Construction of the Turkish National Corpus (TNC)

Yeşim Aksan[1], Mustafa Aksan[1], Ahmet Koltuksuz[2], Taner Sezer[1], Ümit Mersinli[1], Umut Ufuk Demirhan[1], Hakan Yılmazer[1], Özlem Kurtoğlu[1], Gülsüm Atasoy[1], Seda Öz[1], İpek Yıldız[1]

Mersin University[1], Yaşar University[2]

Mersin Üniversitesi Fen-Edebiyat Fakültesi, 33343 Mersin, Turkey; Yaşar Üniversitesi Mühendislik Fakültesi, 35100 İzmir, Turkey

E-mail: yesim.aksan@gmail.com, mustaksan@gmail.com, ahmet.koltusuz@yasar.edu.tr, tanersezerr@gmail.com, umit@mersinli.org, umutufuk@gmail.com, yilmazerhakan@gmail.com, ozlemkurtoglu@mersin.edu.tr, gulsumatasoy@mersin.edu.tr, sedaoz@mersin.edu.tr, iyildiz@mersin.edu.tr

Abstract

This paper addresses theoretical and practical issues experienced in the construction of *Turkish National Corpus* (TNC). TNC is designed to be a balanced, large scale (50 million words) and general-purpose corpus for contemporary Turkish. It has benefited from previous practices and efforts for the construction of corpora. In this sense, TNC generally follows the framework of British National Corpus, yet necessary adjustments in corpus design of TNC are made whenever needed. All throughout the process, different types of open-source software are used for specific tasks, and the resulting corpus is a free resource for non-commercial use. This paper presents TNC's design features, web-based corpus management system, carefully planned workflow and its web-based user-friendly search interface.

Keywords: Turkish National Corpus, corpus construction, corpus linguistics

1. Introduction

This paper addresses theoretical and practical issues experienced in the construction of *Turkish National Corpus* (TNC). Since corpus building is a time consuming and labour-intensive process, design decisions for TNC required careful preparation. Following previously tested practices, the design of TNC benefited from previous efforts and practices. The overall framework of British National Corpus (BNC) is adopted in construction of TNC with modifications in a number of metadata criteria while reducing the corpus size strictly proportionally.

In the construction of the corpus, a special workflow is devised to benefit from available resources and existing facilities. A corpus management system is developed to gather, maintain, process and monitor data. Open-source software is preferred at all stages of corpus construction and data management. The end result is a free resource corpus for non-commercial use, covering a period of 20 years (1990-2009). Unique with its size of 50 million words, balance of the corpus is achieved through a wide range of text categories it covers. The design principles and their particular applications are discussed in detail in following sections of the paper.

2. Design features of TNC

TNC is the product of a project aiming at building a relatively large scale, representative, general corpus of contemporary Turkish. The size of the corpus is 50 million words. It consists of samples of textual data (98%) across a wide variety of genres covering a period of 20 years (1990-2009). 2% of TNC consists of transcribed spoken data. The distribution of number of words in the corpus is determined proportionally for each text domain, time and medium of text following the model of BNC. In other words, the size of BNC is reduced to size of TNC by preserving the original quantificational distribution.

A balanced general corpus contains texts from a wide range of genres, and text chunks for each genre are sampled proportionally for inclusion in a corpus (Biber, 1993). The measures of balance and representativeness are considered to be scalar and it is underscored that there is no scientific measure for corpus balance (Leech, 2007; McEnery & Hardie, 2012). Corpus-builders mainly adopt an existing corpus model when building their corpus. British National Corpus (Aston & Burnard, 1998) is generally accepted as a balanced corpus, and it has been followed in the construction of the American National Corpus, the Korean National Corpus, the Polish National Corpus, and the Russian Reference Corpus (McEnery, Xino & Tono, 2006). In the construction of TNC we too generally followed the framework of BNC.

In selecting written texts for TNC, representativeness of the corpus is achieved through balance and sampling of Turkish or varieties of contemporary Turkish. Accordingly, written texts included in TNC are selected using three criteria: text *domain*, *time* and *medium* of text.

Text domain in BNC covers two major types as *imaginative* and *informative*. Here, domain refers to main subject field. For example, imaginative domain consists of fiction (novel, short story, poem, drama). Informative domain includes samples from social sciences (e.g. sociology, linguistics), art (e.g. architecture, design), commerce-finance (e.g. business, industry), belief-thought (e.g., religion, philosophy), world affairs (e.g., history, politics), applied science

(e.g., engineering, computing), natural-pure science (e.g., maths, physics), and leisure (e.g., food, travel, gardening). Same as in the BNC, imaginative texts constituted 19% and informative texts %81 of TNC.

Time refers to the period of text production. All texts that finally are sampled in TNC are published in period between 1990 and 2009. In this respect, distribution of the size of texts in each year is decided in terms of relative representation for each domain and medium. Thus, for example, texts from academic journals are represented for each year in almost equal number of words.

Medium of text refers to type of text production. Written samples are collected from books, periodicals (newspapers, magazines, and scientific journals), published (annual reports) or unpublished documents (student essays, e-mails, and blogs), and texts written-to-be spoken such as, news broadcasts, screen plays. Table 1 demonstrates the proportion of written component of TNC sampled according to domain and medium.

Domain	%	Medium	%
Imaginative	19	Book	58
Social Science	16	Periodicals	32
Art	7	Miscellaneous published	5
Commerce/ finance	8	Miscellaneous unpublished	3
Belief and thought	4	To-be-spoken	2
World affairs	20		
Applied science	8		
Natural science	4		
Leisure	14		

Table 1: Composition of written component of TNC

Transcriptions from spoken data constitute 2% of TNC's database, which involves spontaneous, every day conversations and speeches collected in particular communicative settings, such as meetings, lectures. Spoken component of TNC contains a total of 1 million words. 500.000 words come from orthographic transcriptions of every day conversation and its relevant medium and 500.000 of them are orthographic transcriptions of context-governed speeches. In determining the metadata fields for spoken data, TNC follows criteria defined by both spoken component of BNC and STC (Ruhi et al., 2010).

3. Corpus management system and workflow for TNC

Practical work on corpus construction followed a series of successive steps: (1) capturing data: obtaining texts, transcripts of spoken language, and speech recordings; (2) computerizing data: downloading available electronic texts, scanning, keyboarding and transcribing spoken data; (3) encoding metadata: recording metadata information of written texts and transcribed spoken data; (4) annotating corpus data: part-of-speech annotation; (5) developing a search interface: a user-friendly, web-based graphical user interface that would ease access to the corpus data; (6) releasing the corpus: releasing a beta version for local testing, and then first release for national and international use.

For capturing written data in TNC, a list of sampling units is prepared from best sellers, prizewinners, and books-in-print lists, among others. Course-books are selected to represent different academic fields and, periodicals and journals are selected among those that are supposed to represent broadest interests and orientations. A more detailed description of data selection and computerizing procedures is found in Aksan (2009) and Aksan and Aksan (2009).

In computerizing the data, selected texts with different file formats are converted to UTF-8 encoded text formats after Optical Character Recognition (OCR) process. A web-based corpus management system was developed to process and monitor data coming from the OCR, keyboarding, existing electronic texts or speech transcriptions. The system also made corpus creation easy, transparent and stable for research team members and non-experts working in the construction of the corpus. In TNC corpus management system, first metadata of computerized files are encoded to a relational (MySQL) database. Then, textual data and transcribed speeches are controlled manually on the basis of spell checking and editing guidelines developed by the expert linguists of research team. To secure an error free corpus database in TNC, each text sample was rigorously checked and rechecked. The workflow of TNC design sets up a number of checkpoints along the course, which is shown in Figure 1.

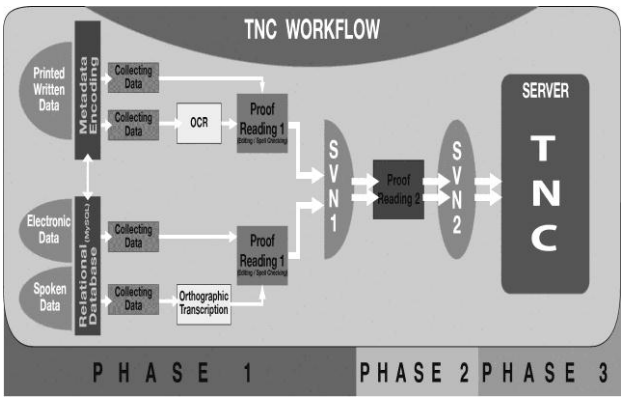

Figure 1: TNC workflow

Given that TNC database covers almost 5.000 entries, the subversion (SVN) control is used to monitor and process data and check revision history of the files. TNC has two major SVN repositories. The first repository

contains the first version of the written texts, spoken texts that are proofread by data entry operators. The second repository is the main part of the corpus database, and only research team members have file permissions to reread, modify, and delete the text files.

4. Annotation in TNC

As for the analysis and annotation of morphemes in TNC texts, a Turkish module for NooJ (Silberztein, 2003) was developed. This module has been one of the major derivative products of the TNC creation process. Following a root-driven, non-stochastic, rule-based approach, lexicon files are built from scratch since available taggers for Turkish vary in their tagsets and in their lexical entries. Additional modules to annotate only derivational or inflectional affixes are also formed with a pre-defined tagset. Preliminary rule-based disambiguation attempts based on a well-documented ambiguity set (Aksan et al., 2011) are currently in progress. Below are sample annotations from NooJ_TR module (Aksan & Mersinli, 2011).

geldi

<gel,V+past>

yapabilecekler

<yap,V+bfa+Va1+futr+pl>

evdekilerdenmiş

<ev,N+loc+kiA+pl+abl+perf>

gidiyorlardı

<git,V+bfi+imprf+pl+past>

günün en güzel saatleri

<gün,N+gen> <en,N-AV> <güzel,AJ> <saat,N+pl+p3s>

kitabın kapağının rengi

<kitap,N+gen> <kapak,N+p3s+bfn+gen> <renk,N+p3s>

Türk dilinin zenginlikleri

<Türk,NP> <dil,N+p3s+bfn+gen> <zenginlik,N+pl+gen>

After building a fully disambiguated training corpus with NooJ_TR, further supervised/stochastic disambiguation tools are tested for tagging all TNC texts. Lack of standardisation in a common tagset for Turkish prevented the developers from using available taggers or pre-tagged corpora for training and testing stochastic tools. "Which affixes should be annotated?" and "which affixes should be embedded in the lexicon files and why?" are questions yet to be answered for Turkish. Available taggers do not provide a linguistically motivated documentation on derivational and/or inflectional affixes they analyze.

5. TNC interface

5.1 Search interface

TNC has a platform-independent, user-friendly, web-based graphical user interface for making queries. To achieve this end, a relational database in MySQL with MyISAM storage engine is formed. The database includes all TNC texts, their metatextual information (detailed specification of the categories and/or sections of text) and full-text indicies to optimize search performance. Search performance is a challenging task during the implementation of such interfaces considering the database size.

A PHP, HTML, CSS and JavaScript based front-end is designed to allow users to perform queries by adding filtering criteria based on the metadata available in the database. In other words, users can define restrictions to generate concordance outputs from specific domains, genres, publication dates, authors, text types etc. (see Figure 2). The interface also allows ready-made restrictions on the position of the search term in a word-form as a simpler way of using wildcards. In other words, this function allows affix or root-form searches in TNC.

To illustrate the features of the interface, for instance, users can search for "*iyor*" by selecting "ending with" in the "Position" menu. Additional criteria can be defined by selecting "Imaginative prose" under the "Domain" menu, "Female" under "Sex of author" menu and "1995 to 1998" under "Publication date" menu. By providing the above filtering criteria, users can search for verbs inflected with progressive affix "*iyor*" in Turkish only in fiction texts written by female authors between the years 1995 and 1998. Filtering can also be extended to include specifications on genre, derived text type, media, sample, audience and types of author. Case-sensitivity and window span can also be specified through TNC interface.

It is not required for users to know any special query syntax to make corpus queries. With minimal web interface elements, such as radio buttons and drop-down lists, the application provides a user-friendly environment for researchers.

5.2 Search results

Concordance window displays search results in keyword-in-context-mode. The interface allows users to navigate through concordance output by providing instant access from any concordance line to larger context of the search item and to a bibliographical record describing the source of the cited concordance line (see Figure 3). The interface facilitates manipulation of the concordance output. The user can sort the query result alphabetically on any of five positions to the left of right of the search item. The user can also export the concordance in XLS file format (MS Excel) having each word on the left and right side of the search item in separate columns to make sorting and pattern matching easier for researchers.

Current release of TNC is a beta version for local testing, and bug reports provided by local users are being generated for future releases.

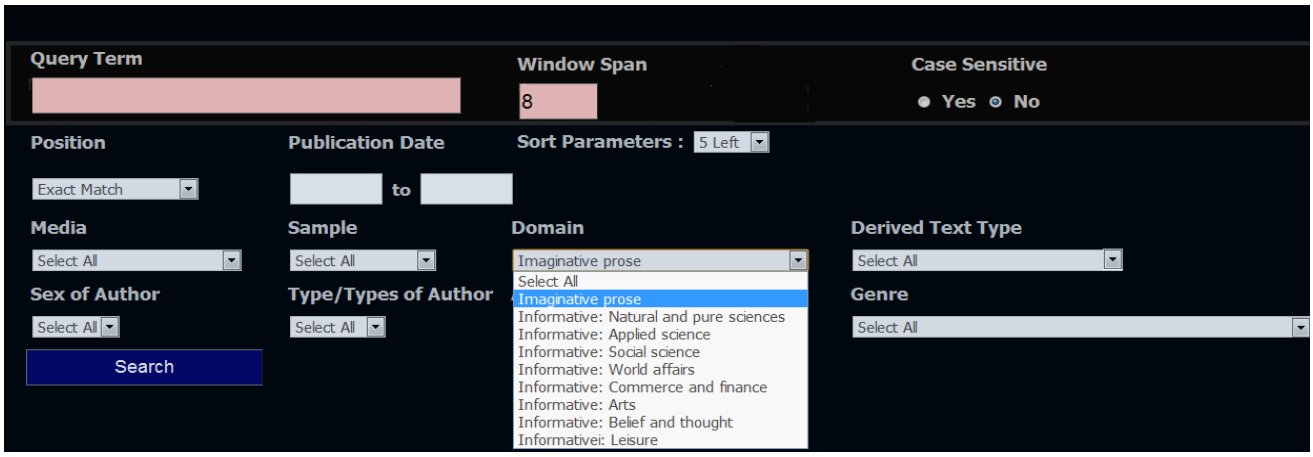

Figure 2: TNC search page

Figure 3: TNC results page

6. Conclusion

TNC is a representative and a well-balanced general reference corpus of contemporary Turkish. The construction of TNC is made possible by contributions of a dedicated team of academics, information technology specialists, programmers, and (under) graduates who are actively involved in its various stages. Careful planning and rigorous adherence to design principles derived from previous practices in corpus construction helped builders at all steps in the process. TNC will serve as an important resource for researchers who are interested in linguistic aspects of Turkish.

7. Acknowledgements

TNC was supported by a research grant from the Scientific and Technological Research Council of Turkey (TÜBİTAK, Grant No: 108K242).

8. References

Aksan, Y., Aksan, M. (2009). Building a national corpus of Turkish: Design and implementation. *Working Papers in Corpus-based Linguistics and Language Education* no. 3, 299-31. Tokyo: TUFS.

Aksan, Y. (2009). Türkçe ulusal derlemi oluşturma: İlkeler ve tasarım. In Y. Özdemir (Ed.), *Proceedings of Mersin Symposium* 19-22 November 2008. Mersin: Güven Ofset, pp. 565- 569.

Aksan, M., Mersinli, Ü. (2011). A Corpus Based Nooj Module for Turkish. In Z. Gavriilidou et al. (Eds.), *Proceedings of the Nooj 2010 International Conference and Workshop*. Komotini, pp. 29-39.

Aksan, Y., Ü. Mersinli, Y. Yaldır, U.U. Demirhan (2011). Türkçe Ulusal Dil Derlemi Projesi Biçimbirim Çalışmalarında Belirsizliklerin Sınıflandırılması ve Dağılımı. Paper presented at the *25th National Linguistic Conference*, 5-7 May 2011, Çukurova University.

Aston, G., Burnad, L. (1998). *The BNC Handbook: Exploring The British National Corpus with SARA*. Edinburgh: Edbinburgh University Press.

Biber, D. (1993). Representativeness in corpus design, *Literary and Linguistic Computing* 8:243-257.

Leech, G. (2007). New resources or just better old ones? In M. Hundt, N. Nesselhauf and C. Biewer (Eds.), *Corpus Linguistics and the Web*, pp. 134-

149. Amsterdam: Rodopi.
McEnery, T., Xiao, R., Tono, Y. (2006) *Corpus-based Language Studies*, Routledge.
McEnery, T., Hardie, A. (2012). *Corpus Linguistics*. Cambridge: Cambridge University Press.
Ruhi, Ş., H. Işık-Güle, Ç. Hatipoğlu, B. Eröz-Tuğa, D. Çokal-Karadaş (2010). Achieving Representativeness Through the Parameters of Spoken Language and Discursive Features: The Case of the Spoken Turkish Corpus. In I. Moskowich-Spiegel Fandino, B. Crespo García, I. Lareo Martín (Eds.), *Language Windowing through Corpora*. *Visualización del lenguaje a través de corpus*. Part II. Universidade da Coruna, pp. 789-799.
Silberztein, M. (2003). NooJ Manual. 24.01.2012. http://www.nooj4nlp.net