Sustaining a Corpus for Spoken Turkish Discourse: Accessibility and Corpus Management Issues

Şükriye Ruhi[a], Betil Eröz-Tuğa[a], Çiler Hatipoğlu[a], Hale Işık-Güler[a], M. Güneş Can Acar[b], Kerem Eryılmaz[a], Hümeyra Can[c], Özlem Karakaş[a], Derya Çokal Karadaş[a]

[a] Middle East Technical University, [b] Ankara University, [c] Hacettepe University

[a] İnönü Blvd., 06531 Ankara, Turkey; [b] Ankara Üniversitesi İletişim Fakültesi, 06590 Cebeci Ankara, Turkey; [c] Hacettepe University, 06800 Beytepe, Ankara, Turkey

E-mail: sukruh@metu.edu.tr, beroz@metu.edu.tr, ciler.hatipoglu@gmail.com, hisik@metu.edu.tr, acargunes@gmail.com, keryilmaz@gmail.com, hmeyra.can@gmail.com, ozlm.krks@gmail.com, deryacokal@gmail.com

Abstract

This paper addresses the issues of the long-term availability of language resources and the financing of resource maintenance in the context of the web-based corpus management system employed in the Spoken Turkish Corpus (STC), which operates with EXMARaLDA. Section 2 overviews the capacities of the corpus management system with respect to its software infrastructure, online presentation, metadata management, and interoperability. Section 3 describes the plan foreseen in STC for sustaining the resource, and dwells on the ethical issues surrounding the conflicting demands of free resources for non-commercial research and resource maintenance.

1. Introduction

A set of intertwined and pressing issues need to be tackled in the production of corpora that aim to be freely available for non-commercial research (cf. Haugh, 2009) over long periods. Based on our experience emerging from the ongoing construction of the Spoken Turkish Corpus (STC) within the Middle East Technical University Spoken Turkish Corpus Project (METU-STC), we highlight the need to develop corpus management systems that are accessible to (non-expert) corpus production and annotation work teams, the crucial role of open source software with interoperability capacities, and the financing of corpora maintenance and development.

Section 2 very briefly overviews the content of STC. Section 3 describes the web-based corpus management system developed for the corpus. In Section 4, we present the current plans for sustaining STC and suggest ways for reconciling the demands of ensuring that language resources are free for non-commercial research exploitation with those of the financial exigencies of resource maintenance.

2. STC: Design Features

STC stems from the first project in Turkey aiming to produce a relatively large-scale, general corpus of spoken Turkish discourse. In its initial stage, the corpus is designed to consist of one million words of present-day face-to-face and mediated interactions in Turkish in both formal and informal communicative settings. It is a multi-modal resource that presents transcriptions in a time-aligned manner with audio and video files. A more detailed description of its design features are found in Çokal Karadaş and Ruhi (2009).

STC employs EXMARaLDA (Extensible Markup Language for Discourse Analysis), which is a system of data models, formats and tools for the production and analysis of spoken language corpora (see, Schmidt (2004, 2005) and for a detailed description of EXMARaLDA). Informed by the transcription conventions in a previous corpus project, "Interpreting in Hospitals", which includes interactions in Turkish in Germany, the corpus is currently being transcribed and annotated with an adapted form of HIAT for utterances, utterance boundaries, pauses, overlaps, repairs, interruptions, and frequently occurring paralinguistic features such as laughing and certain emotive tones (see Schmidt (2008) for an overview of the basics of the current HIAT system). In its adult stage it will be a corpus annotated for morphology, the socio-pragmatic features of Turkish (e.g. address terms, (im)politeness markers, and a selection of speech act realizations), anaphora, and gestures.

The construction of STC is taking place in a work environment where little standardization in spoken language transcription with computer-assisted tools is available (Hatipoğlu and Karakaş, 2010; Işık-Güler & Eröz-Tuğa, 2010). Furthermore there few to no resources providing quantificational data on the production and reception of spoken domains and genres (Ruhi and Can, 2010; Ruhi, Işık-Güler, Hatipoğlu, Eröz-Tuğa & Çokal Karadaş, 2010), which means that basic research in these areas need to proceed concurrently with the production process. The research, annotation, and recorder teams, on the other hand, involve both expert and non-experts: linguists with a specialization in pragmatics and conversation analysis, IT infrastructure experts and programmers, (under)graduates and professionals in language studies and other areas, and volunteers from the general public interested in supporting the corpus production throughout its various stages. Thus two of the foremost priorities of METU-STC were the development of a workflow and corpus management system that could cater to the needs of this type of environment. More detailed descriptions of STC and its workflow are presented in Acar and Eryılmaz (2010).

3. The Web-based Corpus Management System for STC (STC-CMS)

STC-CMS is a web-based system that was developed and is being improved to make the process of the management and the monitoring of corpus production easy, transparent and consistent for team members who are not specialists in the technology of digital architectures. The system is designed with the goal of maximum automation and validation, as well as a clearly defined, traceable workflow, which enables monitoring of the design parameters of the corpus and the progress of the workflows, and the maintenance of consistency in production (see Fig. 1). The system thus also enables an "agile" (Voorman & Gut, 2008) workflow for controlling representativeness, which as underscored by Leech (2007) and Čermák (2009), remains a central issue in spoken corpora production.

As STC employs EXMARaLDA, a central function of STC-CMS is to achieve integration with its tools. STC-CMS performs this by generating EXMARaLDA compatible transcription and corpus metadata files.

3.1 File creation and interoperability

The system enables smooth control of the media and metadata files through a web interface and a relational (MySQL) database for metadata. Contributors submit recordings and metadata through the web forms, where they are validated and added to the database. At that stage, STC-CMS generates the EXMARaLDA compatible transcription files, which makes it possible to use EXMARaLDA tools and formats in STC. When a transcription file is submitted, it is checked into an SVN system for backup measures.

Being an open source system, EXMARaLDA and its associated tools do not pose the risk of being unavailable in future, and they can be sustained by other programmers. When fully checked for system operation features, STC-CMS will function as an open source project for further enhancement of its capacities and use by other resource producers who may wish to contribute to the STC database or who may wish to develop their own.

Using various file and data formats, STC tries to minimize the risk of digital obsolescence. Amongst its notable features, the system allows any subset of the corpus to be defined and published using EXMARaLDA libraries through a password restricted web site, where anyone with a web browser may access the corpus.

The sustainability of STC is also enhanced by employing EXMARaLDA's various export options (see Fig. 1). Transcriptions can be exported to HTML, PDF, RTF, TEI-compliant and XML-based EXMARaLDA formats, which ensure accessibility, long-term archivability and interoperability (see Schmidt (2005) for a detailed description of the relation between EXMARaLDA and TEI formats). The system also harnesses EXMARaLDA's capabilities for exporting to different transcription systems like Praat, ELAN, and TASX Annotator.

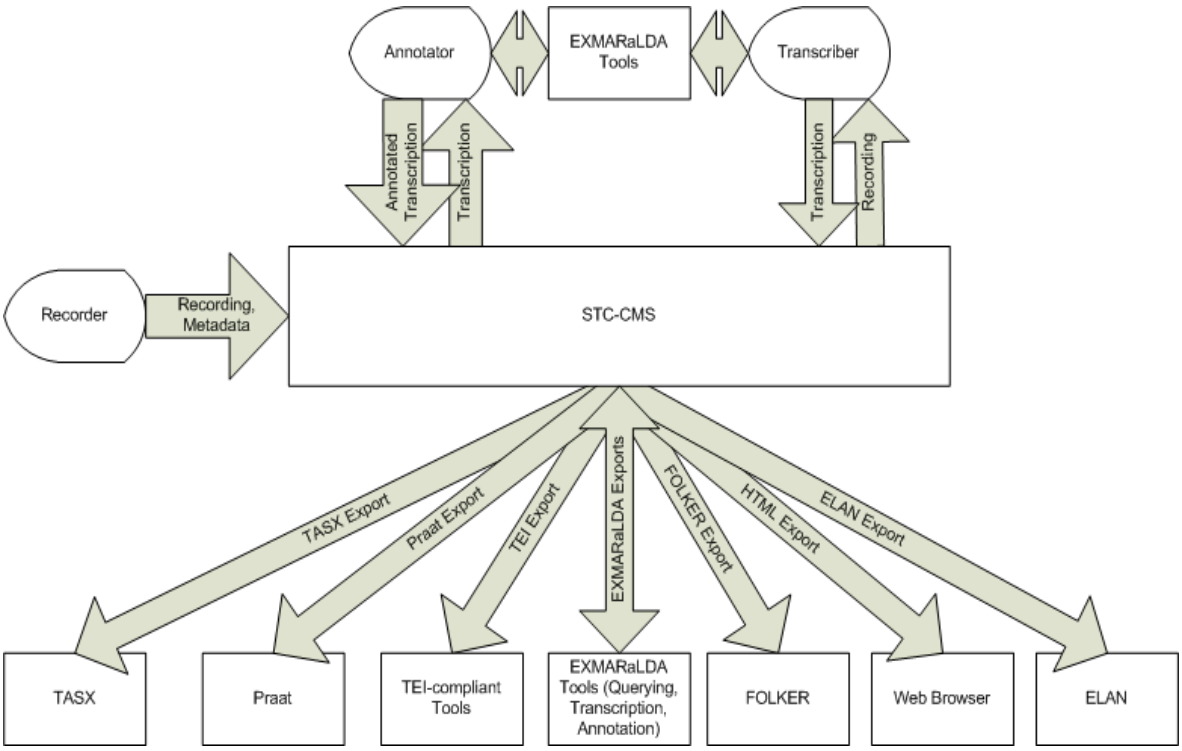

Figure 1: STC workflow and interoperability

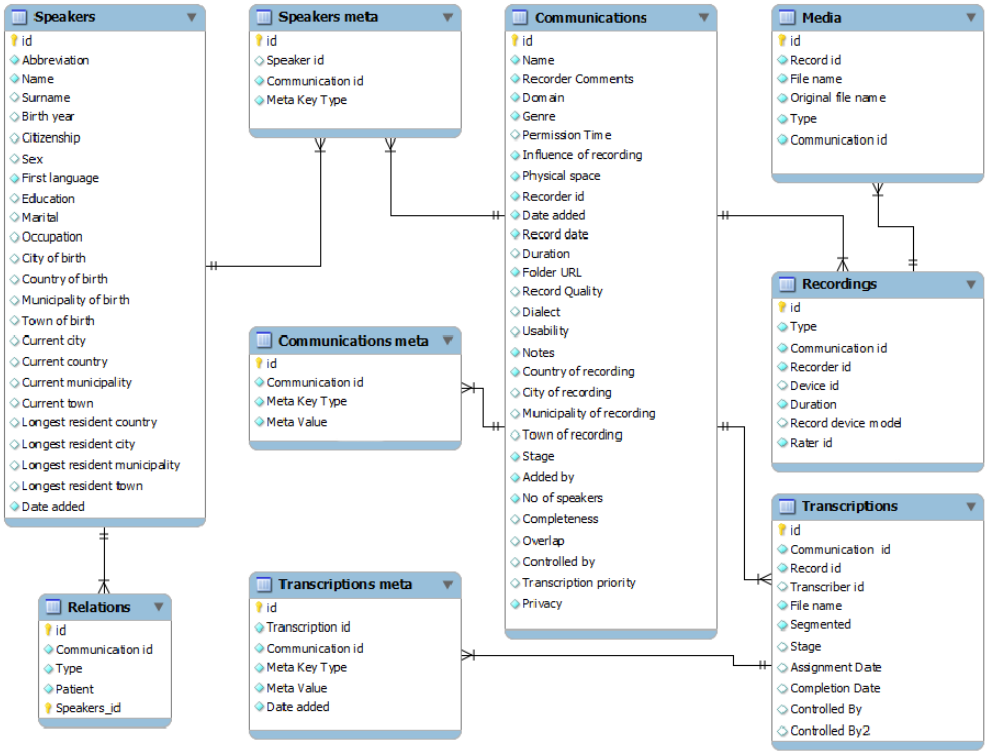

Figure 2: The database structure of STC-CMS

3.2 Metadata in STC

Given the crucial role of standardization in the maintenance of language resources, a few notes on the current state of the metadata in STC are due. The STC metadata fields have been defined through comparisons with the spoken component of BNC, the ISLE Meta Data Initiative (IMDI), sociolinguistic and pragmatics studies in the Turkish context (Ruhi, Işık-Güler, Hatipoğlu, Eröz-Tuğa & Çokal Karadaş, 2010), and the standard fields in COMA – EXMARaLDA's corpus manager tool (see Fig. 2 for an overview of the fields).

In addition to including classificatory and descriptive information on both the recordings and the speakers in the communications, STC follows the practice of providing an overview of the corpus in terms of communication categories, distribution of gender and age (see, e.g., the Spoken Dutch Corpus; Oostdijk, 2000). The METU-STC web site currently presents the overall features and communication types in the DEMO version of STC, along with the projected corpus design, the terms of use and information on copyright holders. Detailed information concerning the corpus design, and the transcription and annotation conventions will be added in its final version.

In regard to the formal properties of the metadata, COMA allows for the addition of Dublin Core (DC) fields to the coma file. Plans are being made to develop a web-based rather than a file-based system for search purposes once a standard metadata format has been decided upon. At present, our experience is that there are considerable differences in guidelines proposed by various spoken language resource initiatives such that an early commitment to any one of these might prove problematic in the long run (see, for example, BNC, IMDI, the Bavarian Archive for Speech Signals, and the parameters discussed in Čermák (2009) for spoken corpora). Given that the purposes for resource production are more varied in the case of spoken language corpora compared to written language corpora, this situation is understandable. In this regard, we find Schmidt's (2010) call for a concerted effort to achieve a "stepwise approximation" between the practices of communities of resource producers, users and language technologists a viable route to be pursued.

4 Accessibility and resource maintenance

STC-CMS and the interoperability capacities of EXMARaLDA are closely linked to the second major issue addressed in the paper: that of ensuring the long-term availability of free resources for non-commercial research. In our context, STC is a product that is being funded over two years by a national institution. It is obvious that this duration is vastly inadequate to achieve the long-term objective of extending the corpus to the size of ten million words. On the positive side, though, the project is hosted by the Dept. of Foreign Language Education at METU, which has a strong incentive to support language research and language resources, and the core research team consists of faculty members at the department. It should thus be possible to maintain the laboratory conditions and the required research activities for the expansion of the

resource. Funding for continued infrastructure maintenance and tool development, for example, will be secured through a variety of long-term projects related to STC.

STC is being constructed with recordings donated by individuals and media institutions. So it is imperative both to protect the royalty rights of contributors and to remain prepared for the possibility of a fluctuating production team. STC is built on the understanding that copyright owners of the various versions of the corpus and its sub-corpora will distribute the corpora freely for non-commercial research purposes. Other uses of the corpus (e.g. materials development in educational settings, NLP commercial applications and products derived therein) will be commercialized for the sole purpose of corpus maintenance and research directly impinging on the development of the corpus. Such commercial uses will be handled through various presentation types at different rates depending on the purpose of commercialization (e.g. internet access and cd/dvd formats; availability of either the whole corpus or sub-corpora; educational vs. non-educational purposes).

To further tackle the challenge of keeping STC a free resource while ensuring its expansion, the present copyright holders are planning to use a combination of CreativeCommons licenses in the forthcoming stages. Amongst the various license options that would allow for expansion of STC it appears that "Attribution Non-Commercial Share Alike" (cc by-nc-sa) provides a practical solution. This option allows for derivative work under the same conditions of the original terms of use. Such multiple availability options may respond to the demands of differing legal strictures and ethical stances both across language resource production communities and across national systems in the sharing and development of resources.

5 Concluding Remarks

Several issues remain to be resolved concerning sustainability, and funding is certainly a pressing issue. However, our experience with STC suggests that standardization in metadata and annotation, and by consequence, the development of tools with interoperability capacities, are by far more crucial in the current state-of-affairs. In this regard, we suggest that collaboration amongst the various stakeholders should involve not only resource producers and experts in digital architectures, but also the user end.

6 Acknowledgements

STC is financed by a research grant from the Turkish Scientific and Technological Research Institution (Türkiye Bilimsel ve Teknolojik Araştırma Kurumu, TÜBİTAK, Grant No. 108K285). We are deeply grateful to Dr. Thomas Schmidt and Dr. Kai Wörner for being a part of the project research team during October 2008–November 2009 and for their ongoing support in the construction and annotation of STC.

7 References

Acar, G.C., Eryılmaz, K. (2010). Sözlü Derlem için Web Tabanlı Yönetim Sistemi. Paper presented at the XXIV. Linguistics Conference, Middle East Technical University.

Bavarian Archive for Speech Signals. http://www.phonetik.uni-muenchen.de/forschung/BITS/TP1/Cookbook/

BNC. www.natcorp.ox.ac.uk

Bührig, K., Bernd, M. Interpreting in Hospitals. http://www.exmaralda.org/corpora/en_sfb_k2.html

CreativeCommons. http://creativecommons.org/

Čermák, F. (2009). Spoken Corpora Design: Their Constitutive Parameters. *International Journal of Corpus Linguistics* 14(1), pp. 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, pp. 311--320.

EXMARaLDA. http://exmaralda.org/

Hatipoğlu, Ç, Karakaş, Ö. (2010). Sözlü Derlem Çeviriyazısını Standart Dil ve Ağıza Göre Ölçünleştirme. Paper presented at XXIV. Linguistics Conference, Middle East Technical University.

Haugh, M. (2009). Designing a Multimodal Component of the Australian National Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian Corpus*. Somerville, MA: Cascadilla Proceedings Project, pp. 74--86.

Işık-Güler, H., Eröz-Tuğa, B. (2010). Çeviriyazıda Geribildirim, Durak, Sesli Duraklama ve Ünlemlerin Ölçünleştirilmesi. Paper presented at XXIV. Linguistics Conference, Middle East Technical University.

ISLE Meta Data Initiative. http://www.mpi.nl/IMDI/

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*. Amsterdam: Rodopi, pp. 133--149.

Middle East Technical University Spoken Turkish Corpus Project. http://std.metu.edu.tr/en/

Oostdijk, N. (2000). Meta-Data in the Spoken Dutch Corpus Project. LREC 2000 Workshop, Athens. http://www.mpi.nl/IMDI/documents/2000%20LREC/oostdijk_paper.pdf

Ruhi, Ş. Can, H. (2010). Sözlü Derlem için Veribilgisi Geliştirme: Bağlam ve Tür Kavramlarına Derlem Dilbilimi Açısından Bir Bakış. Paper presented at XXIV. Linguistics Conference, Middle East Technical University.

Ruhi, Ş., Işık-Güler, H., Hatipoğlu, Ç., Eröz-Tuğa, B., Çokal Karadaş, D. (2010). Achieving Representativeness Through the Parameters of Spoken Language and Discursive Features: The Case of the Spoken Turkish Corpus**.** Paper presented at II. International Conference on Corpus Linguistics, Universidade da Coruña.

Schmidt, T. (2004). Transcribing and Annotating Spoken Language with EXMARaLDA. In *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

Schmidt, T. (2005). Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech. In: Arbeiten zur Mehrsprachigkeit, Folge B 62.

Schmidt, T. (2008). Overview of HIAT transcription conventions. http://www1.uni-hamburg.de/exmaralda/files/HIAT_EN.pdf

Schmidt, T. (2010)**.** Linguistic tool development between community practices and technology standards. Paper to be presented at the "Standardising Policies within eHumanities Infrastructures" Workshop at LREC 2010.

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