreements, paralinguistic features)
  - C*ontextualization cues* (e.g., register changes, code-switching)
  - *Interactional utterances* (e.g., formulaic expressions)
  - *Other pragmatic markers*: Discourse markers, discourse particles, and interjections

PRIORITIES IN PRAGMATIC ANNOTATION

- ODT-STD aims to enable automatic search of pragmatic elements in Turkish. It will therefore give priority to annotation of the following:

  a. Pragmatic markers (e.g., primary and secondary interjections (Norrick 2008), discourse markers and discourse particles)

  b. Discourse deixis (e.g., pronominal *bu* (this), *şu* (this/that))

  c. Overlaps, filled and unfilled pauses, repairs

  d. Discursive formulaic expressions (e.g., thanking formulae; (dis)argeement markers)

  e. (Im)politeness markers (address forms, T/V, tense/aspect)

  f. metalanguage (laughing, puffing, etc.)

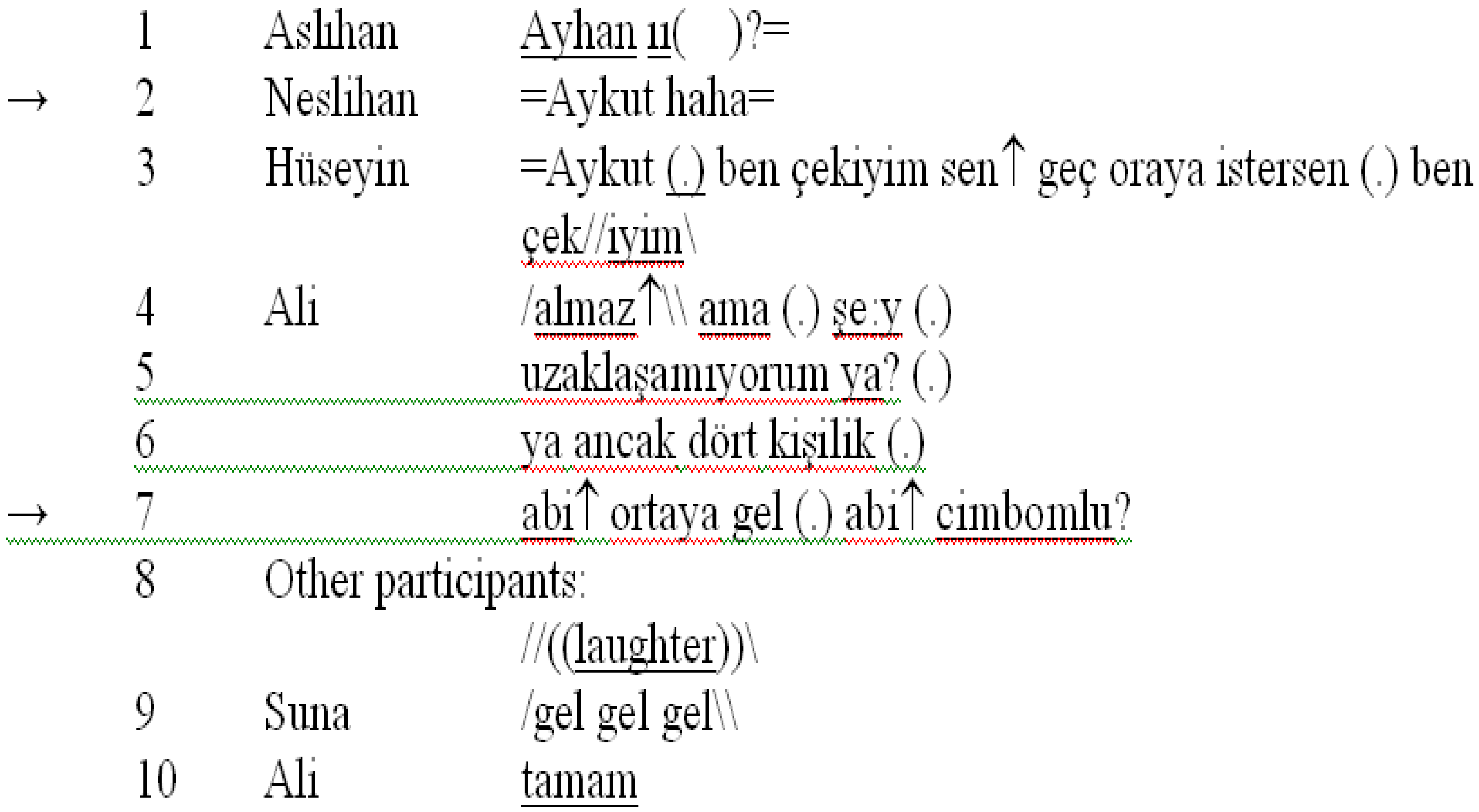
ANNOTATION PRINCIPLES

- The annotation is based on the principle of **least interpretive work** on the part of the transcriber.
  - E.g., When there are overlaps, these 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 preparation of the annotation scheme is making use of the available literature on Turkish discourse. The pilot recordings are also currently being examined to develop it.
- The annotation of pragmatic markers follows a hierarchical coding system.
  - E.g. Discourse markers are annotated for morphology and semantic contribution

EXAMPLE: INTERJECTIONS

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

A VIEW OF ELEMENTS TO BE ANNOTATED (represented according to conversation analysis)



```
     1    Aslıhan     Ayhan ıı(    )?=
→    2    Neslihan    =Aykut haha=
     3    Hüseyin     =Aykut (.) ben çekiyim sen↑ geç oraya istersen (.) ben
                      çek//iyim\
     4    Ali         /almaz↑\\ ama (.) şe:y (.)
     5                uzaklaşamıyorum ya? (.)
     6                ya ancak dört kişilik (.)
→    7                abi↑ ortaya gel (.) abi↑ cimbomlu?
     8    Other participants:
                      //((laughter))\
     9    Suna        /gel gel gel\\
    10    Ali         tamam
```

STRATEGIES in WORK IN PROGRESS

- A compilation of a corpus for spoken Turkish is an endeavour 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 on Turkish.

References

- Allwood, Jens (ed.), 2001. Dialog Coding — Function and Grammar Göteborg Coding Schemas. *Gothenburg Papers in Theoretical Linguistics* 85. Gothenburg University.
- Biber, D. (1993). Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics* 19, 2, 219-241.
- Cattoni, R., Danieli, M., Sandrini, V., Soria, C. (2002). ADAM: The SI-TAL Corpus of Annotated Dialogues. Retrieved from *https://nats-www.informatik.uni-hamburg.de/intern /proceedings/2002/LREC/pdf/237.pdf*
- Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., C. Kowtko,J., H. Anderson, A. (1997). The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics* 23, 1, 13-31.
- Cattoni, R., Danieli, M., Sandrini, V., Soria, C. (2002). ADAM: The SI-TAL Corpus of Annotated Dialogues Retrieved from *https://nats-www.informatik.uni-hamburg.de/ intern/proceedings/2002/ LREC/pdf/237.pdf*
- Norrick, N. (2008) Interjections as pragmatic markers. *Journal of Pragmatics*.
- Voormann, Holger, Ulrike Gut. 2008. Agile Corpus Creation. *Corpus Linguistics and Linguistic Theory* 4, 2, 235–251.

Thank you for listening

☺