CREATING A SUSTAINABLE LARGE CORPUS OF SPOKEN TURKISH FOR MULTIPLE RESEARCH PURPOSES

Şükriye Ruhi
Middle East Technical University
Dept. of Foreign Language Education, Faculty of Education
Dumlupınar Blv., No.1, Üniversiteler Mah., 06800 Ankara, Türkiye
sukruh@metu.edu.tr; sukriyeruhi@gmail.com

Abstract

This paper focuses on the basic issues underlying the creation of sustainable, general purpose, large size corpora of naturally occurring spoken Turkish. It highlights issues relating to standardization in corpora, sustainability and interoperability.

1 Introduction

The possibility of conducting various linguistic research and developing and implementing speech technology tools for a given language largely relies on the availability of large scale corpora for the language. At present, Turkish lacks such a resource. This paper overviews the TÜBİTAK funded research conducted through October 2008 – April 2011 for the creation of the *Spoken Turkish Corpus* at Middle East Technical University (METU) with a view of highlighting a number of issues that need to be tackled first and foremost for the creation of a general purpose, large corpora of present-day spoken Turkish. The paper argues that standardization in transcription, and sustainability and interoperability in terms of corpus tools are key areas of concern for a multiple purpose corpus for Turkish.

The *Spoken Turkish Corpus* (hereafter, STC; http://std.metu.edu.tr), which is a multi-media corpus of naturally occurring contemporary Turkish is under construction at the Department of Foreign Language Education at METU. It is designed to ultimately reach the size of 10 million words, comparable to the size of the spoken component of the British National Corpus. It currently has a raw database of approximately 3 million words of audio and video recordings in a variety of geographical and social settings and domains (e.g., radio and television archives, conversations at the home and workplaces, and lectures). Approximately 440,000 words of these recordings are currently under transcription control, and partial morphological and speech act annotation processing in the corpus management system created for the corpus. Close to 50,000 words have been published in the demo version, which is currently available free of charge for non-profit research purposes (see Ruhi et al. 2010a for details on the design of STC). It is expected that derivative and commercial uses will be subject to charge for the purpose of maintaining resources for the sustainability of the corpus.

STC employs EXMARaLDA, which is an open source system of data models, formats and tools for the production and analysis of spoken language corpora and which allows transcriptions to be viewed in a time-aligned manner with recordings (Schmidt 2004; http://exmaralda.org/). Transcriptions are done with EXMARaLDA's Partitur Editor, using an adapted and revised form of HIAT (Ruhi et al. 2010b). The morphological analysis is being conducted with the freely available TRmorph (Çöltekin 2010), the annotation of requestive speech acts is being implemented with Sextant (Wörner 2009), using a tagset developed by Ş. Ruhi and K. Eryılmaz Ruhi et al. 2011). The web-based search system is currently being finalized, although search is available through Exakt and CoMa, which are two other EXMARaLDA tools.

The Corpus Management System for STC (STC-CMS) is a web-based system that is developed as a part of the STC Project for making the management of corpus production process easy, transparent and consistent. STC-CMS integrates with EXMARaLDA by generating the EXMARaLDA compatible transcription and corpus metadata files, and enhances EXMARaLDA with a web-based system which has been developed for making the management of corpus production and presentation flexible enough for scalable resources and use by non-experts during the production process. The system is designed to enable smooth control of the media and metadata files through a web interface and a relational (MySQL) database for metadata (Acar & Eryılmaz 2010; Ruhi et al. 2010c). STC-CMS will be available for use by other similar corpora construction research through consultation with Ş. Ruhi.

The METU STC Project started its research with ten members. Ş. Ruhi, Ç. Hatipoğlu, B. Eröz-Tuğa and H. Işık-Güler (hisik@metu.edu. tr) worked as research team members with an expertise in linguistics; M.G.C. Acar (acargunes@gmail.com) and K. Eryılmaz (keryilmaz@gmail.com) are currently still working as IT experts and program developers. They also specialise in (cognitive) linguistics and natural language processing. Ö. Karakaş and H. Can, two graduate students in linguistics, worked as metadata and transcription processors. The current research team was and is still being supported with the language technology and linguistic expertise of Dr. T. Schmidt (thomas.schmidt@uni-hamburg.de) and Dr. K. Wörner (kai.woerner@uni-hamburg.de) from Hamburg University.

STC is now continuing as a METU funded research project, with Ş. Ruhi as project director and H. Işık-Güler as research team member.

2 Use and users of STC

STC is designed for use by researchers in the public and private sector in a variety of language related domains, including but not restricted to linguists with a special interest in naturally occurring Turkish discourse, spoken Turkish morphology and syntax, lexicologists, and speech technology experts interested in developing speech recognition tools. Currently, STC's demo version is in use by 45 linguists, first and foreign language teachers, phonologists with an interest in speech recognition, language teaching material development, and cultural studies. It is expected that its end users will diversify in interest as fuller versions of the corpus are published.

In this paper, it is important to highlight the fact that a grammar and lexicon of spoken Turkish is not available for Turkish. Owing to the fact that transcriptions in STC are from recordings of conversations, and not read material, it is highly likely that the resource will be highly useful to, for example, researchers interested in the development of tools for spoken language automatic translation.

2.1 Strengths

One of the major strengths of STC is that despite its current small size, its recording database is rich in attesting to the use of Turkish by a wide range of speakers of varying age and language features. The recordings in the database and transcriptions include both standard Turkish as spoken in major cities of Turkey (e.g., Ankara and İstanbul) and dialectal forms (e.g., dialects of Kastamonu and Muğla). In this respect, STC's future versions will function as true general corpus for Turkish.

To users not particularly involved in using corpora for language research purposes, the metadata of a corpus is a set of features that may easily be disregarded. In STC, the linguistic, sociolinguistic and technical features of each recording (e.g., kind of recorder used, presence of standard or dialectal Turkish, social role relations, etc.) are noted carefully. In addition to these features, STC also records conversational topics and speech acts that are performed in the conversation. These two features enhance the search features of the resource for a variety of research purposes (e.g., lexical search, speech act and discourse structure annotation).

One of the major products of the STC project is the transcription conventions guideline for spoken Turkish (Ruhi et al. 2010b). As is well-known to researchers of spoken language, transcription standardization is a major requirement for further dissemination of corpora to a wider range of end users. The guideline is one step in this direction. Since the recordings include both standard and non-standard Turkish and since the written form of Turkish is not a reflection of its spoken form, STC employs a double tier transcription method that reflects the spoken forms both in their standard orthography and with annotations for dialectal pronunciation, including annotation that describes specific non-prosodic features. This system greatly enhances the use of the corpus for several purposes.

In relation to the transcription work, it is also important to point out that, despite the time-consuming nature of the work, the transcriptions conventions guideline and Partitur Editor have proven to be systems that are easily learnt in their basics by non-experts, as attested by the large number of transcribers who have been working in this stage of the project.

One of the advantages of STC is that it has taken modern, multi-modal corpora as an example to follow. In this respect, the corpus will enable research on gesture and language. Besides being multi-modal, the corpus tools used for the construction of STC are TEI compatible. In this respect, as language technologies change, it will be possible to render data comprehensible to other systems (Schmidt 2004; Ruhi et al. 2010c).

The corpus management system of STC (STC-CMS) has also been a major derivative product of the project. Our experience with STC has shown that it is a system suitable for use by non-experts. The Northern Cyprus Spoken Turkish Project, which is underway at METU-NCC, will soon be using this system for its research. We this expect the system to be useful for other researchers, too.

As mentioned above, the STC project was initially funded by TÜBİTAK. The current project results have been evaluated as having the potential to lead to advanced research in spoken Turkish by the project reviewer.

2.2 Challenges

Like the construction of all large-scale spoken corpora, STC is a time-consuming and expensive enterprise. One of the assets of the project was the availability of volunteers to do recordings from among the university student and administrative population. However, obtaining recordings in workplaces and from radio and television channels are true challenges, and this aspect of the project has only begun to pick up. The research team has observed that the building up of long term relationships enhances trust, which is essential in corpus research in such environments. Despite this difficulty, though, a more demanding stage in the construction of STC is the time-consuming, scientifically challenging, and expensive transcription work.

One of the major findings that has emerged during the construction of STC is the need to standardize the transcription of ‘words’ that are special to spoken Turkish (e.g., backchannels such as *hıı* and *hmm* and fillers such as *e* and *ee* etc.), as such words, far from being semantically and pragmatically empty, are important in the construction of meaning in interaction. Consider, for example, the pragmatic differences between the utterances *ha tamam* and *tamam* as responses to a previous utterance. The presence of dialectal Turkish and the difference between orthography and speech rendering has made the standardization of transcription a long and cyclical process.. However, the double transcription tier has resolved this challenge to a great extent at pesent.

Another challenge is to achieve consistency among different transcribers. In the construction of STC, this is tackled by having different transcribers work at each control stage of a recording transcription.

Our experience in STC has shown to us that transcription, especially in a multi-modal corpus, is the most arduous work and that possibly the longest time and the largest amount of human energy and financial resources in a corpus project should be devoted to this stage. This is essential in several respects, but it is of utmost significance for developing a multi-purpose spoken corpus.

3 Conclusion

As noted in the previous sections, the creation of spoken corpora for languages is time-consuming and expensive work. It is thus of utmost importance to guarantee a long life for such corpora. For this, the STC experience has shown that standardization, maintaining interoperability, and enhancing sustainability through efficient use of resources are significant factors. In this paper, issues that relate to the safeguarding of archives and the rights of resource providers have not been brought to the foreground, but these involve sensitive legal issues and standardized procedures are necessary here too. As a step in enhancing research on Turkish, I would suggest that the formation of a human languages resources research centre which would be supported through the collaboration and funding of various institutions would foster communication and enable the pooling of research efforts and expertise.

Acknowledgment

Research reported in this paper has been supported by TÜBİTAK (Grant no.: 108K283), and METU (Grant no.: BAP-05-03-2011-001).

References

Acar, M. Güneş C., Eryılmaz, K. 2010. Sözlü derlem için web tabanlı yönetim sistemi. *24. Ulusal Dilbilim Kurultayı Bildiri Kitabı*. 17-18 Mayıs 2010. Ankara: ODTÜ, Yabancı Diller Eğitimi Bölümü, 437-443.

British National Corpus. www.natcorp.ox.ac.uk

Çöltekin, Çağrı. 2010. A freely available morphological analyzer for Turkish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation* (LREC2010), Valletta, Malta, May 2010.

Ruhi, Şükriye, Işık-Güler, Hale, Hatipoğlu, Çiler, Eröz-Tuğa, Betil, and Derya Çokal Karadaş. 2010a. Achieving representativeness through the parameters of spoken language and discursive features: The case of the *Spoken Turkish Corpus.* In: I. Moskowich-Spiegel Fandiño et al. (eds.), *Language Windowing through Corpora. Visualización del Lenguaje a Través de Corpus. Part II. A Coruña: Universidade da Coruña, 789-799*.

Ruhi, Şükriye, Hatipoğlu, Çiler, Eröz-Tuğa, Betil, Işık-Güler, Hale. 2010b. *A Guideline for Transcribing Conversations for the Construction of Spoken Turkish Corpora Using EXMARaLDA and HIAT.* ODTÜ-STD: Setmer Basımevi.

Ruhi, Şükriye, Eröz-Tuğa, Betil, Hatipoğlu, Çiler, Işık-Güler, Hale, Acar, M. Güneş C., Eryılmaz, Kerem, Can, Hümeyra, Karakaş, Özlem, and Derya Çokal Karadaş. 2010c. Sustaining a corpus for spoken Turkish discourse: Accessibility and corpus management issues. In *Proceedings of the LREC 2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management*. Paris: ELRA, 44-47. http://www.lrec-conf.org/proceedings/lrec2010/ workshops /W20.pdf# page=52

Ruhi, Şükriye, Schmidt, Thomas, Wörner, Kai, and Kerem Eryılmaz. 2011. Annotating for precision and recall in speech act variation: The case of directives in the Spoken Turkish Corpus. Paper to be presented at GSCL 2011, 28-30 September 2011, Hamburg University.

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.pdfreference stub

Wörner, Kai. 2009. Werkzeuge zur flachen Annotation von Transkriptionen gesprochener Sprache. Bielefeld: Bielefeld University.