『コーパスに基づく言語学教育研究報告』No.3（2009）

Building a National Corpus of Turkish: Design and Implementation[1]

Yeşim Aksan & Mustafa Aksan
(Mersin University)

Abstract

A project to build a national corpus of Turkish is designed by a team of linguists at Mersin University and is funded by the Scientific and Technological Research Council of Turkey. The first large-scale project of its kind, the corpus will be available for users by 2011. This paper will introduce the developmental stages of the project and will discuss its design features.

1. Introduction

Corpus is defined as a collection of texts stored in an electronic database. A corpus represents varieties of spoken or written text types, sampling language in use, providing researchers the most fundamental database upon which they can search various aspects of language.

Rapid developments in the recent computer technology made it possible to access and process huge amount of data and consequently have caused a shift in linguistic research from introspective data to naturally occurring data. This shift ultimately led to the construction and use of a number of well constructed language corpora which in turn help recognize the importance of real data representing those aspects of language that were not possible to observe otherwise. Furthermore, corpus construction processes contributed enourmously to the emergence of a new field in linguistics, namely corpus linguistics and also to the efficiency of the applications of various natural language processing studies.

In the last three decades, 190 different corpora representing 80 different languages have been constructed. Today, most of these corpora are organized under various international consortiums and are available for users for a variety of purposes. Such corpora serve not only for general and applied linguists but also computational linguists and educational linguists alike. Currently, among the corpora of languages available for access there is no such corpus of Turkish.

In this paper we introduce the *Turkish National Corpus* (TNC) project (http://tnc.org.tr), and

[1] This paper has been written in the framework of the project funded by the Scientific and Technological Research Council of Turkey (project no: 108K242).

describe TNC's design and interface features. First part of the paper discusses the selection and sampling procedures of written and spoken data in the construction of the electronic database of TNC. Second part of the paper addresses linguistic and search tools planned to be created for the TNC. More specifically, the individual steps to develop a parts of speech tagger and corpus interface are described.

2. Information on the *Turkish National Corpus* Project

Turkish National Corpus project is designed by a team of linguists from Mersin University and is funded by the Scientific and Technological Research Council of Turkey for a period of three years (2008-2011). It was launched in October 2008 and will run until October 2011. The estimated size of the corpus will be 50 million words and it will cover the time period from 1990 to 2008. Annotation of the corpus will be parts of speech tagging. As for the medium of the corpus, 95% of it will consist of samples compiled from written language and 5% of it will involve samples gathered from spoken Turkish. TNC will have a user-friendly interface and it will be accessed via internet by any web browser. TNC will not be specifically restricted to any particular subject field, genre or register so it will be a *general* corpus. Since it will contain samples of both written and spoken language, it will be a *mixed* corpus. TNC will be a sample corpus because it will be composed of text samples no longer than 45.000 words. It will include imaginative and informative texts representing contemporary use of Turkish of the late 20th century, and with this feature TNC will be a *synchronic* corpus.

In short, the output of the Turkish National Corpus project will be a national corpus which should represent Turkish language in the most comprehensive and balanced way possible in order to allow for all types of researchers to access and derive the relevant information that would serve multiple research purposes.

In building TNC, we will follow the methodology of corpus linguistics in its current state. The activites involved in the construction of TNC are divided into four work packages, and they are illustrated in Figure 1:

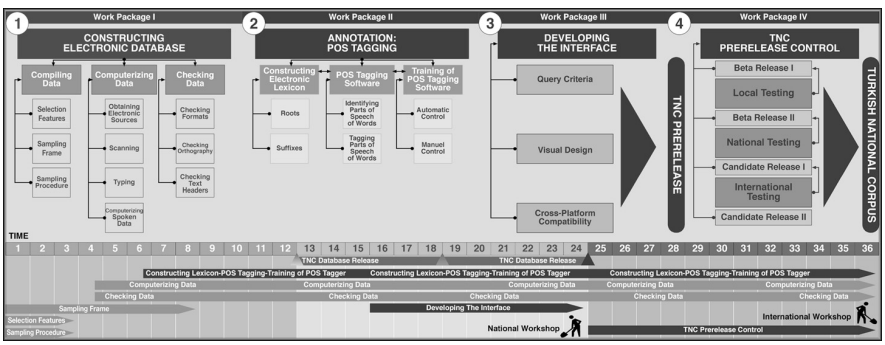

Figure 1 Work packages of the TNC project

Constructing electronic database consists of collection, computerizing and checking of corpus data. The activities of the first workpackage include identification of data sources, defining criteria for sampling, computerizing the data and checking and revising both the corpus data and text headers regularly.

Annotation of the corpus involves the steps of developing a parts of speech tagging software for Turkish. The activities of the second workpackage include the following: text annotation and parts of speech tagging, preparing a lexicon of roots and affixes for parts of speech tagger; developing a parts of speech tagger that would tag at least 5 million lexemes; testing and upgrading the tagger .

In developing the corpus interface, we will consider query criteria, visual design and cross-platform compatibility of the software that will operate the Turkish National Corpus. The activities of the third workpackage include: developing a user friendly graphic interface that would ease access to the corpus; defining query critera and query fields that would appear in the interface in accordance with the fundamental design of the corpus; checking and upgrading compatability of the interface (Windows, Linux, MacOS, SunOS, and Pardus) for different users and different operating systems.

Turkish National Corpus prerelease control will include two beta and two candidate releases for local, national and international testing. The activities of the fourth workpackage include: releasing beta version I of the corpus for local trial; releasing beta version II for local and national use; releasing candidate versions I for local, national and international uses; upgrading the versions of the corpus on the basis of the feedback provided at the end of each trials.

3. Construction of TNC's written and spoken database

The major issue that should be addressed in design of TNC is its representativeness. As Biber maintains (1993: 242) "representativeness refers to the extent to which a sample includes the full range of variability in a population." In other words, representativeness can be achieved through balance and sampling of language or language variety presented in a corpus. A balanced general corpus contains texts from a wide range of genres, and text chunks for each genre are sampled proportionally for the inclusion in a corpus. Texts are selected on the basis of external criteria that are defined situationally irrespective of the distribution of linguistic features. Linguistically defined internal criteria are not considered to be primary parameters for the selection of corpus data (Atkins, Clear and Ostler, 1992). Overall, can we mention a scientific measure for a balanced corpus? The answer is no. The most widely followed approach to corpus balance is that corpus-builders adopt an existing corpus model when building their corpus. *British National Corpus* (BNC) (http://info.ox.ac.uk/bnc) is generally accepted as being a balanced corpus. The BNC model 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 and Tono, 2006).

The TNC project will also follow the BNC model and it will make necessary adjustments whenever a need would arise for a change.

Written text included in Turkish National Corpus will be selected using three criteria: *domain*, *time* and *medium of text*. Domain refers to subject field of the text. Imaginative domain consists of fiction (novel, short story, poem, drama). Informative domain will include samples from social science, art, commerce-finance, belief-thought, world affairs, applied science, natural-pure science, and leisure. Time refers to the period of text production. All the texts should be published between 1990 and 2008. Medium of text refers to type of text production. Written samples are collected from books, periodicals (newspapers, magazines), published or unpublished documents and texts written to be spoken such as, news broadcasts, screen plays.

It has been agreed that selection of texts for a general purpose corpus should take into consideration both received (read and heard) and produced (written and spoken) language (Biber, Conrad, Reppen, 1998; Kennedy, 1998; Meyer, 2002 among others). Where do corpus-builders find published materials representing target population? Catalogues of books, books in print lists, best seller lists, particularly prize winners of, lists of current magazines and periodicals, periodical circulation figures give useful information about the totality of written text produced and received. Yet, there is no single source of information about published material that can provide a satisfactory basis for a sampling frame. A combination of various sources would give useful information about the totality of written text produced and received. They are mainly statistics about books and periodicals that are published, bought or borrowed.

In order to construct the written database of TNC, we have so far identified a list of sampling units, that is, published materials on any subject matter from bestsellers lists, prizewinners of competitions lists (e.g., prizes of *Yunus Nadi, Orhan Kemal Roman, Haldun Taner Öykü*), and from books in print lists. We also try to find library lending statistics to detect widely received written materials published between 1990 -2008. For periodicals, we have decided to collect data from five national newspapers (*Radikal, Milliyet, Zaman, Türkiye, Cumhuriyet*) and magazines (*Birikim, Aktüel, Telepati, Sızıntı, Aksiyon*) each covers a variety of worldviews, ideologies and subject matters. We are also compiling texts from local newspapers.

To obtain representative samples from target population (i.e., published materials) for TNC, we have decided to use stratified random sampling technique (Biber, 1993). We will first divide the whole target population into relatively homogeneous groups. Texts belonging to imaginative and informative domains and text categories or genres that take part under these domains constitute the target population. Upon dividing them into homogenous groups, text samples will be selected from each group randomly. As for sample size, following the BNC model, a target size of 45.000 words will be chosen from books. A convenient break point will be chosen for text samples so that a continuous stretch of discourse from within the whole will be selected. Samples will be taken randomly from the beginning, middle or end of longer texts. Copyright permissions will be obtained from the publishers. Sampling from newspapers will be different. Since newspapers

consist of combination of texts, discrete texts that they contain will be separated and classified individually according to the selection and classification features. For instance, the individual stories in one issue of a newspaper will be grouped according to subject matter, for example as business articles, leisure articles, sports articles.

5% of TNC's database is constituted by spoken data. As employed by BNC, there will be two parts in the spoken component. (i) Demographic part will contain transcriptions of spontaneous natural conversations made by members of public. Equal numbers of men and women recruits will use a personal stereo to record all their conversations over two to seven days, and log details of every conversation recorded (date, time, setting, brief details of other participants) in a notebook. Permission will be obtained form the participants of the conversations upon each recording. Information about the participants, such as age, sex, accent, and occupation will be recorded when available. (ii) Context-governed part will consist of transcriptions of recordings made at specific types of meeting and event. We aim to collect speech recorded in each of the following categories of social context:

Educational-informative events: Lectures, news broadcasts, and classroom discussions

Institutional-public events: Political speeches, parliamentary proceedings, and council meetings

Leisure events: Sport commentaries, club meetings, and after-dinner speeches

3.1. Computerizing and checking data

Re-use of existing electronic texts:

Turkish is relatively well- represented in the internet, that is to say, a sizable language data is already available. Among many, we may count various official documents prepared either by government agencies or private companies representing specific medium of language use, majority of the national newspapers with full contents available dating back to 1998, most of weekly or monthly magazines with various specialized contents, academic texts produced by researchers presented at web pages of about hundred different universities ranging over natural, social and health sciences, hundreds of blog pages and thousands of institutional and personal web pages.

Scanning:

A quick search of the internet for Turkish reveals that potential material available for a corpus unfortunately goes back to late 1990s. This means that for TNC to represent the language "older" than late 1990s, an important amount of published material is yet to be computerized. Hence, computerizing the data will be quite laborious and time-consuming. In this respect, the selection of the published material to be computerized will be important; the project design sets up a number of linguistic and non-linguistic criteria in the selection and computerizing the written and spoken texts.

In the next step, texts coming with different formats will be converted into simple texts and the OCR software will be trained to produce as output the least problematic texts. Most of the

OCR available lack Turkish character set and we expect to be able to produce a working set through training.

Keyboarding:

As one goes back in time, in our case, going back to early 1990s, the written texts available sometimes come in relatively poor quality. When scanned, the output is an image of the text that cannot be converted properly into a text file. In such cases, a team of data entry operators will simply type the texts in question. Typing is further required when the text that is qualified to be represented in the corpus is handwritten.

3.2. Workflow and Checking

The end result of corpus building process should be a collection of text with no or minimal errors. To ensure an error free corpus database, each text sample should be rigorously checked and rechecked. In order to handle the data coming from the OCR and typing, the project design sets up a number of check points along the course which is represented in Figure 2.

Any text coming from data entry operators is first checked for spelling and other relevant coding properties by both human controllers and by specialized (and sometimes modified) computer software. Following this initial checking, the texts are then committed to Subversion repository 1 (SVN), here used as a data management system. SVN software provides the management and control of multiple versions of the same data worked on by members of the corpus team. Any user of the system can recall and recheck the data at any time and can recommit the final version of the error-free data to SVN repository 2.

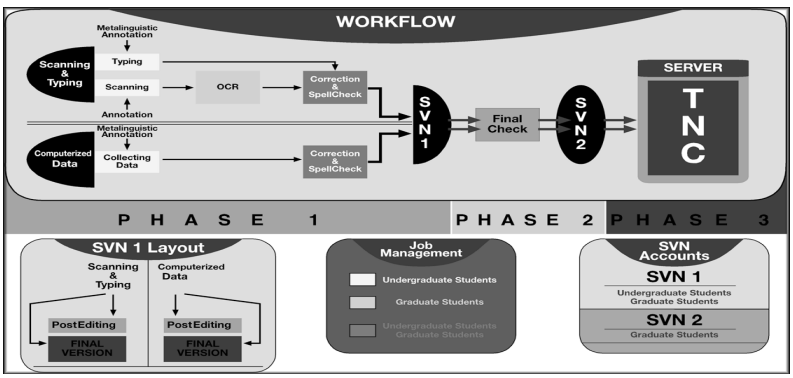

Figure 2 Workflow chart

The data entry operators include undergraduate students and graduate students. The graduate students are also the members of the project team and are further responsible for the control of the data entry process. Since a huge amount of linguistic data is to be processed, the entry process

needs to be tracked properly. To keep control and proper track of the workflow, each data entry operator working at the lab should fill a “yellow card” which encodes the details of his or her activity (see Figure 3). The information in the cards is stored both in printed data files and in electronic data files.

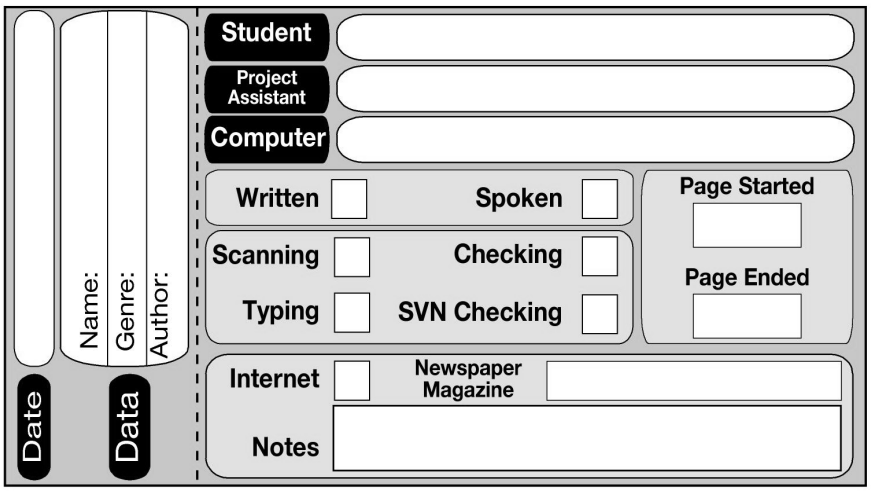

Figure 3 Logging the entries

4. Annotations in TNC

Annotation adds value to a corpus. Corpus annotation is “ a crucial contribution to the benefit a corpus brings, since it enriches the corpus as a source of linguistic information for the future research and development” says Leech (1997:2).

TNC will provide meta-linguistic annotation and grammatical annotation. The annotation process, for the time being, will follow the standards already set by previous corpus building projects. In this respect, as indicated above, we will follow the BNC model. A sample model of text header for a book, representing a meta-linguistic annotation is given below:

Title:	Excerpt from Benim Adım Kırmızı. Sample containing about 38.764 words
Spoken or written:	Written
Number of words:	38.764
Derived text type:	Fiction
Text type: Fiction:	Prose
Publication date:	1998
Age of author:	46
Sex of author:	Male
Type of author:	Sole

Age of audience: Adult
Text domain: Imaginative Fiction
Medium of text: Book
Text sample: Middle sample
Name of author: Orhan Pamuk
Name of text: Benim Adım Kırmızı
Target audience sex: Mixed
Publisher: İletişim
Place of publication: İstanbul
Key words: History, crime, classical arts

This meta-textual information will be added to all text samples to be included in the corpus. The user of the corpus will thus find all the information relevant to the published material and can search for items on the basis of meta-textual information.

The parts-of-speech tagging (POS tag) will be conducted at two different stages. In the first stage, lists of root forms and suffixes will be prepared alongside with a rule set. The second stage will test the outputs of the tagged texts coming from both manually and automatically tagged material.

A rule-based morphological analyzer will work on an electronic lexicon of Turkish, and will run initially on the corpus data of at least 10 million words. The process will obtain the most likely Turkish grammatical categories — such as nouns, verbs, adjectives, adverbs, conjunctions, etc. Simultaneously, a probabilistic analyzer will further run on a corpus data containing manually tagged parts of speech annotation, and will turn the most probable word classes of Turkish.

The results obtained from both analyzers will then be compared for their validity. From these outputs, the morphological analyzer will be updated. To achieve precision in POS tagging software, human post-editing will be conducted on the outputs of software (see Figure 4).

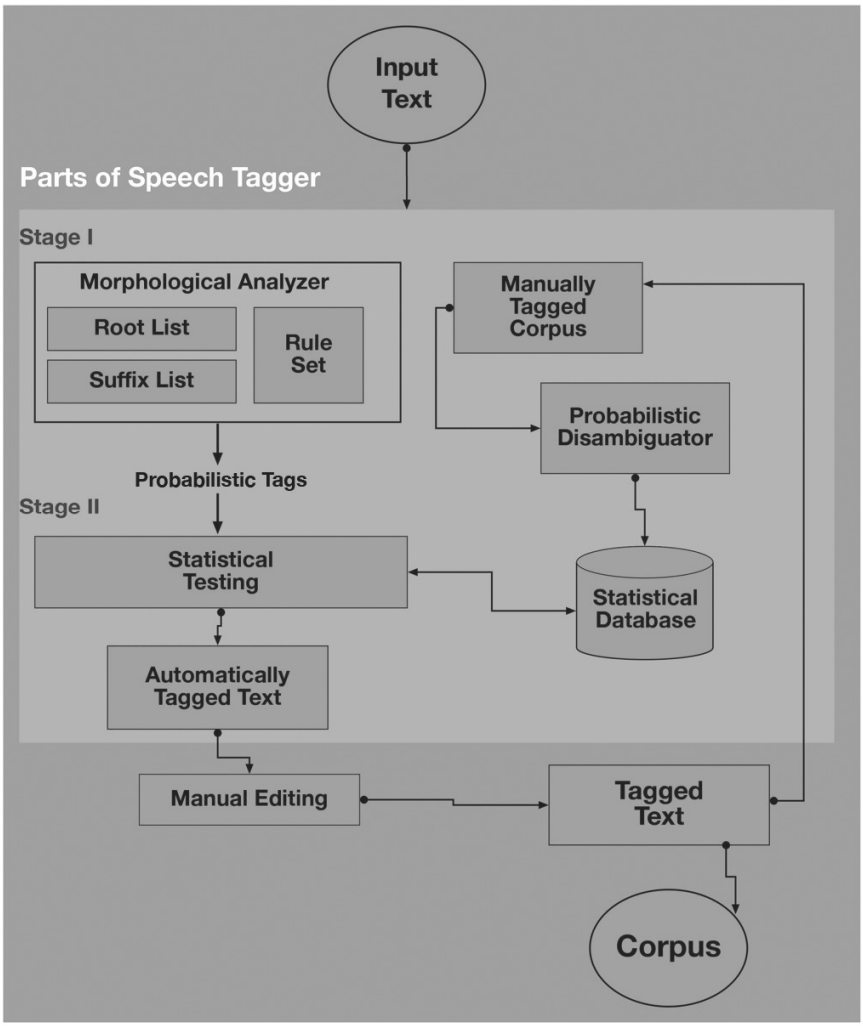

Figure 4 Developing a parts-of-speech tagger for Turkish

5. Developing TNC's interface

The interface will be designed so as to provide the most easily accessible environment for the users of the corpus. In the construction of the interface, the TNC project will concentrate on three aspects of a user-friendly interface: interface features, visual design and cross-platform compatibility of corpus interface.

The interface features of the TNC will be based on features provided by the BNCweb — CQP edition (Smith, Hoffmann, Rayson, 2008). The interface will include features primarily to produce concordance output from the corpus and will also provide a range of other functions, some of

which include, specification of the categories of text, and/or sections of texts; specification about what to search for; navigating through concordance output; facility to access metatextual information and frequency data. To illustrate interface features, for instance, the user of TNC will search for the word aşk 'love' within written or spoken texts, within fiction texts written in the 2000 by female authors or within the headlines of newspaper texts. The user will have a choice of viewing concordance lines within sentence context, or in keyword-in-context mode. The user will compile frequency lists of lexical items based on user-definable criteria such as lemma types, POS-tags.

The visual design of the corpus interface will be user-friendly just as the one created for the BNCweb, as illustrated in Figure 5:

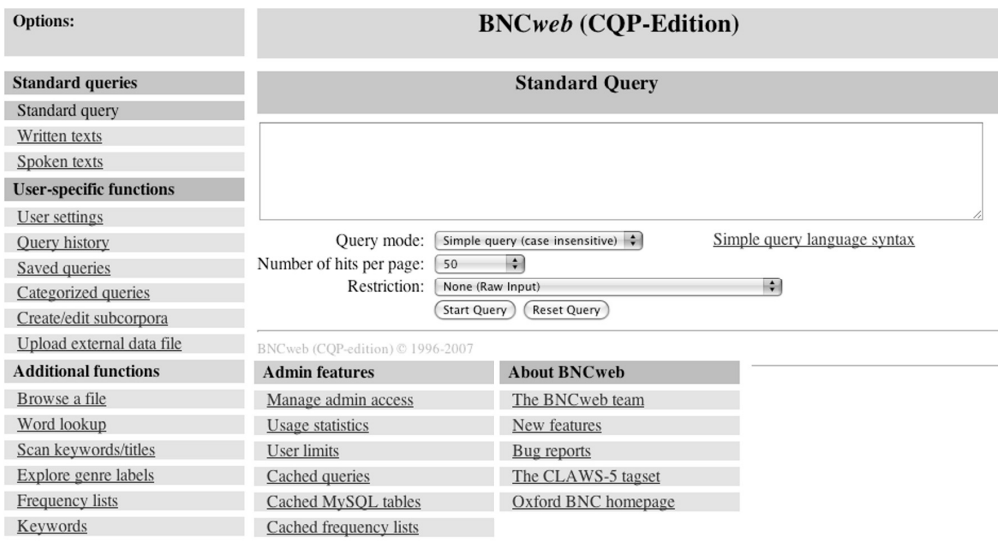

Figure 5 BNCweb interface

Cross-platform compatibility (platform free interface) will be furnished by emphasizing at least three criteria: Character set, texts in the corpus data will be saved in UTF-8 character code; Platform-free corpus interface will be constructed in terms of w3c standards and it will be accessed through any web-browser on any platform; Result set, results of the corpus will be used on any platform and they will be compatible with other software.

Finally, prereleases of the corpus will come at four stages. Two beta and two candidate releases of Turkish National Corpus will be available for potential users before its final release in 2011.

Beta Release I - Local Testing: Bringing out the corpus on the local network for testing.

Beta Release II - National Testing: Upon local testing, revealing the corpus to national users and linguists.

Release Candidate I - International Testing: Upon national testing, making the corpus available to international users.

Release Candidate II - Turkish National Corpus: The final candidate release before the release of the corpus.

6. Conclusion

As underscored by Kawaguchi (2007: 32) “Linguistic corpora furnish a great variety of linguistic research domains and fruitful applications in language education.” In this context, the specific contribution of Turkish National Corpus project will be more evident especially in relation to different fields of research. First of all, both the design and the specific applications of the corpus linguistics methodology in the process of building a new corpus will provide specific feedback to the theory of corpus linguistics. Second, developing language-processing software for an agglutinating language will bring out new challenges for those who are working in the field.

We hope that Turkish National Corpus will function as the basic source of reference for both national and international researchers who are willing to learn more about structural and lexical patterns of Turkish. A well-designed and balanced corpus will also increase the size of Turkish linguistic data in electronic format. Findings of corpus-based studies on Turkish will help develop more Turkish friendly and efficient word processors, internet browsers, and similar other widely used applications. On the practical side, corpus–based research will also contribute to constructing electronic archives, developing standards of electronic communications in Turkish, text summarization, text and web content mining, information extraction and retrieval and developing security systems. Further potential products of a national corpus will be a thorough and comprehensive grammar of Turkish and a dictionary that will represent both the frequencies and usage based meanings of lexemes.

References

Atkins, Sue, Jeremy Clear and Nicholas Ostler. (1992) “Corpus design criteria”, *Literary and Linguistic Computing* 7:1-16.

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

Biber, Douglas, Susan Conrad and Randi Reppen. (1998) *Corpus Linguistics,* Cambridge University Press.

Kennedy, Graeme. (1998) *An Introduction to Corpus Linguistics*, Longman.

Kawaguchi, Yuji. (2007) “Introduction”, in Yuji Kawaguchi et.al (Eds.) *Corpus-based Perspectives in Linguistics,* 31-38, John Benjamins.

Leech, Geoffrey. (1997) “Introducing corpus annotation”, in Roger Garside, Geoffrey Leech and

Anthony McEnery (Eds.) *Corpus Annotation,* 1-18, Longman.
McEnery, Tony, Richard Xiao, Yukio Tono. (2006) *Corpus-based Language Studies*, Routledge.
Meyer, Charles. (2002) *English Corpus Linguistics*, Cambridge University Press.
Smith, Nicholas, Sebastian Hoffmann and Paul Rayson. (2008) "Corpus tools and methods, today and tomorrow: Incorporating linguists' manual annotations", *Literary and Linguistic Computing* 23:163-180.
http://info.ox.ac.uk/bnc
http://www.bncweb.info