Achieving representativeness through the parameters of spoken language and discursive features: the case of the Spoken Turkish Corpus

Şükriye RUHİ*
Hale IŞIK-GÜLER
Çiler HATİPOĞLU
Betil ERÖZ-TUĞA
Derya ÇOKAL KARADAŞ
Middle East Technical University

In this paper we overview the ongoing debate on achieving representativeness in general spoken corpora with the purpose of proposing a model for spoken corpora design and construction workflows. The proposal is illustrated in the context of an ongoing implementation for the Spoken Turkish Corpus, a corpus that will consist of one million words of present-day Turkish spoken in Turkey in its initial stage. The paper proposes a cyclic workflow and design scheme that is based on the principles of an "agile" corpus design and annotation system (Voorman and Gut, 2008), and argues that a three-pronged set of feature criteria, namely, demographic, contextual, and discursive features can be fruitfully combined to monitor and achieve representativeness. The paper discusses the underlying principles in the design scheme and outlines the metadata features of the web-based corpus management system, which utilizes and complements EXMARaLDA tools (Schmidt, 2004) in corpus construction and monitoring.

Key words: *spoken corpus design criteria, representativeness, metadata, discursive features, web-based corpus management*

En la presente ponencia se examina el debate en curso sobre adquirir la representatividad en el corpus hablado general con el objetivo de proponer un modelo para el diseño del corpus hablado y del volumen de trabajo de construcción. La propuesta está ilustrada dentro del marco de la implementación en curso para el Spoken Turkish Corpus , un corpus que estará formado de un millón de palabras de la lengua turca actual hablada en Turquía en su fase inicial. Esta ponencia propone un volumen de trabajo cíclico y un esquema de diseño que está basado sobre los principios de un "agile" diseño de corpus y del sistema de anotación. Voorman y Gut (2008) expone que una serie de criterios de características de tres-condiciones (etapas), a saber, las características demográficas, contextuales, y discursivas pueden estar perfectamente combinadas para monitorizar y conseguir representatividad. Esta ponencia discute los principios subyacentes en el esquema de diseño y traza las características metadata del sistema de gestión de corpus basado en el web que utiliza y complementa EXMARaLDA (Schmidt 2004) en la construcción y de la monitorización del corpus.

Palabras clave: diseño del corpus hablado criterios, representatividad, metadata, características discursivas, gestión de corpus basado en el web

1. INTRODUCTION[1]

Achieving representativeness, balancedness and comparability in corpus construction are three requirements that have and are still engaging scholars in debate as to how best to

[1] This study and STC is financed by a research grant from the Turkish Scientific and Technological Research Institution (Türkiye Bilimsel ve Teknolojik Araştırma Kurumu, TÜBİTAK, Grant No. 108K283). We are deeply grateful to Dr. Thomas Schmidt and Dr. Kai Wörner for their support in the construction of STC.

*Corresponding author. Contact: Dept. of Foreign Language Education, Faculty of Education, Middle East Technical University, İnönü Blvd., 06531 Ankara, Turkey, e-mail: sukruh@metu.edu.tr, sukriyeruhi@gmail.com

approach these issues in terms of theory, methodology, and the dire practicalities of corpus compilation, especially since Biber's (1993) seminal article on representativeness (see, e.g., Leech, 2007; Váradi, 2001). Two central points of this debate concern approaches to sampling (proportional vs. stratified) and the conceptualization of frequency of communication types. Underlying the various debates is the fundamental question: What is it that one expects to achieve with the corpus construction? Is it to produce a resource that lays open a maximal view on language variation (Biber, 1993), or is it to produce a resource in the standard statistical sense of representing language, based on demographic criteria (Váradi, 2001). While the two appear to be in opposition, both goals translate themselves within corpus linguistics into the expectation that one should be able to use the resource to make generalizations about the language (Leech, 2007).

Whilst the dust has certainly not settled, a less frequently broached issue is how to mesh features emerging from demographic, contextual, topical (Crowdy, 1993), and the more newly introduced "situation-governed" categories (Čermák, 2009: 116) within a framework that is responsive to the demands of representativeness. After a very brief overview of proposals in this regard, this paper argues in favor of a three-pronged set of features to achieve representativeness, and illustrates its implementation within the context of the Spoken Turkish Corpus (STC).

2. YARDSTICKS IN SAMPLING

The following sets of criteria and sampling procedures have been proposed and used in corpus compilation (Crowdy, 1993; Čermák, 2009):

1. Demographic
2. Contextual features
3. Topical
4. Situation-governed

The spoken component of BNC, for example, is based on the first three criteria, and the coding of texts according to the second and third criteria is reflected as genres. While the first set of criteria is geared toward representing geographical variation, the second is geared to capture register variation. If one works with factors in the first set, one runs the risk of producing a "skewed" compilation while the second set of criteria would allow for heterogeneity (Leech, 2007: 138). There is thus a certain tug-of-war between the two sets. Although he admits that it is more of an ideal rather than something that can be directly implemented, Leech states that the unit for sampling is the “initiator-text-receiver nexus”, which he refers to as an “ATOMIC COMMUNICATIVE EVENT” (Leech, 2007: 138). Thus, whether one applies proportional or stratified sampling, one needs to consider frequency of reception.

Čermák (2009) introduces another model that is based partly on contextual features in the sense of setting variables, and features that are characteristic of the spoken form of language as opposed to the written form of language. He argues that “a (proto)typical spoken corpus is […] made up of data where specific spoken features, that are not to be found in written corpora, predominate, or are sometimes even exclusively present […]” (Čermák, 2009: 114). Along with the parameter of “awareness” during recording (p. 117), he thus suggests that a prototypical spoken text would have plus values for all twelve parameters:

1. +spoken
2. +dialogue
3. +proximity
4. +equality
5. +private
6. +informal
7. +interactive
8. +present
9. +non-multiple
10. +spontaneous
11. +casual
12. +not aware

(from Čermák, 2007: 118)

This way of approaching spoken corpus design is parallel to the nature of data that has typically formed the empirical bases of research in conversation analysis and discourse analysis, and meets the demands of corpus-based pragmatics, which go beyond what "traditional" corpus linguistics caters for in terms of data structures (see Teubert, 2005; Schmidt & Wörner, 2009).

How is the model to be implemented and monitored, though, in a manner that also takes into consideration both the demographic and the topical dimensions of spoken discourse? In the following, we dwell on the corpus design and corpus management features of STC, which will be the product of a project that started in October 2008 with the aim of producing a general corpus of one million words of present-day Turkish spoken discourse in its initial stage.[2]

3. FEATURES OF THE SPOKEN TURKISH CORPUS

[2] A DEMO version is available for browsing and research purposes via http://std.metu.edu.tr/en/ .

To briefly describe the features of the technical aspects of STC, let us note that it employs EXMARaLDA (Schmidt, 2004), which is an open source software for corpus production that allows for online access to multimodal files. A detailed description of the technological infrastructure of STC is provided in Ruhi, Eröz-Tuğa, Hatipoğlu, Işık-Güler, Acar, Eryılmaz, Can, Karakaş and Çokal Karadaş (2010).

3.1 *Corpus design: metadata and annotation*

Independent of Čermák's study, STC was designed along the above-mentioned parameters. It attempts to monitor and address representativeness through demographic statistical measures, and enhances the monitoring of register variability through a close tracking of topics and speech acts.

Besides constructing a metadata system for domain, interactional goal and speaker features (e.g. age, education and language proficiencies), we maintain that the inclusion of speech acts (Searle, 1973) and conversational topics provides a crucial tool in monitoring the samples according to the tenor and affective tone of communicative events. While enabling future use of the corpus for a variety of research purposes ranging from discourse-level annotation to corpus-based and/or corpus-driven emotion research, these discursive dimensions are significant in tracing what may be the 'hidden' dimensions of the communicative events, which would not be available for the monitoring of the corpus compilation if sampling were based only on contextual and sociopragmatic variables. Naturally, the annotation of speech acts is but one scheme that would serve these purposes, but it renders granularity to the sampling beyond what can be achieved with domain and setting categorization.

Viewed from another perspective, spoken texts are slippery resources of language in terms of domain and setting categorization such that they are spatio-temporally characterized by shifts in interactional goals. A service encounter on a public transportation vehicle or at a shop, for example, can easily turn into a chat. Thus, if a communicative event were to be classified only for its domain of interaction, one would risk the chance of tracing subtle differences within the same domain, and hence, lose track of variability along the formality-informality dimension. In this regard, the simultaneous annotation of topics and speech acts addresses the concern for achieving maximal variability in register.

Other than a proportional sampling approach that controls the demographic dimension, the sampling of recordings is based on the identification of domains of discourse, for which the physical space of the interaction, the social relationships between the participants, the main thrust of the communication (e.g., chatting, transactional, educational, etc.), and the medium of communication are taken into consideration. Table 1 below reflects the design along these dimensions.

	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) Mediated: a) Telephone b) Broadcasts
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 bus, taxi, on the *dolmuş*[3]) 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)

Table 1. Major interactional samples in STC (from Çokal Karadaş and Ruhi 2009: 317)

Taking this layout as a starting point, what we have tried to achieve in STC is a "balanced" corpus. We take Leech's (2007) definition: "a corpus is 'balanced' when the size of its subcorpora (representing particular genres or registers) is proportional to the relative

[3] *dolmuş*: a minibus used for public transportation

frequency of occurrence of those genres in the language's textual universe as a whole. In other words, balancedness equates with proportionality" (p. 4). There have been few attempts, however, to explain what this requirement means, and no serious attempt was ever made to ensure that the genres, in the Brown Corpus or the BNC, for example, were proportional in this sense (ibid.). Balancedness is very difficult to demonstrate, even for very carefully constructed corpora.

For the development of STC, 8 major domains were identified (see Table 2). As will be observed, the major categories are based on social role relationships and the sub-categories are a mixture of topics, goals of interaction and conversational topics.

MAJOR DOMAINS	**MAIN INTERACTIONAL GOAL & MEDIUM**
1. FAMILY MEMBERS & RELATIVES	*chats, cultural events, narratives,* *telephone conversation, educational interaction,* *trips with the family*
2. FRIENDS AND FAMILY	*(same as in 1)*
3. FRIENDS	*(same as in 1)*
4. WORKPLACE COMMUNICATION	*meeting, shopping, workplace chats, telephone conversations, cultural events, work-related dinners* *interviews, appointments*
5. EDUCATION	*lecture in the social sciences, lecture in science, lecture in skills courses, seminars, conferences, panels* *student conferencing, parent-teacher meeting educational panel, interviews for educational programs school trips*
6. SERVICE ENCOUNTERS	institutional, shopping, service encounter on public transport
7. BROADCASTS	*news, news commentary, debate, series & films, sports* *educational, documentary, entertainment, competition* *culinary, health, children's programs*
8. OTHER	*brief encounter, religious discourse (sermons), legal discourse (e.g. court cases)* *political speech, political meeting, other public speeches,* *other public meeting, research*
9. UNCLASSIFIED	

The relative weightings of these domains were computed according to the results gained by small-scale data collection on "what Turkish people do and what type of interactions they hold in a regular day" as well as by consulting available demographic statistics.

Participants were asked to record everything they did, and how many hours in a number of (a) week days and (b) weekends they spend conversing in these domains (e.g. with friends, with colleagues, on the phone in the workplace, etc.) or are a recipient of such conversations (i.e. for broadcast sub-types). Considering the daily engagements of the working population, stay-at home, retired people and students, and researcher intuitions, representative 24-hour breakdown scenarios were created. Based on these average values, the projected weightings of each of the conversational domains/events in terms of hours in the 1 million spoken words in STC were calculated. Using the grid system, the breakdown was also projected on to the seven geographic regions of Turkey, in line with the ratio of the population in the regions. This gave the team slots to be filled according to domain>region>interaction types. Secondary level delimiters on these slots were gender and age.

Initially starting opportunistically, the STC had now reached 86 spoken data collection volunteers around Turkey who have submitted recordings for the corpus. The team closely guides the volunteers according to the grid system on the types of future interactions that need to be recorded.

Due to the nature of spoken discourse, not much value can be arrived at by controlling the length of each sample from a specific interaction sub-type, as written corpora compilers often do. Spoken corpora would lose from its linguistic and socio-pragmatic value if communication types are screened for equaling length and cut for that purpose. For instance, the length of workplace meetings in the Marmara region may be conventionally different than those held in the northern region (Black Sea) owing to socio-cultural traits and values (e.g. longer phatic talk before decision-making). Thus, for STC, no cutting or altering of individual samples collected is implemented beyond that of maintaining the privacy of sensitive information in the name of ensuring proportionality. This procedure will thus make the resource valuable for pragmatics research, which would require that communicative events be recorded in full rather than cut off to maintain proportionality.

The three-pronged scheme in STC is also enhanced by the design of the transcription and annotation scheme. STC takes within its purview a number of features that interactional sociolinguistics (see, e.g., Goffman, 1971) and the field of discourse analysis reveal as being significant in interaction. To keep track of the tenor of the communicative events, STC thus prioritizes the annotation of following pragmatic features:

a. Overlaps, filled and unfilled pauses, repairs

b. Discursive, formulaic expressions (e.g., thanking formulae)

c. (Im)politeness markers (e.g. address forms and T/V forms)

d. Non-prosodic features (e.g. laughing)

e. Gestures[4]

3.2 *Corpus management*

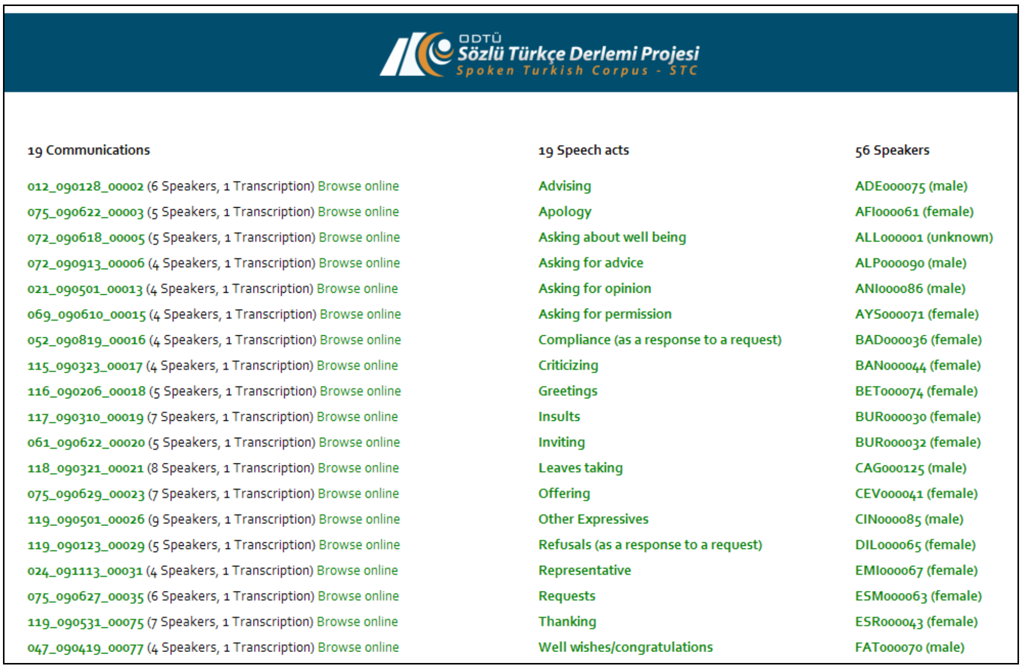
The STC corpus management system enhances EXMARaLDA with a web-based system interface and a relational (MySQL) database for metadata, which has been developed for making the management of corpus production and presentation flexible enough for use by non-experts. In this manner, experts and non-experts can submit annotation on conversational topics and speech acts, and edit them at any stage of the workflow to attain a finer-grained description of the sample. The system thus enables continuous monitoring of the corpus design parameters, with loops at each stage to the upper levels:

1. Annotation scheme of metadata for the samples
2. Entry of samples into the system, along with domain and speaker metadata
3. Transcription and annotation of recording, conversational topics and speech acts

In other words, the system implements an “agile” (Voorman and Gut, 2008) workflow in monitoring representativeness. As the system allows for the construction of sub-corpora at any stage in the workflow, it is possible to produce intra-corpus comparable data using any one of the design features. This enables issues concerning the practicalities of introducing ‘missing’ samples to be handled by using standard statistical measures.

Figure 1 illustrates the main page of the STC DEMO Version, where the three lists include each sample, the speech acts in the corpus and the speaker IDs. By clicking any of the items in the list, one arrives at its metadata. For example, clicking on one of the speech acts allows one to see links to all the samples containing tokens of the speech act.

[4] To be implemented in later stages of the development of the corpus.

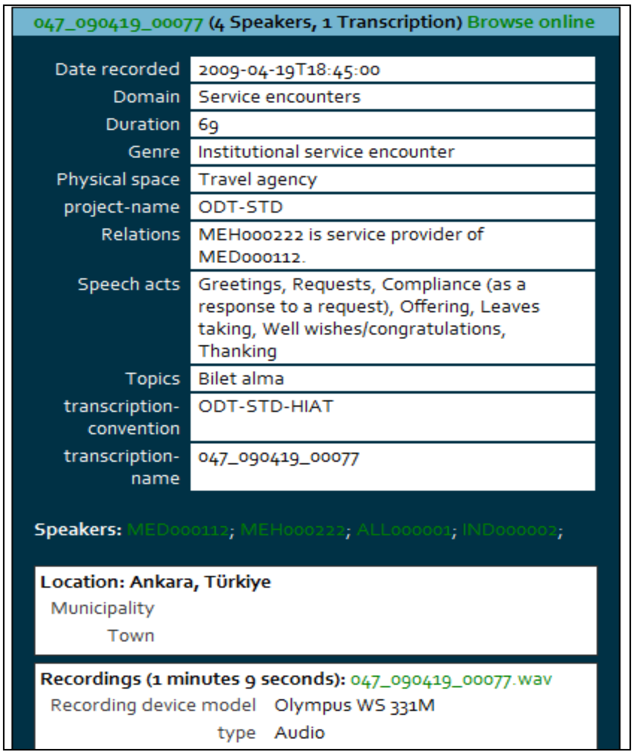


ODTÜ
Sözlü Türkçe Derlemi Projesi
Spoken Turkish Corpus - STC

19 Communications	19 Speech acts	56 Speakers
012_090128_00002 (6 Speakers, 1 Transcription) Browse online	Advising	ADE000075 (male)
075_090622_00003 (5 Speakers, 1 Transcription) Browse online	Apology	AFI000061 (female)
072_090618_00005 (5 Speakers, 1 Transcription) Browse online	Asking about well being	ALL000001 (unknown)
072_090913_00006 (4 Speakers, 1 Transcription) Browse online	Asking for advice	ALP000090 (male)
021_090501_00013 (4 Speakers, 1 Transcription) Browse online	Asking for opinion	ANI000086 (male)
069_090610_00015 (4 Speakers, 1 Transcription) Browse online	Asking for permission	AYS000071 (female)
052_090819_00016 (4 Speakers, 1 Transcription) Browse online	Compliance (as a response to a request)	BAD000036 (female)
115_090323_00017 (4 Speakers, 1 Transcription) Browse online	Criticizing	BAN000044 (female)
116_090206_00018 (5 Speakers, 1 Transcription) Browse online	Greetings	BET000074 (female)
117_090310_00019 (7 Speakers, 1 Transcription) Browse online	Insults	BUR000030 (female)
061_090622_00020 (5 Speakers, 1 Transcription) Browse online	Inviting	BUR000032 (female)
118_090321_00021 (8 Speakers, 1 Transcription) Browse online	Leaves taking	CAG000125 (male)
075_090629_00023 (7 Speakers, 1 Transcription) Browse online	Offering	CEV000041 (female)
119_090501_00026 (9 Speakers, 1 Transcription) Browse online	Other Expressives	CIN000085 (male)
119_090123_00029 (5 Speakers, 1 Transcription) Browse online	Refusals (as a response to a request)	DIL000065 (female)
024_091113_00031 (4 Speakers, 1 Transcription) Browse online	Representative	EMI000067 (female)
075_090627_00035 (6 Speakers, 1 Transcription) Browse online	Requests	ESM000063 (female)
119_090531_00075 (7 Speakers, 1 Transcription) Browse online	Thanking	ESR000043 (female)
047_090419_00077 (4 Speakers, 1 Transcription) Browse online	Well wishes/congratulations	FAT000070 (male)

Figure 1. Main page of STC DEMO Version

Figure 2 illustrates the metadata data for one sample.[5]



047_090419_00077 (4 Speakers, 1 Transcription) Browse online

Date recorded	2009-04-19T18:45:00
Domain	Service encounters
Duration	69
Genre	Institutional service encounter
Physical space	Travel agency
project-name	ODT-STD
Relations	MEH000222 is service provider of MED000112.
Speech acts	Greetings, Requests, Compliance (as a response to a request), Offering, Leaves taking, Well wishes/congratulations, Thanking
Topics	Bilet alma
transcription-convention	ODT-STD-HIAT
transcription-name	047_090419_00077

Speakers: MED000112; MEH000222; ALL000001; IND000002;

Location: Ankara, Türkiye

Municipality	
Town	

Recordings (1 minutes 9 seconds): 047_090419_00077.wav

Recording device model	Olympus WS 331M
type	Audio

Figure 2. Part of a communication metadata

[5] Currently, conversational topics in STC are annotated in Turkish for ease of use by the transcribers. In the final version, metadata and annotation will be retrievable in Turkish and English.

4. CONCLUDING REMARKS

Our experience during the construction of STC is that, owing to the mobility in the population, spoken corpora for Turkish on a much larger scale require one to keep close track of place of birth and length of stay in various locations in order to achieve representativeness in accent and dialects. A further issue is that speakers in the modern world may be diglossic and multilingual. In this regard, education appears to be a more reliable feature in terms of keeping track of expectations in sampling in the Turkish context.

Arguably, conversational topics can be searched to keep track of register and genre variation through word searches, and indeed this could have been an option for STC. But the added value of this annotation has been that one can observe the accumulation of topics even without interim sub-corpora constructions. This has the added value of overviewing the topical range at any time in the workflow. Since the size of the corpus is extremely small at its current stage of development, it remains to be seen whether doing speech act annotation will eventually produce better representativeness. Our experience is that they are rich data for tracing the sociopragmatically significant aspects of language use. Thus speech act metadata are functioning as a higher-order variable in monitoring the heterogeneity of the samples with respect to the interactional parameters. While topic and speech act annotation obviously is time-consuming, the pay-off is considerable, especially in regard to its potential to maintain a corpus-driven approach to corpus construction itself.

REFERENCES

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics, 19* (2), 219-241.

BNC www.natcorp.ox.ac.uk

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing, 8*(4), 259-265.

Čermák, F. (2009). Spoken corpora design: Their constitutive parameters. *International Journal of Corpus Linguistics 14* (1), 113-123.

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

EXMARaLDA. http://exmaralda.org/

Goffman, E. (1971). *Frame Analysis*. New York: Harper and Row.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp.133-149). Amsterdam: Rodopi.

Ruhi, Ş. Eröz-Tuğa, B., Hatipoğlu, Ç., Işık-Güler, H., Acar, G. C., Eryılmaz, K., Can, H., Karakaş, Ö., Çokal Karadaş, D. (2010). Sustaining a Corpus for Spoken Turkish Discourse: Accessibility and Corpus Management Issues. Paper to be presented at *LREC 2010*. Malta, 17-23 May.

Schmidt, T. (2004). Transcribing and annotating spoken language with EXMARaLDA. *Proceedings of the LREC-Workshop on XML Based Richly Annotated Corpora, Lisbon 2004.* Retrieved from http://www1.uni-hamburg.de/exmaralda/files/Paper_LREC.pdf

Schmidt, T. & Wörner, K. (2009). EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, *19* (4), 565-582.

Searle, J. (1976). A classification of illocutionary acts. *Language and Society,* 5, 1-23.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics* 1, 1-13.

Váradi, T. (2001). The linguistic relevance of corpus linguistics. In P. Rayson, A. Wilson, T. McEnery, A. Hardie & S. Khoja (Eds.), *Proceedings of the Corpus Linguistics 2001 Conference.* UCREL Technical Papers (Vol. 13) (pp.587-593). Lancaster: UCREL.

Voormann, H., & Gut, U. (2008). Agile corpus creation. *Corpus Linguistics and Linguistic Theory 4* (2), 235-251.