A corpus-based word frequency list of Turkish: Evidence from the subcorpora of *Turkish National Corpus* project*

Yeşim Aksan** – Yılmaz Yaldır**

Word frequency studies have a central role in various disciplines, such as linguistics, cognitive psychology, natural language processing, computational linguistics. Developments in the computer technologies and information processing help researchers make comprehensive word lists on the basis of digitally constructed language corpora. Since Kucera and Francis's first corpus-based word frequency lists derived from the *Brown Corpus* (1967), a variety of research have been conducted on general or specialized corpora to obtain rank frequency order and distribution of words for different Indo-European languages (Johansson & Hofland 1989; Leech et al. 2001; Baroni et al. 2004; Ha et al. 2006; Davies & Gardner 2010). In Turkish, Göz's dictionary (2003), which is based on a 1 million-word general corpus, is the only work on word frequency. In general, lexical properties of Turkish and, in particular, word frequency lists of text collections representing different registers of Turkish need to be described via corpus-based word frequency lists.

Keeping this necessity in mind, this study has two aims: (1) to produce word frequency lists of Turkish on the basis of two subcorpora, namely the *Corpus of Contemporary Turkish Fiction* and the *Corpus of Contemporary Turkish News Texts*. In this respect, frequency lists of both root types and word classes in Turkish are prepared; (2) to compare these two corpora by using frequency profiling information.

This paper is organized as follows. First we explain basic concepts and review literature of word frequency studies. Then, we describe the construction of two subcorpora used to derive wordlists and explain the steps followed in tokenization and root type mapping scheme on which the token and root counts are based. Finally, we compare rank frequency and word class lists of Turkish Fiction and Turkish News Texts Corpora.

1. Basic Concepts and Word Frequency Studies

The most frequently used terms of the study are *frequency*, *token*, *type*, *lemma* or *headword*, *type / token ratio*, and *standard type / token ratio*.

Frequency essentially refers to a value that specifies the number of occurrences of a particular linguistic item in a corpus. In other words, what is meant by the term frequency is the number of realization of a token, a type or a headword in a corpus or the number that shows how often we come across a particular linguistic element in a given

* *Turkish National Corpus* is supported by a research grant from the Turkish Scientific and Technological Research Institution (TÜBİTAK, Grant No. 108K242).

** Mersin University.

corpus. The frequency of a particular linguistic item can be given either as a numeric value – which is known as raw data – or as a percentage. *Token* refers to a linguistic item that is limited by a space character or a punctuation mark on both sides in a corpus. Let us assume that we find a number of words, such as *kitaplık, adamdan, evlere, gelmiştik, uzak* in a corpus. In this case, each of these words is considered to be a token. However, the relationship between a word and a token is not always straightforward. It is not unusual that some coding systems separate a particular word into two tokens depending on its composition despite the well-established spelling conventions. For example, in English, some words, like *didn't* or *he's*, can be regarded as linguistic items composed of two different tokens. In this specific case, for instance, the *did* and *n't* parts of *didn't* represent different types of linguistic information, corresponding two different tokens for one word. At this stage, it is wise to ask the following question: should we accept numeric symbols or punctuation marks as tokens? In order to avoid confusion at later stages of corpus study, it is reasonable to determine the type of coding system beforehand. In other words, it is rather advisable to specify the types of elements in a corpus that will be accepted as tokens from the very beginning, explicitly.

Any distinct word form making up a given corpus is referred to as *type*. In order to show the relationship between token and type, consider the following example: suppose that we come acrosss the word *evlerimizde* in 8 different places in a given corpus. Then, the item *evlerimizde* is regarded as a single (word) type which is represented by 8 different tokens in the corpus. Now, assume that we have a mini corpus which is composed of only a single sentence, as in the following: *Kitap ve defterleri aynı sırada ve aynı biçimde dizdi.*

Based on our previous definitions, there are 9 tokens in this mini corpus. Since we come across the words *aynı* and *ve* twice in the corpus, the number of different types is 7. Therefore, we can still talk about a single type even if a particular word is repeated many times in our corpus and we will adopt the idea that this single type is represented by a certain number of tokens in the corpus. The number of tokens is never lower than the number of types in a corpus.

This study adopts the following definition of the term *lemma* or *headword*. Lemma is the uninflected basic form of a word type. For instance, let us suppose that the following words are found in our corpus in the specified numbers:

mutluluğundan	(once)
mutluluktan	(2 times)
mutluluklar	(3 times)
mutlulukta	(4 times)
mutluluktu	(2 times)

In that case, we have 12 tokens and 5 different types. In fact, these 5 different types are the inflected forms of the same word, namely *mutluluk*. Therefore, the lemma (headword) *mutluluk* is the word that can represent these 5 types. In a corpus, the total number of lemmas are almost always far less than the total number of types.

Type/token ratio is the value obtained by dividing the total number of types by the

total number of tokens in a corpus. Since the total number of types is almost always far less than the total number of tokens, type/token ratio is invariably less than 1 (one). Usually, this ratio is given as a percentage. Type/token ratio does not reflect the lexical richness of a big corpus with respect to the variety of words found in that corpus. Standard type/token ratio is a better, more objective tool to represent the real variety of words in a corpus. Standard type/token ratio is the average of type/token ratios calculated for all 2000-word corpus files, making up the whole corpus (Baker & Hardie & McEnery 2006).

Since the beginning of the 20th century, long before the widespread usage of computers, a number of impressive frequency studies were carried out by a number of prominent figures in the field. These studies provided statistical information to be used in pedagogical contexts. Thorndike (1921), for example, was one of such influential studies. The study was based upon a corpus of 4,5 million words composed of classical literary texts and child literature. In fact, the principle of vocabulary control, which is very important for the design and redaction of pedagogical reading materials, owes much to this pioneering study by Thorndike. In this context, this principle can be stated in its most simple form in the following way: the most frequent words of a particular language should be taught to foreign learners first.

During 1930s, thanks to the support of the Carnegie Corporation, a number of prominent linguists and language teaching experts, such as Thorndike, West, Palmer, Sapir and Faucett, came together and carried out a series of statistical vocabulary studies. Thorndike and Lorge's (1944) *The Teacher's Word Book of 30,000 Words* and Michael West's *A General Service List of English Words* (1953) were the products of this project.

Upon the developments in computational linguistics dating back to 1950s, studies on corpus linguistics became possible by the increase in data storage capabilities and speed in data processing. In this way, language corpora have become a primary tool for word frequency research. Here, we summarize three corpus based word frequency studies. The first one is *Computational Analysis of Present-Day American English* by Nelson Francis and Henry Kucera (1967) which is based on *Brown Corpus*. *Brown Corpus* is composed of 1 million words drawn from written texts of 1960s American English. This was the first corpus of its type because its compilers employed computers in its preparation stage. It is made up of 500 different written texts, which come from 15 distinct genres. In a way, Brown Corpus has functioned as a model for succeeding corpora prepared in later periods.

The second corpus-based frequency study is Leech, Rayson and Wilson's work (2001) based on the British National Corpus (BNC). BCN is a corpus of 100 million words comprised of British English texts representing 1980s and 1990s. 90 % of the texts were written and 10 % of them were spoken. The written component of the corpus contains texts from a variety of domains and genres. The spoken component of BNC includes dialogues, business meetings, radio programmes, official meetings, etc. recorded by a group of volunteers from different geographical regions and social classes. The part of speech tagging of the corpus is carried out by means of a software called CLAWS. The word frequency lists of this corpus is given in *Word Frequencies of Written and Spoken English: Based on the British National Corpus*. The frequency lists of the subcomponents of this corpus were prepared in two forms so that it is possible to find the frequencies of lemmas

as well as that of word types for the corpus.

The last work is *A Frequency Dictionary of Contemporary American English* (2010) by M. Davies and D. Gardner based on *Corpus of Contemporary American English* made up of more than 400 million words. This dictionary presents us with the most satisfactory word lists. These lists extraordinarily include not only the most frequent words, but also the most frequent collocates. Since the corpus itself is extremely large, the total number of the words found in this corpus is also exceptionally large. Unlike the previous studies, it is possible to find the ranked frequency lists of the first 10 thousand or 20 thousand words. Along with these frequency lists, first 20–30 collocates of all words in the lists are also identified in the study.

As for the frequency studies in Turkish, we see that such studies in Turkish morphology dates back to 1960s. Pierce (1961; 1962) worked on frequency of inflectional and derivational suffixes both in spoken and written Turkish. In his 1961 study, he first built a 140,000-word corpus which is mainly involving the conversations of illiterate factory workers and recorded life stories of illiterate army draftees told by themselves. He found that the most frequent 21 suffixes in Turkish were inflectional suffixes. The first ten of these suffixes are:

Table 1. Pierce (1961) The most frequent inflectional suffixes in spoken Turkish

Rank	Suffix	Sample
1	*-İyor*	*geliyor*
2	*-Di*	*gitti*
3	*-(y)A*	*okula*
4	*-(y)I*	*evi*
5	*-lAr*	*kitaplar*
6	*-(s)I*	*kapısı*
7	*-(y)Im*	*giderim*
8	*-mIş*	*gelmiş*
9	*-(n)In*	*evin*
10	*-DE*	*okulda*

Pierce (1962), on the other hand, is a study on frequencies of Turkish suffixes in written texts. This study used a corpus of 100,000 words from written texts, including military field manuals, course books, poems, religious stories, and selected articles from newspapers and periodicals. Pierce (1962) identified 139 different suffixes in its corpus. Out of the most frequent 29 suffixes, only 4 of them were derivational suffixes. According to the findings of the study, the most frequent 10 suffixes in this corpus of written texts are given below.

Table 2. Pierce (1962) The most frequent inflectional suffixes in written Turkish

Rank	Suffix	Sample
1	*-(y)I*	*evi*
2	*-lAr*	*kitaplar*
3	*-DE*	*okulda*
4	*-(n)In*	*evin*
5	*-(y)I*	*kitabı*
6	*-(y)Im*	*giderim*
7	*-(y)A*	*okula*
8	*-Di*	*gitti*
9	*-DEn*	*evden*
10	*-mIş*	*gelmiş*

Göz (2003) prepared the first frequency dictionary of written Turkish. In doing this, he first compiled a pool of written materials that can represent written Turkish. The genres and their percentages of the materials in the pool can be seen in Table 3.

Table 3. Göz (2003) Genre and distribution of written texts

Genre	Percentage
Press	35
Novel-Story	20
Science	8
Popular Science	9
Fine Arts, Biography	8
Hobby	4
Religion	3
Course Books	3
Miscellaneous	10

After that, proper names were removed from the pool. Then, the total number of tokens were determined by means of a word count software. According to this result, there were 975,141 tokens in the pool. This step was followed by the specification of total number of word types, which was given as 179,861 by Göz (2003). This number reduced to 22,693 as far as the total number of lemmas was concerned. In other words, there were 22,693 headwords in his dictionary.

2. Method

2.1. Database

Two equal size sub-corpora covering a period of 20 years (1990–2009) were constructed from the databases of an ongoing *Turkish National Corpus Project. Corpus of Contemporary Turkish Fiction* (CCTF) is a 1 million-word corpus and it consists of samples from the novels and short stories of contemporary Turkish authors. Out of 200 texts

CCTF contains, 129 of them are novels and 71 of them are short stories. Table 4 shows the types and number of the fiction texts compiled in the construction of the CCTF.

Table 4. Types in the corpus of contemporary Turkish fiction

Types	Number of texts
General fiction	125
Romantic fiction	29
Historical fiction	16
Mystery	12
Humour	11
Adventure	7

Corpus of Contemporary Turkish News Texts (CCTNT) is a 1 million-word corpus. It contains news texts from different sections of five national newspapers which have different ideological point of views. *Cumhuriyet*, *Türkiye*, *Zaman*, *Milliyet* and *Radikal* are the newspapers used in the construction of CCTNT. Representativeness and balance of the two corpora were achieved by including wide range of texts through equally sized samples. Sampling frames and number of text samples compiled in the construction of CCTNT are shown below.

Table 5. Sampling frame and distribution of text samples in CCTNT

Year: 2009	Economy	Social	Sports	Science	Miscellaneous	*Total*
Cumhuriyet	4,450	4,450	4,450	4,450	4,450	22,250
Türkiye	4,450	4,450	4,450	4,450	4,450	22,250
Zaman	4,450	4,450	4,450	4,450	4,450	22,250
Milliyet	4,450	4,450	4,450	4,450	4,450	22,250
Radikal	4,450	4,450	4,450	4,450	4,450	22,250
Total	22,250	22,250	22,250	22,250	22,250	111,250

2.2. Generating word lists from the Corpora

The software NooJ (Silberztein 2003) as a corpus processor was used to generate word lists from fiction and news texts corpora. Since one of our aims is to develop a Turkish module for NooJ, we focus on unique root types. Thus, we mapped each word token onto a root type in this study. To achieve this end, we follow the steps specified below.

1. Token lists for each corpus were extracted with NooJ including frequencies and a case-sensitive word form lists.
2. From these token lists non-words such as single letters and typing/OCR errors were filtered out.
3. Proper nouns, abbreviations, acronyms are extracted from the token lists.
4. Lemmatization – in our case, root type identification – was made through the pre-defined lemma, parts of speech and affix database of Turkish National Corpus Project.
5. The type-frequencies with related root types and lexical categories were matched. For 53.000 different tokens, root type lists were produced semi-automatically. Lexical category of each root type was also pre-tagged.
6. Root type and word class frequencies were computed via Excel.

This paper presents the preliminary results of a statistical study on present-day Turkish lexicon and is based on observed frequencies of single word forms. Followings are all beyond the scope of this word frequency study:

i. statistics on multi-word units such as compounds with light verbs *ol-*, *et-*,
ii. frequencies of affixes and affix combinations of Turkish,
iii. context sensitive disambiguation of homographs such as *için* and related quantitative data,
iv. computing and comparing the frequencies of multiple senses a given root type may have.

3. Preliminary findings

3.1. Root Type / Token Ratio in two corpora

Following the idea of type/token ration, we calculate root type/token ratio. The number of unique root types in each corpus was divided by the number of tokens. While in fiction texts, the ratio is 0,11 in news texts it is 0,09. Corpus of fiction texts contains slightly more root types than the corpus of news texts.

3.2. The 15 top-ranked root types in frequency lists

In both corpora, as is expected, the top ranks are occupied by function words such as *bir* 'a, one', *ve* 'and' and light verbs such as *ol-* 'become', *et-* 'make'. Following these function words, we see polysemous verbs which have many different meanings such as *al-* 'take', *ver-* 'give'. There is a strong correlation between frequency of use and degree of polysemy (Kennedy 1998: 108).

The most striking difference between fiction and news texts is that frequency list of fiction contains two pronouns among the 6 top-ranked words. In the third order, the pronoun *o* 's/he', and in the sixth order the pronoun *ben* 'I' are seen. It appears that genre specific aspects of fiction texts require referential cohesion via pronoun uses. Making

qualitative data analysis is out of the scope of this paper, but the frequency profiling of fiction texts related to pronoun use should be studied in detail.

As is stated by Kennedy (1998: 102), "the more narrowly focused the corpus, the more content words find their way into the higher frequency levels". In our frequency lists of fiction and news texts, we see that content words such as *gör-aç* 'see-open' are assigned rank order twelve, *bak-gör* 'look-see' are assigned rank order thirteen respectively. Lexical semantic analysis of these content words can shed light on register based different uses of them.

Table 6. Rank / frequency profile: The 20 top-ranked root types

Rank	Fiction	Observed Frequency	News	Observed Frequency
1	*bir*	33,673	*ol*	24,847
2	*ol*	20,242	*ve*	21,081
3	*o*	14,844	*bir*	18,879
4	*ve*	11,316	*et*	11,235
5	*bu*	11,194	*bu*	10,985
6	*ben*	10,443	*yap*	9,551
7	*de*	10,363	*al*	6,256
8	*et*	7,750	*ver*	6,216
9	*ne*	7,474	*gel*	6,192
10	*gel*	7,466	*için*	6,067
11	*gibi*	6,492	*ile*	5,281
12	*gör*	5,853	*aç*	4,947
13	*bak*	5,632	*gör*	4,344
14	*baş*	5,592	*ön*	4,295
15	*kendi*	5,342	*bil*	4,29
16	*çok*	5,263	*de*	4,284
17	*ama*	5,261	*bul*	4,271
18	*bil*	5,101	*son*	4,154
19	*sen*	5,019	*çok*	3,895
20	*iç*	5,018	*sür*	3,878

LogLikelihood (LL) ratio of the 10 top-ranked root types in fiction and news corpora is calculated. LL is a measure of significance and it compares the observed and expected values for two datasets (Rayson and Garside 2000). On the basis of the 0,05 significance level, except for the demonstrative pronoun *bu* 'this', the rank frequency of all root types are significant due to the nature of the corpora. For instance, out of 95 cases of every 100 samples, *bir* 'a, one 'will always be used more in fiction when it is compared to news texts.

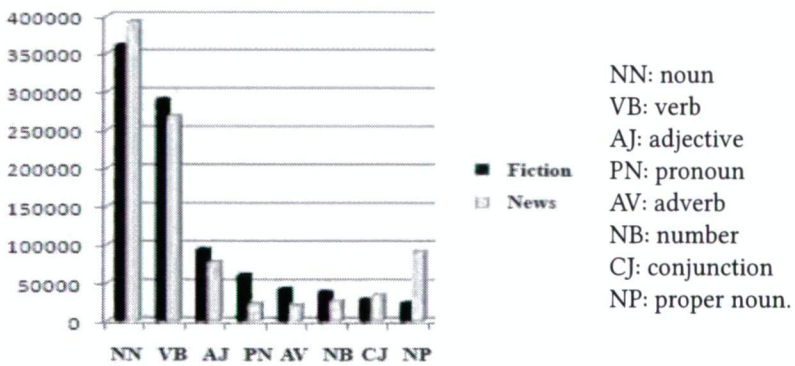
Table 7. LogLikelihood ratio of the 10 top-ranked root types of CCTF compared to CCTN

Root type	Fiction	News	LogLike
Bir	336723	18879	4221,51*
Ol	20242	24847	471,14*
O	14844	2339	8591,5*
Ve	11316	21081	2989,61*
Bu	111194	10985	1,97
Ben	10443	1499	7532,05*
De	10363	4284	2600,93*
Et	7750	11235	643,37*
Ne	7474	2061	3263,91*
Gel	7466	6192	119,01*

4. Part of speech frequencies in two corpora

The frequency and distribution of word classes reflect the nature of the two corpora. The results are similar to the ones identified for the word class rank order in the one-million Brown Corpus and London-Oslo-Bergen Corpus. Nouns are more frequent in the informative prose sections of the both corpora when compared with the imaginative prose: 28.50%–21.77%. We observe the same fact in our study. With the percentages of 39.31, nouns have slightly higher frequency in news texts when compared to fiction. We see a higher proportion of adjectives, pronouns and adverbs in fiction compared to news texts (See Figure 1. below). Similar results are obtained when informative and imaginative sections of the Brown Corpus were compared: pronouns account for 11.94% and adverbs accounts for the 6.72% of imaginative prose section of the corpus. On the other hand, pronouns account for 4.75% and adverbs account for the 4.73% of informative prose section of the Brown Corpus (Francis and Kucera 1982: 547).

Figure 1. Word class frequency distribution in CCTF and CCTNT



5. Conclusion and suggestions

In this paper, we have produced word lists and compared two corpora using frequency profiling. On the basis of token/root type mappings, we produced root type frequency lists and identified frequency and distribution of word classes in the subcorpora derived from the databases of Turkish National Corpus Project.

We would like to make some suggestions for the future studies. In extracting token / root type frequency lists, it is better to work with a part of speech tagged corpus where grammatical categories are assigned to words via an automatic tagger and human post-editing. To obtain high-quality frequency list, it is advisable to avoid ambiguities through automatic annotation of word classes. As a measure of lexical richness, lexical statistics like Zipf's Law (1965) can be used. Lexical density and richness of a language is described by determining frequency distribution, frequency spectra, and standard token/type ratio of corpus-based frequency lists. Finally, it is a necessity for Turkish to prepare a frequency dictionary based on a contemporary general corpus, containing at least 30 million words.

References

Baker, P., Hardie, A. & McEnery, T. 2006. *A glossary of corpus linguistics.* Edinburgh: Edinburgh University Press.

Baroni, M. et al. 2004. Introducing the *La Repubblica* Corpus: A large, annotated, TEI (XML)-compliant corpus of newspaper Italian. *Proceedings of LREC 2004.* ELDA, 1771–1774.

Davies, M. & Gardner, D. 2010. *Frequency dictionary of Contemporary American English.* London: Routledge.

Francis, W. N. & Kucera, H. 1982. *Frequency analysis of English usage. Lexicon and grammar.* Boston: Houghton Mifflin.

Göz, İ. 2003. *Yazılı Türkçenin kelimes sıklığı sözlüğü.* Ankara: Türk Dil Kurumu.

Ha, L. Q. et al. 2006. Zipf and type-token rules for the English, Spanish, Irish and Latin Languages. *Web Journal of Formal, Computational & Cognitive Linguistics* 8.

Johansson, S. & Hofland, K. 1989. *Frequency analysis of English vocabulary and grammar.* Oxford: Clarendon.

Kennedy, G. 1998. *An introduction to corpus linguistics.* London: Longman.

Kucera, H. & Francis, W. N. 1967. *Computational analysis of present-day American English.* Providence: Brown University Press.

Leech, G., Rayson, P. & Wilson, A. 2001. *Word frequencies in written and spoken English: Based on the British National Corpus.* London: Longman.

Pierce, J. E. 1961. A frequency count of Turkish affixes. *Anthropological Linguistics* 3: 9, 31–42

Pierce, J. E. 1962. Frequencies of occurrence for affixes in written Turkish. *Anthropological Linguistics* 4: 6, 30–41.

Rayson, P. & Garside, R. 2000. Comparing corpora using frequency profiling. In: *Proceeedings of the Workshop on Comparing Corpora, 38th Annual Meeting of the Association for Computational Linguistics 1–8 October 2000.* Hong Kong. 1–6.

Silberztein, M. 2003. NooJ manual. *www.nooj4nlp.net*

Thorndike, E. L. 1921. *The teacher's word book.* New York: Teachers College Columbia University.

Thorndike, E. L. & Lorge, I. 1944. *The teacher's word book of 30,000 words.* New York: Columbia University Press.

Turkish National Corpus Project. *www.tnc.org.tr / www.tudd.org.tr*

West, M. 1953. *A general service list of English words.* London: Longman.

Zipf, G. K. 1965. *Human behavior and the principle of least effort.* New York: Hafner.