APPLYING AND COMPARING SMOOTHING TECHNIQUES TO CONTEMPORARY PRINTED TURKISH

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of
Dokuz Eylül University
In Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy in Computer Engineering, Computer Engineering Program

by

Gökhan DALKILIÇ

June, 2004

İZMİR

Ph.D. THESIS EXAMINATION RESULT FORM

We certify that we have read this thesis **"APPLYING AND COMPARING SMOOTHING TECHNIQUES TO CONTEMPORARY PRINTED TURKISH"** completed by **GÖKHAN DALKILIÇ** under supervision of **ASSOCIATIVE PROFESSOR DR. YALÇIN ÇEBİ** and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Assoc. Prof. Dr. Yalçın ÇEBİ
Supervisor

Prof. Dr. Alp KUT
Jury Member
(Thesis Committee Member)

Yrd. Doç. Dr. Zafer Diri
Jury Member
(Thesis Committee Member)

Yrd. Doç. Dr. Banu Diri
Jury Member

Yrd. Doç. Dr. İsa Cakır
Jury Member

Approved by the
Graduate School of Natural and Applied Sciences

Prof. Dr. Cahit HELVACI
Director

ACKNOWLEDGMENTS

I want to thank my advisor Dr. ÇEBİ for his support and to see me like one of his colleagues not like his student. He believed in me with the basic idea of this thesis and always encouraged me to conclude on this thesis.

During this thesis, there are some friends who really helped me to see the end of this project by their encouraging help. I want to thank Tanzer ONURGİL and Rıfat AŞLIYAN for their help on n-gram creation algorithms. I also want to thank future computer scientist and my student Mutlu KAYA for his support of the prototype program for text error correction. I am also grateful to my officemate Şerife SUNGUN who always thought my health like her health with her everyday breakfasts.

I want to thank my parents for listening to my moans and groans about the writing process. They always supported my ideas and believed in my judgments.

İzmir, June 2004 Gökhan DALKILIÇ

ABSTRACT

For speech and optical character recognition, text correction, data encryption, etc. determination of the structural properties of a natural language is essential. These properties can be analyzed under two different categories; morphological and statistical analysis. For statistical analysis, a corpus which is a representative sample of the natural language is needed. Word n-gram frequencies of that corpus can be determined by using suitable algorithms and missing n-grams can be estimated by using smoothing techniques.

In this study, in order to compare and apply smoothing techniques to contemporary Turkish, a corpus named TurCo from which word n-gram frequencies would be investigated, was created. In order to calculate word n-grams, different algorithms were developed and tested.

After finding monogram, bigram, trigram, tetragram and pentagram word lists, their characteristics were analyzed. For generalization, Zipf's Law was applied, and to increase the accuracy in Zipf's Law, Mandelbrot Law was applied by finding the appropriate constants of Mandelbrot.

As the corpus could not be big enough to represent all of the language, smoothing techniques were used to estimate the unseen word n-grams. After the investigation and comparison of smoothing techniques, it was assumed that Back-off technique would give the best result. To apply this technique and to evaluate the results, by using the Minimum Edit Distance method a prototype program was developed, and the results were compared with Microsoft Word XP.

Keywords: Corpus, word n-grams, smoothing, n-gram analysis algorithm, Turkish corpus, Turkish word n-grams

ÖZET

Ses ve optik karakter tanıma, metin düzeltme, veri sıkıştırma, vs. için doğal bir dilin yapısal özelliklerinin belirlenmesi gereklidir. Bu özellikler, morfolojik ve istatistiksel analiz olmak üzere iki ayrı kategoride incelenebilir. İstatistiksel analiz için, doğal dili temsil eden örnek bir külliyata (corpus) ihtiyaç vardır. Bu külliyatın kelime n-gram frekansları, uygun algoritmalar kullanılarak saptanabilir ve eksik olan n-gramlar düzeltme (smoothing) teknikleriyle tahmin edilebilir.

Bu çalışmada, düzeltme tekniklerini karşılaştırmak ve güncel Türkçeye uygulamak amacıyla, kelime n-gram frekanslarının araştırılabileceği TurCo isminde bir külliyat yaratılmıştır. Kelime n-gramlarının hesaplanması için değişik algoritmalar geliştirilmiş ve denenmiştir.

Monogram, digram, trigram, tetragram ve pentagram kelime listeleri bulunduktan sonra özellikleri incelenmiştir. Genelleme yapmak için Zipf Kanunu uygulanmış ve Zipf Kanunu'nun duyarlılığını arttırmak için uygun Mandelbrot sabitleri bulunmuştur.

Külliyat, dilin tümünü temsil edecek kadar büyük olamayacağından, görülmeyen n-gramların tahmini için düzeltme teknikleri kullanılmalıdır. Düzeltme tekniklerinin incelenmesi ve karşılaştırılması sonucunda, Back-off yönteminin en uygun çözümü verebileceği öngörülmüştür. Bu yöntemin uygulanması ve sonuçların değerlendirilmesi için En Kısa Düzeltme Uzaklığı (Minimum Edit Distance) yöntemi de kullanılarak deneysel bir yazılım geliştirilmiş ve sonuçlar Microsoft Word XP ile karşılaştırılmıştır.

Anahtar sözcükler : Külliyat, kelime n-gramları, yumuşatma, n-gram analiz algoritması, Türkçe külliyat, Türkçe kelime n-gramları

CONTENTS

Page

Chapter Three
TURKISH CORPUS AND ITS CHARACTERISTICS

Chapter Four
WORD N-GRAMS

Page

Chapter Five

EVALUATION OF SMOOTHING TECHNIQUES AND DEVELOPING A TEST SOFTWARE

Chapter Six

CONCLUSION

APPENDICES

Page

Page

Page

LIST OF TABLES

Page

Page

LIST OF FIGURES

Page

CHAPTER ONE

INTRODUCTION

1. INTRODUCTION

Since mid 1940s, investigation on structural properties of the natural languages have been carried out by different researchers. Base studies in this field were done by Shannon, Zipf and Good. Shannon analyzed English language and defined entropy and prediction of printed English in his study in 1948 (Shannon, 1948). In his work, Zipf offered a rule which can be accepted for every statistical distribution in 1949 (Zipf, 1949 as cited in Teahan, 1998). Good has proposed a formula in 1953, which can be used for offering some assumptions of the missing parts of a word or sentence (Good, 1953). Because in that era computer technology wasn't developed yet, enough data couldn't be collected and processed. With the development of computer technology, much more data are collected, and depending on Shannon, Zipf and Good's studies, new technologies are developed to process the data obtained.

Natural language structure determination assists speech recognition, optical character recognition, text correction, and data encryption processes. The structure determination process covers two main topics: Morphological Analysis and Statistical Analysis. Morphological analysis covers the investigation of word types (verb, noun, adjective, adverb, etc.), the plural form of the word, dividing the word into morphemes and understanding whether the morpheme is a root, suffix or prefix. Morphological analysis can be supported by statistical analysis to estimate the missing parts of a text.

Statistical analysis can be mainly divided into two parts: Letter Analysis and Word Analysis. Letter Analysis includes letter n-gram frequencies, relationship between letters such as letter positions according to each other, consonant and vowel letter placements, etc. Word analysis mainly includes the investigation of number of

letters in a word, the order of the letters in a word, word n-gram frequencies, word orders in a sentence.

When the frequencies of the words are listed in a descending order, and ranked starting from one, the multiplication of the frequency and the rank will be constant. This is called Zipf's Law, and it's a general law for different kinds of observations. To improve this approach for natural languages, Mandelbrot added some constants to Zipf's Law (Choi, 2000).

Smoothing techniques are used to make some estimation for the unseen n-grams by using the known n-grams, after the determination of the statistical properties. There are different methods of smoothing, which are *Add-one Smoothing*, *Good-Turing Estimation* and *Back-Off Technique*. In order to determine structural properties, and process smoothing techniques, a corpus, which is a special collection of textual material collected according to a certain set of criteria (Manning & Schütze, 2000) that will represent the natural language, should be created. The corpus can be used both for morphological and statistical analysis. A corpus is defined as "Balanced", and "Unbalanced". It can be balanced by taking samples of all different topics like Medicine, Law, Spoken Language, etc. which will make that corpus a representative of the language. But, in fact it is not possible to take samples from all topics. Instead of taking small pieces from different areas, a large unbalanced corpus will be better because it will consist of lots of words. When working on letter analysis, small sized corpora like 1.4 million letters are used (Dalkılıç, 2001), but for word analysis much bigger corpora are needed.

There are some general analysis that can be applied to a corpus like n-gram analysis, Number of Different Words (NODW) and Different Word Usage Ratio (DWUR) that can also give the general characteristics of the corpus. N-gram analysis is one of the common statistical methods carried on a corpus. In n-gram analysis, firstly n-gram frequencies are calculated by using the corpus. Each letter or word can be counted from a corpus to create the list of most frequently used words or letters. But also, words or letters can be counted by grouping two by two (consequent 2

words or letters) or three by three (consequent 3 words or letters). If it is counted by each word or letter, it is called monogram (1-gram), if it is counted two by two, it is called bigram (digram or 2-gram). This can go to n (n-gram).

By using n-grams, language model probabilities can be estimated and used in speech recognition systems (Nadas, 1984). N-gram analysis can be used in correcting words by detecting non-words. It can be mixed with pattern matching and strings that don't appear in a given word list can be determined. It is also useful for OCR (Optical Character Recognition) (Kukich, 1992), data compression and encryption.

The main goal of this study is to apply and compare smoothing techniques on contemporary Turkish. These techniques can be examined by the help of a corpus sampling the language. For this reason a corpus of Turkish will be created during this study.

Previous studies made on other languages and Turkish, corpora of different languages, and the general concepts of n-grams, Zipf's Law, Mandelbrot's Law, and Smoothing techniques were investigated and given in Chapter 2.

After analyzing the previous works, it was seen that a large corpus is needed to work on word analysis. In Chapter 3, the details of the created Turkish corpus (TurCo) which has 50,111,828 words and ~362MB, were given. Efficient algorithms were needed to find the word n-gram frequencies from such a big corpus. In chapter 3, the details of word n-gram analyzing algorithms and their comparison were given.

In Chapter 4, the details of the word n-grams were given. Besides n-gram analysis, the correspondency between a given language and the commonly accepted rules, such as Zipf's Law, which is one of the most common laws for different kinds of observations, should also have been investigated in Chapter 4.

As a corpus cannot be big enough to find all the word n-grams, smoothing techniques are used to estimate the unknown word n-grams. In Chapter 5, smoothing

techniques were compared for Turkish. And also, smoothing techniques were applied by creating a prototype program for Turkish spelling error detection and correction. Prototype program was compared with Microsoft Word XP by running the same examples, and it was determined that the prototype program had better results.

CHAPTER TWO

GENERAL CONCEPTS

2. GENERAL CONCEPTS

2.1. Corpus

A natural language has written and spoken parts. As all the language cannot be collected to make analysis, a corpus that is large enough to sample the language should be created. This corpus can be a sample of contemporary written language or just a collection of old books and old documents or telephone conversations to represent spoken language.

There are lots of corpora created for different languages. Some of them gave importance to the quality, and some to the quantity of the corpus (Church & Mercer, 1993). After assembling the corpus, different analysis can be carried out, such as morphological and statistical analysis. Morphological analysis includes the analysis like investigation of word types, division of the word into morphemes, etc (Jurafsky & Martin, 2000). Statistical analysis consists of word length distribution, different word usage statistics, n-gram analysis for both letters (Shannon, 1951) and words (Jurafsky & Martin, 2000) etc. By using the results of these analysis, the corpus can be used in cryptanalytical procedures, character recognition operations, and it is always important for spelling corrections (Church & Gale, 1991b).

Brown corpus, British National corpus, the Bank of English and English Gigaword corpus are some of the samples of the English corpora and their details are given in section 2.1.1. Czech National corpus, Croatian National corpus, PAROLE, French corpus, COSMAS are some of the corpora for different kinds of languages like Czech, Croatian, French, Belgian French, Catalan, Danish, Dutch, Finnish, German, Greek, Irish, Italian, Norwegian, Portuguese, and Swedish. Their details are

given in section 2.1.2. Koltuksuz, YTÜ, Dalkilic and METU corpus are some Turkish corpora as explained in detail in section 2.1.3.

2.1.1. English Corpora

Brown corpus: This corpus was first assembled in 1963-1964 at Brown University. In 1964, it had 1 million words with 61,805 different words and in a later edition in 1992, the new Brown corpus had 583 million words with 293,181 different words (Jurafsky & Martin, 2000).

British National Corpus (BNC): It has over 100 million words (100,106,008) of modern British English, 90% of it, is a written part including extracts from newspapers, journals, academic books, school and university essays, and 10% spoken part includes a large amount of unscripted informal conversation. This is a project of Oxford University Press also including some other members. It was completed in 1994 and it was released in February 1995 (WEB_1).

The Bank of English: It had 450 million words with over half million different words in January 2002. It has speech and writing. The written part contains books, newspapers, magazines, letters, etc. and the spoken part includes speech from BBC World Service radio broadcasts, and the American National Public Radio, meetings, conversations, etc. The data are either collected from electronic environment or from scanning some books. In 1991, COBUILD and The University of Birmingham created the corpus. The collection of text was started in 1980 (WEB_2).

English Gigaword: It is an English corpus having 1,756,504,000 words and 4,111,240 documents. It is a product of Linguistic Data Consortium. It includes data from Agence France Press English Service, Associated Press Worldstream English Service, The New York times Newwire Service and Xinhua News Agency English Service. It is sold for 2500$ (WEB_3).

2.1.2. Corpora of Other Languages

The Czech National Corpus (CNC): It has synchronous and diachronic parts. Some parts of the synchronous are: Database and dictionaries (Electronic databases and dictionaries), SYN2000 (Balanced representative of contemporary written Czech and contains about 100 million words), ORAL (Spoken Czech). Some parts of diachronic are: The bank of diachronic Czech (2,000,000 words of transcribed texts; 100,000 words of transliterated texts; 200,000 words of dialect texts) (WEB_4).

Croatian National Corpus: It has 30 million words and 9,156,446 tokens as of February 24th, 2003. It includes older and contemporary text. It is available through Croatian Academic Research Network (WEB_5).

PAROLE: It is a multilingual corpora of Belgian French, Catalan, Danish, Dutch, English, French, Finnish, German, Greek, Irish, Italian, Norwegian, Portuguese and Swedish. It has 20,000 entries per language. All the texts are younger than 1970 (WEB_6).

French corpus: It has 20,093,099 words from books (3,267,409 words from CD-ROM), newspapers (13,856,763 words from Le Monde newspaper), periodicals (942,963 words from HERMES and CNRS-Infos), etc. (2,025,964 words) (WEB_7).

COSMAS (Corpus Search Management Analysis System): It is a German corpus having 1,903,000,000 running words. "Running" means new words are added each day. Only 1181 million words are available to public because of copyright restrictions. It is a product of "Institut für Deutsche Sprache, Mannheim" (WEB_8).

2.1.3. Turkish Corpora

Koltuksuz Corpus: It is the one of the Turkish corpora that is used for letter statistics and to find out some of the characteristics of Turkish language. It has 6,095,457 characters and formed of 24 Novels and Stories of 22 different authors.

These Novels and Stories were typed into the computer by a data entry group (Koltuksuz, 1995).

YTÜ Corpus: This corpus was created for morphology based data compression study. It has 4,263,847 characters from 14 different documents: 3 Novels, 1 PhD Thesis, 1 Transcription, and 9 Articles (Diri, 2000).

Dalkilic Corpus: It was created for letter statistics and to define the characteristics of Turkish language like Koltuksuz corpus. It has 1,473,738 characters taken from Hurriyet newspaper web archive (01/01/1998 – 06/01/1998 mainpage and 01/01/1998 – 06/30/1998 authors) (Dalkılıç, 2001).

Dalkilic Corpus: It is the combination of all the previous Turkish corpora, Koltuksuz, YTÜ and Dalkilic corpora, with a size of 11,749,977 characters (Dalkılıç & Dalkılıç, 2001).

METU Turkish Corpus: It is a collection of 2 million words of post-1990 written Turkish samples having 7262 grammatical sentences (WEB_9).

TurCo: The corpus created during this study. It has a capacity of 362.449MB, and 50,111,828 words. TurCo is nearly the half size of the British National Corpus (WEB_1), more than the Croatian Natural Corpus (WEB_5) which is 60% of TurCo, but not as big as English Gigaword (WEB_3). The details are given in chapter 3.

There are also other corpora for Turkish like ~2.2M Words (Güngör, 1995).

2.2. N-Grams

Statistical analysis is an alternative way to morphological analysis in linguistic, and finding the N-grams is one of the essential part of the statistical analysis. In an N-gram model, previous N-1 letters or N-1 words are used to predict the Nth letters

or Nth word. This study is based on word N-grams, and after this point, whenever just *n-gram* is used, it means *word n-gram*.

In a natural language, the words used in spoken language are less than the words used in written text, and also because of differences of dialects, and for other reasons, the word structure of spoken language differs from the written language. This makes it hard to identify the number of words in a natural language (Jurafsky & Martin, 2000).

Morphological analysis is based on the structure of the language like the words generated from some roots, the position of the verb, noun, etc. in the sentence. For an agglutinative language like Turkish, morphological analysis (Oflazer et al., 1994) needs a complementary analysis, for applications such as OCR, speech recognition, etc., where the blanks can be filled by using some clues, like statistical analysis.

N-gram word analysis is based on the probability of the words. If each word can follow the other word with the same probability in a sentence, then all the words will have the same probability for all positions in a sentence and monograms may be enough to calculate bigrams, trigrams, etc. But, each word has a different probability in different positions, so this approach is never true for a natural language. A way of predicting an unknown word is to look at the previous words that come before it. The previous word (bigram), previous 2 words (trigram) or more (n-gram where n>3) can be used for prediction. To make this analysis, some sample text, which is called corpus and explained in section 2.1, is needed. Although, a corpus of size 11.5MB for Turkish (Dalkılıç, 2001) is good enough to generate letter n-grams, for a word n-gram analysis much bigger corpus like TurCo is a prerequisite.

First step of the n-grams is *monograms*. Word definition in TurCo is the sequence of characters of Turkish alphabet between two space characters. This word definition also gives the definition of monograms.

A natural language has some rules like in Turkish: "The sentence ends with a verb", which makes the probability of some words to be nearly zero in some positions in the sentence.

When estimating the next word in the sentence, the probabilities of word sequences (sentences) have to be examined. A sentence which is correct according to the grammar structure, may have a low probability of existence. Like the sentence "*Bir arabayı vurdum*" (I shot a car) is correct according to the grammar rules, but has a very low probability to be seen.

The probability of a letter in a sentence is given in Formula 2.1 (Shannon, 1951; Garrett, 2001).

$$\text{Probability} = \frac{\text{Number of "selected letter" from the text}}{\text{Total number of letters in the text}} \quad \textbf{(2.1)}$$

Like letters, the probability of a word in a sentence is given in Formula 2.2.

$$\text{Probability} = \frac{\text{Number of "selected word" from the text}}{\text{Total number of words in the text}} \quad \textbf{(2.2)}$$

The estimation of the probabilities used in Formula 2.1 and 2.2 are called Maximum Likehood Estimation (MLE). The general formula for MLE probability for a given n-gram is given in Formula 2.3 (Jurafsky & Martin, 2000).

$$P_i = \frac{C_i}{N}, \text{ i} = 1, 2, \ldots, \text{t} \quad \textbf{(2.3)}$$

Where C is the count, N is the total number of n-grams.

The results of MLE have some probabilities being 0. The results can be used for segmentation-based recognition of free-format handwriting text (pen-based input) where a character sequence is segmented first, then each segment is recognized individually (Senda & Yamada, 2001).

The bigram model approximates the probability of a word given all the previous words P ($W_n \mid W_1^{n-1}$) (P is the probability and W_1^{n-1} means string of word $W_1 \ldots W_{n-1}$). A bigram is also called as a first order Markov model (it looks one token into the past), a trigram is called a second order Markov model. The basic bigram model is like a simple Markov chain which has one state for each word. Markov chain is a weighted finite-state automation. In general it is given in Formula 2.4 which shows the general equation for n-gram approximation to the conditional probability of the next word in a sequence (Jurafsky & Martin, 2000).

$$P(W_n|W_1^{n-1}) \approx P(W_n|W_{n-N+1}^{n-1}) \qquad \textbf{(2.4)}$$

By using Formula 2.4, when the probabilities of all the previous words are given, by using only the probability of the previous N words, probability of a word (w_n) can be approximated (Jurafsky & Martin, 2000).

When "Bugün de her zamanki gibi gidiyorum" (I am going today as usual) is taken as an example, for the bigram model, instead of using P (gidiyorum | Bugün de her zamanki gibi), P (gidiyorum | gibi) can be used, which is called Markov assumption. The probability of the word "gidiyorum" can be calculated when it follows the word "gibi". For trigram model P (gidiyorum | zamanki gibi), is used which is the probability whenever "gidiyorum" comes after "zamanki gibi".

There are differences between Turkish and English especially in word ordering and word formation like the differences between Russian and English (Whittaker & Woodland, 2003). Turkish is an agglutinative language where words can be created by adding successive suffixes. For example, the word "Osmanlılaştıramayabileceklerimizden mişsiniz" has the root ("Osman") and suffixes: Osman-lı-laş-tır-ama-yabil-ecek-ler-imiz-den-miş-siniz. The translation to English is "(behaving) as if you were of those whom we might consider not converting into an Ottoman" (Oflazer et al., 1994). There would be 15 words in English when 1 word was translated. This may be an extreme case, but there are lots of words in Turkish that are represented more than 1 word in English.

2.3. *Zipf's Law*

George Zipf, a psycholinguist, stated in "*Human behavior and the principle of least effort*" book in 1949 that word frequencies and lots of other observations, follow hypergeometric laws (Zipf, 1949 as cited in Teahan, 1998 and Choi, 2000). After analyzing the word frequencies for the novel *Ulysses* manually, Zipf claimed his best known "Zipf's Law" (Quan Ha et al., 2002).

2.3.1. Zipf Approach

According to Zipf, when the probability list of the words is sorted in descending order and numbered from 1, 2, to n, which are called the *rank* (*r*), then the product of *rank* and the *probability* (*P(r)*) of the corresponding word will be *constant* (μ) for each word (Choi, 2000). This is called Zipf's Law and given in Formula 2.5.

$$P(r) = \frac{\mu}{r}, \text{r} = 1, 2, \ldots, \text{n} \qquad \textbf{(2.5)}$$

Using Formula 2.5 and the definition of P(r), Formula 2.6 can be generated.

$$P(r) = \frac{F(r)}{\text{Number of Words}} = \frac{\mu}{r} \Rightarrow F(r) = \frac{\mu * \text{Number of Words}}{r}, \text{r} = 1, 2, \ldots, \text{n} \qquad \textbf{(2.6)}$$

where *F(r)* is the frequency of the word which has rank r. As μ, μ **Number of Words (NOW)* will also give a constant.

Word frequency and rank distribution graph for Brown corpus, the most known corpus for English language, is given in Figure 2.1 (Witten & Bell, 1990). The straight line shows Zipf's Law, the other dotted points are the actual values. As seen from Figure 2.1, they nearly match.

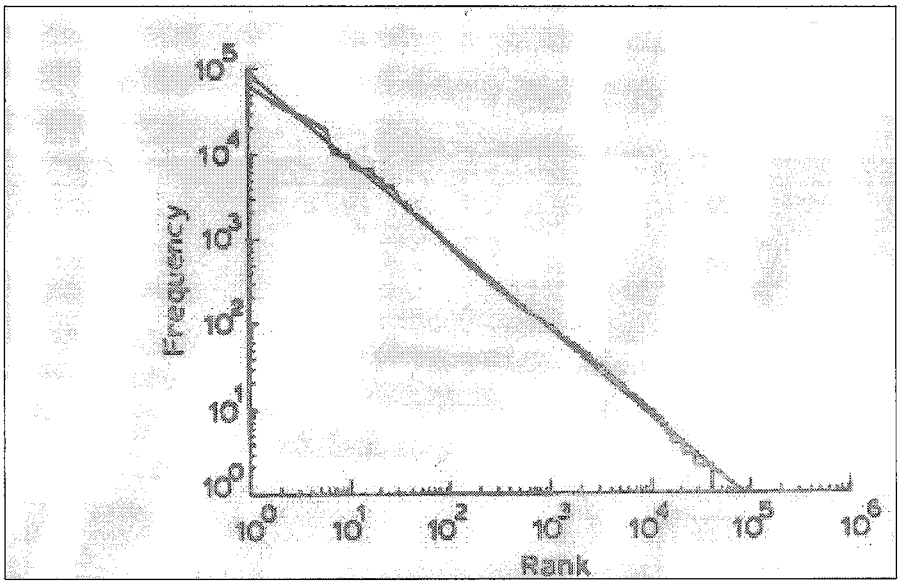


Figure 2.1 Word frequency data from the Brown Corpus and Zipf distribution

2.3.2. Mandelbrot Approach

Zipf's Law is too general to apply for n-gram distributions (Witten & Bell, 1990) (Choi, 2000). To improve the fitness of the Zipf's Law distribution, Mandelbrot added two new parameters, and Formula 2.5 changed to Formula 2.7 (Witten & Bell, 1990).

$$P(r) = \frac{\mu}{(c+r)^B} \quad , r = 1, 2, \ldots, n \qquad \textbf{(2.7)}$$

Where, r:rank; c and B are Mandelbrot constants; and $\mu=(c+r)^B*P(r)$ or $\mu=((c+r)^B*f(r))/NOW$.

Turkish and Korean are similarities according to morphological structures. They are both agglutinative languages. Because of this similarity and the information found (Choi, 2000), Korean is selected for comparison. For Korean, lots of postpositions can be added to the root, and each combination creates a different word, lots of words can be generated like Turkish (Choi, 2000).

The rank-probability graph with real values and Mandelbrot distribution plotted for Korean language is given in Figure 2.2 (Choi, 2000). The curve shows Mandelbrot distribution and the dotted points are the real values.

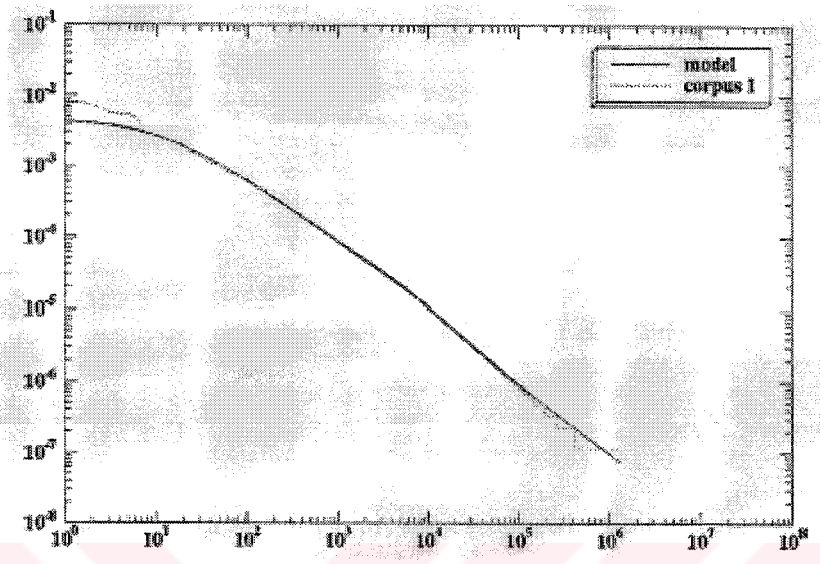


Figure 2.2 Rank probability graph for Korean Corpus (Choi, 2000)

As seen in Figure 2.2, the model graph, generated by using Formula 2.7, nearly matches the actual values, except for the starting and ending values where it has much more fluctuations.

2.3.2.1 Determination of Mandelbrot Constants

In statistics, after plotting the points, in order to find the formula of the distribution, which can help to predict the unknown values, regression analysis can be used. Regression analysis defines the relation between the plotted points and the formula that is used for the estimation of the points. R^2 value, which is called "*coefficient of determination*" and shows the closeness of the estimated values with the actual data, can be calculated.

R^2 is the proportion of the variance in dependent variable y attributable to the variance in independent variable x as shown in Formula 2.8. As R^2 gets larger, by introducing the independent variable x, the more total variation in the dependent variable y reduced. If there is a perfect match between two variables, then R^2=1. If R^2=0, then none of the variation in y is being explained by x. So, the closeness to 1,

the greater the linear association between x and y. For any regression data set, R^2 shows how close the raw regression data points are to the theoretical data (Rabbins & Daneman, 2002).

$$R^2 = \frac{\left(\sum(x-\bar{x})(y-\bar{y})\right)^2}{\sum(x-\bar{x})^2\sum(y-\bar{y})^2} \quad \textbf{(2.8)}$$

For the determination of c and B values, coefficient of determination value is used to find the best c and B values by assigning different c and B values.

2.4. Smoothing Techniques

For n-gram analysis by increasing n value, a bigger corpus is needed. Because the corpus is finite, some of the n-grams of the language won't appear in the list generated from the corpus, and it will end up with a sparse matrix. When it is shown with an example:

Example corpus="Bugün de her zamanki gibi gidiyorum"

Where the bigram probability is analyzed, and the results are given in Table 2.1.

$$P_{Bugün\ \text{de}} = \frac{1}{5},\ P_{de\ \text{her}} = \frac{1}{5},\ P_{her\ \text{zamanki}} = \frac{1}{5},\ P_{\text{zamanki gibi}} = \frac{1}{5},\ P_{gibi\ \text{gidiyorum}} = \frac{1}{5}$$

Table 2.1 Unsmoothed bigram values

	Bugün	de	her	zamanki	gibi	eve	gidiyorum
Bugün	0	1	0	0	0	0	0
de	0	0	1	0	0	0	0
her	0	0	0	1	0	0	0
zamanki	0	0	0	0	1	0	0
gibi	0	0	0	0	0	1	0
eve	0	0	0	0	0	0	1
gidiyorum	0	0	0	0	0	0	0

Some n-grams that can appear in the language, will have zero frequency and zero probability. For example, it is important in adaptive statistical text compression, because for the appearance of new words always a small part of the code space is reserved (Witten & Bell, 1991).

To assign some non-zero probabilities to these values is called smoothing. Smoothing is an approximation and like MLE it doesn't give exact results, but it has an advantage of having probabilities for the unseen n-grams.

Many smoothing techniques are defined such as; Cat-Cal (Church & Gale, 1989), Deleted Estimation (Church & Gale, 1991a), Parametric Empirical Bayes' method (Nadas, 1984 as referred in Katz, 1987), Add-One smoothing, Good-Turing Discounting, and Back-off technique (Jurafsky & Martin, 2000). Add-One smoothing, Good-Turing Discounting, and Back-off technique are most frequently used and the details of them are given respectively in 2.4.1, 2.4.2 and 2.4.3. The comparison of Add-One smoothing, Good-Turing and Back-off algorithms can be found in (Huang & Yu, 2001). Back-off algorithm can be improved by using some distributions optimized for the backing off method (Kneser & Ney, 1995). Kneser's optimized back-off algorithm can also be improved (Martin et al., 1999).

2.4.1. Add-One Smoothing

The simplest way of smoothing is Add-one smoothing which is also called as Laplace's Law. In this technique, 1 is added to all the counts, so the n-gram tokens having 0 count will be 1. As the count of each n-gram token is increased by one, for probability calculation, the number of different n-gram tokens is added to the denominator. It is the simplest smoothing, and doesn't give good results and not used frequently (Jurafsky & Martin, 2000).

Add-one smoothed probability is given in Formula 2.9 (Jurafsky & Martin, 2000).

$$P_i^* = \frac{C_i + 1}{N + V}, \text{i} = 1, 2, \ldots, \text{t} \qquad \textbf{(2.9)}$$

The difference between Formula 2.3 (MLE) and Formula 2.9 is the 1 in the numerator and V in the denominator. Where 1 show the add-one increase and V is the vocabulary size (number of different n-gram tokens).

When one is added to the count of the all unseen bigrams Table 2.1 changes to Table 2.2 (number of different n-gram tokens=49).

Table 2.2 Smoothed bigram values

	Bugün	de	her	zamanki	gibi	eve	gidiyorum
Bugün	1	2	1	1	1	1	1
de	1	1	2	1	1	1	1
her	1	1	1	2	1	1	1
zamanki	1	1	1	1	2	1	1
gibi	1	1	1	1	1	2	1
eve	1	1	1	1	1	1	2
gidiyorum	1	1	1	1	1	1	1

$P_{Bugün\ \text{de}} = \frac{2}{5+49} = \frac{2}{54}$, $P_{de\ \text{her}} = \frac{2}{54}$, $P_{her\ \text{zamanki}} = \frac{2}{54}$, $P_{\text{zamanki}\ \text{gibi}} = \frac{2}{54}$, $P_{gibi\ \text{gidiyorum}} = \frac{2}{54}$

and for the unseen bigram it is: $P_{unseen\ \text{bigram}} = \frac{1}{54}$.

2.4.2. Good Turing Discounting

This algorithm is based on the idea that the count of the things seen once can be used to estimate the count of the things never seen, and is more complex than Add-one smoothing.

By using Turing's original idea, Good introduced another probability estimates of items called as Good-Turing algorithm (Good, 1953) and then some more study was done on his algorithm (Church & Gale, 1991a). This algorithm assumes that the distribution of the probability estimates of items is binomial.

In Good-Turing Discounting, the smoothed count c* is calculated as in Formula 2.10 (Jurafsky & Martin, 2000).

$$c^* = (c+1)\frac{N_{c+1}}{N_c},\ i = 1, 2, \ldots, t \qquad \textbf{(2.10)}$$

Where N_c is the number of n-grams that occur c times.

For example, N_0 is the number of n-grams that never occurred, N_1 is the number of n-grams that occurred only once. The square of Number of different words ($NODW^2$) gives the theoretical maximum number of bigrams. This is theoretical because the combination of all the words cannot be used to form word pairs as the language has some rules (For example; a sentence starts with a subject). Number of the appeared bigrams is subtracted from this number to find never non-appeared bigrams. By using Good-Turing smoothing, curve-fitting functions and model combinations, a general language model for information retrieval can be generated (Song & Croft, 1999).

2.4.3. Back-off

Back-off technique, which was introduced by Katz (Katz, 1987), is much better than the others, because it doesn't directly make approximations. In this technique, if there is a zero frequency n-gram, then (n-1)-grams will be used. For example; If a particular trigram is not found, then bigrams; if bigrams is not found, then monograms are used to fill the sparse matrix as given in Formula 2.11. So, the n-gram model will be built by using (n-1), (n-2) until when monograms are reached in a recursive structure (Jurafsky & Martin, 2000).

$$\hat{P}(w_i|w_{i-2}w_{i-1}) = \begin{cases} \tilde{P}(w_i|w_{i-2}w_{i-1}), & \text{if } C(w_{i-2}w_{i-1}w_i) > 0 \\ \alpha_1\tilde{P}(w_i|w_{i-1}), & \text{if } C(w_{i-2}w_{i-1}w_i) = 0 \text{ and } C(w_{i-1}w_i) > 0 \\ \alpha_2\tilde{P}(w_i), & \text{otherwise} \end{cases} \quad \textbf{(2.11)}$$

The total probability of $\hat{P}(w_i|w_{i-2}w_{i-1})$ must be 1, so a new term $\tilde{P}(w_i|w_{i-2}w_{i-1})$ is introduced. Now, the total probability of $\tilde{P}(w_i|w_{i-2}w_{i-1})$, $\alpha_1\tilde{P}(w_i|w_{i-1})$, $\alpha_2\tilde{P}(w_i)$ will be 1, by using the constants α_1, α_2 (Katz, 1987).

2.5. Minimum Edit Distance

Minimum Edit Distance (MED) algorithm is used to find the similarity between two strings. MED algorithm gives a value which is the minimum number of

insertion, deletion and substitution operations needed to transform one string into another. In the simplest way, each of these operations has a cost of 1 (Jurafsky & Martin, 2000).

MED is computed using a table-driven method to solve large problems by combining solutions to small problems which is called dynamic programming. The algorithm of MED by using dynamic programming is given as (WEB_10):

```
EDITDISTANCE(A[1..m],B[1..n]):
 for  i←1 to m
    Edit[i,0]←i
 for j←1 to m
    Edit[0,j]←j
 for i←1 to m
    for j←1 to n
     if A[i]=B[j]
      Edit[i,j]←min{ Edit[i-1,j]+1, Edit[i,j-1]+1, Edit[i-1,j-1]}
     Else
      Edit[i,j]←min{ Edit[i-1,j]+1, Edit[i,j-1]+1, Edit[i-1,j-1]+1}
Return Edit[m,n]
```

Edit[i,j] is defined as the edit distance between the first i characters of the first string, and the first j characters of the second string. From an empty string to a string of j characters, j insertions are needed Edit[0,j]=j. From a string of i characters to an empty string, i deletions are needed (Edit[i,0]=i).

For each of insertion, deletion and substitution, in general:

Insertion: Edit[i,j]=Edit[i-1,j]+1

Deletion: Edit[i,j]=Edit[i,j-1]+1

Substitution: If the characters are the same Edit[i,j]=Edit[i-1,j-1], if not Edit[i,j]=Edit[i-1,j-1]+1.

In this function, the value of each cell is calculated with a function using the values of the adjacent cells, and getting the minimum. Edit[m,n] gives the edit distance between the two entire strings.

If the characters in the two strings are the same, the position of the character is shown as bold in Figure 2.3. Deletion operation is shown as horizontal arrow, insertion operation is shown as vertical arrow and substitution operation is shown as diagonal arrow. As a result the edit distance between ALGORITHM and ALTRUISTIC is six.

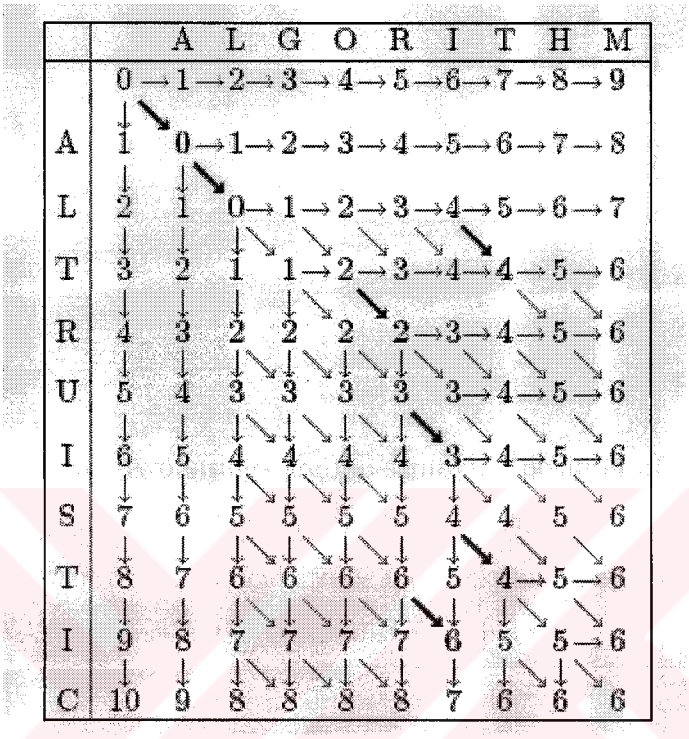


Figure 2.3 Edit Distance Table (WEB_10)

CHAPTER THREE

TURKISH CORPUS AND ITS CHARACTERISTICS

3. TURKISH CORPUS AND ITS CHARACTERISTICS

3.1. Turkish Corpus – TurCo

3.1.1. Sources of TurCo

TurCo is a contemporary Turkish text corpus, consists of text data collected from 10 different websites, Novels and Stories, and Turkey's National Program for the European Union. The name of the websites and the dates on which they were downloaded are given in Table 3.1.

Most parts (98.11%) of TurCo were collected from websites. 1.89% of the corpus is Novels and Stories, and the name of the author, the name of the novel or story and the size of each piece are given in Appendix D. 0.32% of the corpus is Turkey's National Program for the European Union and it was downloaded from http://www.abgs.gov.tr.

Table 3.1 Download dates of the websites

Part #	Property	Name	Download Date
1	Web Site	www.tbmm.gov.tr	November 2nd, 2001
2	Web Site	www.stargazete.com.tr	November 10th, 2001
3	Web Site	www.hurriyet.com.tr	October 29th, 2001
4	Text	Novels and Stories	Various
5	Web Site	www.die.gov.tr	November 2nd, 2001
6	Web Site	www.arabul.com	October 29th, 2001
7	Web Site	www.pcmagazine.com.tr	October 29th, 2001
8	Web Site	www.bilimteknoloji.com.tr	October 29th, 2001
9	Text	Turkey's National Program for EU	March 19th, 2001
10	Web Site	www.lazland.com	November 9th, 2001
11	Web Site	www.yeniasir.com.tr	October 29th, 2001
12	Web Site	www.pankitap.com	October 29th, 2001

www.abgs.gov.tr TurCo 'dan neden cikarildi

3.1.2. Processing Collected Data and Corpus Organization

File format of all pages taken from WEB was HTML, so these files should have been transferred to plain text format by removing HTML tags. To remove HTML tags, each file was opened by using Internet Explorer and saved as text file, but most of the websites include hundreds of small web pages, and to process an individual page is easier than processing hundreds of pages.

For this reason, after downloading, by using the *copy* command in DOS, pages in each site have been combined into separate files each having ~30 MB capacity, and than these ~30 MB capacity HTML files were converted to text files. 30 MB files were used because after several experiments, it was found out that Internet Explorer 6.0 crashes with files greater than 30 MB. During these processes, a computer having Intel Pentium III processor, 512 MB RAM, and Windows XP operating system was used.

The next step after creating the text files is filtration. The punctuation marks and special characters including non-Turkish characters were removed from these files, and new files had only 29 Turkish characters[1] (21 consonants and 8 vowels) and space character. Each *word* can be identified by the string of Turkish characters between two spaces. As the words and word groups would be counted from the corpus, in order to process easily, each letter was changed to lower case. Processing was first carried out on small examples to see the accuracy of the algorithm. Some errors had been corrected while working with small examples, and then the program was run for the real data. This filtering process is given in the following pseudo code:

```
  /* Gets the corpus as the input and returns the filtered corpus as
output */
  function FILTER (corpus) returns fcorpus
     n←LENGTH (corpus)
     c←LOWCASE (corpus)
     alphabet←"abcçdefgğhıijklmnoöprsştuüvyz "

     /* Filter the corpus. At the end it will contain the
characters in Turkish alphabet */
```

[1] Appendix E

```
    for i from 0 to n do
          if c[i] in alphabet then fcorpus←fcorpus+c[i]
    end do
end function
```

In order to obtain a large scale corpus, balancing can be neglected. Thus, many words from many different sources can be included into the corpus. It may not be possible to properly balance the corpus (Church & Mercer, 1993). "Only a large corpus of natural language enables us to identify recurring patterns in the language and to observe collocational and lexical restrictions accurately..." from (Hanks, 1990, p.36) as referred in (Church & Mercer, 1993, p.15). TurCo has been formed from 12 different sources having different properties and sizes and has a total capacity of more than 362 MB as given in Table 3.2.

In Table 3.2, NOW (token) shows the total "Number of Words" in each file As a total, TurCo includes 50,111,828 words.

Table 3.2 NOW, corpora file size and distribution %

Part #	Name	NOW	Corpora Files' Sizes[2] (MB)	Distribution (%)
1	www.tbmm.gov.tr	23,396,817	170.747	46.69
2	www.stargazete.com.tr	9,746,093	69.103	19.45
3	www.hurriyet.com.tr	9,415,716	69.140	18.79
4	Turkish Novels and Stories	4,668,306	33.571	1.89
5	www.die.gov.tr	948,116	6.387	9.32
6	www.arabul.com	753,571	4.994	1.50
7	www.pcmagazine.com.tr	527,757	3.722	1.05
8	www.bilimteknoloji.com.tr	203,620	1.450	0.41
9	Turkey's National Program for EU	160,562	1.249	0.32
10	www.lazland.com	135,519	0.954	0.27
11	www.yeniasir.com.tr	96,857	0.707	0.19
12	www.pankitap.com	58,894	0.425	0.12
	TOTAL	50,111,828	362.449	100.00

The biggest part in the corpus is Part #1 with 23,396,817 words, which forms 46.69% of the corpus. This part includes official session reports of Turkish Parliament, and represents both written and spoken Turkish. The first three sites form 84.93% of the corpus.

[2] Includes only Turkish alphabet and space character

3.2. Algorithms for N-gram Analysis

Different algorithms were developed and applied onto the corpus in order to determine the word n-grams. They were classified as *without* and *with* databases and their performance were compared with each other.

3.2.1. Algorithms without Databases

Three algorithms; algorithm with a linked list structure (3.2.1.1), algorithm with an index structure (3.2.1.2), and algorithm using virtual corpus (3.2.1.3), were developed and tested under this case.

3.2.1.1 Algorithm with a Linked List Structure

In this algorithm (ALG-1), all the words are read one by one from the corpus file and loaded into the memory until the end of the file. A linked list structure is used to hold the values in the memory. Each node of the linked list has the word itself and a count of that word. As a starting point, the first word is loaded into the memory with a count value of 1. When the second word is read, it is compared with the first word in the memory. If they are the same, then the count is increased by 1, else the second word is added as the next node with a count value of 1.

Whenever a new word comes (which wasn't loaded to the memory yet), comparison of the new word with the words in the memory is needed, whether to increase one of the counts or add a new node, which is a time consuming job. For small files it works fine (For 42,208 different words it takes 122 seconds). But, the linked list structure gets bigger in the memory as the different number of the word size gets hundreds of thousands (for bigrams, handling 2,864,577 different words takes more than a day). The pseudo code is given as following:

```
/* Gets the filtered corpus and n in n-gram as the inputs and
returns the head of the sorted n-gram linked list structure */
  function NGRAM_COUNT_1 (fcorpus,n) returns head_ngram
   /* Create a node structure for ngram including token and count */
      struct ngram
            token
            count
            next
```

```
      end struct
      i←0
      token←"NULL"
      new head_ngram ["NULL",0,NULL]
      while not eof (fcorpus) do
            for i from 0 to n do
            token←token+read 1 word from fcorpus ( Read the next
word each time)
            end do
            found←0
            ngram=head_ngram
            while ngram<>NULL and found=0 do
                  if ngram.token=token
                  then
                        ngram.count←ngram.count+1
                        found←1
                  else ngram←ngram.next
                  end if
            end do //while do
            if found=0 then
                  previous_ngram←ngram
                  new ngram
                  ngram.next←NULL
                  previous_ngram.next←ngram
                  ngram← previous_ngram
                  ngram.token←token
                    ngram.count←1
                  ngram.next←NULL
            end if
      end do //while not do
   end function
```

The developed computer program was run on a Compaq ML 350 having 512 MB RAM, dual Intel Pentium III/500 CPU with the Windows 2000 Advanced Server Operating System. The selection and organization of the data files were completed after ~70 hours of computing time. During this computation, an unsigned 16-bit integer was assigned for word counts. After 60 hours of computation it was observed that count values of some words like "VE" were greater than 65,536, so that a 32-bit unsigned integer was used.

3.2.1.2 Algorithm with an Index Structure

To overcome the problems of ALG-1, a new algorithm (ALG-2) was created which uses indexing and partially loading the corpus to the memory. It has the same node structure, the word and the count value. Instead of putting all the words into the memory, it first loads the memory the words having the initial letter "a". Also, an index structure is created by using the second, third and fourth letters. For example a

word like "adam" (which means "man" in English) is selected, then "dam" will be used for indexing. First, the letters are converted to numbers to indicate the position in a triple array. Each element of the array points out to a linked list structure. These numbers are 4,0,15 for "dam" which shows the positions of the letters in Turkish alphabet starting with 0 and ending with 29. So, if a word "adam" comes then (4)(0)(15) is the position of the array. As there are 29 different letters, the array positions start with (0)(0)(0) (aaa) and ends with (28)(28)(28) (zzz). For each position, there will be a linked list structure. For example if "adama" is the first element of the linked list then "adamb" (if any) will be the second element.

This linked list structure is the same as the first algorithm with some advantages:

- Only a portion of the corpus is loaded into the memory (if each letter has the same probability then 1/29 of the corpus is loaded into the memory) that makes use of the memory more efficient.
- Because of indexing, the CPU is used more efficient.

The pseudo code of the algorithm is given as following:

```
/* Gets the filtered corpus as the input and writes the tokens and
the corresponding frequencies to a file */
  function NGRAM_COUNT_2mono (fcorpus)
      Create a matrix start[29,29,29]
      struct monogram
            word
            count
            next
      end struct

      alphabet←"abcçdefgğhıijklmnoöprsştuüvyz"
      for i from 0 to 28 do
       while not eof (fcorpus) do
         word←word+ read 1 word from fcorpus ( Read the next word
each time)
         found←0
            if word[0]=alphabet[i] then
                  index1←position of word[1] in alphabet
                  index2←position of word[2] in alphabet
                  index3←position of word[3] in alphabet
                        monogram← start[index1,index2,index3]
                        while monogram<>NULL and found=0 do
                              if monogram.word=word
                              then
                                    monogram.count←monogram.count+1
```

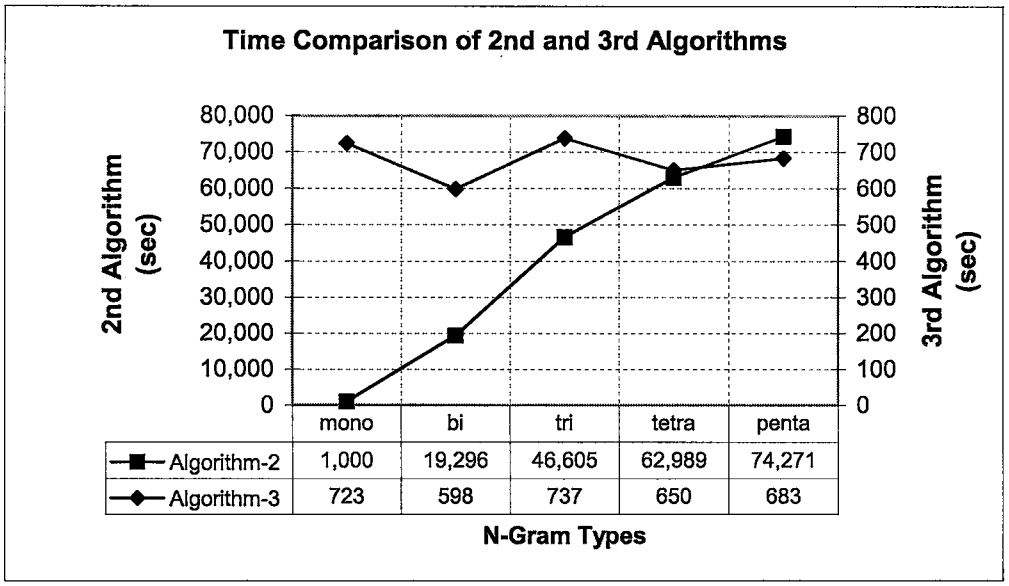


Figure 3.3 Time comparison of ALG-2 and ALG-3

3.3. Word Distribution by Sites and Categories

On the corpus, word-counting operations have also been carried out in order to determine the different words and their occurrence frequencies in the corpus. Number of Different Words (NODW), Site Different Word Ratio (SDWR) and Different Word Usage Ratio (DWUR) values are given in Table 3.8.

Table 3.8 NODW, SDWR and DWUR values by sites

Part #	Category	Web Sites	NODW	SDWR	DWUR
1	1	www.tbmm.gov.tr	342,544	49.88	1.46
2	2	www.stargazete.com.tr	255,024	37.13	2.62
3	2	www.hurriyet.com.tr	99,432	14.48	1.06
4	7	Turkish Novels and Stories	309,030	4.50	6.62
5	4	www.die.gov.tr	20,760	0.30	2.19
6	5	www.arabul.com	42,208	6.15	5.60
7	3	www.pcmagazine.com.tr	46,743	6.81	8.86
8	3	www.bilimteknoloji.com.tr	29,228	4.26	14.35
9	4	Ulusal Program	13,103	1.91	8.16
10	6	www.lazland.com	37,057	5.40	27.34
11	2	www.yeniasir.com.tr	25,294	3.68	26.11
12	3	www.pankitap.com	14,633	2.13	24.85
TurCo			686,804	100.00	1.37

NODW (token type) shows the "Number of Different Words" each part has. As each part of the corpus may have some common words, a total of 686,804 different words (TNDW: Total Number of Different Words) were obtained from the total corpus. Each part of the corpus may have some common words.

Turkish Novels and Stories (Part #4), which forms 1.89% of the corpus, is in second order with the different word count value 309,030 which is 90% of NODW of Part #1 (Table 3.8). Part #4 was collected from different novel and story authors (110 authors), so it is obvious that they use much more words than ordinary people. Although Turkish language is extremely rich, most of the words are not used by ordinary people.

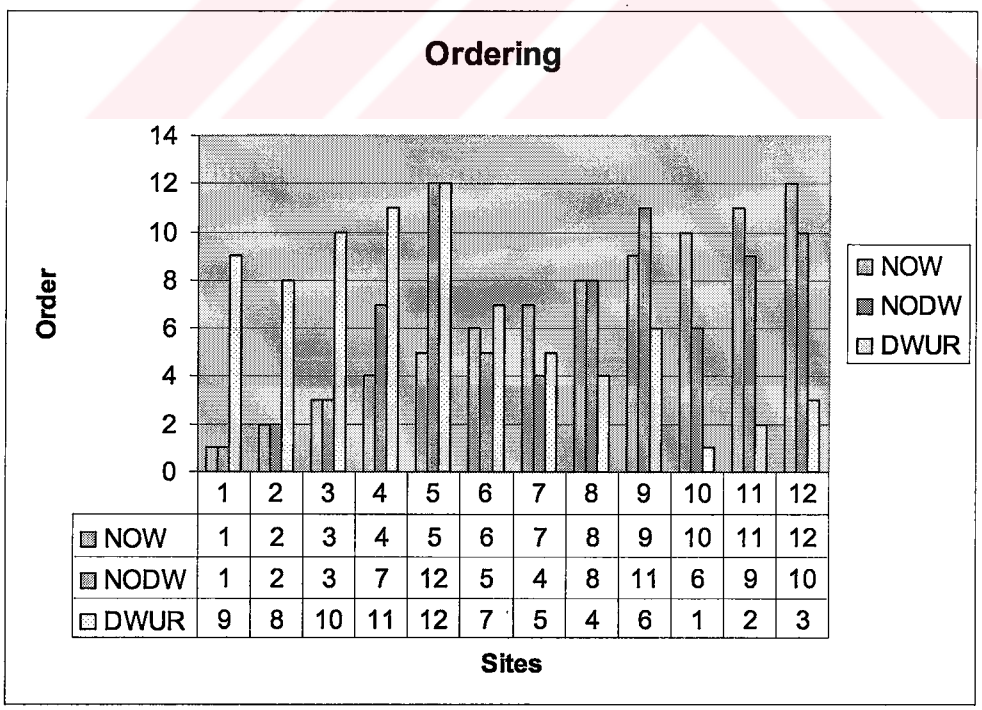
SDWR values, which represent the different word ratio of each part, have been calculated as in Formula 3.1.

$$SDWR = NODW / TNDW \qquad (3.1)$$

DWUR represents the ratio of different words used in each part and has been calculated as in Formula 3.2.

$$DWUR = NODW / NOW \qquad (3.2)$$

The sites used in the corpus were ordered by their NOW, NODW and DWUR values. These orderings are given in Figure 3.4.



	1	2	3	4	5	6	7	8	9	10	11	12
NOW	1	2	3	4	5	6	7	8	9	10	11	12
NODW	1	2	3	7	12	5	4	8	11	6	9	10
DWUR	9	8	10	11	12	7	5	4	6	1	2	3

Figure 3.4 The ordering of the parts of the Corpus

The sites have also been classified into 7 categories by their contents. These are:

- Category-1 : Written and Spoken Language
- Category-2 : Newspapers
- Category-3 : Magazines, Books
- Category-4 : Technical Documents
- Category-5 : Search Engine
- Category-6 : Entertainment
- Category-7 : Novels and Stories

Since the corpora sizes of categories differ from each other, such as Category-1 has 23,396,817 words and Category-4 has 160,562 words, meaningful results cannot be expected after the categorical comparisons. For this reason, categories were not compared with each other.

The biggest part (Part #1) in this corpus is in Category-1, and has 23,396,817 words which forms 46.69% of the corpus as shown in Table 3.7. This part was also classified alone, because of its representation of both written and spoken Turkish. At the same time, this part has the biggest NODW value (342,544), and SDWR with 49.88% (Table 3.8). Despite its greatness, this part has the fourth smallest DWUR ratio with the value of 1.46%, which means that same words are used much. Corpus generated for this part includes official session records of Turkish parliament.

The smallest part in this corpus is Part #12 that is in Category-3 and has a NOW value with 58.894 forming 0.12% of the corpus (Table 3.7). This part also has the third smallest NODW and SDWR values (13,103 and 2.13%), and the third biggest DWUR value of 24.85% (Table 3.8), which means that same words are used less. Part #12 is a bookstore site and includes book summaries, book chapters and some comments about books.

Part #10, which is in Category-6 and having NOW value of 135.519 forming 0.27% of the corpus, has the biggest DWUR value (27.34%). This part was classified alone as an entertainment part into Category-6. NODW and SDWR values of this

part are respectively 37,057 and 5.40%. This part contains anecdotes from Turkish culture.

Part #5, which is the fifth biggest part in this corpus having 948,116 NOW value forming 9.32% of the corpus, has the smallest DWUR value (0.22%). The NODW and SDWR values of this part are respectively 2076 and 0.30%. This part is a website including technical documents and in Category-4.

As seen in Figure 3.4, DWUR value does not depend on the NOW and NODW values. At the same time, this value also does not depend on the part category. The sites in the same category may have different DWUR values such as sites 2, 3 and 11 in the Category-2, and sites 7, 8 and 12 in the Category-3.

When the spoken and written languages have been taken together, the DWUR value shows great differences between categories, such as Category-1 and Category-6. DWUR value depends on the part contents, especially depends on the ratio of written language to the spoken language. When written language begins to take greater parts, the DWUR value decreases. It can be seen in Category-1, which is Turkish Parliament's part, which includes many written documents, and in Category-6, which consists of mainly spoken language, including slang words.

General different word usage ratio (DWUR) of the corpus was calculated as 1.37%. DWUR values show great differences depending on the part content from 1.06% to 27.34%. Content of the part has great effect on the DWUR value. The increase in the written part causes a decrease in the DWUR value. Because, the variety of the words in spoken language is more than the written language.

3.4. Number of Different Words in TurCo

3.4.1. Theoretically Expected Word Counts

For TurCo, NODW is 686,804 where TNDW is 50,111,828 that gives the number of word monograms. If it's expected that, all the words can come after the other, the number of different word bigrams can be maximum 686,804 * 686,804 = 471,699,734,416 which is much more than TurCo. This can be formulized as in Formula 3.3. All theoretical values are calculated by using Formula 3.3 and given in Table 3.9.

Bi = Mono * Mono
Tri = Mono * Mono * Mono
Tetra = Mono * Mono * Mono * Mono
Penta = Mono * Mono * Mono * Mono * Mono **(3.3)**

Because of the rules of the language, it is not easy to calculate bigrams. The actual size that is found for the number of different bigrams is 11,998,590 which is 0.0025% of the theoretical maximum number of different bigrams.

Table 3.9 Theoretical NODW values

Part #	NOW	Mono	Bi	Tri	Tetra	Penta
1	2.3397E+07	3.4254E+05	1.1734E+11	4.0193E+16	1.3768E+22	4.7161E+27
2	9.7461E+06	2.5502E+05	6.5037E+10	1.6586E+16	4.2298E+21	1.0787E+27
3	9.4157E+06	9.9432E+04	9.8867E+09	9.8306E+14	9.7747E+19	9.7192E+24
4	4.6683E+06	3.0903E+05	9.5500E+10	2.9512E+16	9.1202E+21	2.8184E+27
5	9.4812E+05	2.0760E+04	4.3098E+08	8.9471E+12	1.8574E+17	3.8560E+21
6	7.5357E+05	4.2208E+04	1.7815E+09	7.5194E+13	3.1738E+18	1.3396E+23
7	5.2776E+05	4.6743E+04	2.1849E+09	1.0213E+14	4.7738E+18	2.2314E+23
8	2.0362E+05	2.9228E+04	8.5428E+08	2.4969E+13	7.2979E+17	2.1330E+22
9	1.6056E+05	1.3103E+04	1.7169E+08	2.2496E+12	2.9477E+16	3.8624E+20
10	1.3552E+05	3.7057E+04	1.3732E+09	5.0887E+13	1.8857E+18	6.9880E+22
11	9.6857E+04	2.5294E+04	6.3979E+08	1.6183E+13	4.0933E+17	1.0354E+22
12	5.8894E+04	1.4633E+04	2.1412E+08	3.1333E+12	4.5849E+16	6.7091E+20
TurCo	5.0112E+07	6.8680E+05	4.7170E+11	3.2397E+17	2.2250E+23	1.5281E+29

3.4.2. Number of Word N-gram Values

Actual NOW and NODW values for monograms to pentagrams are given in Table 3.10. The first column shows part numbers, second column shows NOW and other columns show NODW values.

However, the NODW value of the TurCo for the monogram is 686,804, this value increases to 11,998,590 for bigram, 24,173,546 for trigram, 29,089,011 for tetragram and 30,943,233 for pentagram as given in Figure 3.5. These values are extremely lower than the expected theoretical maximum number of different n-grams.

Table 3.10 NOW and NODW values for n-gram analysis

Part #	NOW	NODW				
		Mono	Bi	Tri	Tetra	Penta
1	**23,396,817**	342,544	5,986,479	12,678,061	15,777,420	17,097,808
2	**9,746,093**	255,024	3,357,730	5,855,471	6,663,058	6,908,635
3	**9,415,716**	99,432	605,695	845,748	920,656	958,681
4	**4,668,306**	309,030	2,864,577	4,306,896	4,589,170	4,632,697
5	**948,116**	20,760	99,322	157,321	197,490	230,977
6	**753,571**	42,208	157,629	230,740	276,883	317,419
7	**527,757**	46,743	231,623	317,316	344,765	356,928
8	**203,620**	29,228	94,075	112,352	117,096	119,719
9	**160,562**	13,103	56,158	82,999	98,646	110,554
10	**135,519**	37,057	103,339	117,098	119,301	120,252
11	**96,857**	25,294	74,394	85,032	86,578	87,069
12	**58,894**	14,633	39,252	45,592	47,397	48,440
TurCo	**50,111,828**	686,804	11,998,590	24,173,546	29,089,011	30,943,233

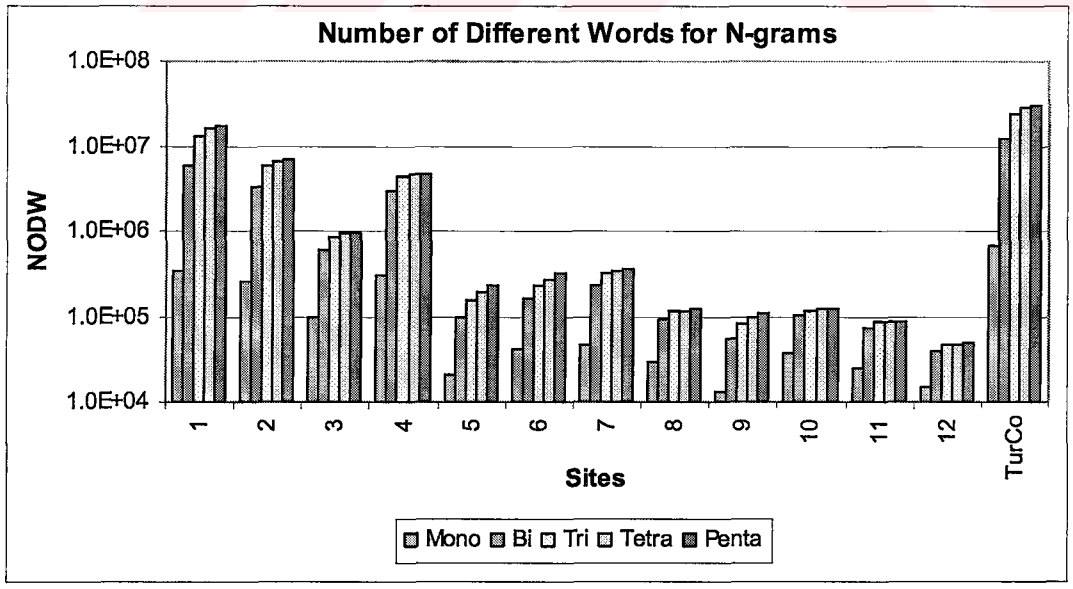


Figure 3.5 Number of different words for n-grams

The increase ratios of NODW values for each type of analysis when compared with monograms and with (n-1)-gram are calculated and given in Table 3.11, Figure 3.6 and 3.7. For the increase from monogram to bigram, the trend graphs for each element of the corpus can be found in Appendix C. Also, the formula of each graph is given.

Table 3.11 Increase ratios of NODW values

Part #	Based on previous N-gram (%)				Based on Monogram (%)			
	1 → 2	2 → 3	3 → 4	4 → 5	1 → 2	1 → 3	1 → 4	1 → 5
1	1,647.65	111.78	24.45	8.37	1,647.65	3,601.15	4,505.95	4,891.42
2	1,216.63	74.39	13.79	3.69	1,216.63	2,196.05	2,512.72	2,609.01
3	509.16	39.63	8.86	4.13	509.16	750.58	825.92	864.16
4	826.96	50.35	6.55	0.95	826.96	1,293.68	1,385.02	1,399.11
5	378.43	58.39	25.53	16.96	378.43	657.81	851.30	1,012.61
6	273.46	46.38	20.00	14.64	273.46	446.67	556.00	652.04
7	395.52	37.00	8.65	3.53	395.52	578.85	637.58	663.60
8	221.87	19.43	4.22	2.24	221.87	284.40	300.63	309.60
9	328.59	47.80	18.85	12.07	328.59	533.44	652.85	743.73
10	178.86	13.31	1.88	0.80	178.86	215.99	221.94	224.51
11	194.12	14.30	1.82	0.57	194.12	236.17	242.29	244.23
12	168.24	16.15	3.96	2.20	168.24	211.57	223.90	231.03
TurCo	1,647.02	101.47	20.33	6.37	1,647.02	3,419.72	4,135.42	4,405.39

As seen from Table 3.11 and Figure 3.6, the NODW value for the TurCo increases by 1,647.02% from monogram to bigram, 101.47% from bigram to trigram, 20.33% from trigram to tetragram and 6.37% from tetragram to pentagram. These values show that the increase ratio exponentially decreases from monograms to other n-grams as n increases.

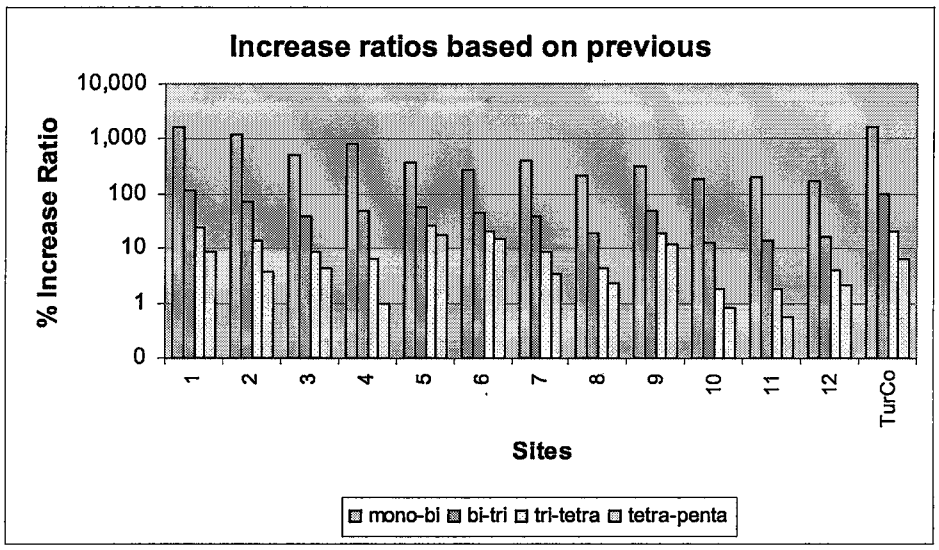


Figure 3.6 Increase ratios of NODW based on previous value

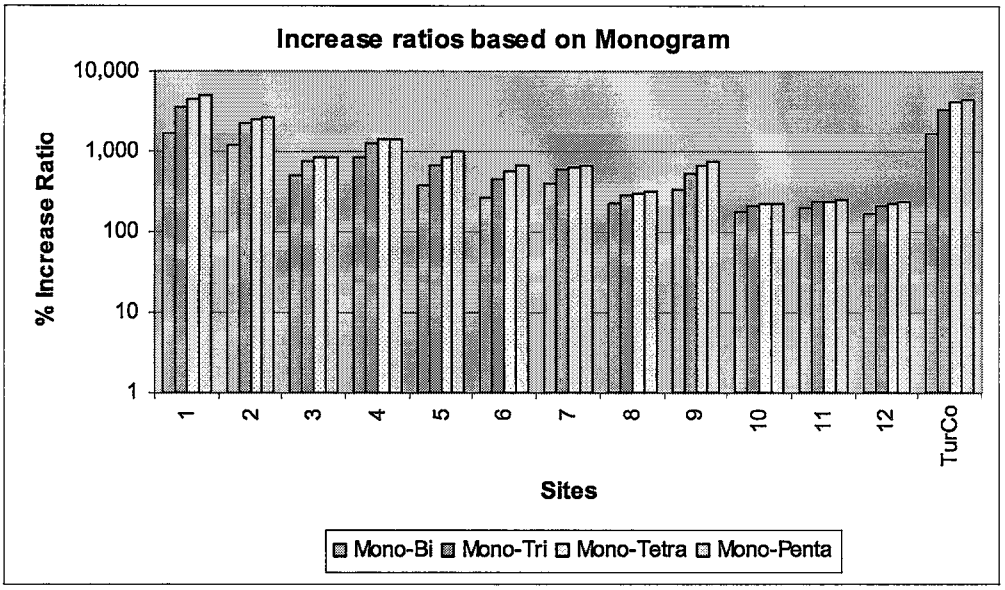


Figure 3.7 Increase ratios of NODW based on previous monograms

When the increase ratios based on monograms are considered, it can be seen from Figure 3.7 that the ratios increase logarithmically, with the increase ratios from monogram to bigram 1,647.02%, to trigram 3,419.72%, to tetragram 4,135.42% and to pentagram 4,405.39%.

3.5. Determination of the Most Frequently Used Words

After creating the monograms, the frequency distributions were also calculated. The monograms are sorted in an ascending order according to their frequency distribution and by using the top 100 of them, Figure 3.8 is created. Until the 15th word, the cumulative frequency distribution shows an exponential trend. After that, this trend tends to go on as linear. The first 15 words form 10.01% and the first 100 words form the 22.47% of the corpus.

```
                                    found←1
                                else monogram←monogram.next
                                end if
                            end do // while do
                            if found=0 then
                                previous_monogram←monogram
                                new monogram
                                monogram.next←NULL
                                previous_monogram.next←monogram
                                monogram← previous_ monogram
                                monogram.token←word
                                monogram.count←1
                                monogram.next←NULL
                            end if
                end if //if word[0]
            end do //while not do

    /* Write the sorted list to a file. Each element in the file will
    contain the token and its frequency */
                for j from 0 to 28 do
                    for k from 0 to 28 do
                        for l from 0 to 28 do
                            monogram ←start[j, k, l]
                            while monogram <>NULL do
                                write monogram.token and
monogram.count to file
                                monogram ← monogram.next
                            end do
                            start[j, k, l]←NULL
                        end do //for l
                    end do //for k
                end do //for j
            end do //for i
    end function
```

This algorithm works fine for monograms (only words where n=1). But, when word pairs (bigrams where n=2) are used again it starts to need lots of memory, because there will be lots of nodes to store the word pairs in the memory. Average word size is 6.23 characters for Turkish. For bigrams, this becomes 6.23 x 2 + 1 (space character) = 13.482. For n-grams this value is 6.23 x n + 1. So, more efficient method is needed for n-grams where n>1. The pseudo code is given as following:

```
    /* Gets the filtered corpus and n in n-gram as the inputs and
writes the tokens and the corresponding frequencies to a file */
    function NGRAM_COUNT_2ngram (fcorpus,n)
        Create a matrix start[29,29,29]
        struct ngram
                token
                count
                next
        end struct
```

```
      alphabet←"abcçdefgğhıijklmnoöprsştuüvyz"
      for i from 0 to 28 do
            for j from 0 to 28 do
                  while not eof (fcorpus) do
/* Create the token from the corpus. For example if n=2, it's a
bigram, then read 2 consecutive words and concatenate them to form
the token */
                  for k from 0 to n do
                        token←token+ read 1 word from fcorpus (
Read the next word each time)
                  end do
                  found←0
                  if token[0]=alphabet[i] and token[1]=alphabet[j]
then

/* Change a letter to the corresponding number by using alphabet.
For example a is 0, b is 1,…, z is 28 */
                        index1←position of token[2] in alphabet
                        index2←position of token[3] in alphabet
                        index3←position of token[4] in alphabet
                        ngram← start[index1,index2,index3]

/* If the searched ngram is already in the linked list then just
increase the count of that node by 1 or find the end of the linked
list to add a new node */
                        while ngram<>NULL and found=0 do
                              if ngram.token=token
                              then
                                    ngram.count←ngram.count+1
                                    found←1
                              else ngram←ngram.next
                              end if
                        end do

/* Create a node and add it to the linked list structure */
                        if found=0 then
                              previous_ngram←ngram
                              new ngram
                              ngram.next←NULL
                              previous_ngram.next←ngram
                              ngram← previous_ngram
                              ngram.token←word
                              ngram.count←1
                              ngram.next←NULL
                        end if
                  end if //if token[0]
                 end do //while not do

/* Write the sorted list to a file. Each element in the file will
contain the token and its frequency */
                  for k from 0 to 28 do
                        for l from 0 to 28 do
                              for m from 0 to 28 do
                                    ngram ←start[k, l, m]
                                    while ngram<>NULL do
                                          write ngram.token and
ngram.count to file
```

```
                                        ngram ← ngram.next
                                    end do //while do
                                    start[k, l,m]←NULL
                                end do //while m
                            end do //while l
                        end do //while k

            end do //for j
        end do //for i
    end function
```

To make the algorithm more efficient, for n-grams greater then monograms, 1/841 of the word pairs are loaded into the memory by looking at the first two letters. So, first word pairs starting with "aa" are loaded by counting the number of occurrences, then "ab" and it goes on till "zz". There is no change for the indexing method.

This algorithm was coded using a computer running Windows XP, having 512MB RAM, and a Pentium4 2.2GHz CPU, and for bigrams with a corpus size 4.994 MB the program runs for 135 seconds where the first method needs 1137 seconds. But, for large-scale corpus, this enhancement is not enough. In Table 3.3 and 3.4, time needed to run this algorithm for each part of the corpus is given in seconds, the speed of the algorithm in KB/sec and also log10 values of this speed is given for each part of the corpus. These values were calculated for monograms, bigrams, trigrams, tetragrams and pentagrams. In Figure 3.1, the log10 values of the speed are also given.

Table 3.3 Time values of the second algorithm for monogram and bigram

Part #	Size	Monogram			Bigram		
	MB	time(sec)	KB/sec	log10(e)	time(sec)	KB/sec	log10(e)
1	170.747	389	449.47	2.65	8147	21.46	1.33
2	69.103	138	512.76	2.71	1960	36.10	1.56
3	69.140	100	707.99	2.85	1159	61.09	1.79
4	33.571	91	377.77	2.58	1114	30.86	1.49
5	6.387	14	467.16	2.67	128	51.10	1.71
6	4.994	14	365.28	2.56	135	37.88	1.58
7	3.722	14	272.24	2.43	109	34.97	1.54
8	1.450	8	185.60	2.27	57	26.05	1.42
9	1.249	6	213.16	2.33	45	28.42	1.45
10	0.954	8	122.11	2.09	49	19.94	1.30
11	0.707	7	103.42	2.01	42	17.24	1.24
12	0.425	5	87.04	1.94	30	14.51	1.16
TurCo	362.449	1000	371.15	2.57	19296	19.23	1.28

Table 3.4 Time values of the second algorithm for tri, tetra and pentagram

Part #	Trigram			Tetragram			Pentagram		
	time(sec)	KB/sec	log10(e)	time(sec)	KB/sec	log10(e)	time(sec)	KB/sec	log10(e)
1	20,359	8.59	0.93	28,293	6.18	0.79	33,787	5.17	0.71
2	3,197	22.13	1.35	3,941	17.96	1.25	4,336	16.32	1.21
3	1,183	59.85	1.78	1,367	51.79	1.71	1,307	54.17	1.73
4	1,769	19.43	1.29	1,963	17.51	1.24	2,116	16.25	1.21
5	139	47.05	1.67	148	44.19	1.65	155	42.20	1.63
6	173	29.56	1.47	234	21.85	1.34	294	17.39	1.24
7	121	31.50	1.5	127	30.01	1.48	130	29.32	1.47
8	59	25.17	1.4	61	24.34	1.39	62	23.95	1.38
9	51	25.08	1.4	53	24.13	1.38	56	22.84	1.36
10	51	19.15	1.28	52	18.79	1.27	52	18.79	1.27
11	44	16.45	1.22	44	16.45	1.22	45	16.09	1.21
12	31	14.04	1.15	32	13.60	1.13	32	13.60	1.13
TurCo	46,605	7.96	0.9	62,989	5.89	0.77	74,271	5.00	0.70

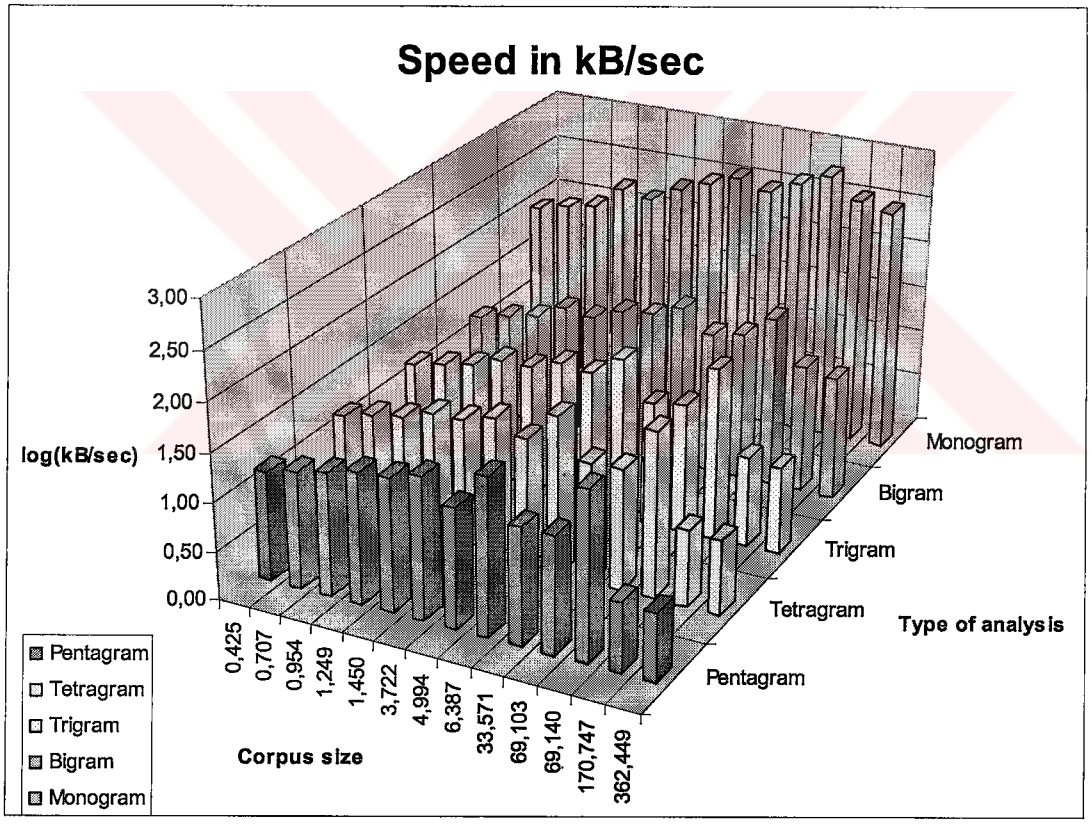


Figure 3.1 Log10 graph for Algorithm 2

3.2.1.3 Algorithm using Virtual Corpus

This algorithm (ALG-3) is the fastest of all. For this algorithm, a virtual corpus consists of token pointers to the original corpus, is created (Chunyu & Yorick, 1998), and is used for the counting operation of n-gram tokens.

In the first step, the entire corpus is put into a character array in the memory without making any changes (This is the original corpus). On this step, the algorithm requires large amount of memory.

In the second step, another array is used to create the index (Virtual corpus). On this array, each element is a pointer to the first array that shows the position of the first letter of each word. Instead of using an array, linked list structure can also be preferred.

For example, the first array which includes the sentence "buraya da oraya da gitmem" ("I don't go neither here nor there") can be thought as a small corpus. The elements of the second array will be: 0, 7, 10, 16, and 19 that show the starting point of each word in the corpus. Then, the second array will be sorted. After sorting alphabetically in an ascending order by using Quicksort algorithm, the array elements will be 0, 7, 16, 19, and 10. For n-grams, the virtual corpus is sorted by each n-gram token. Whether for monogram or more, until this point, the algorithm is the same for all n-gram structures.

The last step is to count the n-gram tokens by using the virtual corpus. The first and second elements of the virtual corpus are taken and the corresponding n-gram tokens are searched from the first array. Tokens found from the real corpus are compared to see whether they are same or not. If the first token is not the same as the second one, this means the token appears only once in the corpus and its count is "1". If the two tokens are the same, its count is increased by "1" and the second token is compared with the third element. It goes on like that till the end of the array. After finishing the counting operation for all the corpus, word counts are written to a file

directly. For the example given above, the count results will be: buraya 1, da 2, gitmem 1, oraya 1. The pseudo code is given as following:

```
/* Gets two element of the real corpus, compares them, if the first
one is smaller returns 1 else returns 0 */
  function firstoneissmaller (first,second) returns result
      if length(corpus[first])<length(corpus[second]) then
            size←length(corpus[first])
      else
            size←length(corpus[second])
      end if
      for i from 1 to size do
            if corpus[first]<corpus[second] then
                  return 1
            else if (corpus[first]=corpus[second]) then
                  first←first+1
                  second←second+1
            else
                  return 0
            end if
      end do //for i
      return 1
  end function

/* Gets the filtered corpus and n in n-gram as the inputs and writes
the tokens and the corresponding frequencies to a file */
  function WORD_COUNT_3 (fcorpus,n)
      corpus←fcorpus
      iv←0 //Index of virtual corpus
      ic←0 //Index of real corpus
      vcorpus[iv]←0

      /* Create Virtual Corpus using Real Corpus */
      while not eo(corpus) do
      while corpus[ic]<>" " do
            ic←ic+1
      end do
      iv←iv+1
      vcorpus[iv]←ic
      end do

/* Sort the elements of the Virtual Corpus using the Real Corpus. If
n is 1 just use words, if n is 2 use 2 consecutive words to form the
token, … Do not make any changes to the Real Corpus */
      vcorpus←quicksort (corpus[vcorpus],n) using firstoneissmaller
function
      i←0
      count←1

/* Compare the elements of the corpus by using Virtual Corpus. As
the Virtual Corpus gives the sorted list, same tokens come one after
the other. */
      while not eo(vcorpus) do
      while corpus[vcorpus[i]],n= corpus[vcorpus[i+1]],n do
                  count←count+1
            i←i+1
```

```
    end do
          write corpus[vcorpus[i]],n and count to file
    count←1
    end do
end function
```

This algorithm was coded using a computer running Windows XP, having 512MB RAM, and a Pentium4 2.2GHz CPU.

For bigrams with a corpus size 4.94 MB this program needs only 6 seconds where the corresponding value is 135 seconds in the second algorithm. The values smaller than 1 are denoted with "*" in Table 3.5 and 3.6. In Table 3.5, time, speed and log10 values of the speed for monograms and bigrams, and in Table 3.6 the corresponding values of trigrams, tetragrams and pentagrams are given, and speed values of all n-grams are given in Figure 3.2.

Table 3.5 Time values of the third algorithm for monogram and bigram

Part #	Size	Monogram			Bigram		
	MB	time(sec)	KB/sec	log10(e)	time(sec)	KB/sec	log10(e)
1	170.747	163	1072.67	3.03	180	971.36	2.99
2	69.103	59	1199.35	3.08	70	1010.88	3.00
3	69.140	57	1242.09	3.09	60	1179.99	3.07
4	33.571	25	1375.07	3.14	33	1041.72	3.02
5	6.387	4	1635.07	3.21	4	1635.07	3.21
6	4.994	3	1704.62	3.23	6	852.31	2.93
7	3.722	2	1905.66	3.28	2	1905.66	3.28
8	1.450	*	*	*	1	1484.80	3.17
9	1.249	*	*	*	*	*	*
10	0.954	*	*	*	*	*	*
11	0.707	*	*	*	*	*	*
12	0.425	*	*	*	*	*	*
TurCo	362.449	723	513.34	2.71	598	620.65	2.79

Table 3.6 Time values of the third algorithm for tri, tetra and pentagram

Part #	Trigram			Tetragram			Pentagram		
	time(sec)	KB/sec	log10(e)	time(sec)	KB/sec	log10(e)	time(sec)	KB/sec	log10(e)
1	196	892.07	2.95	215	813.23	2.91	215	813.23	2.91
2	86	822.81	2.92	76	931.07	2.97	79	895.72	2.95
3	62	1141.93	3.06	68	1041.17	3.02	67	1056.71	3.02
4	34	1011.08	3.00	35	982.19	2.99	39	881.45	2.95
5	4	1635.07	3.21	5	1308.06	3.12	5	1308.06	3.12
6	5	1022.77	3.01	4	1278.46	3.11	4	1278.46	3.11
7	2	1905.66	3.28	3	1270.44	3.10	5	762.27	2.88
8	1	1484.80	3.17	1	1484.80	3.17	1	1484.80	3.17
9	*	*	*	*	*	*	1	1278.98	3.11
10	*	*	*	*	*	*	*	*	*
11	*	*	*	*	*	*	*	*	*
12	*	*	*	*	*	*	*	*	*
TurCo	737.00	503.59	2.70	650	571.00	2.76	683	543.41	2.74

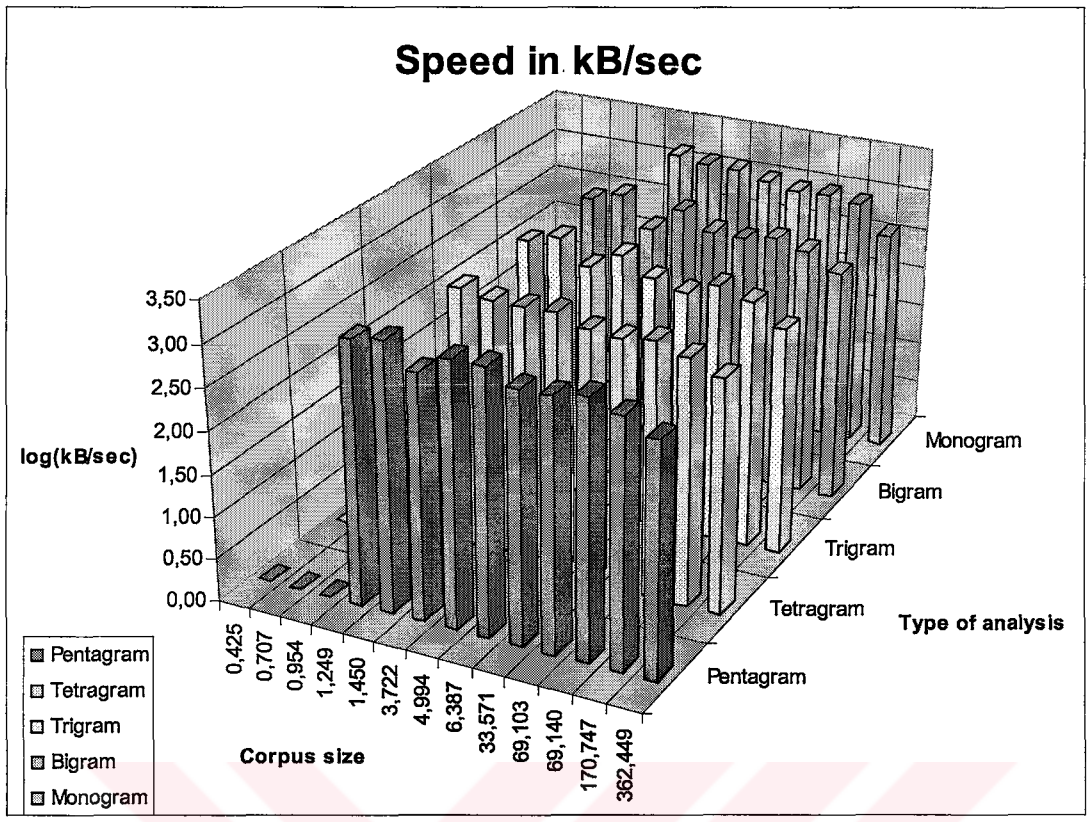


Figure 3.2 Speed in Log10 graph for Algorithm 3

3.2.2. Algorithms with Databases

For the implementation of the first algorithm with databases, MySQL was used as database, and PHP was used as programming language under Linux environment. In this algorithm, the entire corpus was loaded into a table word by word. The table has two fields: word and the count. First word is loaded from the corpus and put into the database table with a count 1. Then, the next word is taken and compared with the first word in the table. If they are same, the count is increased by one, else it is added as a new word with a count 1. With a Pentium 3-800MHz computer having 256MB RAM, with the corpus size of ~362MB, it took ~95 hours for trigrams and ~65 hours for bigrams which is nearly as slow as the first algorithm (using the linked list structure).

The second algorithm with databases was implemented by using MySQL as database, and C as programming language. By using threads in C programming language under Linux environment, the n-grams were counted in parallel. Different

number of threads was tried. This method was found to be faster than the method using MySQL and PHP without threads, but it is slower than the second algorithm denoted in section 3.2.1.2 (indexing and partially loading the corpus). It took days to count the n-grams for bigrams and trigrams with a Pentium 3-800MHz computer having 256MB RAM.

For both algorithms with databases, indexes were created, but it was seen that the size of the index files are as big as data files, so the indexing was found inefficient.

3.2.3. Comparison of the Algorithms

Because of its ineffectiveness, the n-gram analysis could not been carried out with the first algorithm. Only ALG-2 and ALG-3 were compared, since all the other algorithms were too slow that takes even days to count n-grams.

There are some advantages of the ALG-3:

- Speed: It is fast,
- Storage Space: The count values of n-gram tokens which needs too much space for large scale corpora, don't need to be stored,
- Virtual Corpus: As the index, which is composed of integer numbers, is sorted, it is faster than changing the places of all n-gram tokens (needs string operations) in the memory,
- CPU Usage: The second algorithm uses 99% of the memory nearly all the time. The average CPU usage of the third algorithm is only 13%.

The only disadvantage of the ALG-3 is the usage of the memory. ALG-2 uses a memory space between 8 and 11MB for ~362MB corpus, but the ALG-3 needs more than 400MB of memory. For a large-scale corpus, a method like in the ALG-2 can be adapted to the ALG-3 if enough memory space doesn't exist. The corpus can be loaded to the memory partially (for example first the n-gram tokens starting with letter A).

Time values obtained for sites in the corpus for different n-grams are given in Table 3.7. When speed of these algorithms compared with each other for the total corpus, it can be seen that second algorithm requires 1.38 times more time than third algorithm for monograms (1,000 seconds to 723 seconds), 32.27 for bi- (19,296 sec to 598 sec), 63.23 for tri- (46,605 sec to 737 sec), 96.91 for tetra- (62,989 sec to 650 sec) and 108.74 for penta-grams (74,271 sec to 683 sec). This can also be seen from Figure 3.3.

Table 3.7 N-gram analysis times in seconds for sites in the corpus

Part #	Mono		Bi		Tri		Tetra		Penta	
	Alg-2	Alg-3	Alg-2	Alg-3	Alg-2	Alg-3	Alg-2	Alg-3	Alg-2	Alg-3
1	389	163	8,147	180	20,359	196	28,293	215	33,787	215
2	100	57	1,159	60	1,183	62	1,367	68	1,307	67
3	138	59	1,960	70	3,197	86	3,941	76	4,336	79
4	91	25	1,114	33	1,769	34	1,963	35	2,116	39
5	14	4	128	4	139	4	148	5	155	5
6	14	3	135	6	173	5	234	4	294	4
7	14	2	109	2	121	2	127	3	130	5
8	8	*[3]	57	1	59	1	61	1	62	1
9	6	*	45	*	51	*	53	*	56	1
10	8	*	49	*	51	*	52	*	52	*
11	7	*	42	*	44	*	44	*	45	*
12	5	*	30	*	31	*	32	*	32	*
TOTAL	**1,000**	**723**	**19,296**	**598**	**46,605**	**737**	**62,989**	**650**	**74,271**	**683**

The most important thing in the ALG-3 is the time requirement for the analysis. Especially for small-scale corpora, time value stays below 1 second, so they cannot be shown on the table.

When "n" in n-gram analysis increases, the ALG-3 gets more advantageous. When n=5 it is even ~108 times faster than the ALG-2. This gives the advantage of running the ALG-3 even for n>5. The written program was designed to handle $0<n<11$. For English the analysis was generally done for $1 \leq n \leq 3$ (O'Boyle et al., 1996) (Witten & Bell, 1990).

[3] *:Time value is below 1 second.

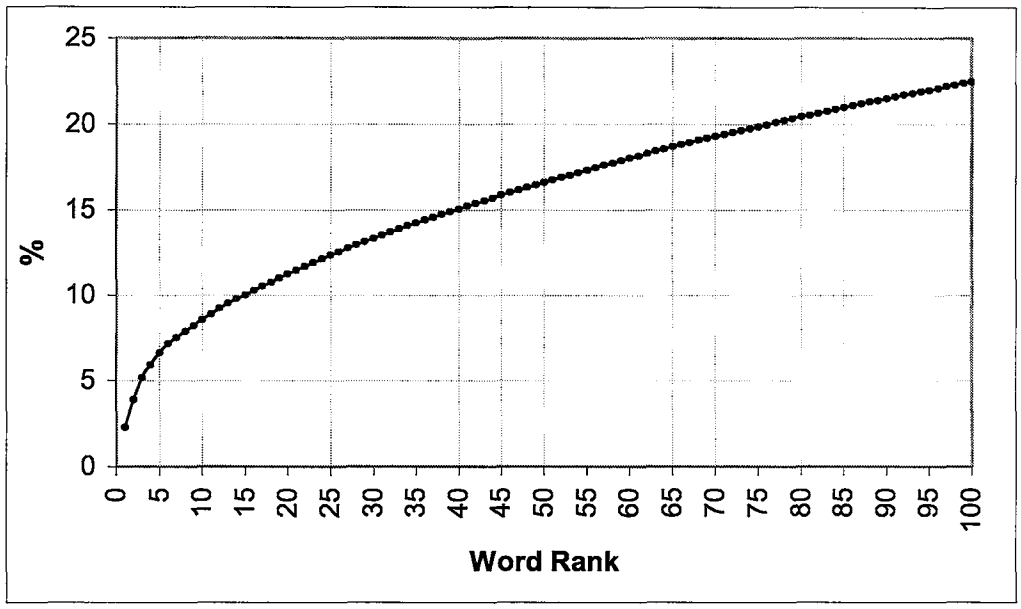


Figure 3.8 Cumulative frequency distribution of the top 100 words (Most to least used)

The first 7 words and their distributions inside the categories are given in Table 3.12. Since the eighth most commonly used word in the corpus was "Türkiye", which is a noun and it is specific to this corpus, only the first 7 words will be examined in this section.

The seven most frequently used words in TurCo are not "NOUN", instead they are adverbs, conjunctions, adjectives and prepositions. Words that are in Turkish top 7 are in English top 53 as shown in Table 3.13 (Garett, 2001). Depending on these results, it is said to be, "*Most frequently used words are not nouns in natural languages*".

Table 3.12 Distribution of the first 7 most frequently used words in TurCo

	1	2	3	4	5	6	7	8	9	10	11	12	TurCo
Word	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**
VE	2.92	1.50	1.47	3.07	4.08	5.70	1.90	2.15	4.05	2.20	2.13	2.60	2.27
BİR	1.56	1.47	1.26	1.95	1.61	5.67	1.88	2.00	3.08	1.64	1.96	2.48	1.60
BU	1.53	1.00	1.13	1.43	0.97	5.67	1.40	1.33	2.04	1.25	1.16	1.29	1.29
DA	0.68	0.95	0.79	0.91	0.81	5.66	1.24	0.88	2.01	1.19	1.06	0.94	0.72
DE	0.68	0.83	0.77	0.87	0.76	3.92	0.92	0.86	1.80	0.72	1.01	0.91	0.71
İÇİN	0.65	0.68	0.57	0.58	0.68	2.13	0.84	0.79	1.64	0.63	0.66	0.68	0.52
İLE	0.61	0.64	0.55	0.51	0.68	1.42	0.81	0.70	1.62	0.61	0.62	0.58	0.40
TOTAL	**8.63**	**7.07**	**6.54**	**9.33**	**9.58**	**30.19**	**9.00**	**8.70**	**16.24**	**8.25**	**8.60**	**9.47**	**7.52**

Table 3.13 Comparison of Turkish with English

Turkish	Rank	%	English	Rank	%
VE	1	2.27	AND	6	1.82
BİR	2	1.60	A, AN, ONE	5, 39, 41	1.95, 0.32, 0.32
BU	3	1.29	THIS	14	0.69
DA	4	0.72	TOO	53	0.19
DE	5	0.71	TOO	53	0.19
İÇİN	6	0.52	FOR	9	1.17
İLE	7	0.40	WITH	20	0.5

During the analysis on the corpus, words having 1 to 5 letters have been determined and first 10 are given in Table 3.14. These words form 15.96% of TurCo and they are listed in Table 3.14 are in top 100 most frequently used words. This shows that frequently used words have different lengths in Turkish language.

Table 3.14 Distribution of first 10, 1,2,3,4,5 letter words

1 Letter		2 Letters		3 Letters		4 Letters		5 Letters	
Word	**%**	**Word**	**%**	**Word**	**%**	**Word**	**%**	**Word**	**%**
A	0.35	VE	2.27	BİR	1.60	İÇİN	0.52	SAYIN	0.32
N	0.34	BU	1.29	İLE	0.40	OLAN	0.28	BÜYÜK	0.23
E	0.25	DA	0.72	ÇOK	0.26	DAHA	0.25	HABER	0.21
İ	0.23	DE	0.71	SON	0.20	YENİ	0.24	KADAR	0.20
O	0.22	EN	0.25	NİN	0.18	GİBİ	0.21	DÜNYA	0.19
R	0.17	NE	0.23	HER	0.18	GÖRE	0.16	SONRA	0.19
M	0.16	IN	0.22	AMA	0.17	SPOR	0.16	GENEL	0.16
L	0.16	İN	0.20	VAR	0.16	ÖZEL	0.15	KANUN	0.16
S	0.14	YA	0.15	GÜN	0.15	HAVA	0.14	SANAT	0.14
K	0.13	Kİ	0.15	İSE	0.14	VEYA	0.13	KABUL	0.14
Total	2.15		6.19		3.44		2.24		1.94
Grand Total: 15.96									

These results can be used to correct a paragraph when some words cannot be recognized such as in OCR and cryptanalysis operations. As the top 10, 1 to 5 letter words are given in Table 3.14, these results can be used for estimation of mistyped words. Suppose that the following paragraph is given:

"*Asya ve Rusya* krizlerinden *sonra* yaşanan depremler *ve* 11 *Eylül* krizine işaret *eden* Ecevit, "Hükümet *bu* durumda, *bu* aşamada görevi devraldı. *Üçlü* koalisyon hükümeti hiçbir şekilde klasik anlamda politikacılık *tavrı* içinde olmadı. *Çok* objektif biçimde, nesnel *bir* şekilde *her* eleştiriyi *göze* alarak *bir tutum* izledi. *Hiç* istifimizi bozmadan *büyük bir* kararlılıkla işimizi sürdürdük." *dedi.* "[4]

[4]Randomly selected from www.milliyet.com.tr economy section, 03.28.2002.

If each word is counted in the paragraph, and sorted according to their # of occurrences, Table 3.15 will be formed. Next step is to compare this table with Table 3.14.

Table 3.15 Word analysis results of the paragraph

2 Letters		3 Letters		4 Letters		5 Letters	
Word	**# of**	**Word**	**# of**	**Word**	**# of**	**Word**	**# of**
VE	2	BİR	3	ASYA	1	RUSYA	1
BU	2	HER	1	ÜÇLÜ	1	SONRA	1
		ÇOK	1	EDEN	1	EYLÜL	1
		HİÇ	1	GÖZE	1	TAVRI	1
				DEDİ	1	TUTUM	1
						BÜYÜK	1

As shown in Table 3.15, there is no *1 letter word* in the selected text. The total number of 2-5 letter words is 21, and 11 of them are located in Table 3.14. For this text, it is obvious that 52.38% of the 1-5 letter words can be estimated by using the values in Table 3.14.

3.6. Word Length Distribution

In Table 3.16 and Figure 3.9, word length distributions of only the first 23 values are given, because they form 99.9517% of all the words which is most of TurCo. Depending on the values in Table 3.16, average word length of Turkish was found as 6.23 letters. Average word length value was found as 6.13 by using a 11.5MB corpus on a previous study (Dalkılıç & Dalkılıç, 2001). Although the size of the corpus was increased by 31.5 times, the average word length value remains nearly the same.

There are 1,520,281 one-letter words makes 3.03% of the corpus TurCo. With the increasing number of letters in the word, this value increases. 5-letter words with the count of 7,548,447 and a percentage of 15.06% has the highest rank. After this value, it decreases as the number of letters in the word increases, and after 13 letters it gets lower than 1%.

When cumulative distributions are examined, the ratio of the words having at most 4 letters is 30.43%, by the addition of the 5-letter and 6-letter words this value increases to 45.49%, and to 57.92% respectively. This value is 69.21% for the words

having at most 7 letters, but after 5-letters the acceleration decreases. Words having more than 15 letters and less than 24 letters form only 0.56% of the corpus TurCo. Each n-letter words having more than 23 letters (n>23) are less than 0.01%, so that they are not shown in Table 3.16. Words having more than 23 letters forms 0.048% of the corpus TurCo.

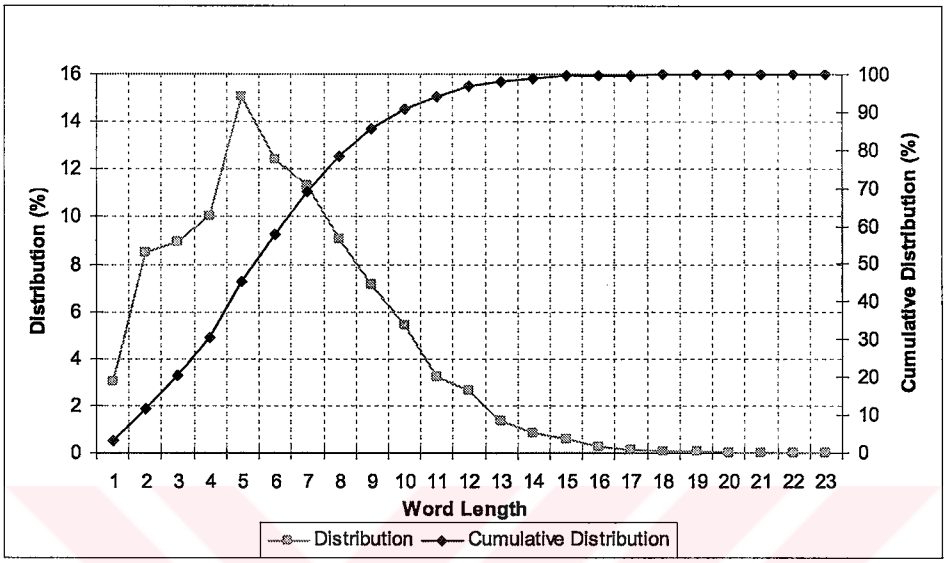


Figure 3.9 Word length distribution

Table 3.16 Word length distribution

Word Length (Letter)	# of Words	Distribution (%)	Cumulative Distribution (%)
1	1,520,281	3.0338	3.0338
2	4,240,703	8.4625	11.4963
3	4,461,703	8.9035	20.3997
4	5,026,935	10.0314	30.4312
5	7,548,447	15.0632	45.4944
6	6,227,019	12.4262	57.9206
7	5,656,719	11.2882	69.2088
8	4,524,240	9.0283	78.2371
9	3,583,148	7.1503	85.3874
10	2,714,246	5.4164	90.8038
11	1,623,411	3.2396	94.0434
12	1,315,454	2.6250	96.6684
13	685,737	1.3684	98.0368
14	408,846	0.8159	98.8527
15	277,798	0.5544	99.4070
16	122,709	0.2449	99.6519
17	66,701	0.1331	99.7850
18	37,215	0.0743	99.8593
19	22,243	0.0444	99.9037
20	11,803	0.0236	99.9272
21	4,821	0.0096	99.9368
22	4,423	0.0088	99.9457
23	3,019	0.0060	99.9517
Total	50,087,621	99.9517	

3.7. Different Word Usage Ratio

Different Word Usage Ratio (DWUR) value, calculated as in Formula 3.2, is one of the characteristics of a corpus, and represents the ratio of different words used in the corpus. This value is also a factor that represents the word duplication characteristic of a corpus. DWUR values can help us to identify the necessary size of the corpus for the word n-gram. DWUR values calculated for each part of TurCo for n-gram where $1 \leq n \leq 5$ and given in Table 3.17.

DWUR value for the TurCo for monogram analysis is 1.37%, and this value increases to 23.94% for bigram, 48.24% for trigram, 58.05% for tetragram and 61.75% to pentagram analysis. These values show that the DWUR ratio increases logarithmically for TurCo as given in Figure 3.10.

Table 3.17 DWUR values for n-gram analysis

Part #	Ngram Types				
	Mono (%)	Bi (%)	Tri (%)	Tetra (%)	Penta (%)
1	1.46	25.59	54.19	67.43	73.08
2	2.62	34.45	60.08	68.37	70.89
3	1.06	6.43	8.98	9.78	10.18
4	6.62	61.36	92.26	98.30	99.24
5	2.19	10.48	16.59	20.83	24.36
6	5.60	20.92	30.62	36.74	42.12
7	8.86	43.89	60.13	65.33	67.63
8	14.35	46.20	55.18	57.51	58.80
9	8.16	34.98	51.69	61.44	68.85
10	27.34	76.25	86.41	88.03	88.73
11	26.11	76.81	87.79	89.39	89.89
12	24.85	66.65	77.41	80.48	82.25
Total	1.37	23.94	48.24	58.05	61.75

The lowest DWUR values are for Part #3, which is a newspaper part and data of this part were taken from Internet, so it includes more duplicated words, such as menu items repeated on each page, links etc. For this reason, DWUR ratio of this part shows linearity instead of increasing logarithmically.

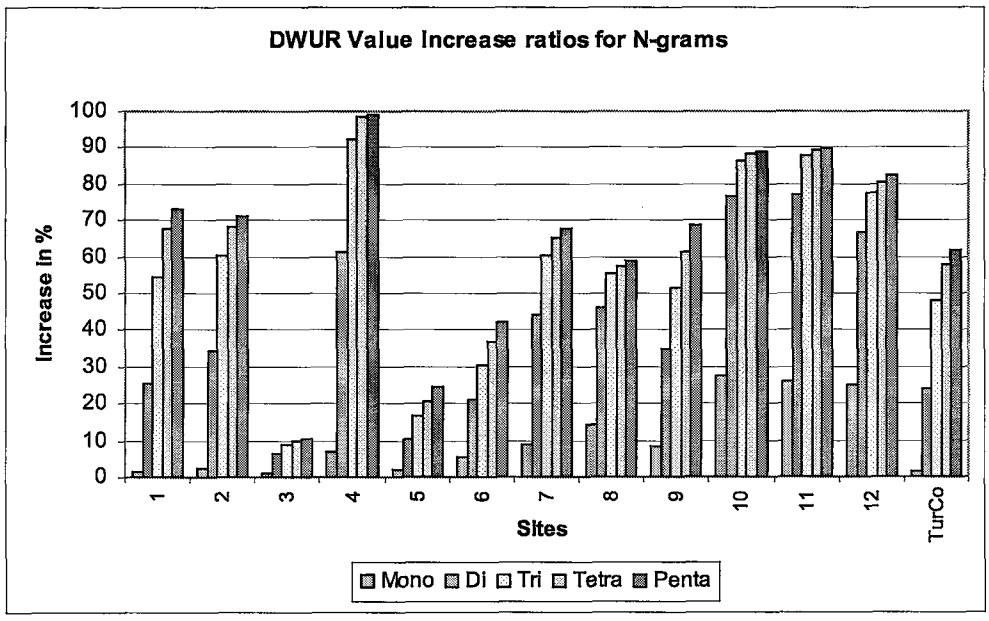


Figure 3.10 DWUR value increase ratios

The effect of duplicating items is easily observed when the Part #4 is examined. This part includes Novels and Stories from different authors in Turkish language, this means, most of the unwanted and unnecessary duplications are eliminated. But, because of the lack of the digital Novels and Stories in Turkish, only ~4.67M NOW could be collected for that part of the corpus. So, the DWUR ratios of this part show logarithmical tendency and close to 100% (99.24%) for pentagrams that also shows that size of the sample isn't enough to make word n-gram analysis where n is higher.

CHAPTER FOUR

WORD N-GRAMS

4. WORD N-GRAMS

By using ALG-3 given in section 3.2.1.3, n-gram analysis for TurCo and its components is done. In this section, these results are given and compared with English.

Because of the agglutinative nature of Turkish, suffixes can be added at the end of the words and by adding suffix after suffix lots of words can be formed from only 1 word, so the different word counts in Turkish can be strictly higher than English, but lower than Korean.

4.1. Comparing Turkish and English Word N-grams

When the results obtained from n-gram analysis have been examined, it is seen that, they show the unbalanced property of the corpus, because of the duplicating menu items taken from web sites and included in the corpus.

The total frequency of the first 100 words of English forms 42% of a total of 100,237 different words (Choi, 2000). In addition to English, for Korean, the total frequency of the first 2782 words forms 42% of a total of 1,344,018 different words (Choi, 2000), and for Turkish the total frequency of the first 537 words forms 42% of the total 686,804 different words.

From Table 4.1, the total frequency of the first 995 monograms forms 50% of the all monograms, and this number is 357,862 for bigrams, 3,450,151 for trigrams, 8,066,198 for tetragrams and 11,774,644 for pentagrams. As n is small in n-gram, total frequency of small percentage of the n-gram forms 50% of all the n-grams. In Table 4.1, this is also given for 10% and 20%. Same information is also given for

Novels & Stories in Table 4.2. For Novels & Stories, trigram, tetragram and pentagram values are much closer to each other than TurCo, this is directly related to the NODW values.

Table 4.1 Number of types and their percentage in TurCo

n-gram	10%	20%	50%	NODW
Mono	15	77	995	686,804
Bi	860	8,016	357,862	11,998,590
Tri	4,570	90,692	3,450,151	24,173,546
Tetra	11,306	235,672	8,066,198	29,089,011
Penta	21,050	352,596	11,774,644	30,943,233

Table 4.2 Number of types and their percentage in Novels & Stories

n-gram	10%	20%	50%	NODW
Mono	9	56	1448	309,030
Bi	2,823	22,332	530,424	2,864,577
Tri	138,400	572,253	1,972,743	4,306,896
Tetra	387,696	854,527	2,255,018	4,589,170
Penta	431,224	898,054	2,298,545	4,632,697

Because of its reliability, representative property and the size, Novels and Stories is compared with the results obtained from Brown corpus (Teahan, 1998).

Top 25 results taken from monogram, bigram and trigram analysis for Novels and Stories are given in Table 4.3. In Table 4.3, the frequency values are sorted in a decreasing order and the first 25 elements are given. For each element of the corpus, first 20 highest monogram, bigram, trigram, tetragram and pentagram frequency lists are given in Appendix A. For Novels & Stories and TurCo these lists are for the first 50 elements.

When the total frequencies of first 25 values are considered, it is seen that the total ratio forms 15.13% of all the monograms, decreases to 1.08% for bigrams, and to 0.14% for trigrams. The most frequently used word in Novels and Stories is "bir". At the same time, this word is the second most frequently used word in TurCo. This word is used with the word "şey" as "bir şey" and takes second place in bigrams. These two words together are used with the word "başka" as "başka bir şey" and take the first place in the trigrams. This shows the durability and the reliability of the analysis.

Table 4.3 Top 25 of Word, bigram, and trigrams analysis of Novels and Stories

Rank	Word	Freq.	%	Bigram	Freq.	%	Trigram	Freq.	%
1	bir	143,536	3.08	ya da	6,789	0.15	başka bir şey	906	0.019
2	ve	91,089	1.95	bir şey	5,020	0.11	ne var ki	597	0.013
3	bu	66,601	1.43	başka bir	2,538	0.05	ali rıza bey	360	0.008
4	da	42,402	0.91	büyük bir	2,500	0.05	bir yandan da	327	0.007
5	de	40,814	0.87	ne kadar	2,437	0.05	bir süre sonra	318	0.007
6	için	27,230	0.58	ve bu	2,032	0.04	ne olursa olsun	279	0.006
7	ne	23,782	0.51	o kadar	1,945	0.04	bir an önce	257	0.006
8	gibi	22,783	0.49	ben de	1,910	0.04	ne yazık ki	244	0.005
9	daha	22,025	0.47	böyle bir	1,880	0.04	her şeyden önce	235	0.005
10	o	21,778	0.47	o zaman	1,836	0.04	ya da bir	227	0.005
11	çok	18,872	0.40	hem de	1,793	0.04	albay aureliano buendia	217	0.005
12	sonra	17,549	0.38	her zaman	1,609	0.03	bir kez daha	213	0.005
13	her	16,856	0.36	bu kadar	1,548	0.03	jose arcadio buendia	200	0.004
14	ama	15,801	0.34	bir gün	1,514	0.03	her ne kadar	197	0.004
15	kadar	15,296	0.33	hiçbir şey	1,460	0.03	mustafa kemal paşa	196	0.004
16	ya	13,879	0.30	yeni bir	1,460	0.03	ittihat ve terakki	194	0.004
17	ki	13,171	0.28	belki de	1,445	0.03	en ufak bir	193	0.004
18	olan	12,317	0.26	değil mi	1,423	0.03	mustafa kemal in	192	0.004
19	ile	12,264	0.26	ne de	1,404	0.03	mustafa kemal atatürk	191	0.004
20	olarak	12,225	0.26	o da	1,392	0.03	bunun içindir ki	190	0.004
21	bütün	11,419	0.24	sonra da	1,352	0.03	bir şey değildir	188	0.004
22	zaman	11,348	0.24	her şeyi	1,350	0.03	bir şey yok	180	0.004
23	diye	11,306	0.24	daha çok	1,296	0.03	her zamanki gibi	169	0.004
24	ben	11,044	0.24	bir daha	1,292	0.03	kısa bir süre	169	0.004
25	dedi	10,883	0.23	gibi bir	1,282	0.03	sovyet rusya nın	169	0.004
Total		706,270	15.13		50,507	1.08		6608	0.142

Until the 25th word ("dedi") in the monogram analysis, there are no verbs or nouns; instead they are adverbs, conjunctions and adjectives. Also it is the same for the first 25 values of bigrams, but this situation changes for trigrams. The person names such as "ali rıza bey", "albay aureliano buendia", "jose arcadio buendia", "mustafa kemal paşa", "mustafa kemal atatürk", who are the heroes in some Novels and Stories, are repeated much more.

From Table 4.4, the monograms and bigrams, until the 25th word in Novels and Stories, are similar to the first 24 values of the Brown corpus, which means the both corpora include adverbs, conjunctions and adjectives, but not nouns or verbs in the top 24 rank.

Table 4.4 Word statistics from the Brown Corpus (Teahan, 1998)

Rank	Word	Freq.	%	Bigram	Freq.	%	Trigram	Freq.	%
1	the	69,967	6.83	of the	9,764	0.95	one of the	403	0.04
2	of	36,451	3.56	in the	6,142	0.6	the united states	340	0.03
3	and	28,919	2.82	to the	3,509	0.34	as well as	238	0.02
4	to	26,218	2.56	on the	2,490	0.24	some of the	179	0.02
5	a	23,539	2.3	and the	2,274	0.22	out of the	176	0.02
6	in	21,406	2.09	for the	1,861	0.18	the fact that	167	0.02
7	that	10,783	1.05	to be	1,722	0.17	i don t	162	0.02
8	is	10,092	0.99	at the	1,664	0.16	the end of	149	0.01
9	was	9,815	0.96	with the	1,542	0.15	part of the	146	0.01
10	he	9,797	0.96	of a	1,506	0.15	it was a	144	0.01
11	for	9,492	0.93	it is	1,474	0.14	there was a	142	0.01
12	it	9,090	0.89	in a	1,434	0.14	it is not	136	0.01
13	with	7,285	0.71	from the	1,420	0.14	to be a	134	0.01
14	as	7,249	0.71	that the	1,404	0.14	there was no	132	0.01
15	his	6,999	0.68	by the	1,371	0.13	of the united	129	0.01
16	on	6,764	0.66	it was	1,301	0.13	the u s	127	0.01
17	be	6,385	0.62	he was	1,088	0.11	there is a	126	0.01
18	s	6,284	0.61	as a	1,001	0.1	a number of	123	0.01
19	i	5,934	0.58	with a	917	0.09	in order to	120	0.01
20	at	5,381	0.53	he had	903	0.09	most of the	114	0.01
21	by	5,345	0.52	is a	876	0.09	it is a	114	0.01
22	this	5,144	0.5	of his	810	0.08	members of the	110	0.01
23	had	5,131	0.5	is the	808	0.08	end of the	108	0.01
24	not	4,614	0.45	for a	795	0.08	and in the	108	0.01
25	are	4,386	0.43	was a	788	0.08	of the new	107	0.01

4.2. Zipf's Law

According to Zipf's Law, *f*r* is constant (Choi, 2000). In order to check if a language obeys Zipf's Law, word rank-frequency values of that language can be sampled. Some samples of word monogram values of English language are given in Table 4.5.

The data in Table 4.5 nearly matches Zipf's Law. The first three values and the values in the middle (r=90,100,200) tend to bulge. Except those values f*r is approximately constant as the Zipf's Law states (Manning & Schütze, 2000).

Table 4.5 f*r values on Tom Sawyer (Manning & Schütze, 2000)

Word	Freq.(f)	Rank (r)	f*r	Word	Freq.(f)	Rank (r)	f*r
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5325	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8200	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
oh	116	90	10440	could	2	4000	8000
two	104	100	10400	applausive	1	8000	8000

For the Novels & Stories, f*r increases until the rank 200. Zipf's Law holds between 200 and 50,000. After 50,000, it decreases linearly as seen in Table 4.6. There is a bit off between 90,000-150,000. As a whole, from Figure 4.1, it is like a second degree polynomial as given in Formula 4.1.

$$y = -1006.2x^2+42648x+52359 \qquad (4.1)$$

Line with diamonds shows the actual values, the other line shows the matching polynomial with R^2=0.9004.

Table 4.6 f*r values on word monograms of Novels&Stories from TurCo

Word	Freq.(f)	Rank (r)	f*r	Word	Freq.(f)	Rank (r)	f*r
bir	143536	1	143536	ayhan	558	900	502200
ve	91089	2	182178	yaşayan	510	1000	510000
bu	66601	3	199803	davet	274	2000	548000
o	21778	10	217780	kalacak	186	3000	558000
olarak	12225	20	244500	bulunmak	141	4000	564000
kendi	9000	30	270000	yapısına	69	8000	552000
başka	7466	40	298640	edemezdi	55	10000	550000
karşı	6056	50	302800	gerektiğin	25	20000	500000
benim	5649	60	338940	ruiz	15	30000	450000
arasında	5000	70	350000	torunlarına	10	40000	400000
e	4564	80	365120	gözlemlerle	8	50000	400000
ise	4058	90	365220	batıcılık	6	60000	360000
biri	3542	100	354200	çıkarılıyor	5	70000	350000
ye	1976	200	395200	arkadaştır	4	80000	320000
konusunda	1448	300	434400	tartındı	3	90000	270000
boş	1167	400	466800	roles	3	100000	300000
tatlı	960	500	480000	hüviyetini	2	150000	300000
köy	808	600	484800	yağmurça	1	153633	153633
ifade	710	700	497000	yanıtlamakta	1	160000	160000
arcadio	626	800	500800				

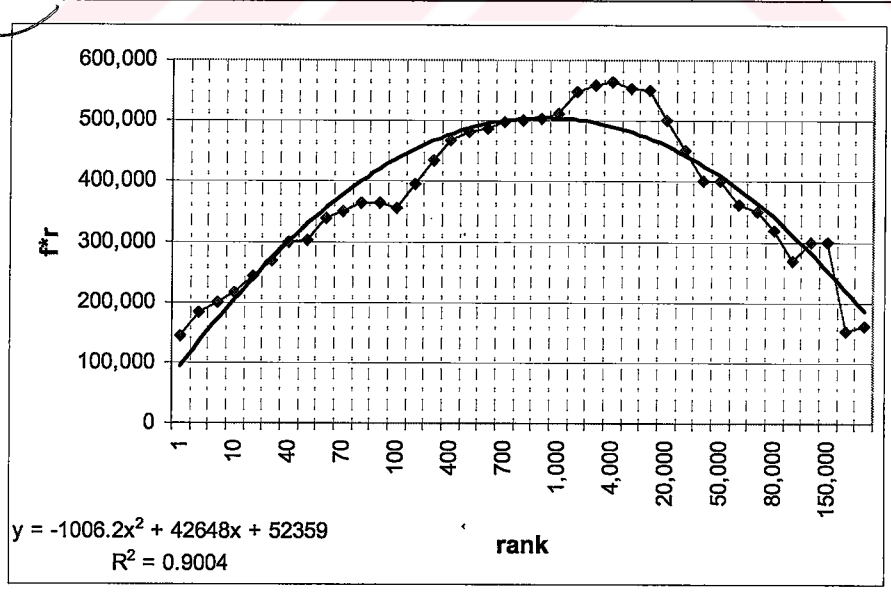


Figure 4.1 f*r value graph on word monograms of Novels&Stories from TurCo

When the same analysis is done for Turkish on TurCo, it is seen that, f*r increases until the rank 300, and Zipf's Law holds between 300 and 4000. After 4000, it decreases linearly as seen in Table 4.7. As a whole, from Figure 4.2, it is like a second degree polynomial as given in Formula 4.2. Dotted points show the actual values, the curve shows the matching polynomial with R^2=0.9277.

$$y= -15228x^2+601655x+91471 \qquad \textbf{(4.2)}$$

Table 4.7 f*r values on word monograms of TurCo

Word	Freq.(f)	Rank (r)	f*r	Word	Freq.(f)	Rank (r)	f*r
ve	1137582	1	1137582	sağlamak	6354	1000	6354000
bir	803553	2	1607106	başarı	3205	2000	6410000
bu	646620	3	1939860	önergeler	2100	3000	6300000
n	169675	10	1696750	koşulları	1587	4000	6348000
büyük	116978	20	2339560	incelediğini	712	8000	5696000
dünya	96287	30	2888610	böyledir	552	10000	5520000
göre	81062	40	3242480	yapılabileceği	228	20000	4560000
ki	74659	50	3732950	gittiğimiz	131	30000	3930000
durumu	69049	60	4142940	giyimli	87	40000	3480000
ilk	59241	70	4146870	önlenmiştir	62	50000	3100000
olduğu	56229	80	4498320	istihdamdaki	46	60000	2760000
bugün	52326	90	4709340	kaydetme	36	70000	2520000
diye	46981	100	4698100	dönebilen	29	80000	2320000
milyar	28999	200	5799800	durmamış	24	90000	2160000
sohbet	20733	300	6219900	egem	20	100000	2000000
eb	15967	400	6386800	sekseni	10	150000	1500000
mu	12778	500	6389000	görmemişsiniz	5	200000	1000000
saygıyla	10967	600	6580200	güçleştireceği	2	300000	600000
halde	9295	700	6506500	gönülaydın	1	393709	393709
savaş	8166	800	6532800	götürebiliyorum	1	400000	400000
geniş	7200	900	6480000				

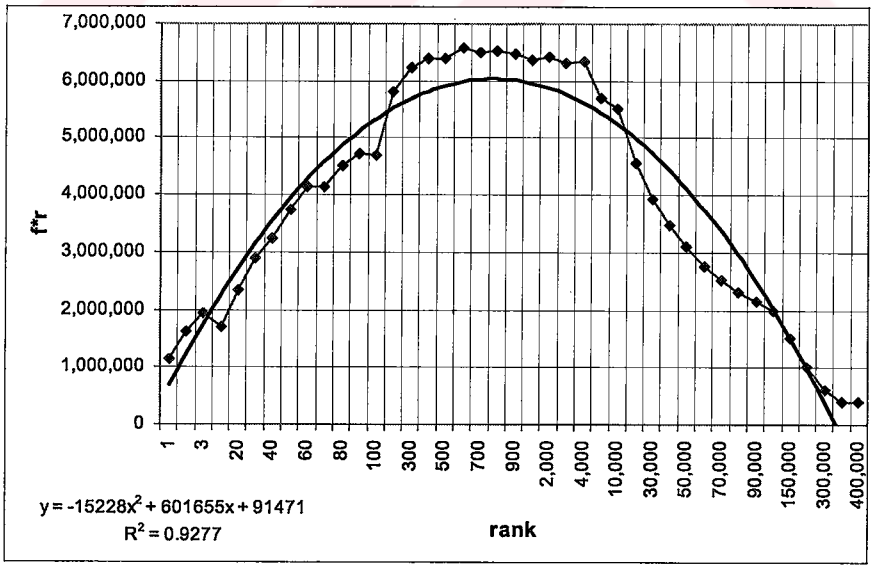


Figure 4.2 f*r value graph on word monograms of TurCo

In order to get accurate results, all word monograms, bigrams, and trigrams should be used. To show this graphically, rank-frequency graphs can be drawn. The rank-frequency graph of monograms (types), bigrams and trigrams for English from Brown Corpus is given in Figure 4.3 (Teahan, 1998). As the values on the graph cannot show linearity, English can not be said to have a "*perfect match*" with Zipf's Law.

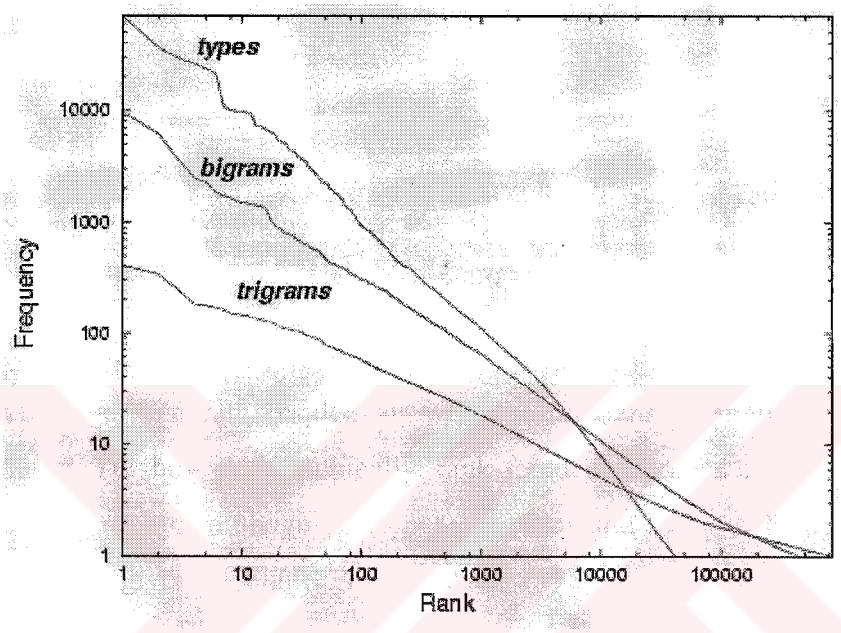


Figure 4.3 Rank frequency data for word n-grams (Brown Corpus) (Teahan, 1998)

The graphs for two different corpora of Turkish are given in Figure 4.4 and Figure 4.5. The rank-frequency graph of Novels and Stories is given as Figure 4.4.

The monogram, bigram and trigram graphs are said to be similar to their corresponding for English, and they obey Zipf's Law, but not perfectly. As the corpus size is not big enough to obtain more accurate results, the tetragram and pentagram graphs do not show obeyence to Zipf's Law.

The rank-frequency graph of TurCo, which is the main corpus with a word count of 50,111,828, is given as Figure 4.5. The monogram, bigram, trigram and tetragram graphs are said to be obeying Zipf's Law with some bulges, but not perfectly.

Rank frequency values for TurCo (Figure 4.5) are similar to the values of Brown Corpus (Figure 4.3). Rank frequency graph of each piece of TurCo are given in Appendix B.

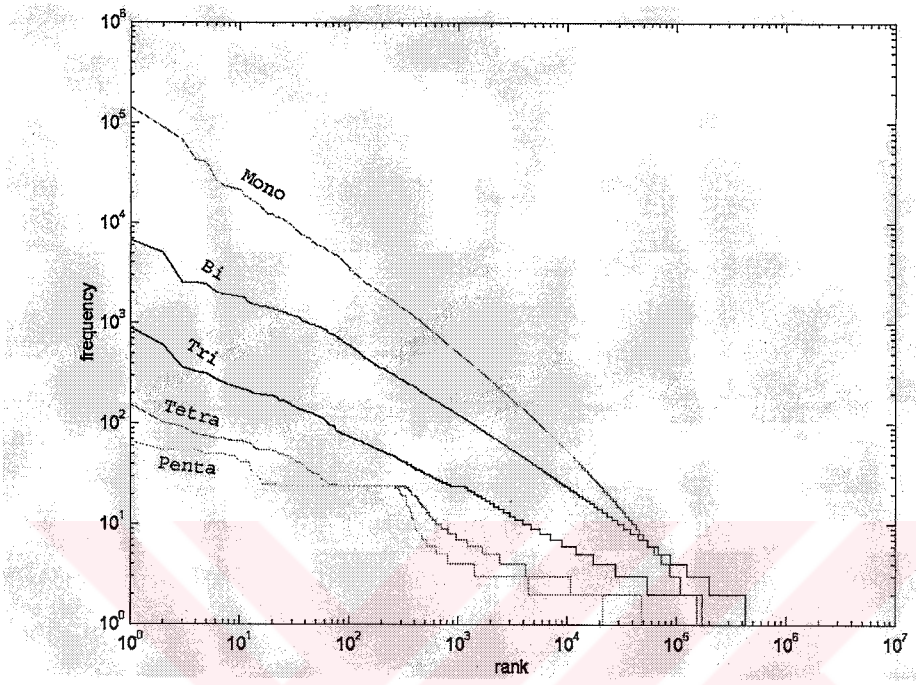


Figure 4.4 Rank frequency data for Novels & Stories

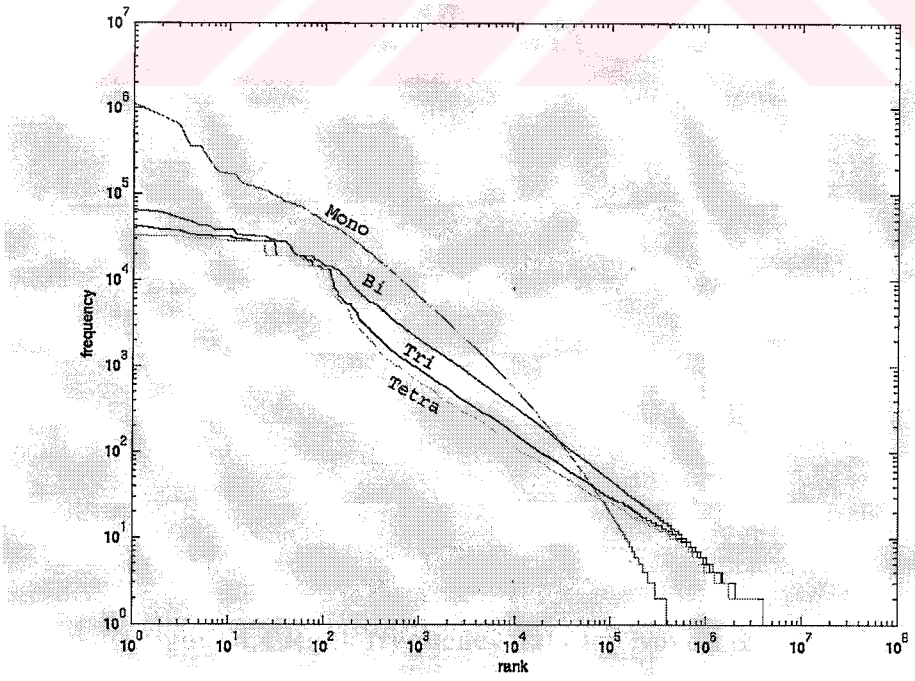


Figure 4.5 Rank frequency data for TurCo

4.3. *Determination of Mandelbrot Constants*

Mandelbrot added some parameters to Zipf's Law to make a close approximation to the real values as shown in Formula 4.3 (Teahan, 1998).

$$P(r) = \frac{\mu}{(c+r)^B} \text{ , r = 1, 2, ..., n} \qquad \textbf{(4.3)}$$

Where μ=(r*f(r))/NOW or μ=r*P(r). According to Mandelbrot B>1 in all cases (Choi, 2000). Miller ((Miller et al., 1957) as cited in (Witten & Bell, 1990)) found the values of the constants for English as in Formula 4.4.

$$P(r) = \frac{0.11}{(0.54+r)^{1.06}} \text{ , r = 1, 2, ..., n} \qquad \textbf{(4.4)}$$

In order to determine the c and B values for Turkish, the similarity between the graphs of actual and by Formula 2.6 created values should be investigated. This similarity can be better determined by using statistical methods. According to Formula 4.3 first μ is needed to be calculated, and then the c and B values in the same formula.

By using average f*r and NOW, μ is calculated as 0.063, with an average $f*r$ of 292,440 (Formula 4.5).

$$\mu = \frac{Average\ of\ f*r}{\text{NOW}} = \frac{292{,}440}{4{,}668{,}306} = 0.063 \cdot \qquad \textbf{(4.5)}$$

For Turkish, different c and B values were tried to find the most appropriate R^2 values, and the best 40 R^2 and the corresponding c and B values are given in Table 3. For Novels and Stories, c is found between 0.26 and 0.35 with an average of 0.30, and B is found as 0.90, both of which have the best R^2 value with 0.9958.

Table 4.8 Mandelbrot constants

Novels & Stories			TurCo		
C	B	R^2	C	B	R^2
0.35	0.90	0.9958	0.28	0.78	0.9883
0.33	0.90	0.9958	0.27	0.78	0.9883
0.30	0.90	0.9958	0.29	0.78	0.9882
0.29	0.90	0.9958	0.28	0.79	0.9882
0.27	0.90	0.9958	0.29	0.79	0.9881
0.26	0.90	0.9958	0.28	0.77	0.9881
0.38	0.90	0.9957	0.30	0.77	0.9880
0.37	0.90	0.9957	0.30	0.80	0.9878
0.25	0.90	0.9957	0.27	0.80	0.9878
0.25	0.90	0.9957	0.25	0.80	0.9878
0.24	0.90	0.9957	0.35	0.80	0.9877
0.40	0.90	0.9956	0.39	0.80	0.9876
0.39	0.90	0.9956	0.40	0.80	0.9875
0.42	0.90	0.9955	0.30	0.76	0.9875
0.43	0.90	0.9954	0.42	0.80	0.9874
0.42	0.91	0.9954	0.43	0.80	0.9873
0.42	0.89	0.9954	0.26	0.75	0.9873
0.20	0.90	0.9954	0.46	0.80	0.9871
0.20	0.90	0.9954	0.27	0.75	0.9871
0.44	0.91	0.9953	0.40	0.77	0.9868
0.44	0.90	0.9953	0.30	0.75	0.9867
0.43	0.91	0.9953	0.53	0.80	0.9865
0.43	0.89	0.9953	0.53	0.83	0.9860
0.44	0.89	0.9952	0.55	0.85	0.9849
0.44	0.92	0.9951	0.53	0.85	0.9849
0.42	0.87	0.9949	0.61	0.85	0.9848
0.50	0.90	0.9948	0.60	0.85	0.9848
0.15	0.80	0.9945	0.59	0.85	0.9848
0.40	0.85	0.9942	0.50	0.85	0.9848
0.50	0.95	0.9941	0.45	0.85	0.9847
0.40	0.95	0.9938	0.65	0.85	0.9846
0.20	0.80	0.9934	0.70	0.85	0.9844
0.30	0.95	0.9929	0.40	0.85	0.9844
0.27	0.80	0.9918	0.30	0.85	0.9834
0.20	0.95	0.9912	0.50	0.75	0.9830
0.30	0.80	0.9910	0.53	0.75	0.9823
0.40	1.10	0.9778	0.70	0.90	0.9807
0.40	0.70	0.9549	0.61	0.90	0.9802
0.60	0.70	0.9444	0.45	0.90	0.9785
0.10	1.20	0.9438	0.40	0.90	0.9777

In order to define the Zipf's Law formula with Mandelbrot constants c and B, the best fitted (R^2 =0.9958) curve's c and B values, which are given in Table 4.8, are put into the Formula 4.3 and by using the μ value in Formula 4.5, Formula 4.6 is obtained.

$$P(r) = \frac{0.063}{(0.30 + r)^{0.90}} \quad , r = 1, 2, \ldots, n \, . \qquad \textbf{(4.6)}$$

The graph, obtained for Novels and Stories after adding Mandelbrot constants to Zipf's Law, is given in Figure 4.6.

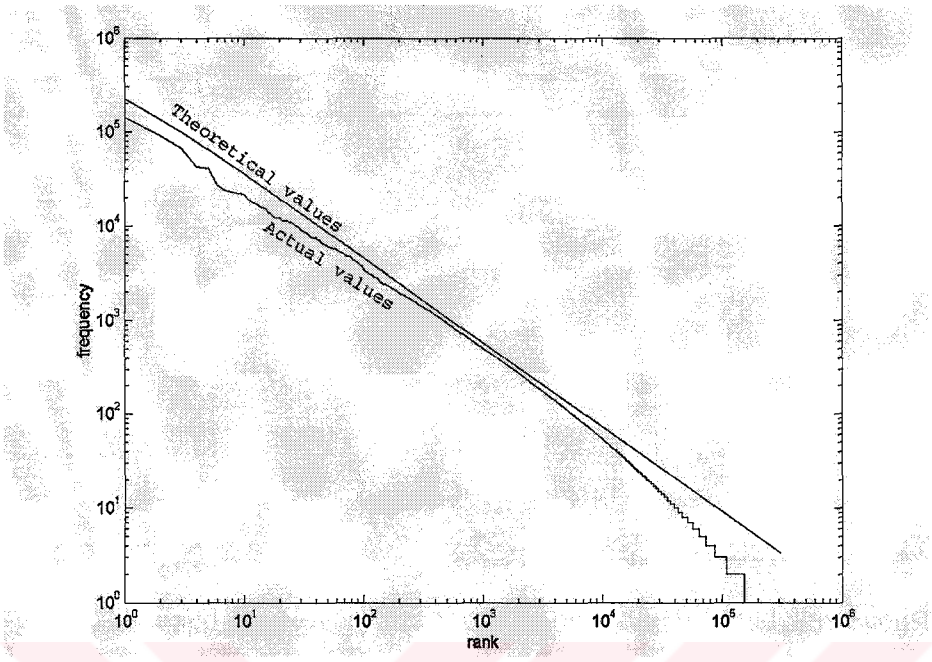


Figure 4.6 Graph of actual and theoretical values with c=0.30, B=0.90 for Turkish Novels and Stories

For TurCo, c is found as 0.27, and B is found as 0.78 with an R^2 value of 0.9883, as given in Table 4.8. By substituting the values found for TurCo with the values in formulae 4.5 and 4.6, the formulae 4.7 and 4.8 were obtained.

$$\mu = \frac{Average\ of\ f*r}{\text{NOW}} = \frac{1{,}176{,}000}{50{,}111{,}828} = 0.024 \cdot \qquad \textbf{(4.7)}$$

$$P(r) = \frac{0.024}{(0.27 + r)^{0.78}} \text{ , r = 1, 2, ..., n .} \qquad \textbf{(4.8)}$$

The graph, obtained for TurCo after adding Mandelbrot constants to Zipf's Law, is given in Figure 4.7.

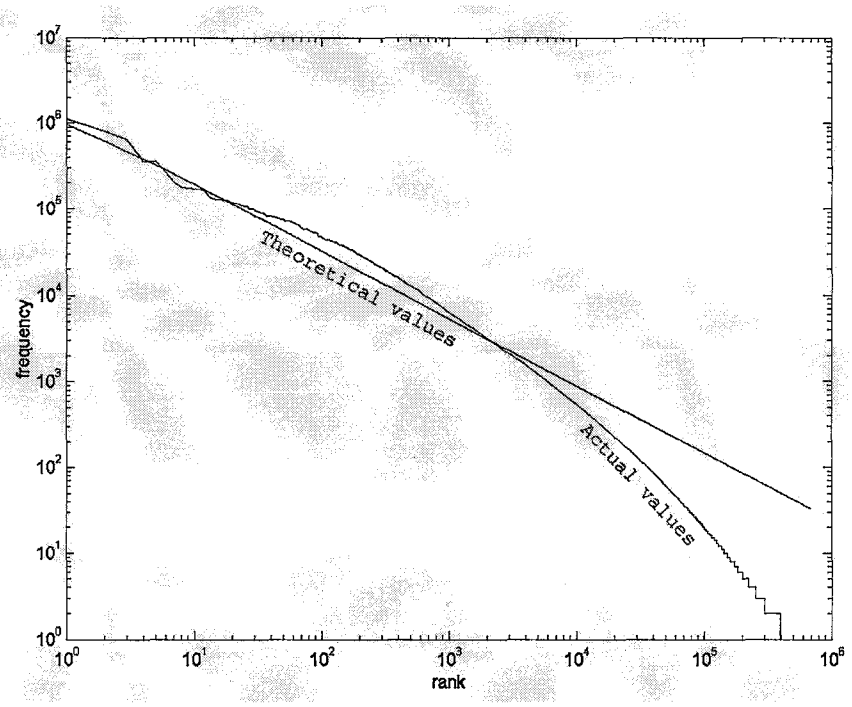


Figure 4.7 Graph of actual and theoretical values with c=0.27, B=0.78 for TurCo

When Figure 4.6 and 4.7, and their corresponding R^2 values are compared, a better match is found for Novels and Stories. The corpus for Figure 4.6 (Novels and Stories) includes fewer duplicated words than the corpus TurCo. At the same time, TurCo includes different types of text, which causes strong stair effect at the beginning ($10^{0.5}$ thru 10^1) and end (>10^5) of the rank, and a bow more then Figure 4.6 between 10^1 -10^4. When both corpora are compared to the rank 10^4, it can be said that Figure 4.6 shows better match to Zipf's Law with Mandelbrot correction more than Figure 4.7.

CHAPTER FIVE

EVALUATION OF SMOOTHING TECHNIQUES AND DEVELOPING A TEST SOFTWARE

5. EVALUATION OF SMOOTHING TECHNIQUES AND DEVELOPING A TEST SOFTWARE

5.1. Evaluation of Smoothing Techniques

By using Maximum Likehood Estimation (MLE), and counting n-grams in TurCo, the n-gram probability values were calculated. As all the n-grams cannot be seen in a single corpus, some of the unseen n-grams get zero values, and a sparse matrix is formed (Section 2.4). For a language, there are a limited number of frequent n-grams (Table 4.1), but there is infinite number of rarely seen n-grams, and it seems nearly impossible to form a corpus of having all the n-grams (Manning & Schütze, 2000). To fill the zero values in the sparse matrix, smoothing techniques are commonly used.

Generally smoothing techniques use discounting methods, where the probability of each seen event is discounted to create a probability for the unseen event, because total probability must be 1 every time. As described in Section 2.4, the most commonly used smoothing techniques are Add-One Smoothing, Good-Turing Smoothing and Back-Off techniques.

5.1.1. Add-One Smoothing

Add-One smoothing misses the mark for unseen bigrams by up to three orders of magnitude, and has larger errors than the maximum likehood estimator (Gale & Church, 1994). When Add-One smoothing and Good-Turing smoothing is compared,

Add-One is correct only if the ratio of unseen n-grams to observed types equals the ratio of all types to sample size, but there isn't any relation between the sample size and the unseen n-grams. So, as there is an accurate estimator like Good-Turing, there is no need to use Add-One smoothing (Gale & Church, 1994).

5.1.2. Good-Turing Discounting

Good-Turing is a much better approximation for smoothing word n-grams. For English, the counts of bigrams and trigrams (c) and number of counts (N_c) by using Good-Turing approach and Formula 2.7 are given in Table 5.1 (Teahan, 1998). For example there are 339,808 bigrams appeared only once, and 48,548 appeared twice.

Table 5.1 Smoothed frequency estimates based on Good-Turing (Brown corpus)

Bi			Tri		
c	**Nc**	**c***	**c**	**Nc**	**c***
0	1,722,665,025	0.000197	0	71,499,211,862,625	1.08E-08
1	338,808	0.287	1	775,545	0.112
2	48,548	1.157	2	43,378	0.852
3	18,721	2.074	3	12,325	1.742
4	9,707	3.079	4	5,368	2.652
5	5,978	3.924	5	2,847	3.692
6	3,910	4.947	6	1,752	4.767
7	2,763	6.095	7	1,193	5.687
8	2,105	6.708	8	848	6.315
9	1,569	8.120	9	595	7.782
10	1,274	9.144	10	463	8.102

For count 0, the value of N_c was calculated by using monograms. For bigrams, this is *total number of words * total number of words* that gives the theoretical maximum number of bigrams. For c=0 in bigrams N_c was calculated by finding total number of bigrams and subtracting this number from theoretical maximum number of bigrams. From Table 5.1, this value is 1,722,665,025 for bigrams (Teahan, 1998). For trigrams, this is calculated as "theoretical number of trigrams (total number of monograms)^3 minus total number of trigrams" and the result is found as 71,499,211,862,625.

The last column for both bigram and trigrams is c* which shows the Good-Turing reestimates of corresponding count (c) values. For bigrams, c* values are much closer to c values.

The same calculations were done for TurCo. For c=0, theoretical values of bigrams, trigrams, tetragrams and pentagrams were used from Table 3.9. To find N_c of c=0 for bigrams, the sum of all N_c values were calculated and this value is subtracted from the theoretical value of bigrams. The same method is used for trigrams, tetragrams and pentagrams.

The Good-Turing estimates for Part #4 of TurCo (Novels & Stories) were calculated as mentioned above and the results are given in Table 5.2 and Figure 5.1. When bigrams and trigrams for each c value ($0 \leq c \leq 10$) are compared, it can be seen that, the corresponding c* decreases from bigram to trigrams. Except c=5 and 9 from trigrams to tetragrams, and except c=5,7,9 from tetragrams to pentagrams, for the first 10 values, the decrease tendency is seen again.

Good-Turing estimates for TurCo were also calculated and given in Table 5.3 and Figure 5.2. There are much more exceptions when compared with Novels and Stories, such as, for trigrams c=5,7,9 and for tetragrams c=3,5,7,9. But, in general as n in n-gram increases, c* increases with some fluctuations. There are much more exceptions in Table 5.3 because, as commented before, Novels and Stories is a better example for sampling the language than TurCo. So, in general, c* decreases from n=1 to n=∞. As n increases the number of exceptions increases also.

As Back-off method gives better results (Katz, 1987), Back-off is selected as the test environment and a prototype program was generated.

Table 5.2 Good Turing estimates for Novels & Stories

Bi			Tri			Tetra			Penta		
c	Nc	c*	c	Nr	c*	c	Nc	c*	c	Nc	c*
0	95499540900	2.54993E-05	0	2.95122E+16	1.4E-10	0	9.12016E+21	4.98E-16	0	2.8184E+27	1.64E-21
1	2435170	0.188205341	1	4135517	0.056417	1	4540856	0.016478	1	4611168	0.007421
2	229156	0.987755066	2	116657	0.697352	2	37413	0.53436	2	17110	0.527586
3	75450	1.919469848	3	27117	1.517277	3	6664	1.085834	3	3009	0.79096
4	36206	2.843865658	4	10286	2.524791	4	1809	2.059149	4	595	1.470588
5	20593	3.949691643	5	5194	3.601848	5	745	3.736913	5	175	4.32
6	13556	4.797137799	6	3118	4.429442	6	464	3.801724	6	126	2.888889
7	9290	5.895371367	7	1973	5.721237	7	252	5.079365	7	52	6
8	6846	6.784837862	8	1411	6.282778	8	160	5.56875	8	39	4.153846
9	5161	7.829877931	9	985	7.370558	9	99	8.181818	9	18	8.888889
10	4041	8.653551101	10	726	7.939394	10	81	6.382716	10	16	5.5

Table 5.3 Good Turing estimates for TurCo

Bi			Tri			Tetra			Penta		
c	Nc	c*	c	Nc	c*	c	Nc	c*	c	Nc	c*
0	4.717E+11	1.8E-05	0	3.23965E+17	6.23E-11	0	2.22501E+23	1.15E-16	0	1.52814E+29	1.81E-22
1	8509583	0.336636	1	20173152	0.192652	1	25619485	0.135129	1	27684722	0.117103
2	1432318	1.14832	2	1943196	0.938817	2	1730973	0.853447	2	1620988	0.831689
3	548253	2.296057	3	608102	2.244841	3	492431	2.36159	3	449386	2.474033
4	314705	2.881413	4	341273	2.432788	4	290730	2.162711	4	277949	2.042101
5	181359	4.371131	5	166049	4.586369	5	125753	5.007006	5	113520	5.263055
6	132124	4.992515	6	126927	4.694131	6	104941	4.58611	6	99577	4.551473
7	94233	7.29859	7	85116	8.867146	7	68753	10.37558	7	64746	10.85238
8	85971	7.023624	8	94342	6.590257	8	89169	6.267772	8	87831	6.170919
9	67092	9.969743	9	69082	11.50473	9	62099	12.6556	9	60222	13.03942
10	66889	7.047736	10	79477	5.002227	10	78590	3.908436	10	78526	3.549792

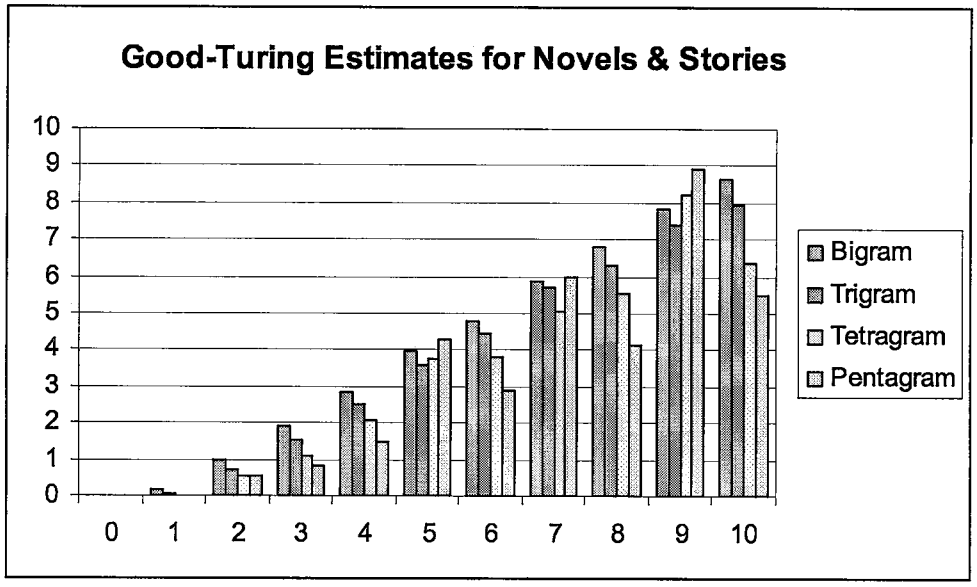


Figure 5.1 Good-Turing Estimates for Novels & Stories

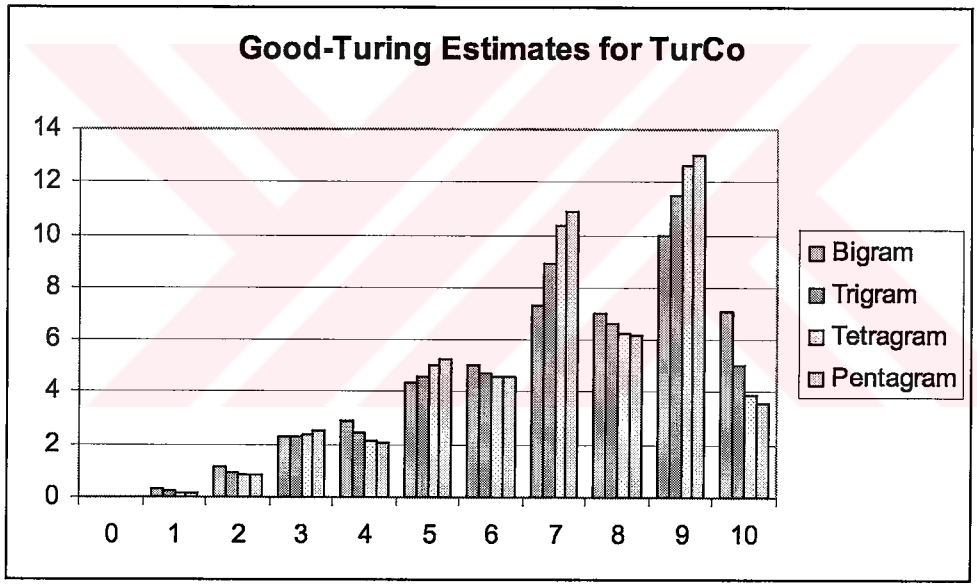


Figure 5.2 Good-Turing Estimates for TurCo

5.1.3. Back-off

This technique is much better than the previous ones because it doesn't directly make approximations. For example when it starts from trigrams, it also uses bigrams and monograms, if needed to fill the sparse matrix as given in Formula 2.11. By using these values, the smoothing is much more close to reality.

As these statistics were already created, a prototype program was developed to test the back-off technique.

5.2. Prototype Program using Back-off Technique

The aim of the prototype program is to find text errors and suggest corrections depending the concepts of Back-off technique, and at the same time it checks the validity of Back-off technique by using a large scale corpus (TurCo). The program was developed by using C programming language and text-indexing method under Windows environment. In order to increase the accuracy of the program, Minimum Edit Distance (MED) algorithm was also included.

This program can be thought as a combination of spell and grammar checker in Microsoft (MS) Word. To achieve this task, n-gram word lists (monogram, bigram, and trigram), obtained from Turkish Corpus (TurCo), are used.

Since the main goal of this program is not to create a text editor with different functions, the program has a text input screen and options menu with spell checking and suggesting corrections functions for Turkish by using TurCo. For the sake of simplicity, the user can type words without punctuations. The main screen of the program is given as Figure 5.3.

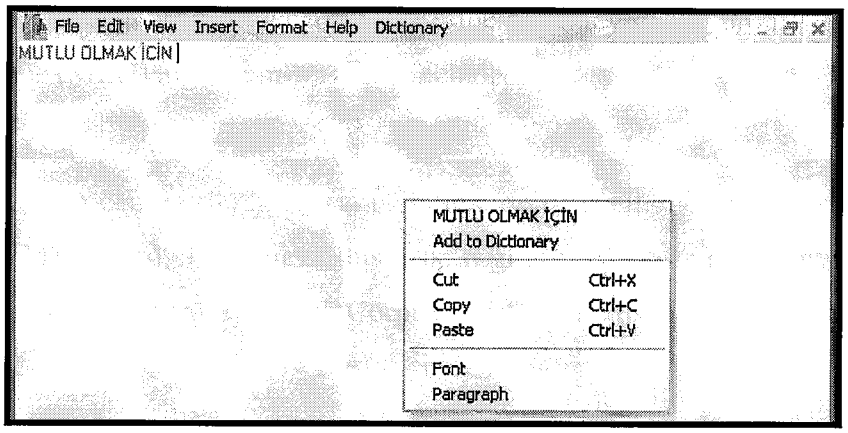


Figure 5.3 Prototype program screenshot

During the run of the prototype program, when spell checking is required to be carried out on the written text, at least 3 words are needed. Before spell checking, *Matching Criteria, N-gram Model,* and *Maximum # of suggestions* (Figure 5.4) options have to be selected. These options are found in *Options* Window, under *Format* menu item. In order to achieve spell checking, selecting an n-gram model is essential. After selecting the options, since the trigram model is common element of the n-gram options, the program tries to correct only the last word of the three words, by using the information obtained from previous two words.

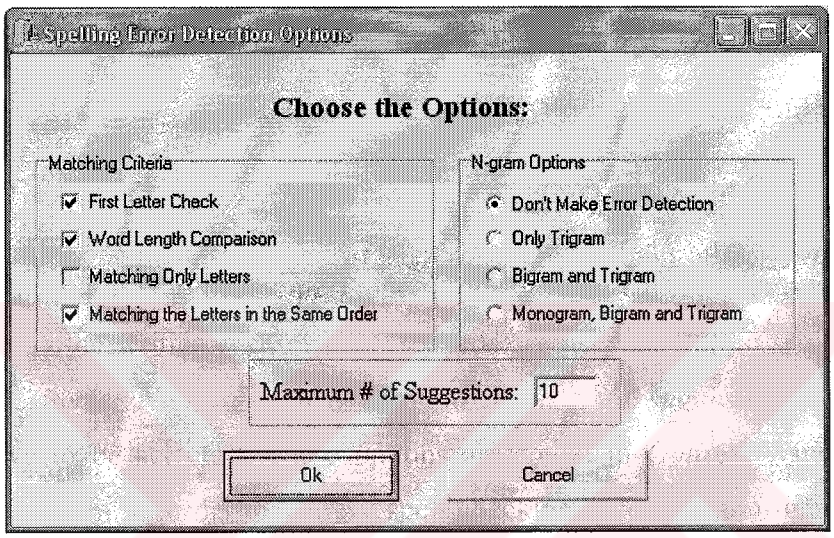


Figure 5.4 Prototype program "options" screenshot

5.2.1. Matching Criteria

When the first two words in trigram list or only the first word in the bigram list is matched with the written text, the list of the last word suggestions are generated with some predefined options. These options are to increase the accuracy of the match and not to list user lots of meaningless words.

i) First Letter Check: When the user writes "ALİ EVE GLEİD", and the match of the first two words (ALİ EVE) is found from the trigram list, the words to be suggested must begin with "G". When the bigram list is searched, then "EVE" must be the match, and again the words beginning with "G" will be suggested. When the monogram list is searched, then the words beginning with "G" will be suggested.

ii) Word Length Comparison: When generating the suggestion list to the user, it checks the length of the last word. The length of the last written word must be the same as the word in the list. When the user writes "ALİ EVE GLEİD", the suggestion list may include "GELDİ", "KALDI", "GİTTİ", which have the same word length as "GLEİD".

iii) Matching Only Letters: When the user writes "ALİ EVE GLEİD", in the suggestion list there should be "GELDİ", which is one of the meaningful words that can be created using the same letters. This option checks the words to have the same letters whether in the same order or not.

iv) Matching the Letters in the Same Order: In this option, the words are compared letter by letter, and if they have the same letters in the same order with at least 50% match, then that word is selected. This means, the length of the user typed word must be at least half of the selected word. With this option when "ALİ EVE GALDİ" is written, "ALİ EVE GELDİ" will be found where both last words have the same letters except the second letter.

To obtain closest match, combination of these rules should be used. The rules *i* and *ii* can be used together, as two words can have the same first letter and the same number of letters. But the rules *iii* and *iv* cannot be used together, because the assumption of the matching only letters is different than matching the letters in the same order. So, combination of all rules can be shown as: $i \ \& \ ii \ \& \ (iii \mid iv)$.

5.2.2. Selecting N-gram Model

By using *Options* window of the program, a user can select one of the models to be applied for spell checking process, which are:

i) Only Trigrams: In this option, only trigram list is searched. This means the first two words will be searched in the trigram list for a match. When first two words matched, a list of trigrams is obtained and depending on the matching criteria, which are previously selected by the user, the appropriate ones will be selected for the last

word. If no match is found, then the program will end with "no solution". This can be called as "Non-smoothing".

ii) Bigrams and Trigrams: In this option, bigram and trigram lists are searched. If no match is found for the first two words in the trigram list, then bigram list is searched for the match of the first word. When first word matched, a list of bigrams is obtained and according to the matching criteria, the appropriate ones will be selected for the second word of the bigram. This can be called as "Partial Back-off Smoothing".

iii) Monogram, Bigram and Trigram: In this option, monogram, bigram and trigram lists are searched. If there is no match for the trigram and bigram lists, then according to the matching criteria, the monogram list is searched for a word. This option can be called as "Full Back-off Smoothing".

5.2.3. Indexing

When using n-gram lists, the biggest concern is the speed of searching through lists. TurCo has 24,173,546 elements in trigram list, 11,998,590 elements in bigram list, and 686,804 elements in monogram list, resulting 36,858,940 elements. For the worst case the search space is nearly 37 million elements. The program should interact with user, so it has to be fast.

For searching, an index methodology is used like in ALG-2 in section 3.2.2. For monograms, an index according to the first letter is created. As the list of the monograms is alphabetically sorted, the index for Novels & Stories monograms is given in Table 5.4 under mono caption. Each number indicates the positions of the words starting with corresponding letter in the monogram file as bytes.

For bigrams and trigrams, a cascaded index structure is used. First, an index of the "first three letters" is formed. This is the first index (Table 5.4). Because the size of the first index is high, a second index is formed for the first index.

The second index (Table 5.4) is the index of the *First index* by using the first letters ("A", "B", ..., "Z"). This index shows the position of each letter in first index file in bytes. By using the first index, the word part "GEZ" is searched beginning from "GAA" which is the starting point of the word beginning with the letter "G" and has the position of 66,444 in the first index file. When "GEZ" is found in the first index, it points the position of itself in the bigram list which is 83,356,028 in bytes. For trigrams, this index structure is also used for and two indexes were created.

5.2.4. Experiments with Prototype Program

The easiest way to test a program is to compare it with a well-known spell checker for Turkish. The commonly used word processor is MS Word XP, which has spell and grammar checkers for English, and spell checker for Turkish.

Table 5.4 Monogram and Bigram Index Files for Novels & Stories

Mono		Bigram First Index		Bigram Second Index	
A	0	A	0	A	0
B	744329	AA	172781	B	13524
C	1501271	AAA	174205	C	23394
Ç	1675329	AAB	177285	Ç	32466
D	1953284	...		D	37191
E	2649476	BA	16924276	E	47292
F	3028468	BAA	16936040	F	59052
G	3202229	BAB	16936866	G	66444
Ğ	3775486	...		Ğ	75306
H	3776854	GEV	83356028	H	76314
I	4136434	GEY	83370966	I	84546
İ	4179888	GEZ	83376669	İ	92568
J	4600870	GF	83476568	J	103614
K	4620906	...		K	107793
L	5709384	ZZA	235842085	L	117516
M	5810505	ZZE	235842140	M	125370
N	6293060	ZZN	235842284	N	135324
O	6422733	ZZZ	235842295	O	143472
Ö	6666474			Ö	152901
P	6852686			P	157731
R	7128684			R	165816
S	7292868			S	173754
Ş	8127700			Ş	184275
T	8232428			T	188244
U	8826258			U	198828
Ü	9022707			Ü	207459
V	9102149			V	211680
Y	9266591			Y	217518
Z	9905549			Z	224553

To test the accuracy of the prototype program, the results obtained by using two different kinds of corpora are compared with the results obtained from MS Word XP. The suggested alternatives of the MS Word XP and the prototype program using Novels & Stories, and using TurCo are given in Table 5.5.

The prototype program gives a listing of the suggested words to the user. The total number of words to be listed is selected as a default value of 10, which means the first 10 words matching the criteria are listed. This value can be increased from the options menu for precision and can be decreased not to confuse the user.

Table 5.5 Suggestion Comparisons of MS Word XP with both Corpora

Text containing error	**Suggestions of**		
	MS Word XP	**Novels & Stories**	**TurCo**
MUTLU OLMAK İCİN	İÇİN CİN İDİN İĞİN İKİN	MUTLU OLMAK İÇİN	MUTLU OLMAK İÇİN
EVE ADIMINI ATMAZSCAKTIR	*No Suggestions*	EVE ADIMINI ATMAYACAKTIR	EVE ADIMINI ATMAYACAKTIR
DAĞLAR KADAR BÜYÜKDÜR	BÜYÜRDÜR BÜYÜMDÜR BÜYÜNDÜR BÜYÜKTÜR BÜYÜKÜDÜR	BÖYLEDİR BÜYÜKTÜR ...	DAĞLAR KADAR BÜYÜKTÜR
BİR ABD FİRMASÜNDEN	*No Suggestions*	FİLDİŞİNDEN FARMAŞONLAR FARLARINDAN FURMASINLAR FORMÜLÜNDEN FİRMASINDAN ...	BİR ABD FİRMASINDAN
MUHTEMEL AFETLERDE CSN	SN CAN CİN MSN	CAN CUN ...	MUHTEMEL AFETLERDE CAN
GÜNÜN ADAMI OLMU	OLCU OLDU OLLU OLSU OLUMU	OLAN OLDU OLUR OLSUN OLMAK OLSA OLUP OLMUŞ ...	GÜNÜN ADAMI OLMUŞ

When the first sentence in Table 5.5, "MUTLU OLMAK İCİN", where "İCİN" is wrong spelled, is considered; it is found that MS Word XP caught the correct answer with the first suggestion, and gave 4 more suggestions, but none of them could form a meaningful sentence. While MS Word XP offers many words, the prototype program suggests only one sentence, which is the correct answer. The matching criteria "First Letter Check" and "Word Length Comparison" are used in this example. Since the correct word is found by using only these criteria, there is no need to use others. Also, the program found "MUTLU OLMAK İÇİN" word list from trigram lists.

For the second sentence, "EVE ADIMINI ATMAZSCAKTIR", where "ATMAZSCAKTIR" is wrong spelled, there is no suggestion from MS Word XP, but there is again one suggestion from the program. The same criteria used for the first example are used. Again, there are no other criteria needed to be used, because the correct word is found by using only first two criteria.

For the first and second sentences, the program finds the same solutions regardless of the corpus.

For the third sentence, while using Novels and Stories and the options "First Letter Check" and "Word Length Comparison" and "Matching the Letters in the Same Order", the program suggests lots of words, because it can't find the trigram from the trigram list and also the bigram from the bigram list, so it gives the list of monograms where there are much more possibilities. Depending on this situation, the correct word "BÜYÜKTÜR" is found in the list as the second suggestion, while it exists in the fourth row of suggestions of MS Word XP.

When the TurCo is used, the result can be achieved easier, while the answer is only one sentence, which is "DAĞLAR KADAR BÜYÜKTÜR", and taken from trigram list.

The situations are nearly the same for the fourth and fifth examples. While using Novels & Stories, and the same matching criteria as the third example, the correct suggestions for the fourth example is the sixth, and for the fifth example is the first one. Better results are obtained for both examples while using prototype program rather than using MS Word XP, which even does not have any suggestion for the fourth example.

When TurCo is used, the results can be achieved easier as in example 3. The answers include only one sentence for each, which are taken from trigram list, respectively "BİR ABD FİRMASINDAN", and "MUHTEMEL AFETLERDE CAN".

For the sixth sentence, "First Letter Check" and "Matching the Letters in the Same Order" options are used. While using Novels & Stories, the program suggests lots of words, as in previous three examples, it can't find the trigram from the trigram list and also the bigram from the bigram list, so it gives the list of monograms where there are much more possibilities. Depending on this situation, the correct word "OLMUŞ" is found in the list as the eighth suggestion, while it does not exist in the suggestions of MS Word XP.

When the TurCo is used for the sixth sentence, the result can be achieved easier, while the answer is only one sentence, which is "GÜNÜN ADAMI OLMUŞ", and taken from trigram list.

Corpus size effect is observed on these examples. When TurCo is used for the n-gram list, the prototype program perfectly finds the correct answer as seen in Table 5.5. While the corpus size increases, and includes more word n-grams, the result can be achieved easier.

5.2.5. Enhancement by MED Algorithm

To increase the accuracy of the prototype program, and to achieve accurate results by using small sized corpora, Minimum Edit Distance (MED) algorithm can be used. MED describes the minimum number of editing operations (insertion, deletion,

substitution), needed to transform one string to another, between two strings (Jurafsky & Martin, 2000). The prototype program is enhanced by using this method with the operating costs of each "insertion", "deletion" and "substitution" assumed as "1".

A modification to MED was proposed by Levenshtein (Jurafsky & Martin, 2000). According to this approach, a substitution can be defined as a combination of deletion and insertion operations, so each substitution has to be counted as "2".

While using Novels and Stories, the third, fourth, fifth and sixth examples have more than one suggestion. For these examples MED algorithm can be used to determine the nearest solution.

For the third example (DAĞLAR KADAR BÜYÜKDÜR), the prototype program catches the correct answer in the second suggestion. In order to change "BÜYÜKDÜR" to "BÜYÜKTÜR", only one operation is needed to substitute "D" by "T" (MED Value=1). While changing the word "BÜYÜKDÜR" to "BÖYLEDİR", four operations are needed to substitute "Ü" to "Ö", "Ü" to "L", "K" to "E", and "Ü" to "İ" (MED Value=4). As a result, "BÜYÜKTÜR" has smaller MED value, so it can be selected as the correct answer. It can be easily seen that, the fourth example gives the best result with the word "FİRMASINDAN", which is the correct word, with a MED value of 2.

The fifth example cannot be enhanced by using MED algorithm, as the MED values of both alternatives are the same (MED Value=1).

The worst results are obtained in the sixth example. The results for the example 6 (GÜNÜN ADAMI OLMU), by using MED techniques, are given in Table 5.6. Because of the same MED values, there is a need for further enhancement and modified MED by Levenshtein is also applied

The lowest MED values for example 6 are obtained for "OLDU" and "OLMUŞ", both having the same MED value as 1 where substitution value is 1. MED values for the same words were calculated as respectively 2 and 1, by using the substitution value as 2. As a result, the word "OLMUŞ" is selected as the correct word.

Table 5.6 MED Values of the Example 6

Word	MED Value (Substitution=1)	MED Value (Substitution=2)
OLAN	2	4
OLDU	1	2
OLUR	2	2
OLSUN	2	3
OLMAK	2	3
OLSA	2	4
OLUP	2	2
OLMUŞ	1	1

CHAPTER SIX

CONCLUSION

6. CONCLUSION

Determination of a language's structure covers the main topics: Morphological analysis and Statistical analysis. In this study, structure of Turkish was analyzed on the basis of statistical analysis from the point of view of word analysis. Word analysis is part of the statistical analysis which covers the investigation of number of letters in a word, the order of letters in a word, and word orders in a sentence, word n-gram frequencies. Zipf's Law can be used for the estimation of word n-gram frequencies. Smoothing techniques are useful for the unknown n-gram frequencies.

The corpus used in this study having ~362MB capacity and including 50,111,828 words, is the biggest known corpus created for the Turkish language. As it includes different types of documents related with written and spoken Turkish, it is said to be a diverse and unbalanced corpus.

Working on a large scale corpus requires much CPU and memory power, and sufficient algorithms. During this study, different algorithms were developed to count word n-grams. It was seen that, programs using databases like MySQL are too slow for this kind of operations because of the general nature of databases. By using a specific and suitable algorithm, lots of time can be gained, and more detailed analysis can be done.

Number of Different Words (NODW) and Different Word Usage Ratio (DWUR) are some of the characteristics of a corpus. NODW values increase strictly from monograms to the bigrams, but the increase ratio shows an exponential characteristic from monograms to pentagrams.

Average word length of Turkish is 6.23 letters (TurCo). 5-letter words have the highest percentage of 15.06. 1 to 10 letter words form 90.80% of TurCo. The most frequently used 100 words form the 22.47% of the corpus. The most frequently used words (top 7) in Turkish corpus are not "NOUN". In English, these words can be found in top 53 (Garett, 2001).

Turkish, a Ural-Altaic language, is one of the key languages of the world (Schultz & Waibel, 1997). It is an agglutinative language like Korean (Choi, 2000) where lots of words can be created by adding suffixes to words. For example, the longest words from TurCo are: "etkileşimleştirdiklerimizden, görevlendirilebilmektedirler, toplumsallaştırılmadıklarını". They have the roots "etkileş", "görev", and "toplum". All the other parts of the words are the suffixes. Having this kind of long words formed by adding lots of suffixes makes the Turkish rich, and that's why the number of different words are higher than English.

Turkish obeys Zipf's Law with some modifications to Mandelbrot's Law (Appendix B). For Turkish, B value in Mandelbrot's Law was found smaller than 1, like Korean (Choi, 2000) unlike English. Information temperature (1/B), as defined by Mandelbrot, shows the variation of a vocabulary. As 1/B increases, the vocabulary more varies. The agglutinative behavior of Turkish makes the curve (Figure 4.7) to bow like Korean (Choi, 2000).

After going over different smoothing techniques like Add-one, Good-Turing and Back-off, Back-Off technique was selected, and a prototype program, based on this technique and MED algorithm, was prepared. The results obtained from the program and MS Word XP were compared, and it was observed that more accurate results had been obtained from the prototype program.

The results of the prototype program with and without the MED algorithm were compared, and it was seen that MED algorithm added significant improvements as giving more accurate results. MED algorithm can be used regardless the smoothing technique chosen.

The success of the prototype program lies with the specific algorithms developed and the corpus size (TurCo). TurCo includes some non-Turkish words and spelling errors, and they can be eliminated, but this process may not be useful, because in Turkish lots of non-Turkish words are used and some spelling errors are done. Used corpus mainly consists of web sites, so the hyperlinks and duplicated words, such as menu items repeated on each page, can also be eliminated.

The detailed results obtained from the corpus analysis can be additionally used in author identification, grammar checking and correction in Turkish. The results of the analysis can also be used in cryptanalytical procedures, text compression (Dalkılıç, 2001), character recognition operations and, is always important for spelling corrections (Church & Gale, 1991a). Also they can be helpful for checking and correcting errors in fax-to-voice devices and phonetic-transcription that are used for disabled persons (Kukich, 1992).

In the future, the positions of the words in the sentence can be investigated and the results can be used to increase the ratio of text correction. A large-scale balanced corpus can be created and unwanted duplicated items coming from web sites can be eliminated. Corpus size can be increased by adding new sites including novels, technical papers, written reports, thesis etc. At the same time, with the classification of these documents, different corpora of different fields can be generated, e.g. medical corpus, engineering corpus, etc.

REFERENCES

Choi, S.W. (2000). Some Statistical Properties and Zipf's Law in Korean Text Corpus. Journal of Quantitative Linguistics, 7:1, pp. 19-30.

Chunyu, K., & Yorick W. (1998). The virtual corpus approach to deriving ngram statistics from large scale corpora. Proceedings of International Conference on Chinese Information Processing Conference, Beijing.

Church, K., and Gale, W. (1989). Enhanced Good-Turing and Cat-Cal: Two New Methods for Estimating Probabilities of English Bigrams. Second Darpa Workshop on Speech and Natural Language.

Church, K. & Gale, W. (1991a). A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probability of English Bigrams. Computer Speech and Language, 5, pp.19-54

Church, K. & Gale, W. (1991b). Probability Scoring for Spelling Correction. Statistics and Computing, pp.93-103.

Church, K. & Mercer, R. (1993). Introduction to the Special Issue on Computational Linguistics Using Large Corpora. Computational Linguistics, 19:1, pp. 1-24.

Dalkılıç, G. (2001). Some Statistical Properties of Contemporary Printed Turkish and A Text Compression Application. MSc Thesis. International Computing Institute, Ege University.

Dalkılıç, M.E. & Dalkılıç, G. (2001). Some Measurable Language Characteristics of Printed Turkish. Proc. of the XVI. International Symposium on Computer and Information Sciences, pp. 217-224.

Diri, B. (2000). A Text Compression System Based on the Morphology of Turkish Language. Proc. of the XV International. Symposium on Computer and Information Sciences, pp. 12-23.

Gale, W. A. & Church, K. W. (1994). What is wrong with adding one?. Corpus-based Research into Language, pp. 189-198.

Garett, P. (2001). Making Breaking Codes, Prentice Hall, pp. 31-36.

Good, I. J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. Biometrika, 40: ¾, pp. 237-264.

Güngör, T. (1995). Computer Processing of Turkish: Morphological and Lexical Investigation. PhD. Dissertation. Computer Engineering Dept., Boğaziçi University.

Manning, C. D. & Schütze H. (2000). Foundations of Statistical Natural Language Processing, The MIT press, pp. 119.

Hanks, P. (1990). Evidence and Intuition in Lexicography, in J. Tomaszczyk and B. Lewandowska - Tomaszczyk (eds.). Meaning and Lexicography, p. 36.

Huang F. & Yu M. (2001). Analyzing the properties of smoothing methods for language models. IEEE International Conference on Systems, Man, and Cybernetics, Vol:1, pp. 512-517.

Jurafsky, D. & Martin, J.H. (2000). Speech and Language Processing, Prentice Hall, pp. 193-199.

Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustic, Speech, and Signal Processing, 35:3, pp. 400-401.

Kneser, R. & Ney, H. (1995). Improved backing-off for M-gram language modeling. International Conference on Acoustics, Speech, and Signal Processing, 1, pp. 181-184.

Koltuksuz, A. (1995). Simetrik Kriptosistemler için Türkiye Türkçesinin Kriptanalitik Ölçütleri. PhD. Dissertation, Computer Eng. Dept., Ege University, İzmir, Turkey.

Kukich K. (1992). Technique for automatically correcting words in text. Periodical Issue Article of ACM Press, pp.377-439.

Le Quan Ha, Sicilia-Garcia E. I., Ji Ming; Smith F. J. (2002). Extension of Zipf's Law to Words and Phrases. The 17th International Conference on Computational Linguistics.

Martin, S.C., Ney, H. & Zaplo, J. (1999). Smoothing methods in maximum entropy language modeling. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1, pp. 545-548

Miller G.A., Newman E.B., Fiedman, E.A. (1957). Some effects of intermittent silence. American Journal of Psychology, 70, pp. 311-313.

Nadas, A. (1984). Estimation of probabilities in the language model of the IBM speech recognition system. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32:4, pp. 859-861.

O'Boyle, P., Ming, J., McMahon, J., & Smith, F.J. (1996). Improving n-gram models by incorporating enhanced distributions. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1:7, pp. 168-171.

Oflazer, K., Göçmen E., & Bozşahin C. (1994). An Outline of Turkish Morphology. Report on Turkish Natural Language Processing Initiative Project.

Robbins J.L., Daneman J.C. (2002). The Best Known Statistic – Coefficient of Determination (r2): What it does and does not do in Regression Analysis. National Estimator Fall 2002. pp. 18-27.

Schultz, T. & Waibel, A. (1997). Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets. Proceedings of EuroSpeech '97.

Senda, S., & Yamada, K. (2001). A maximum-likelihood approach to segmentation-based recognition of unconstrained handwriting text. Sixth International Conference on Document Analysis and Recognition. pp. 184 -188.

Shannon, C.E. (1948). A Mathematical Theory of Communication. The Bell System Technical Journal. vol:27, pp.379-423, pp. 623-656.

Shannon, C.E. (1951). Prediction and Entropy of Printed English. The Bell System Technical Journal, 30:1, pp. 50-64.

Song F. & Croft W. B. (1999). A General Language Model for Information Retrieval, 316-321 ACM Press Proceedings, pp.316-321.

Teahan, W.J. (1998). Modeling English Text. Ph.D. Dissertation, The Univ. of Waikato, New Zealand, pp. 50-52, 64-69.

Whittaker, E.W.D. & Woodland P. C. (2003, January). Language modeling for Russian and English using words and classes. Computer Speech & Language, pp.87-104.

Witten, I. H. & Bell, T.C. (1990). Source models for natural language text. Int J Man-Machine Studies, 32:5, pp. 545-579.

Witten, I. H. & Bell, T.C. (1991). The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression, IEEE Transactions on Information Theory, 37:4, pp.1085-1094

Zipf, G.K. (1949). Human behavior and the principle of least effort. Reading, MA. Addison-Wesley Publishing. USA.

WEB_1, (2003). British National Corpus (BNC). http://www.hcu.ox.ac.uk/BNC/what/index.html, 02/01/2003.

WEB_2, (2003). The Bank of English. http://www.cobuild.collins.co.uk/boe_info.html, 02/01/2003.

WEB_3, (2003). English Gigaword. http://www.ldc.upenn.edu/Catalog/ Catalog Entry. jsp?catalogId=LDC200 3T05, 02/01/2003.

WEB_4, (2003). The Czech National Corpus (CNC). http://ucnk.ff.cuni.cz/english/index.html, 02/01/2003.

WEB_5, (2003). Croatian National Corpus. http://www.hnk.ffzg.hr/corpus. htm, 02/01/2003.

WEB_6, (2003). PAROLE. http://www.elda.fr/cata/text/doc/parole.html, 02/01/2003.

WEB_7, (2003). French Corpus. http://www.elda.fr/cata/text/W0020.html, 02/01/2003.

WEB_8, (2003). COSMAS, German Corpus. http://corpora.ids-mannheim.de/~cosmas/, 02/01/2003.

WEB_9, (2004). METU Corpus. http://www.ii.metu.edu.tr/~corpus/corpus.html, 04/14/2004.

WEB_10, (2004). Dynamic Programming Lecture Notes. http://www-courses.cs.uiuc.edu/~cs373u/, 04/26/2004.

APPENDICES

A. First 20 Word N-Gram (1≤n≤5) Frequencies

A.1. Arabul

Rank	Mono	Freq.	%	Bi		Freq.	%	Tri			Freq.	%
1	yeni	42971	5.702	yeni	pencerede	42751	5.673	yeni	pencerede	aç	42750	5.673
2	pencerede	42752	5.673	pencerede	aç	42750	5.673	com	yeni	pencerede	16204	2.150
3	aç	42750	5.673	com	yeni	16204	2.150	tr	yeni	pencerede	6619	0.878
4	http	42683	5.664	arası	kayıtlar	10602	1.407	net	yeni	pencerede	5563	0.738
5	com	29567	3.924	tr	yeni	6629	0.880	arabul	un	seçtikleri	5322	0.706
6	ve	16069	2.132	com	tr	5754	0.764	basketbol	gezi	rehberleri	5322	0.706
7	tr	10722	1.423	cep	telefonu	5612	0.745	cep	telefonu	eb	5322	0.706
8	arası	10624	1.410	net	yeni	5563	0.738	e	gazeteler	göçmenlik	5322	0.706
9	kayıtlar	10602	1.407	eb	programlama	5409	0.718	eb	programlama	magazin	5322	0.706
10	net	7952	1.055	arabul	un	5338	0.708	gazeteler	göçmenlik	cep	5322	0.706
11	eb	7843	1.041	e	gazeteler	5335	0.708	gezi	rehberleri	e	5322	0.706
12	e	6542	0.868	gezi	rehberleri	5334	0.708	göçmenlik	cep	telefonu	5322	0.706
13	cep	5763	0.765	rehberleri	e	5324	0.707	rehberleri	e	gazeteler	5322	0.706
14	programlama	5726	0.760	basketbol	gezi	5322	0.706	seçtikleri	basketbol	gezi	5322	0.706
15	telefonu	5616	0.745	gazeteler	göçmenlik	5322	0.706	telefonu	eb	programlama	5322	0.706
16	rehberleri	5582	0.741	göçmenlik	cep	5322	0.706	un	seçtikleri	basketbol	5322	0.706
17	magazin	5562	0.738	programlama	magazin	5322	0.706	arası	kayıtlar	arabul	5295	0.703
18	basketbol	5538	0.735	seçtikleri	basketbol	5322	0.706	kayıtlar	arabul	un	5286	0.701
19	arabul	5532	0.734	telefonu	eb	5322	0.706	magazin	ana	sayfa	5272	0.700
20	un	5479	0.727	un	seçtikleri	5322	0.706	programlama	magazin	ana	5272	0.700

A.1. Arabul (Cont'd)

Rank	Tetra				Freq.	%
1	com	yeni	pencerede	aç	16204	2.150
2	tr	yeni	pencerede	aç	6619	0.878
3	net	yeni	pencerede	aç	5563	0.738
4	arabul	un	seçtikleri	basketbol	5322	0.706
5	basketbol	gezi	rehberleri	e	5322	0.706
6	cep	telefonu	eb	programlama	5322	0.706
7	e	gazeteler	göçmenlik	cep	5322	0.706
8	gazeteler	göçmenlik	cep	telefonu	5322	0.706
9	gezi	rehberleri	e	gazeteler	5322	0.706
10	göçmenlik	cep	telefonu	eb	5322	0.706
11	rehberleri	e	gazeteler	göçmenlik	5322	0.706
12	seçtikleri	basketbol	gezi	rehberleri	5322	0.706
13	telefonu	eb	programlama	magazin	5322	0.706
14	un	seçtikleri	basketbol	gezi	5322	0.706
15	arası	kayıtlar	arabul	un	5286	0.701
16	kayıtlar	arabul	un	seçtikleri	5286	0.701
17	eb	programlama	magazin	ana	5272	0.700
18	programlama	magazin	ana	sayfa	5272	0.700
19	pencerede	aç	arası	kayıtlar	4966	0.659
20	yeni	pencerede	aç	arası	4966	0.659

Rank	Penta					Freq.	%
1	arabul	un	seçtikleri	basketbol	gezi	5322	0.7062
2	basketbol	gezi	rehberleri	e	gazeteler	5322	0.7062
3	cep	telefonu	eb	programlama	magazin	5322	0.7062
4	e	gazeteler	göçmenlik	cep	telefonu	5322	0.7062
5	gazeteler	göçmenlik	cep	telefonu	eb	5322	0.7062
6	gezi	rehberleri	e	gazeteler	göçmenlik	5322	0.7062
7	göçmenlik	cep	telefonu	eb	programlama	5322	0.7062
8	rehberleri	e	gazeteler	göçmenlik	cep	5322	0.7062
9	seçtikleri	basketbol	gezi	rehberleri	e	5322	0.7062
10	un	seçtikleri	basketbol	gezi	rehberleri	5322	0.7062
11	arası	kayıtlar	arabul	un	seçtikleri	5286	0.7015
12	kayıtlar	arabul	un	seçtikleri	basketbol	5286	0.7015
13	eb	programlama	magazin	ana	sayfa	5272	0.6996
14	telefonu	eb	programlama	magazin	ana	5272	0.6996
15	yeni	pencerede	aç	arası	kayıtlar	4966	0.6590
16	pencerede	aç	arası	kayıtlar	arabul	4965	0.6589
17	aç	arası	kayıtlar	arabul	un	4961	0.6583
18	com	tr	yeni	pencerede	aç	4594	0.6096
19	kayıt	site	arası	kayıtlar	siteler	3259	0.4325
20	cjb	net	yeni	pencerede	aç	2984	0.3960

A.2. Bilim Teknoloji

Rank	Mono	Freq.	%	Bi		Freq.	%	Tri			Freq.	%
1	bir	4369	2.146	bilim	ve	414	0.203	bilim	ve	teknoloji	372	0.183
2	ve	4073	2.000	ve	teknoloji	375	0.184	uçak	ve	uzay	234	0.115
3	bu	2706	1.329	on	ednesday	304	0.149	uzay	mars	araştırmaları	223	0.110
4	on	1782	0.875	on	tuesday	290	0.142	ve	uzay	tek	223	0.110
5	by	1750	0.859	ve	uzay	255	0.125	araştırmaları	uzay	teleskopları	217	0.107
6	uzay	1602	0.787	on	sunday	252	0.124	bilim	dünyada	türkiyede	217	0.107
7	bilim	1423	0.699	on	thursday	238	0.117	bilim	tarihi	teknoloji	217	0.107
8	de	1067	0.524	uçak	ve	237	0.116	dünyada	türkiyede	bilim	217	0.107
9	da	1026	0.504	on	monday	236	0.116	elektronik	bilgisayar	uçak	217	0.107
10	için	976	0.479	uzay	teleskopları	233	0.114	genel	sağlık	genetik	217	0.107
11	am	964	0.473	mars	araştırmaları	227	0.111	genel	teknoloji	uzay	217	0.107
12	daha	944	0.464	ev	teknolojileri	226	0.111	genetik	kanser	diyabet	217	0.107
13	teknoloji	920	0.452	genel	teknoloji	225	0.111	kanser	diyabet	genel	217	0.107
14	genel	821	0.403	on	saturday	225	0.111	mars	araştırmaları	uzay	217	0.107
15	pm	817	0.401	bilim	tarihi	224	0.110	otomotiv	ev	teknolojileri	217	0.107
16	çok	806	0.396	uzay	mars	223	0.110	sağlık	genetik	kanser	217	0.107
17	ile	787	0.387	uzay	tek	223	0.110	tarihi	teknoloji	elektronik	217	0.107
18	olarak	764	0.375	sağlık	genetik	222	0.109	teknoloji	elektronik	bilgisayar	217	0.107
19	a	760	0.373	teknoloji	elektronik	222	0.109	teknoloji	uzay	mars	217	0.107
20	ne	680	0.334	türkiyede	bilim	221	0.109	teleskopları	genel	sağlık	217	0.107

A.2. Bilim Teknoloji (Cont'd)

Rank	Tetra				Freq.	%
1	uçak	ve	uzay	tek	223	0.110
2	araştırmaları	uzay	teleskopları	genel	217	0.107
3	bilim	dünyada	türkiyede	bilim	217	0.107
4	bilim	tarihi	teknoloji	elektronik	217	0.107
5	dünyada	türkiyede	bilim	tarihi	217	0.107
6	genel	sağlık	genetik	kanser	217	0.107
7	genel	teknoloji	uzay	mars	217	0.107
8	genetik	kanser	diyabet	genel	217	0.107
9	mars	araştırmaları	uzay	teleskopları	217	0.107
10	sağlık	genetik	kanser	diyabet	217	0.107
11	tarihi	teknoloji	elektronik	bilgisayar	217	0.107
12	teknoloji	elektronik	bilgisayar	uçak	217	0.107
13	teknoloji	uzay	mars	araştırmaları	217	0.107
14	teleskopları	genel	sağlık	genetik	217	0.107
15	türkiyede	bilim	tarihi	teknoloji	217	0.107
16	uzay	mars	araştırmaları	uzay	217	0.107
17	uzay	teleskopları	genel	sağlık	217	0.107
18	ev	teknolojileri	malzeme	genel	216	0.106
19	malzeme	genel	teknoloji	uzay	216	0.106
20	otomotiv	ev	teknolojileri	malzeme	216	0.106

Rank	Penta					Freq.	%
1	araştırmaları	uzay	teleskopları	genel	sağlık	217	0.1066
2	bilim	dünyada	türkiyede	bilim	tarihi	217	0.1066
3	bilim	tarihi	teknoloji	elektronik	bilgisayar	217	0.1066
4	dünyada	türkiyede	bilim	tarihi	teknoloji	217	0.1066
5	genel	sağlık	genetik	kanser	diyabet	217	0.1066
6	genel	teknoloji	uzay	mars	araştırmaları	217	0.1066
7	mars	araştırmaları	uzay	teleskopları	genel	217	0.1066
8	sağlık	genetik	kanser	diyabet	genel	217	0.1066
9	tarihi	teknoloji	elektronik	bilgisayar	uçak	217	0.1066
10	teknoloji	uzay	mars	araştırmaları	uzay	217	0.1066
11	teleskopları	genel	sağlık	genetik	kanser	217	0.1066
12	türkiyede	bilim	tarihi	teknoloji	elektronik	217	0.1066
13	uzay	mars	araştırmaları	uzay	teleskopları	217	0.1066
14	uzay	teleskopları	genel	sağlık	genetik	217	0.1066
15	ev	teknolojileri	malzeme	genel	teknoloji	216	0.1061
16	malzeme	genel	teknoloji	uzay	mars	216	0.1061
17	otomotiv	ev	teknolojileri	malzeme	genel	216	0.1061
18	teknolojileri	malzeme	genel	teknoloji	uzay	216	0.1061
19	bilgisayar	uçak	ve	uzay	tek	215	0.1056
20	elektronik	bilgisayar	uçak	ve	uzay	215	0.1056

A.3. Devlet İstatistik Enstitüsü

Rank	Mono	Freq.	%	Bi		Freq.	%	Tri			Freq.	%
1	ve	38647	4.076	b	c	4739	0.500	bir	önceki	yılın	2619	0.276
2	göre	15241	1.608	bir	önceki	4364	0.460	erkek	kad	n	2247	0.237
3	ile	9183	0.969	say	s	4225	0.446	a	b	c	2202	0.232
4	bir	7652	0.807	a	b	2795	0.295	önceki	yılın	aynı	1867	0.197
5	i	7185	0.758	önceki	yılın	2622	0.277	b	c	d	1452	0.153
6	yılı	6485	0.684	kad	n	2579	0.272	önceki	aya	göre	1393	0.147
7	b	6415	0.677	yılın	aynı	2316	0.244	aynı	ayına	göre	1360	0.143
8	k	6355	0.670	erkek	kad	2247	0.237	devlet	istatistik	enstitüsü	1311	0.138
9	c	6018	0.635	l	k	2155	0.227	bir	önceki	aya	1267	0.134
10	s	5791	0.611	ayına	göre	2061	0.217	yılın	aynı	ayına	1162	0.123
11	tablo	5244	0.553	en	yüksek	1461	0.154	sa	l	k	1089	0.115
12	l	5156	0.544	aya	göre	1460	0.154	ekonomik	ve	sosyal	1032	0.109
13	a	4969	0.524	ve	sosyal	1457	0.154	istatistik	enstitüsü	başkanlığı	1020	0.108
14	ise	4937	0.521	c	d	1456	0.154	ve	sosyal	göstergeler	1015	0.107
15	önceki	4749	0.501	önceki	aya	1411	0.149	temel	ekonomik	ve	1008	0.106
16	say	4694	0.495	imalat	sanayi	1379	0.145	eline	geçen	fiyatlar	992	0.105
17	kişi	4671	0.493	aynı	ayına	1375	0.145	ula	t	rma	944	0.100
18	e	4479	0.472	temel	yıllı	1327	0.140	aynı	dönemine	göre	874	0.092
19	artış	4424	0.467	devlet	istatistik	1318	0.139	çiftçinin	eline	geçen	859	0.091
20	ayında	4381	0.462	istatistik	enstitüsü	1315	0.139	yılın	aynı	dönemine	830	0.088

A.3. Devlet İstatistik Enstitüsü (Cont'd)

Rank	Tetra				Freq.	%
1	bir	önceki	yılın	aynı	1864	0.197
2	bir	önceki	aya	göre	1249	0.132
3	yılın	aynı	ayına	göre	1159	0.122
4	önceki	yılın	aynı	ayına	1066	0.112
5	devlet	istatistik	enstitüsü	başkanlığı	1017	0.107
6	ekonomik	ve	sosyal	göstergeler	1015	0.107
7	temel	ekonomik	ve	sosyal	1008	0.106
8	çiftçinin	eline	geçen	fiyatlar	853	0.090
9	yılın	aynı	dönemine	göre	823	0.087
10	erkek	kad	n	erkek	799	0.084
11	kad	n	erkek	kad	799	0.084
12	n	erkek	kad	n	799	0.084
13	elektrik	gaz	ve	su	785	0.083
14	lokanta	pastane	ve	otel	724	0.076
15	eline	geçen	fiyatlar	indeksi	636	0.067
16	göre	bir	önceki	yılın	636	0.067
17	içki	ve	tütün	giyim	624	0.066
18	ve	tütün	giyim	ve	624	0.066
19	e	lence	ve	kültür	617	0.065
20	çe	itli	mal	ve	613	0.065

Rank	Penta					Freq.	%
1	bir	önceki	yılın	aynı	ayına	1063	0.1121
2	önceki	yılın	aynı	ayına	göre	1063	0.1121
3	temel	ekonomik	ve	sosyal	göstergeler	1008	0.1063
4	erkek	kad	n	erkek	kad	799	0.0843
5	kad	n	erkek	kad	n	799	0.0843
6	n	erkek	kad	n	erkek	752	0.0793
7	içki	ve	tütün	giyim	ve	624	0.0658
8	e	lence	ve	kültür	e	613	0.0647
9	l	k	ula	t	rma	613	0.0647
10	sa	l	k	ula	t	613	0.0647
11	ayakkab	konut	ev	e	yas	612	0.0645
12	çe	itli	mal	ve	hizmetler	612	0.0645
13	e	yas	sa	l	k	612	0.0645
14	ev	e	yas	sa	l	612	0.0645
15	konut	ev	e	yas	sa	612	0.0645
16	ve	ayakkab	konut	ev	e	612	0.0645
17	yas	sa	l	k	ula	612	0.0645
18	da	içki	ve	tütün	giyim	611	0.0644
19	e	itim	lokanta	pastane	ve	611	0.0644
20	g	da	içki	ve	tütün	611	0.0644

A.4. Hürriyet

Rank	Mono	Freq.	%	Bi		Freq.	%	Tri			Freq.	%
1	ve	137943	1.465	bize	ulaşın	37957	0.403	arşiv	bize	ulaşın	37815	0.402
2	bir	118489	1.258	son	dakika	37920	0.403	yazarlar	kültür	sanat	32516	0.345
3	bu	106150	1.127	arşiv	bize	37815	0.402	politika	dünya	ekonomi	32493	0.345
4	da	74010	0.786	kültür	sanat	34741	0.369	gündem	politika	dünya	32462	0.345
5	de	72667	0.772	özel	dosyalar	32666	0.347	durumu	astronet	televizyon	32255	0.343
6	için	54076	0.574	gündem	politika	32546	0.346	hava	durumu	astronet	32255	0.343
7	son	52250	0.555	politika	dünya	32518	0.345	magazin	özel	dosyalar	32254	0.343
8	dünya	49151	0.522	dünya	ekonomi	32516	0.345	özel	dosyalar	gezi	32254	0.343
9	tüm	44488	0.472	yazarlar	kültür	32516	0.345	kültür	sanat	magazin	30035	0.319
10	en	41548	0.441	spor	yaşam	32461	0.345	sanat	magazin	özel	29774	0.316
11	dakika	40701	0.432	hava	durumu	32319	0.343	dünya	ekonomi	spor	29329	0.311
12	ekonomi	40499	0.430	astronet	televizyon	32255	0.343	ekonomi	spor	yaşam	29262	0.311
13	gün	40065	0.426	durumu	astronet	32255	0.343	arşivim	mesaj	grupları	18989	0.202
14	bize	39478	0.419	tüm	haberler	32255	0.343	grupları	sohbet	yardım	18989	0.202
15	spor	39191	0.416	dosyalar	gezi	32254	0.343	mesaj	grupları	sohbet	18989	0.202
16	yardım	39117	0.415	magazin	özel	32254	0.343	sohbet	yardım	üyelik	18989	0.202
17	özel	38511	0.409	sanat	magazin	30035	0.319	dakika	arşivim	mesaj	18988	0.202
18	ulaşın	37957	0.403	ekonomi	spor	29358	0.312	son	dakika	arşivim	18988	0.202
19	arşiv	37935	0.403	mesaj	grupları	20228	0.215	üyelik	site	haritası	18988	0.202
20	kültür	37922	0.403	üye	olmak	19025	0.202	yardım	üyelik	site	18988	0.202

A.4. Hürriyet (Cont'd)

Rank	Tetra				Freq.	%
1	gündem	politika	dünya	ekonomi	32439	0.345
2	hava	durumu	astronet	televizyon	32255	0.343
3	magazin	özel	dosyalar	gezi	32254	0.343
4	yazarlar	kültür	sanat	magazin	30012	0.319
5	kültür	sanat	magazin	özel	29774	0.316
6	sanat	magazin	özel	dosyalar	29774	0.316
7	politika	dünya	ekonomi	spor	29306	0.311
8	dünya	ekonomi	spor	yaşam	29239	0.311
9	arşivim	mesaj	grupları	sohbet	18989	0.202
10	grupları	sohbet	yardım	üyelik	18989	0.202
11	mesaj	grupları	sohbet	yardım	18989	0.202
12	dakika	arşivim	mesaj	grupları	18988	0.202
13	sohbet	yardım	üyelik	site	18988	0.202
14	son	dakika	arşivim	mesaj	18988	0.202
15	yardım	üyelik	site	haritası	18988	0.202
16	ana	sayfa	son	dakika	18830	0.200
17	arama	arşiv	bize	ulaşın	18830	0.200
18	arşiv	bize	ulaşın	yardım	18830	0.200
19	arşivim	arama	arşiv	bize	18830	0.200
20	astronet	televizyon	insan	kaynakları	18830	0.200

Rank	Penta					Freq.	%
1	kültür	sanat	magazin	özel	dosyalar	29774	0.3162
2	sanat	magazin	özel	dosyalar	gezi	29774	0.3162
3	yazarlar	kültür	sanat	magazin	özel	29774	0.3162
4	gündem	politika	dünya	ekonomi	spor	29252	0.3107
5	politika	dünya	ekonomi	spor	yaşam	29216	0.3103
6	arşivim	mesaj	grupları	sohbet	yardım	18989	0.2017
7	mesaj	grupları	sohbet	yardım	üyelik	18989	0.2017
8	dakika	arşivim	mesaj	grupları	sohbet	18988	0.2017
9	grupları	sohbet	yardım	üyelik	site	18988	0.2017
10	sohbet	yardım	üyelik	site	haritası	18988	0.2017
11	son	dakika	arşivim	mesaj	grupları	18988	0.2017
12	ana	sayfa	son	dakika	tüm	18830	0.2000
13	arama	arşiv	bize	ulaşın	yardım	18830	0.2000
14	arşiv	bize	ulaşın	yardım	copyright	18830	0.2000
15	arşivim	arama	arşiv	bize	ulaşın	18830	0.2000
16	astronet	televizyon	insan	kaynakları	arşivim	18830	0.2000
17	bilim	teknoloji	yazarlar	kültür	sanat	18830	0.2000
18	dakika	tüm	haberler	gündem	politika	18830	0.2000
19	dosyalar	gezi	piyasanet	hava	durumu	18830	0.2000
20	durumu	astronet	televizyon	insan	kaynakları	18830	0.2000

A.5. Lazland

Rank	Mono	Freq.	%	Bi		Freq.	%	Tri			Freq.	%
1	bir	2982	2.200	e	mail	445	0.328	arkadaşınızın	e	mail	432	0.319
2	ve	2227	1.643	arkadaşınızın	e	432	0.319	e	mail	adresi	219	0.162
3	bu	1699	1.254	bu	yazıyı	301	0.222	paylaşmak	istemez	misiniz	218	0.161
4	rumuz	1616	1.192	istemez	misiniz	222	0.164	adınız	ve	soyadınız	217	0.160
5	de	979	0.722	adınız	ve	219	0.162	adınızı	ve	soyadınızı	216	0.159
6	da	852	0.629	mail	adresi	219	0.162	adresini	de	yazıp	216	0.159
7	ne	827	0.610	paylaşmak	istemez	218	0.161	arkadaşlarınıza	gönderin	adınız	216	0.159
8	ama	802	0.592	mail	adresini	217	0.160	arkadaşlarınızla	paylaşmak	istemez	216	0.159
9	o	752	0.555	ve	soyadınız	217	0.160	de	yazıp	gönder	216	0.159
10	ben	636	0.469	adınızı	ve	216	0.159	e	mail	adresini	216	0.159
11	e	603	0.445	adresi	bu	216	0.159	gönder	butonuna	basınız	216	0.159
12	kadar	592	0.437	adresini	de	216	0.159	gönderin	adınız	ve	216	0.159
13	için	583	0.430	arkadaşlarınıza	gönderin	216	0.159	istemez	misiniz	adınızı	216	0.159
14	çok	574	0.424	arkadaşlarınızla	paylaşmak	216	0.159	mail	adresi	bu	216	0.159
15	seni	474	0.350	butonuna	basınız	216	0.159	mail	adresini	de	216	0.159
16	mail	472	0.348	de	yazıp	216	0.159	misiniz	adınızı	ve	216	0.159
17	daha	458	0.338	gönder	butonuna	216	0.159	soyadınız	arkadaşınızın	e	216	0.159
18	gibi	445	0.328	gönderin	adınız	216	0.159	soyadınızı	yazın	arkadaşınızın	216	0.159
19	ya	442	0.326	misiniz	adınızı	216	0.159	ve	soyadınız	arkadaşınızın	216	0.159
20	arkadaşınızın	434	0.320	soyadınız	arkadaşınızın	216	0.159	ve	soyadınızı	yazın	216	0.159

A.5. Lazland (Cont'd)

Rank	Tetra				Freq.	%
1	adınız	ve	soyadınız	arkadaşınızın	216	0.159
2	adınızı	ve	soyadınızı	yazın	216	0.159
3	adresini	de	yazıp	gönder	216	0.159
4	arkadaşınızın	e	mail	adresi	216	0.159
5	arkadaşınızın	e	mail	adresini	216	0.159
6	arkadaşlarınıza	gönderin	adınız	ve	216	0.159
7	arkadaşlarınızla	paylaşmak	istemez	misiniz	216	0.159
8	de	yazıp	gönder	butonuna	216	0.159
9	e	mail	adresi	bu	216	0.159
10	e	mail	adresini	de	216	0.159
11	gönderin	adınız	ve	soyadınız	216	0.159
12	istemez	misiniz	adınızı	ve	216	0.159
13	mail	adresini	de	yazıp	216	0.159
14	misiniz	adınızı	ve	soyadınızı	216	0.159
15	paylaşmak	istemez	misiniz	adınızı	216	0.159
16	soyadınız	arkadaşınızın	e	mail	216	0.159
17	soyadınızı	yazın	arkadaşınızın	e	216	0.159
18	ve	soyadınız	arkadaşınızın	e	216	0.159
19	ve	soyadınızı	yazın	arkadaşınızın	216	0.159
20	yazın	arkadaşınızın	e	mail	216	0.159

Rank	Penta					Freq.	%
1	adınız	ve	soyadınız	arkadaşınızın	e	216	0.1594
2	adınızı	ve	soyadınızı	yazın	arkadaşınızın	216	0.1594
3	adresini	de	yazıp	gönder	butonuna	216	0.1594
4	arkadaşınızın	e	mail	adresi	bu	216	0.1594
5	arkadaşınızın	e	mail	adresini	de	216	0.1594
6	arkadaşlarınıza	gönderin	adınız	ve	soyadınız	216	0.1594
7	arkadaşlarınızla	paylaşmak	istemez	misiniz	adınızı	216	0.1594
8	de	yazıp	gönder	butonuna	basınız	216	0.1594
9	e	mail	adresini	de	yazıp	216	0.1594
10	gönderin	adınız	ve	soyadınız	arkadaşınızın	216	0.1594
11	istemez	misiniz	adınızı	ve	soyadınızı	216	0.1594
12	mail	adresini	de	yazıp	gönder	216	0.1594
13	misiniz	adınızı	ve	soyadınızı	yazın	216	0.1594
14	paylaşmak	istemez	misiniz	adınızı	ve	216	0.1594
15	soyadınız	arkadaşınızın	e	mail	adresi	216	0.1594
16	soyadınızı	yazın	arkadaşınızın	e	mail	216	0.1594
17	ve	soyadınız	arkadaşınızın	e	mail	216	0.1594
18	ve	soyadınızı	yazın	arkadaşınızın	e	216	0.1594
19	yazın	arkadaşınızın	e	mail	adresini	216	0.1594
20	basınız	dostlarınız	da	sizin	kahkalarınıza	165	0.1218

A.6. Pankitap

Rank	Mono	Freq.	%	Bi		Freq.	%	Tri			Freq.	%
1	bir	1529	2.596	pan	yayıncılık	329	0.559	eb	tasarımı	ebi	199	0.338
2	ve	1461	2.481	eb	tasarımı	199	0.338	hakkı	saklıdır	eb	199	0.338
3	bu	761	1.292	hakkı	saklıdır	199	0.338	her	hakkı	saklıdır	199	0.338
4	da	555	0.942	her	hakkı	199	0.338	saklıdır	eb	tasarımı	199	0.338
5	de	533	0.905	saklıdır	eb	199	0.338	pan	yayıncılık	her	198	0.336
6	pan	398	0.676	tasarımı	ebi	199	0.338	yayıncılık	her	hakkı	198	0.336
7	yayıncılık	342	0.581	yayıncılık	her	198	0.336	kitaptan	bir	bölüm	173	0.294
8	müzik	337	0.572	bir	bölüm	174	0.295	pan	yayıncılık	tan	95	0.161
9	her	326	0.554	kitaptan	bir	173	0.294	yayıncılık	tan	yayınlanmış	89	0.151
10	için	270	0.458	kapak	grafiği	95	0.161	yazarın	pan	yayıncılık	89	0.151
11	olarak	267	0.453	yayıncılık	tan	95	0.161	kapak	grafiği	fatih	74	0.126
12	türk	245	0.416	fiyatı	tl	93	0.158	grafiği	fatih	durmuş	71	0.121
13	tl	232	0.394	tan	yayınlanmış	89	0.151	bir	bölüm	yazar	66	0.112
14	ile	214	0.363	yazarın	pan	89	0.151	bölüm	yazar	biyografisi	66	0.112
15	gibi	212	0.360	fatih	durmuş	87	0.148	ebi	kitaptan	bir	57	0.097
16	in	210	0.357	yazar	biyografisi	83	0.141	ebi	yazarın	pan	57	0.097
17	çok	209	0.355	grafiği	fatih	74	0.126	tasarımı	ebi	kitaptan	57	0.097
18	tasarımı	207	0.351	birinci	basım	68	0.115	tasarımı	ebi	yazarın	57	0.097
19	ebi	201	0.341	bölüm	yazar	66	0.112	tan	yayınlanmış	kitabı	44	0.075
20	hakkı	201	0.341	ya	da	58	0.098	durmuş	fiyatı	tl	41	0.070

A.6. Pankitap (Cont'd)

Rank	Tetra				Freq.	%
1	hakkı	saklıdır	eb	tasarımı	199	0.338
2	her	hakkı	saklıdır	eb	199	0.338
3	saklıdır	eb	tasarımı	ebi	199	0.338
4	pan	yayıncılık	her	hakkı	198	0.336
5	yayıncılık	her	hakkı	saklıdır	198	0.336
6	pan	yayıncılık	tan	yayınlanmış	89	0.151
7	yazarın	pan	yayıncılık	tan	89	0.151
8	kapak	grafiği	fatih	durmuş	71	0.121
9	bir	bölüm	yazar	biyografisi	66	0.112
10	kitaptan	bir	bölüm	yazar	66	0.112
11	eb	tasarımı	ebi	kitaptan	57	0.097
12	eb	tasarımı	ebi	yazarın	57	0.097
13	ebi	kitaptan	bir	bölüm	57	0.097
14	ebi	yazarın	pan	yayıncılık	57	0.097
15	tasarımı	ebi	kitaptan	bir	57	0.097
16	tasarımı	ebi	yazarın	pan	57	0.097
17	yayıncılık	tan	yayınlanmış	kitabı	44	0.075
18	fatih	durmuş	fiyatı	tl	39	0.066
19	grafiği	fatih	durmuş	fiyatı	33	0.056
20	sayfa	tl	kapak	grafiği	31	0.053

Rank	Penta					Freq.	%
1	hakkı	saklıdır	eb	tasarımı	ebi	199	0.3379
2	her	hakkı	saklıdır	eb	tasarımı	199	0.3379
3	pan	yayıncılık	her	hakkı	saklıdır	198	0.3362
4	yayıncılık	her	hakkı	saklıdır	eb	198	0.3362
5	yazarın	pan	yayıncılık	tan	yayınlanmış	89	0.1511
6	kitaptan	bir	bölüm	yazar	biyografisi	66	0.1121
7	eb	tasarımı	ebi	kitaptan	bir	57	0.0968
8	eb	tasarımı	ebi	yazarın	pan	57	0.0968
9	ebi	yazarın	pan	yayıncılık	tan	57	0.0968
10	saklıdır	eb	tasarımı	ebi	kitaptan	57	0.0968
11	saklıdır	eb	tasarımı	ebi	yazarın	57	0.0968
12	tasarımı	ebi	kitaptan	bir	bölüm	57	0.0968
13	tasarımı	ebi	yazarın	pan	yayıncılık	57	0.0968
14	pan	yayıncılık	tan	yayınlanmış	kitabı	44	0.0747
15	ebi	kitaptan	bir	bölüm	yazar	43	0.0730
16	grafiği	fatih	durmuş	fiyatı	tl	33	0.0560
17	kapak	grafiği	fatih	durmuş	fiyatı	33	0.0560
18	sayfa	tl	kapak	grafiği	fatih	31	0.0526
19	tl	kapak	grafiği	fatih	durmuş	31	0.0526
20	pan	yayıncılık	tan	yayınlanmış	diğer	30	0.0509

A.7. *PCMagazin*

Rank	Mono	Freq.	%	Bi		Freq.	%	Tri			Freq.	%
1	ve	10046	1.904	sabit	diskinize	2421	0.459	sağ	fare	düğmesine	1283	0.243
2	bir	9915	1.879	indo	s	1799	0.341	gidin	sağ	fare	1203	0.228
3	bu	7379	1.398	sağ	fare	1402	0.266	üzerine	gidin	sağ	1203	0.228
4	için	6565	1.244	tıklayın	ve	1303	0.247	fare	düğmesine	tıklayın	1201	0.228
5	eb	4860	0.921	fare	düğmesine	1289	0.244	pc	magazine	online	1201	0.228
6	e	4455	0.844	almak	için	1250	0.237	sabit	diskinize	kaydedin	1201	0.228
7	com	4290	0.813	dosya	adı	1245	0.236	animated	gif	arsivi	1198	0.227
8	dosya	3146	0.596	pc	magazine	1228	0.233	animated	gifleri	eb	1198	0.227
9	s	3116	0.590	düğmesine	tıklayın	1219	0.231	arsivi	pc	magazine	1198	0.227
10	sabit	2749	0.521	sizler	için	1213	0.230	bu	animated	gifleri	1198	0.227
11	ile	2620	0.496	üzerine	gidin	1212	0.230	dan	ebmaster	lara	1198	0.227
12	de	2470	0.468	gidin	sağ	1207	0.229	dosyayı	sabit	diskinize	1198	0.227
13	diskinize	2421	0.459	dosyayı	sabit	1204	0.228	düğmesine	tıklayın	ve	1198	0.227
14	net	2149	0.407	ve	dosyayı	1204	0.228	eb	sayfalarınızda	kullanarak	1198	0.227
15	gif	2120	0.402	eb	sayfalarınızda	1202	0.228	ebmaster	lara	sizler	1198	0.227
16	düğmesine	1991	0.377	diskinize	kaydedin	1201	0.228	gif	arsivi	pc	1198	0.227
17	indo	1886	0.357	için	üzerine	1201	0.228	gifleri	eb	sayfalarınızda	1198	0.227
18	pc	1870	0.354	magazine	online	1201	0.228	için	toplamış	olduğumuz	1198	0.227
19	da	1818	0.344	animated	gif	1200	0.227	kullanarak	içeriğini	zenginleştirebilirsiniz	1198	0.227
20	a	1796	0.340	ebmaster	lara	1200	0.227	lara	sizler	için	1198	0.227

A.7. PCMagazin (Cont'd)

Rank	Tetra				Freq.	%
1	gidin	sağ	fare	düğmesine	1203	0.228
2	üzerine	gidin	sağ	fare	1203	0.228
3	sağ	fare	düğmesine	tıklayın	1201	0.228
4	animated	gif	arsivi	pc	1198	0.227
5	animated	gifleri	eb	sayfalarınızda	1198	0.227
6	arsivi	pc	magazine	online	1198	0.227
7	bu	animated	gifleri	eb	1198	0.227
8	dan	ebmaster	lara	sizler	1198	0.227
9	eb	sayfalarınızda	kullanarak	içeriğini	1198	0.227
10	ebmaster	lara	sizler	için	1198	0.227
11	gif	arsivi	pc	magazine	1198	0.227
12	gifleri	eb	sayfalarınızda	kullanarak	1198	0.227
13	için	toplamış	olduğumuz	bu	1198	0.227
14	lara	sizler	için	toplamış	1198	0.227
15	magazine	online	dan	ebmaster	1198	0.227
16	olduğumuz	bu	animated	gifleri	1198	0.227
17	online	dan	ebmaster	lara	1198	0.227
18	pc	magazine	online	dan	1198	0.227
19	sayfalarınızda	kullanarak	içeriğini	zenginleştirebilirsiniz	1198	0.227
20	sizler	için	toplamış	olduğumuz	1198	0.227

Rank	Penta					Freq.	%
1	üzerine	gidin	sağ	fare	düğmesine	1203	0.2279
2	animated	gif	arsivi	pc	magazine	1198	0.2270
3	animated	gifleri	eb	sayfalarınızda	kullanarak	1198	0.2270
4	arsivi	pc	magazine	online	dan	1198	0.2270
5	bu	animated	gifleri	eb	sayfalarınızda	1198	0.2270
6	dan	ebmaster	lara	sizler	için	1198	0.2270
7	eb	sayfalarınızda	kullanarak	içeriğini	zenginleştirebilirsiniz	1198	0.2270
8	ebmaster	lara	sizler	için	toplamış	1198	0.2270
9	gif	arsivi	pc	magazine	online	1198	0.2270
10	gifleri	eb	sayfalarınızda	kullanarak	içeriğini	1198	0.2270
11	için	toplamış	olduğumuz	bu	animated	1198	0.2270
12	lara	sizler	için	toplamış	olduğumuz	1198	0.2270
13	magazine	online	dan	ebmaster	lara	1198	0.2270
14	olduğumuz	bu	animated	gifleri	eb	1198	0.2270
15	online	dan	ebmaster	lara	sizler	1198	0.2270
16	pc	magazine	online	dan	ebmaster	1198	0.2270
17	sizler	için	toplamış	olduğumuz	bu	1198	0.2270
18	toplamış	olduğumuz	bu	animated	gifleri	1198	0.2270
19	almak	için	üzerine	gidin	sağ	1197	0.2268
20	diskinize	almak	için	üzerine	gidin	1197	0.2268

A.8. Novels & Stories

Rank	Mono	Freq.	%	Bi		Freq.	%
1	bir	143536	3.075	ya	da	6789	0.145
2	ve	91089	1.951	bir	şey	5020	0.108
3	bu	66601	1.427	başka	bir	2538	0.054
4	da	42402	0.908	büyük	bir	2500	0.054
5	de	40814	0.874	ne	kadar	2437	0.052
6	için	27230	0.583	ve	bu	2032	0.044
7	ne	23782	0.509	o	kadar	1945	0.042
8	gibi	22783	0.488	ben	de	1910	0.041
9	daha	22025	0.472	böyle	bir	1880	0.040
10	o	21778	0.467	o	zaman	1836	0.039
11	çok	18872	0.404	hem	de	1793	0.038
12	sonra	17549	0.376	her	zaman	1609	0.034
13	her	16856	0.361	bu	kadar	1548	0.033
14	ama	15801	0.338	bir	gün	1514	0.032
15	kadar	15296	0.328	hiçbir	şey	1460	0.031
16	ya	13879	0.297	yeni	bir	1460	0.031
17	ki	13171	0.282	belki	de	1445	0.031
18	olan	12317	0.264	değil	mi	1423	0.030
19	ile	12264	0.263	ne	de	1404	0.030
20	olarak	12225	0.262	o	da	1392	0.030
21	bütün	11419	0.245	sonra	da	1352	0.029
22	zaman	11348	0.243	her	şeyi	1350	0.029
23	diye	11306	0.242	daha	çok	1296	0.028
24	ben	11044	0.237	bir	daha	1292	0.028
25	dedi	10883	0.233	gibi	bir	1282	0.027
26	en	10576	0.227	bir	de	1278	0.027
27	iki	10064	0.216	her	şey	1246	0.027
28	şey	9842	0.211	da	bir	1205	0.026
29	büyük	9545	0.204	bir	şekilde	1202	0.026
30	kendi	9000	0.193	bir	süre	1187	0.025
31	hiç	8889	0.190	de	bir	1179	0.025
32	onu	8549	0.183	daha	iyi	1164	0.025
33	içinde	8481	0.182	olduğu	gibi	1154	0.025
34	onun	8298	0.178	daha	fazla	1151	0.025
35	değil	7915	0.170	ve	bir	1131	0.024
36	iyi	7701	0.165	için	bir	1109	0.024
37	var	7581	0.162	mustafa	kemal	1058	0.023
38	in	7580	0.162	de	bu	1046	0.022
39	nin	7524	0.161	bir	biçimde	1043	0.022
40	başka	7466	0.160	bir	an	1024	0.022
41	mi	7186	0.154	daha	da	1024	0.022
42	böyle	7141	0.153	diye	sordu	1009	0.022
43	bile	7084	0.152	herhangi	bir	984	0.021
44	bana	6758	0.145	hiçbir	zaman	970	0.021
45	ona	6571	0.141	küçük	bir	970	0.021
46	nasıl	6498	0.139	aynı	zamanda	966	0.021
47	olduğunu	6400	0.137	için	de	964	0.021
48	olduğu	6398	0.137	bu	da	959	0.021
49	fakat	6216	0.133	onun	için	938	0.020
50	karşı	6056	0.130	ne	var	922	0.020

A.8. Novels & Stories (Cont'd)

Rank	Tri			Freq.	%
1	başka	bir	şey	906	0.019
2	ne	var	ki	597	0.013
3	ali	rıza	bey	360	0.008
4	bir	yandan	da	327	0.007
5	bir	süre	sonra	318	0.007
6	ne	olursa	olsun	279	0.006
7	bir	an	önce	257	0.006
8	ne	yazık	ki	244	0.005
9	her	şeyden	önce	235	0.005
10	ya	da	bir	227	0.005
11	albay	aureliano	buendia	217	0.005
12	bir	kez	daha	213	0.005
13	jose	arcadio	buendia	200	0.004
14	her	ne	kadar	197	0.004
15	mustafa	kemal	paşa	196	0.004
16	ittihat	ve	terakki	194	0.004
17	en	ufak	bir	193	0.004
18	mustafa	kemal	in	192	0.004
19	mustafa	kemal	atatürk	191	0.004
20	bunun	içindir	ki	190	0.004
21	bir	şey	değildir	188	0.004
22	bir	şey	yok	180	0.004
23	her	zamanki	gibi	169	0.004
24	kısa	bir	süre	169	0.004
25	sovyet	rusya	nın	169	0.004
26	böyle	bir	şey	166	0.004
27	bir	başka	deyişle	162	0.003
28	her	ikisi	de	158	0.003
29	işte	o	zaman	158	0.003
30	ama	yine	de	150	0.003
31	bir	kere	daha	150	0.003
32	ne	de	olsa	149	0.003
33	ya	da	bu	144	0.003
34	diye	karşılık	verdi	142	0.003
35	daha	büyük	bir	139	0.003
36	o	zamana	kadar	138	0.003
37	ya	da	daha	136	0.003
38	o	kadar	çok	135	0.003
39	ve	hem	de	134	0.003
40	başka	bir	deyişle	133	0.003
41	bir	şey	değil	131	0.003
42	m	kemal	in	131	0.003
43	bir	an	için	130	0.003
44	ingiltere	ve	fransa	129	0.003
45	tam	bu	sırada	129	0.003
46	bunun	için	de	127	0.003
47	büyük	millet	meclisi	125	0.003
48	bu	yüzden	de	124	0.003
49	daha	önce	de	124	0.003
50	çok	büyük	bir	122	0.003

A.8. Novels & Stories (Cont'd)

Rank	Tetra				Freq.	%
1	başka	bir	şey	değildir	155	0.0033
2	mustafa	kemal	atatürk	ün	102	0.0022
3	şu	ya	da	bu	93	0.0020
4	ne	var	ki	bu	78	0.0017
5	başka	bir	şey	değildi	75	0.0016
6	bana	öyle	geliyor	ki	72	0.0015
7	türkiye	büyük	millet	meclisi	72	0.0015
8	bir	aşağı	bir	yukarı	69	0.0015
9	sofia	de	la	piedad	68	0.0015
10	santa	sofia	de	la	67	0.0014
11	kısa	bir	süre	sonra	66	0.0014
12	ali	rıza	bey	in	64	0.0014
13	ne	pahasına	olursa	olsun	63	0.0013
14	albay	aureliano	buendia	nin	56	0.0012
15	ajansı	basın	ve	yayıncılık	55	0.0012
16	albay	gerineldo	mar	uez	55	0.0012
17	basın	ve	yayıncılık	a	55	0.0012
18	haber	ajansı	basın	ve	55	0.0012
19	ve	yayıncılık	a	ş	55	0.0012
20	mustafa	kemal	paşa	nın	54	0.0012
21	b	b	i	l	53	0.0011
22	daha	açık	bir	deyişle	53	0.0011
23	a	ş	baskı	çağdaş	50	0.0011
24	ş	baskı	çağdaş	matbaacılık	50	0.0011
25	yayıncılık	a	ş	baskı	50	0.0011
26	gün	haber	ajansı	basın	48	0.0010
27	ittihat	ve	terakki	cemiyeti	48	0.0010
28	yeni	gün	haber	ajansı	48	0.0010
29	her	zaman	olduğu	gibi	47	0.0010
30	milli	eğitim	bakanlığı	nca	45	0.0010
31	her	gün	biraz	daha	44	0.0009
32	jose	arcadio	buendia	nin	43	0.0009
33	baskı	çağdaş	matbaacılık	yayıncılık	41	0.0009
34	çağdaş	matbaacılık	yayıncılık	ltd	41	0.0009
35	matbaacılık	yayıncılık	ltd	şti	41	0.0009
36	ne	var	ne	yok	41	0.0009
37	başka	bir	şey	olmayan	40	0.0009
38	başka	bir	şey	değil	39	0.0008
39	hiçbir	şey	olmamış	gibi	39	0.0008
40	barış	içinde	birarada	yaşama	38	0.0008
41	din	kültürü	ve	ahlak	36	0.0008
42	genç	sokrates	evet	yabancı	34	0.0007
43	her	iki	taraf	da	34	0.0007
44	ya	da	daha	fazla	34	0.0007
45	efl	k	ve	buğdan	33	0.0007
46	kültürü	ve	ahlak	bilgisi	33	0.0007
47	ali	riza	bey	e	32	0.0007
48	bir	ya	da	iki	32	0.0007
49	ii	inci	dünya	savaşı	32	0.0007
50	ittihat	ve	terakki	nin	32	0.0007

A.8. Novels & Stories (Cont'd)

Rank	Penta					Freq.	%
1	santa	sofia	de	la	piedad	67	0.0014
2	ajansı	basın	ve	yayıncılık	a	55	0.0012
3	basın	ve	yayıncılık	a	ş	55	0.0012
4	haber	ajansı	basın	ve	yayıncılık	55	0.0012
5	a	ş	baskı	çağdaş	matbaacılık	50	0.0011
6	ve	yayıncılık	a	ş	baskı	50	0.0011
7	yayıncılık	a	ş	baskı	çağdaş	50	0.0011
8	gün	haber	ajansı	basın	ve	48	0.0010
9	yeni	gün	haber	ajansı	basın	48	0.0010
10	baskı	çağdaş	matbaacılık	yayıncılık	ltd	41	0.0009
11	çağdaş	matbaacılık	yayıncılık	ltd	şti	41	0.0009
12	ş	baskı	çağdaş	matbaacılık	yayıncılık	41	0.0009
13	din	kültürü	ve	ahlak	bilgisi	33	0.0007
14	bir	o	yana	bir	bu	30	0.0006
15	o	yana	bir	bu	yana	30	0.0006
16	feuerbach	ve	klasik	alman	felsefesinin	25	0.0005
17	günümüz	türkçesine	uyarlanmıştır	yayına	hazırlayan	25	0.0005
18	ig	feuerbach	ve	klasik	alman	25	0.0005
19	lud	ig	feuerbach	ve	klasik	25	0.0005
20	türkçesine	uyarlanmıştır	yayına	hazırlayan	egemen	25	0.0005
21	ve	klasik	alman	felsefesinin	sonu	25	0.0005
22	yayına	hazırlayan	egemen	berköz	dizgi	25	0.0005
23	alınmış	ve	çeviri	dili	günümüz	24	0.0005
24	anlama	gücünü	o	yapıtlar	oranında	24	0.0005
25	anlama	ve	duymada	ilk	aşama	24	0.0005
26	anlatımı	olan	sanat	yapıtlarının	benimsenmesidir	24	0.0005
27	anlatımın	düşünce	öğeleri	en	zengin	24	0.0005
28	aracı	olan	yazı	ve	onun	24	0.0005
29	artırması	canlandırması	ve	yeniden	yaratması	24	0.0005
30	aşacak	bir	sağlamlık	ve	yaygınlığı	24	0.0005
31	aşama	insan	varlığının	en	somut	24	0.0005
32	aydınlanma	devrimi	nde	dünya	klasiklerinin	24	0.0005
33	aydınlanma	kitaplığı	kazandırmak	istedik	bu	24	0.0005
34	aydınlarına	şükran	duyuyorum	onların	çabalarıyla	24	0.0005
35	bakımdan	çeviri	etkinliğini	sistemli	ve	24	0.0005
36	bakımdan	önemli	ve	uygarlık	davamız	24	0.0005
37	baskısı	temel	alınmış	ve	çeviri	24	0.0005
38	başlayan	türk	aydınlanma	devrimi	nde	24	0.0005
39	başlayarak	milli	eğitim	bakanlığı	nca	24	0.0005
40	benim	sadık	yarim	kara	topraktır	24	0.0005
41	benimsenmesidir	sanat	dalları	içinde	edebiyat	24	0.0005
42	beş	katı	büyük	olmak	üzere	24	0.0005
43	beş	yıl	içinde	hiç	değilse	24	0.0005
44	biçimde	yönetmek	onun	genişlemesine	ilerlemesine	24	0.0005
45	bilgi	ve	emeklerini	esirgemeyen	türk	24	0.0005
46	bir	aydınlanma	kitaplığı	kazandırmak	istedik	24	0.0005
47	bir	biçimde	yönetmek	onun	genişlemesine	24	0.0005
48	bir	çeviri	kitaplığımız	olacaktır	özellikle	24	0.0005
49	bir	düşünce	düzeyinde	demektir	bu	24	0.0005
50	bir	etkisi	vardır	bu	etkinin	24	0.0005

A.9. Star Gazete

Rank	Mono	Freq.	%	Bi		Freq.	%	Tri			Freq.	%
1	bir	145802	1.496	kültür	sanat	29998	0.308	haber	ekonomi	dünya	29025	0.298
2	ve	143722	1.475	derin	haber	29414	0.302	bölümler	ana	sayfa	28195	0.289
3	bu	97867	1.004	bilim	teknoloji	29193	0.300	derin	haber	magazin	28174	0.289
4	da	92480	0.949	haber	ekonomi	29084	0.298	tv	derin	haber	28167	0.289
5	de	80538	0.826	magazin	kadın	29078	0.298	spor	yazarlar	bilim	28150	0.289
6	haber	66102	0.678	ekonomi	dünya	29028	0.298	yazarlar	bilim	teknoloji	28150	0.289
7	için	62195	0.638	ana	sayfa	28199	0.289	ana	sayfa	haber	28090	0.288
8	ile	49628	0.509	bölümler	ana	28195	0.289	sayfa	haber	ekonomi	28090	0.288
9	ın	42791	0.439	haber	magazin	28174	0.289	kültür	sanat	tv	28086	0.288
10	internet	38166	0.392	tv	derin	28167	0.289	magazin	kadın	spor	28086	0.288
11	in	36909	0.379	spor	yazarlar	28151	0.289	haber	magazin	kadın	28065	0.288
12	dünya	34308	0.352	yazarlar	bilim	28150	0.289	sanat	tv	derin	28058	0.288
13	spor	33806	0.347	sayfa	haber	28094	0.288	kadın	spor	yazarlar	28042	0.288
14	magazin	32847	0.337	kadın	spor	28086	0.288	dünya	kültür	sanat	27981	0.287
15	kadın	32410	0.333	sanat	tv	28086	0.288	ekonomi	dünya	kültür	27981	0.287
16	ana	32345	0.332	dünya	kültür	27982	0.287	bilim	teknoloji	hava	14138	0.145
17	bilim	32063	0.329	hava	durumu	27859	0.286	teknoloji	hava	durumu	14138	0.145
18	çok	31853	0.327	teknoloji	hava	14138	0.145	bilim	teknoloji	internet	13892	0.143
19	tv	31836	0.327	teknoloji	internet	13895	0.143	hava	durumu	astroloji	8184	0.084
20	teknoloji	31646	0.325	durumu	astroloji	8184	0.084	internet	hava	durumu	7895	0.081

A.9. Star Gazete (Cont'd)

Rank	Tetra				Freq.	%
1	tv	derin	haber	magazin	28167	0.289
2	spor	yazarlar	bilim	teknoloji	28150	0.289
3	ana	sayfa	haber	ekonomi	28090	0.288
4	bölümler	ana	sayfa	haber	28090	0.288
5	derin	haber	magazin	kadın	28065	0.288
6	haber	magazin	kadın	spor	28065	0.288
7	kültür	sanat	tv	derin	28058	0.288
8	sanat	tv	derin	haber	28058	0.288
9	magazin	kadın	spor	yazarlar	28042	0.288
10	kadın	spor	yazarlar	bilim	28041	0.288
11	sayfa	haber	ekonomi	dünya	28033	0.288
12	dünya	kültür	sanat	tv	27981	0.287
13	ekonomi	dünya	kültür	sanat	27981	0.287
14	haber	ekonomi	dünya	kültür	27981	0.287
15	bilim	teknoloji	hava	durumu	14138	0.145
16	yazarlar	bilim	teknoloji	hava	14138	0.145
17	yazarlar	bilim	teknoloji	internet	13476	0.138
18	bilim	teknoloji	internet	hava	7895	0.081
19	teknoloji	internet	hava	durumu	7895	0.081
20	bilim	teknoloji	internet	ekler	4892	0.050

Rank	Penta					Freq.	%
1	bölümler	ana	sayfa	haber	ekonomi	28090	0.2882
2	derin	haber	magazin	kadın	spor	28065	0.2880
3	kültür	sanat	tv	derin	haber	28058	0.2879
4	sanat	tv	derin	haber	magazin	28058	0.2879
5	tv	derin	haber	magazin	kadın	28058	0.2879
6	kadın	spor	yazarlar	bilim	teknoloji	28041	0.2877
7	magazin	kadın	spor	yazarlar	bilim	28041	0.2877
8	ana	sayfa	haber	ekonomi	dünya	28033	0.2876
9	haber	magazin	kadın	spor	yazarlar	28021	0.2875
10	ekonomi	dünya	kültür	sanat	tv	27981	0.2871
11	haber	ekonomi	dünya	kültür	sanat	27981	0.2871
12	sayfa	haber	ekonomi	dünya	kültür	27981	0.2871
13	dünya	kültür	sanat	tv	derin	27960	0.2869
14	spor	yazarlar	bilim	teknoloji	hava	14138	0.1451
15	yazarlar	bilim	teknoloji	hava	durumu	14138	0.1451
16	spor	yazarlar	bilim	teknoloji	internet	13476	0.1383
17	bilim	teknoloji	internet	hava	durumu	7895	0.0810
18	yazarlar	bilim	teknoloji	internet	hava	7895	0.0810
19	yazarlar	bilim	teknoloji	internet	ekler	4847	0.0497
20	bilim	teknoloji	hava	durumu	astroloji	3970	0.0407

A.10. TBMM

Rank	Mono	Freq.	%	Bi		Freq.	%	Tri			Freq.	%
1	ve	683740	2.922	türkiye	büyük	38012	0.162	türkiye	büyük	millet	37419	0.160
2	bir	365748	1.563	büyük	millet	37981	0.162	büyük	millet	meclisi	30210	0.129
3	bu	358734	1.533	sayın	başkan	33207	0.142	millet	meclisi	başkanlığına	13933	0.060
4	n	159675	0.682	millet	meclisi	30657	0.131	başkanlığa	geliş	tarihi	13640	0.058
5	sayın	158773	0.679	soru	önergesi	25215	0.108	yazılı	soru	önergesi	12955	0.055
6	de	151213	0.646	n	n	22861	0.098	sayın	başkan	değerli	10904	0.047
7	da	142301	0.608	sıralarından	alkışlar	21613	0.092	önergesi	başkanlığa	geliş	10365	0.044
8	başkan	114552	0.490	başkan	sayın	21100	0.090	soru	önergesi	başkanlığa	9948	0.043
9	türkiye	111540	0.477	değerli	milletvekilleri	20157	0.086	plan	ve	bütçe	9793	0.042
10	için	106286	0.454	ve	bu	18792	0.080	sözlü	soru	önergesi	9425	0.040
11	a	103940	0.444	say	n	16643	0.071	başkan	değerli	milletvekilleri	9311	0.040
12	olarak	101592	0.434	yazılı	soru	16246	0.069	sunuyorum	kabul	edenler	8655	0.037
13	milletvekili	87327	0.373	sayın	milletvekilleri	15558	0.066	sıralarından	alkışlar	başkan	8288	0.035
14	ile	85540	0.366	kabul	edilmiştir	15003	0.064	etmeyenler	kabul	edilmiştir	8206	0.035
15	i	81715	0.349	kanun	teklifi	14902	0.064	ilişkin	sorusu	ve	8093	0.035
16	kanun	79283	0.339	grubu	adına	14665	0.063	oylarınıza	sunuyorum	kabul	7929	0.034
17	r	78610	0.336	ba	kan	14653	0.063	kabul	edenler	etmeyenler	7810	0.033
18	olan	71415	0.305	hakkında	kanun	14610	0.062	alkışlar	başkan	teşekkür	7643	0.033
19	büyük	69955	0.299	kabul	edenler	14379	0.061	başkan	sayın	milletvekilleri	7447	0.032
20	l	67811	0.290	türkiye	nin	14317	0.061	partisi	grubu	adına	7349	0.031

A.10. TBMM (Cont'd)

Rank	Tetra				Freq.	%
1	türkiye	büyük	millet	meclisi	30025	0.13
2	büyük	millet	meclisi	başkanlığına	13923	0.06
3	önergesi	başkanlığa	geliş	tarihi	10320	0.04
4	soru	önergesi	başkanlığa	geliş	9948	0.04
5	sayın	başkan	değerli	milletvekilleri	8950	0.04
6	oylarınıza	sunuyorum	kabul	edenler	7919	0.03
7	yazılı	soru	önergesi	başkanlığa	7610	0.03
8	bakanından	sözlü	soru	önergesi	6490	0.03
9	kabul	edenler	etmeyenler	kabul	6257	0.03
10	sıralarından	alkışlar	başkan	teşekkür	5916	0.03
11	kabul	edenler	kabul	etmeyenler	5852	0.03
12	mikrofon	otomatik	cihaz	tarafından	5730	0.02
13	otomatik	cihaz	tarafından	kapatıldı	5709	0.02
14	edenler	etmeyenler	kabul	edilmiştir	5530	0.02
15	bakanından	yazılı	soru	önergesi	5433	0.02
16	sunuyorum	kabul	edenler	kabul	4784	0.02
17	değişiklik	yapılması	hakkında	kanun	4488	0.02
18	ve	plan	ve	bütçe	4444	0.02
19	enerji	ve	tabii	kaynaklar	4438	0.02
20	cihaz	tarafından	kapatıldı	başkan	4343	0.02

Rank	Penta					Freq.	%
1	türkiye	büyük	millet	meclisi	başkanlığına	13918	0.0595
2	soru	önergesi	başkanlığa	geliş	tarihi	9904	0.0423
3	yazılı	soru	önergesi	başkanlığa	geliş	7610	0.0325
4	mikrofon	otomatik	cihaz	tarafından	kapatıldı	5706	0.0244
5	kabul	edenler	etmeyenler	kabul	edilmiştir	5530	0.0236
6	sunuyorum	kabul	edenler	kabul	etmeyenler	4783	0.0204
7	bakanından	yazılı	soru	önergesi	başkanlığa	4538	0.0194
8	otomatik	cihaz	tarafından	kapatıldı	başkan	4319	0.0185
9	oylarınıza	sunuyorum	kabul	edenler	kabul	4253	0.0182
10	büyük	millet	meclisi	başkanlığına	aşağıdaki	3819	0.0163
11	oylarınıza	sunuyorum	kabul	edenler	etmeyenler	3467	0.0148
12	önergesi	türkiye	büyük	millet	meclisi	3439	0.0147
13	içindekiler	türkiye	büyük	millet	meclisi	3425	0.0146
14	cevabı	türkiye	büyük	millet	meclisi	3258	0.0139
15	okutuyorum	türkiye	büyük	millet	meclisi	3183	0.0136
16	alkışlar	başkan	teşekkür	ediyorum	sayın	3110	0.0133
17	kabul	edenler	kabul	etmeyenler	kabul	3094	0.0132
18	millet	meclisi	başkanlığına	aşağıdaki	sorularımın	2916	0.0125
19	büyük	millet	meclisi	başkanlığına	ilgi	2841	0.0121
20	türkiye	büyük	millet	meclisi	kanun	2794	0.0119

A.11. Ulusal Program

Rank	Mono	Freq.	%	Bi		Freq.	%	Tri			Freq.	%
1	ve	6503	4.050	tarihli	komisyon	2355	1.467	tarihli	komisyon	kararı	1192	0.742
2	tarihli	4941	3.077	tarihli	konsey	1964	1.223	tarihli	komisyon	tüzüğü	1040	0.648
3	ilişkin	3271	2.037	ile	ilgili	1375	0.856	tarihli	konsey	tüzüğü	760	0.473
4	eec	3227	2.010	komisyon	kararı	1213	0.755	tarihli	konsey	direktifi	754	0.470
5	konsey	2894	1.802	komisyon	tüzüğü	1050	0.654	konsey	direktifi	eec	579	0.361
6	ec	2630	1.638	aralık	tarihli	952	0.593	komisyon	tüzüğü	ec	555	0.346
7	ile	2608	1.624	tüzüğü	ec	925	0.576	aralık	tarihli	konsey	453	0.282
8	komisyon	2505	1.560	konsey	direktifi	898	0.559	temmuz	tarihli	komisyon	433	0.270
9	sayılı	2105	1.311	konsey	tüzüğü	848	0.528	tarihli	konsey	kararı	416	0.259
10	kararı	2085	1.299	eec	sayılı	756	0.471	komisyon	tüzüğü	eec	388	0.242
11	tüzüğü	2056	1.281	temmuz	tarihli	736	0.458	aralık	tarihli	komisyon	372	0.232
12	için	1849	1.152	tüzüğü	eec	734	0.457	eec	sayılı	konsey	368	0.229
13	ilgili	1797	1.119	haziran	tarihli	696	0.433	haziran	tarihli	konsey	354	0.220
14	avrupa	1198	0.746	direktifi	eec	676	0.421	konsey	tüzüğü	ec	354	0.220
15	direktifi	1191	0.742	sayılı	konsey	659	0.410	konsey	tüzüğü	eec	336	0.209
16	bir	1097	0.683	kararı	ec	548	0.341	komisyon	kararı	sadece	319	0.199
17	aralık	1043	0.650	konsey	kararı	465	0.290	ilişkin	aralık	tarihli	299	0.186
18	hakkında	995	0.620	oj	l	421	0.262	komisyon	kararı	ec	268	0.167
19	üye	862	0.537	kararı	eec	418	0.260	komisyon	kararı	eec	259	0.161
20	olarak	855	0.533	kasım	tarihli	379	0.236	haziran	tarihli	komisyon	258	0.161

A.11. Ulusal Program (Cont'd)

Rank	Tetra				Freq.	%
1	tarihli	komisyon	tüzüğü	ec	552	0.344
2	tarihli	konsey	direktifi	eec	524	0.326
3	tarihli	komisyon	tüzüğü	eec	383	0.239
4	tarihli	konsey	tüzüğü	ec	338	0.211
5	tarihli	konsey	tüzüğü	eec	324	0.202
6	tarihli	komisyon	kararı	sadece	319	0.199
7	tarihli	komisyon	kararı	ec	262	0.163
8	tarihli	komisyon	kararı	eec	249	0.155
9	metin	aea	ile	uyumludur	231	0.144
10	temmuz	tarihli	komisyon	tüzüğü	211	0.131
11	kararı	metin	aea	ile	207	0.129
12	komisyon	kararı	metin	aea	207	0.129
13	tarihli	komisyon	kararı	metin	207	0.129
14	aralık	tarihli	konsey	tüzüğü	204	0.127
15	temmuz	tarihli	komisyon	kararı	202	0.126
16	aea	ile	uyumludur	ec	196	0.122
17	eec	sayılı	konsey	direktifinin	187	0.116
18	aralık	tarihli	komisyon	kararı	182	0.113
19	aralık	tarihli	komisyon	tüzüğü	178	0.111
20	tarihli	konsey	kararı	ec	172	0.107

Rank	Penta					Freq.	%
1	kararı	metin	aea	ile	uyumludur	207	0.1289
2	komisyon	kararı	metin	aea	ile	207	0.1289
3	tarihli	komisyon	kararı	metin	aea	206	0.1283
4	metin	aea	ile	uyumludur	ec	196	0.1221
5	konsey	direktifi	eec	üye	devletlerin	122	0.0760
6	tarihli	konsey	direktifi	eec	üye	122	0.0760
7	aralık	tarihli	konsey	tüzüğü	ec	114	0.0710
8	temmuz	tarihli	komisyon	tüzüğü	ec	110	0.0685
9	aralık	tarihli	konsey	direktifi	eec	108	0.0673
10	eec	sayılı	konsey	direktifinin	maddesinin	100	0.0623
11	ilişkin	üye	devlet	mevzuatlarının	yakınlaştırılmasına	100	0.0623
12	araçların	ve	treylerlerin	tip	onayına	98	0.0610
13	motorlu	araçların	ve	treylerlerin	tip	98	0.0610
14	onayına	ilişkin	üye	devlet	mevzuatlarının	98	0.0610
15	tip	onayına	ilişkin	üye	devlet	98	0.0610
16	treylerlerin	tip	onayına	ilişkin	üye	98	0.0610
17	üye	devlet	mevzuatlarının	yakınlaştırılmasına	yönelik	98	0.0610
18	ve	treylerlerin	tip	onayına	ilişkin	98	0.0610
19	bendinde	yer	alan	hükme	uygun	97	0.0604
20	c	bendinde	yer	alan	hükme	97	0.0604

A.12. Yeni Asır

Rank	Mono	Freq.	%	Bi		Freq.	%	Tri			Freq.	%
1	ve	2062	2.129	ı	ı	141	0.146	ı	ı	ı	114	0.118
2	bir	1902	1.964	izmir	in	132	0.136	bulvarı	no	çankaya	85	0.088
3	bu	1121	1.157	yeni	asır	113	0.117	gaziosmanpaşa	bulvarı	no	85	0.088
4	de	1028	1.061	bu	hafta	102	0.105	açılış	sayfası	yap	84	0.087
5	da	982	1.014	tel	fa	89	0.092	arşiv	sinema	sanal	84	0.087
6	için	637	0.658	yazı	dizisi	89	0.092	asır	gaziosmanpaşa	bulvarı	84	0.087
7	izmir	599	0.618	bulvarı	no	85	0.088	bugün	yazı	dizisi	84	0.087
8	ne	445	0.459	de	bugün	85	0.088	can	can	cumartesi	84	0.087
9	çok	408	0.421	gaziosmanpaşa	bulvarı	85	0.088	çankaya	izmir	tel	84	0.087
10	in	393	0.406	izmir	tel	85	0.088	de	bugün	yazı	84	0.087
11	olarak	335	0.346	no	çankaya	85	0.088	dış	ekonomi	yazarlar	84	0.087
12	ile	324	0.335	açılış	sayfası	84	0.087	dizisi	arşiv	sinema	84	0.087
13	en	317	0.327	arşiv	sinema	84	0.087	ekle	açılış	sayfası	84	0.087
14	yeni	305	0.315	asır	gaziosmanpaşa	84	0.087	ekonomi	yazarlar	spor	84	0.087
15	gibi	296	0.306	bugün	yazı	84	0.087	izmir	tel	fa	84	0.087
16	daha	294	0.304	can	can	84	0.087	kent	sayfa	siyasi	84	0.087
17	yıl	284	0.293	can	cumartesi	84	0.087	kullanılanlara	ekle	açılış	84	0.087
18	o	274	0.283	çankaya	izmir	84	0.087	künye	reklam	sık	84	0.087
19	her	266	0.275	dış	ekonomi	84	0.087	no	çankaya	izmir	84	0.087
20	sonra	254	0.262	dizisi	arşiv	84	0.087	reklam	sık	kullanılanlara	84	0.087

A.12. Yeni Asır (Cont'd)

Rank	Tetra				Freq.	%
1	ı	ı	ı	ı	88	0.091
2	gaziosmanpaşa	bulvarı	no	çankaya	85	0.088
3	açılış	sayfası	yap	yeni	84	0.087
4	arşiv	sinema	sanal	sergi	84	0.087
5	asır	gaziosmanpaşa	bulvarı	no	84	0.087
6	bugün	yazı	dizisi	arşiv	84	0.087
7	bulvarı	no	çankaya	izmir	84	0.087
8	çankaya	izmir	tel	fa	84	0.087
9	de	bugün	yazı	dizisi	84	0.087
10	dış	ekonomi	yazarlar	spor	84	0.087
11	dizisi	arşiv	sinema	sanal	84	0.087
12	ekle	açılış	sayfası	yap	84	0.087
13	ekonomi	yazarlar	spor	can	84	0.087
14	kent	sayfa	siyasi	dış	84	0.087
15	kullanılanlara	ekle	açılış	sayfası	84	0.087
16	künye	reklam	sık	kullanılanlara	84	0.087
17	no	çankaya	izmir	tel	84	0.087
18	reklam	sık	kullanılanlara	ekle	84	0.087
19	sanal	sergi	telefonlar	eczaneler	84	0.087
20	sayfa	siyasi	diş	ekonomi	84	0.087

Rank	Penta					Freq.	%
1	açılış	sayfası	yap	yeni	asır	84	0.0867
2	arşiv	sinema	sanal	sergi	telefonlar	84	0.0867
3	asır	gaziosmanpaşa	bulvarı	no	çankaya	84	0.0867
4	bugün	yazı	dizisi	arşiv	sinema	84	0.0867
5	bulvarı	no	çankaya	izmir	tel	84	0.0867
6	de	bugün	yazı	dizisi	arşiv	84	0.0867
7	dış	ekonomi	yazarlar	spor	can	84	0.0867
8	dizisi	arşiv	sinema	sanal	sergi	84	0.0867
9	ekle	açılış	sayfası	yap	yeni	84	0.0867
10	ekonomi	yazarlar	spor	can	can	84	0.0867
11	gaziosmanpaşa	bulvarı	no	çankaya	izmir	84	0.0867
12	kent	sayfa	siyasi	dış	ekonomi	84	0.0867
13	kullanılanlara	ekle	açılış	sayfası	yap	84	0.0867
14	künye	reklam	sık	kullanılanlara	ekle	84	0.0867
15	no	çankaya	izmir	tel	fa	84	0.0867
16	reklam	sık	kullanılanlara	ekle	açılış	84	0.0867
17	sanal	sergi	telefonlar	eczaneler	astroloji	84	0.0867
18	sayfa	siyasi	dış	ekonomi	yazarlar	84	0.0867
19	sayfası	yap	yeni	asır	gaziosmanpaşa	84	0.0867
20	sık	kullanılanlara	ekle	açılış	sayfası	84	0.0867

A.13. TurCo

Rank	Mono	Freq.	%	Bi		Freq.	%
1	ve	1137582	2.270	kültür	sanat	65045	0.130
2	bir	803553	1.604	hava	durumu	60214	0.120
3	bu	646620	1.290	ana	sayfa	52562	0.105
4	da	359790	0.718	bilim	teknoloji	48064	0.096
5	de	355645	0.710	yeni	pencerede	42755	0.085
6	için	262514	0.524	pencerede	aç	42751	0.085
7	ile	201530	0.402	türkiye	büyük	38250	0.076
8	türkiye	178250	0.356	büyük	millet	38235	0.076
9	a	174593	0.348	son	dakika	38113	0.076
10	n	169675	0.339	bize	ulaşın	37958	0.076
11	olarak	169252	0.338	arşiv	bize	37815	0.075
12	sayın	162298	0.324	yazarlar	kültür	33508	0.067
13	olan	138142	0.276	sayın	başkan	33259	0.066
14	çok	130576	0.261	özel	dosyalar	32666	0.065
15	daha	127223	0.254	gündem	politika	32546	0.065
16	en	126108	0.252	dünya	ekonomi	32531	0.065
17	e	125444	0.250	politika	dünya	32518	0.065
18	başkan	123932	0.247	spor	yaşam	32461	0.065
19	yeni	117818	0.235	tüm	haberler	32261	0.064
20	büyük	116978	0.233	astronet	televizyon	32255	0.064
21	i	114872	0.229	durumu	astronet	32255	0.064
22	ne	112944	0.225	dosyalar	gezi	32254	0.064
23	ın	109902	0.219	magazin	özel	32254	0.064
24	o	108746	0.217	ekonomi	dünya	32203	0.064
25	haber	107021	0.214	millet	meclisi	30936	0.062
26	gibi	105592	0.211	sanat	magazin	30035	0.060
27	kadar	99802	0.199	derin	haber	29414	0.059
28	son	99242	0.198	ekonomi	spor	29388	0.059
29	in	99233	0.198	haber	ekonomi	29084	0.058
30	dünya	96287	0.192	magazin	kadın	29078	0.058
31	sonra	93742	0.187	bölümler	ana	28195	0.056
32	milletvekili	91049	0.182	haber	magazin	28174	0.056
33	nin	89958	0.180	tv	derin	28167	0.056
34	her	89503	0.179	spor	yazarlar	28152	0.056
35	r	87374	0.174	yazarlar	bilim	28150	0.056
36	ama	85254	0.170	sayfa	haber	28094	0.056
37	m	82411	0.164	kadın	spor	28086	0.056
38	genel	81941	0.164	sanat	tv	28086	0.056
39	kanun	81274	0.162	dünya	kültür	27993	0.056
40	göre	81062	0.162	türkiye	nin	27509	0.055
41	kültür	80565	0.161	ve	bu	26128	0.052
42	spor	80251	0.160	ya	da	25952	0.052
43	l	78570	0.157	soru	önergesi	25259	0.050
44	var	78485	0.157	türkiye	de	23908	0.048
45	gün	76961	0.154	n	n	22944	0.046
46	özel	76846	0.153	sıralarından	alkışlar	21613	0.043
47	ya	75647	0.151	başkan	sayın	21100	0.042
48	ekonomi	75410	0.150	mesaj	grupları	20228	0.040
49	ilgili	74922	0.150	değerli	milletvekilleri	20158	0.040
50	ki	74659	0.149	önceki	gün	19617	0.039

A.13. TurCo (Cont'd)

Rank	Tri			Freq.	%
1	yeni	pencerede	aç	42751	0.085
2	arşiv	bize	ulaşın	37815	0.075
3	türkiye	büyük	millet	37606	0.075
4	yazarlar	kültür	sanat	33508	0.067
5	politika	dünya	ekonomi	32493	0.065
6	gündem	politika	dünya	32462	0.065
7	durumu	astronet	televizyon	32255	0.064
8	hava	durumu	astronet	32255	0.064
9	magazin	özel	dosyalar	32254	0.064
10	özel	dosyalar	gezi	32254	0.064
11	büyük	millet	meclisi	30440	0.061
12	kültür	sanat	magazin	30035	0.060
13	sanat	magazin	özel	29774	0.059
14	dünya	ekonomi	spor	29329	0.059
15	ekonomi	spor	yaşam	29262	0.058
16	haber	ekonomi	dünya	29025	0.058
17	bölümler	ana	sayfa	28195	0.056
18	derin	haber	magazin	28174	0.056
19	tv	derin	haber	28167	0.056
20	spor	yazarlar	bilim	28150	0.056
21	yazarlar	bilim	teknoloji	28150	0.056
22	ana	sayfa	haber	28090	0.056
23	sayfa	haber	ekonomi	28090	0.056
24	kültür	sanat	tv	28086	0.056
25	magazin	kadın	spor	28086	0.056
26	haber	magazin	kadın	28065	0.056
27	sanat	tv	derin	28058	0.056
28	kadın	spor	yazarlar	28042	0.056
29	dünya	kültür	sanat	27981	0.056
30	ekonomi	dünya	kültür	27981	0.056
31	arşivim	mesaj	grupları	18989	0.038
32	grupları	sohbet	yardım	18989	0.038
33	mesaj	grupları	sohbet	18989	0.038
34	sohbet	yardım	üyelik	18989	0.038
35	dakika	arşivim	mesaj	18988	0.038
36	son	dakika	arşivim	18988	0.038
37	üyelik	site	haritası	18988	0.038
38	yardım	üyelik	site	18988	0.038
39	şifre	üye	olmak	18986	0.038
40	ana	sayfa	son	18831	0.038
41	arama	arşiv	bize	18830	0.038
42	arşivim	arama	arşiv	18830	0.038
43	astronet	televizyon	insan	18830	0.038
44	bilim	teknoloji	yazarlar	18830	0.038
45	bize	ulaşın	yardım	18830	0.038
46	dakika	tüm	haberler	18830	0.038
47	dosyalar	gezi	piyasanet	18830	0.038
48	gezi	piyasanet	hava	18830	0.038
49	haberler	gündem	politika	18830	0.038
50	insan	kaynakları	arşivim	18830	0.038

A.13. TurCo (Cont'd)

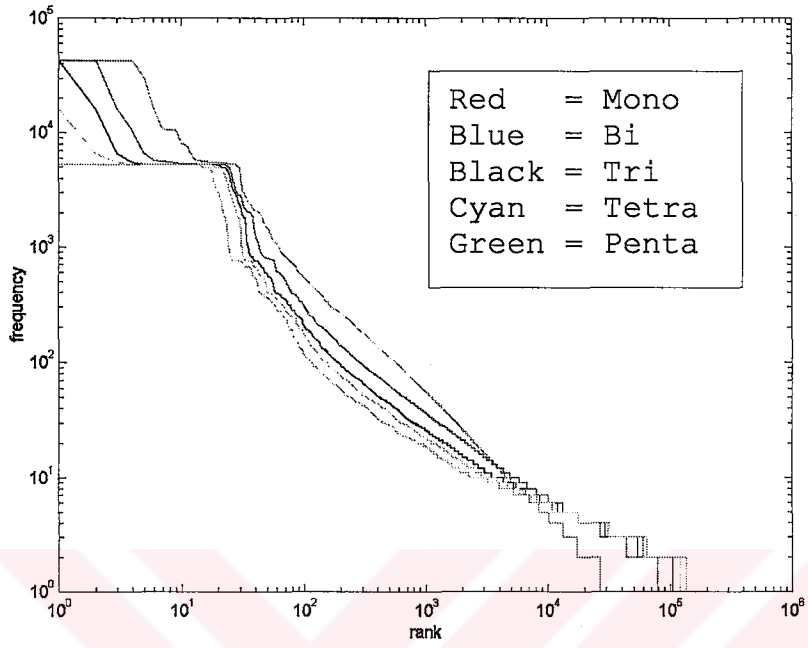
Rank	Tetra				Freq.	%
1	gündem	politika	dünya	ekonomi	32439	0.065
2	hava	durumu	astronet	televizyon	32255	0.064
3	magazin	özel	dosyalar	gezi	32254	0.064
4	türkiye	büyük	millet	meclisi	30195	0.060
5	yazarlar	kültür	sanat	magazin	30012	0.060
6	kültür	sanat	magazin	özel	29774	0.059
7	sanat	magazin	özel	dosyalar	29774	0.059
8	politika	dünya	ekonomi	spor	29306	0.058
9	dünya	ekonomi	spor	yaşam	29239	0.058
10	tv	derin	haber	magazin	28167	0.056
11	spor	yazarlar	bilim	teknoloji	28150	0.056
12	ana	sayfa	haber	ekonomi	28090	0.056
13	bölümler	ana	sayfa	haber	28090	0.056
14	derin	haber	magazin	kadın	28065	0.056
15	haber	magazin	kadın	spor	28065	0.056
16	kültür	sanat	tv	derin	28058	0.056
17	sanat	tv	derin	haber	28058	0.056
18	magazin	kadın	spor	yazarlar	28042	0.056
19	kadın	spor	yazarlar	bilim	28041	0.056
20	sayfa	haber	ekonomi	dünya	28033	0.056
21	dünya	kültür	sanat	tv	27981	0.056
22	ekonomi	dünya	kültür	sanat	27981	0.056
23	haber	ekonomi	dünya	kültür	27981	0.056
24	arşivim	mesaj	grupları	sohbet	18989	0.038
25	grupları	sohbet	yardım	üyelik	18989	0.038
26	mesaj	grupları	sohbet	yardım	18989	0.038
27	dakika	arşivim	mesaj	grupları	18988	0.038
28	sohbet	yardım	üyelik	site	18988	0.038
29	son	dakika	arşivim	mesaj	18988	0.038
30	yardım	üyelik	site	haritası	18988	0.038
31	ana	sayfa	son	dakika	18830	0.038
32	arama	arşiv	bize	ulaşın	18830	0.038
33	arşiv	bize	ulaşın	yardım	18830	0.038
34	arşivim	arama	arşiv	bize	18830	0.038
35	astronet	televizyon	insan	kaynakları	18830	0.038
36	bilim	teknoloji	yazarlar	kültür	18830	0.038
37	bize	ulaşın	yardım	copyright	18830	0.038
38	dakika	tüm	haberler	gündem	18830	0.038
39	dosyalar	gezi	piyasanet	hava	18830	0.038
40	durumu	astronet	televizyon	insan	18830	0.038
41	ekonomi	spor	yaşam	bilim	18830	0.038
42	gezi	piyasanet	hava	durumu	18830	0.038
43	haberler	gündem	politika	dünya	18830	0.038
44	insan	kaynakları	arşivim	arama	18830	0.038
45	kaynakları	arşivim	arama	arşiv	18830	0.038
46	özel	dosyalar	gezi	piyasanet	18830	0.038
47	piyasanet	hava	durumu	astronet	18830	0.038
48	sayfa	son	dakika	tüm	18830	0.038
49	son	dakika	tüm	haberler	18830	0.038
50	spor	yaşam	bilim	teknoloji	18830	0.038

A.13. TurCo (Cont'd)

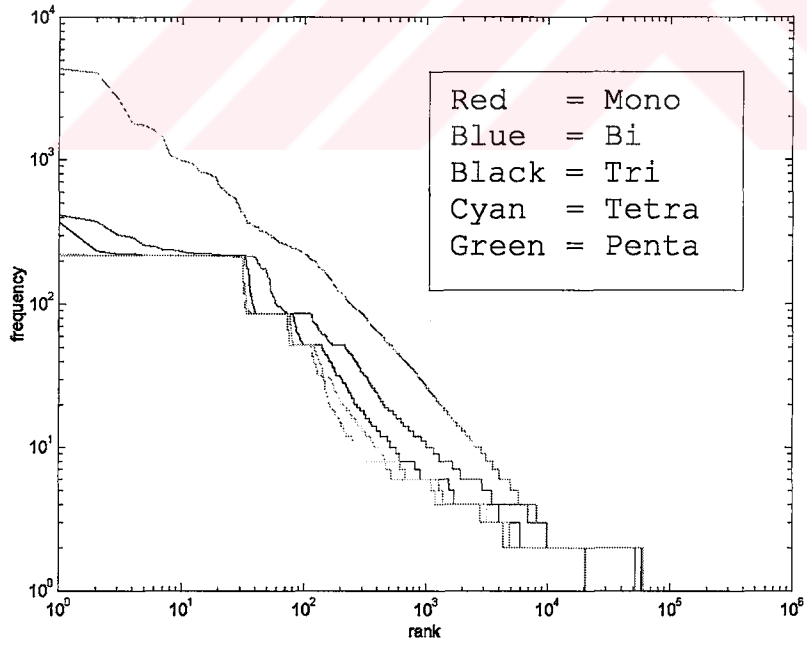
Rank	Penta					Freq.	%
1	kültür	sanat	magazin	özel	dosyalar	29774	0.059
2	sanat	magazin	özel	dosyalar	gezi	29774	0.059
3	yazarlar	kültür	sanat	magazin	özel	29774	0.059
4	gündem	politika	dünya	ekonomi	spor	29252	0.058
5	politika	dünya	ekonomi	spor	yaşam	29216	0.058
6	bölümler	ana	sayfa	haber	ekonomi	28090	0.056
7	derin	haber	magazin	kadın	spor	28065	0.056
8	kültür	sanat	tv	derin	haber	28058	0.056
9	sanat	tv	derin	haber	magazin	28058	0.056
10	tv	derin	haber	magazin	kadın	28058	0.056
11	kadın	spor	yazarlar	bilim	teknoloji	28041	0.056
12	magazin	kadın	spor	yazarlar	bilim	28041	0.056
13	ana	sayfa	haber	ekonomi	dünya	28033	0.056
14	haber	magazin	kadın	spor	yazarlar	28021	0.056
15	ekonomi	dünya	kültür	sanat	tv	27981	0.056
16	haber	ekonomi	dünya	kültür	sanat	27981	0.056
17	sayfa	haber	ekonomi	dünya	kültür	27981	0.056
18	dünya	kültür	sanat	tv	derin	27960	0.056
19	arşivim	mesaj	grupları	sohbet	yardım	18989	0.038
20	mesaj	grupları	sohbet	yardım	üyelik	18989	0.038
21	dakika	arşivim	mesaj	grupları	sohbet	18988	0.038
22	grupları	sohbet	yardım	üyelik	site	18988	0.038
23	sohbet	yardım	üyelik	site	haritası	18988	0.038
24	son	dakika	arşivim	mesaj	grupları	18988	0.038
25	ana	sayfa	son	dakika	tüm	18830	0.038
26	arama	arşiv	bize	ulaşın	yardım	18830	0.038
27	arşiv	bize	ulaşın	yardım	copyright	18830	0.038
28	arşivim	arama	arşiv	bize	ulaşın	18830	0.038
29	astronet	televizyon	insan	kaynakları	arşivim	18830	0.038
30	bilim	teknoloji	yazarlar	kültür	sanat	18830	0.038
31	dakika	tüm	haberler	gündem	politika	18830	0.038
32	dosyalar	gezi	piyasanet	hava	durumu	18830	0.038
33	durumu	astronet	televizyon	insan	kaynakları	18830	0.038
34	dünya	ekonomi	spor	yaşam	bilim	18830	0.038
35	ekonomi	spor	yaşam	bilim	teknoloji	18830	0.038
36	gezi	piyasanet	hava	durumu	astronet	18830	0.038
37	haberler	gündem	politika	dünya	ekonomi	18830	0.038
38	hava	durumu	astronet	televizyon	insan	18830	0.038
39	insan	kaynakları	arşivim	arama	arşiv	18830	0.038
40	kaynakları	arşivim	arama	arşiv	bize	18830	0.038
41	magazin	özel	dosyalar	gezi	piyasanet	18830	0.038
42	özel	dosyalar	gezi	piyasanet	hava	18830	0.038
43	piyasanet	hava	durumu	astronet	televizyon	18830	0.038
44	sayfa	son	dakika	tüm	haberler	18830	0.038
45	son	dakika	tüm	haberler	gündem	18830	0.038
46	spor	yaşam	bilim	teknoloji	yazarlar	18830	0.038
47	teknoloji	yazarlar	kültür	sanat	magazin	18830	0.038
48	televizyon	insan	kaynakları	arşivim	arama	18830	0.038
49	tüm	haberler	gündem	politika	dünya	18830	0.038
50	yaşam	bilim	teknoloji	yazarlar	kültür	18830	0.038

B. Zipf's Law Graphics

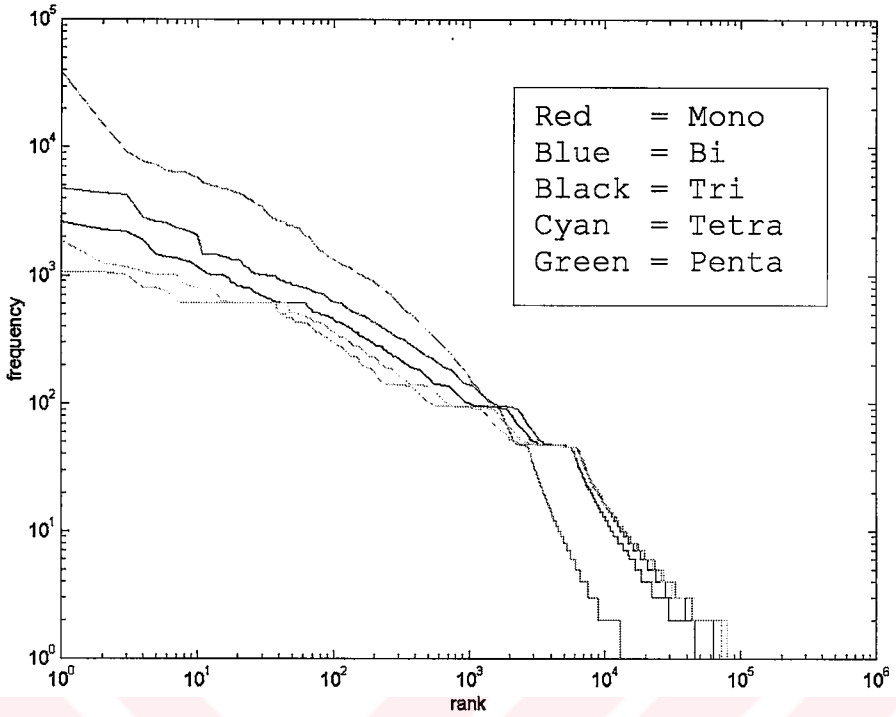
B.1. Arabul



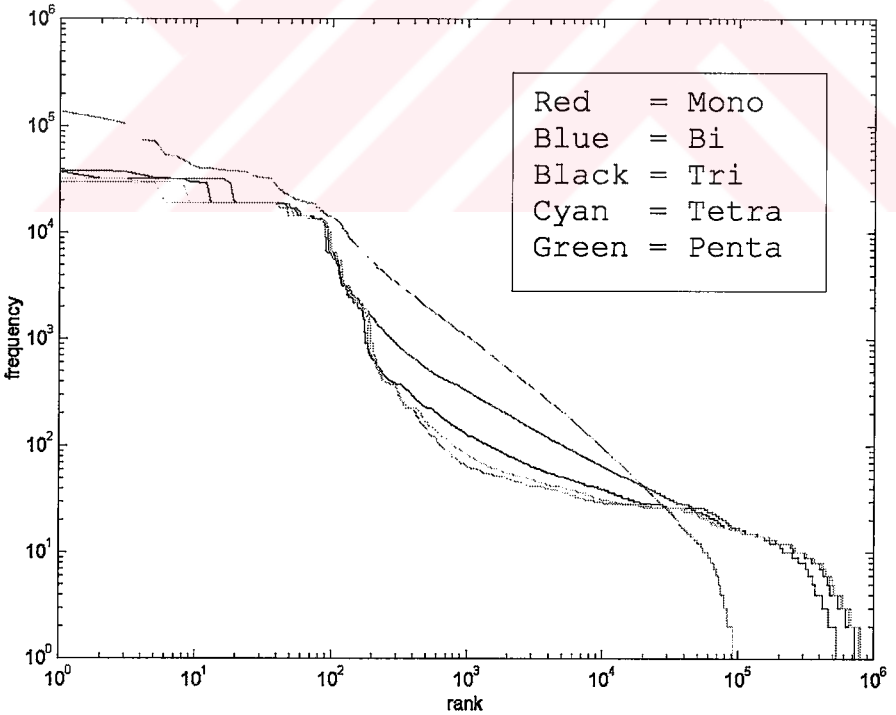
B.2. Bilim Teknoloji



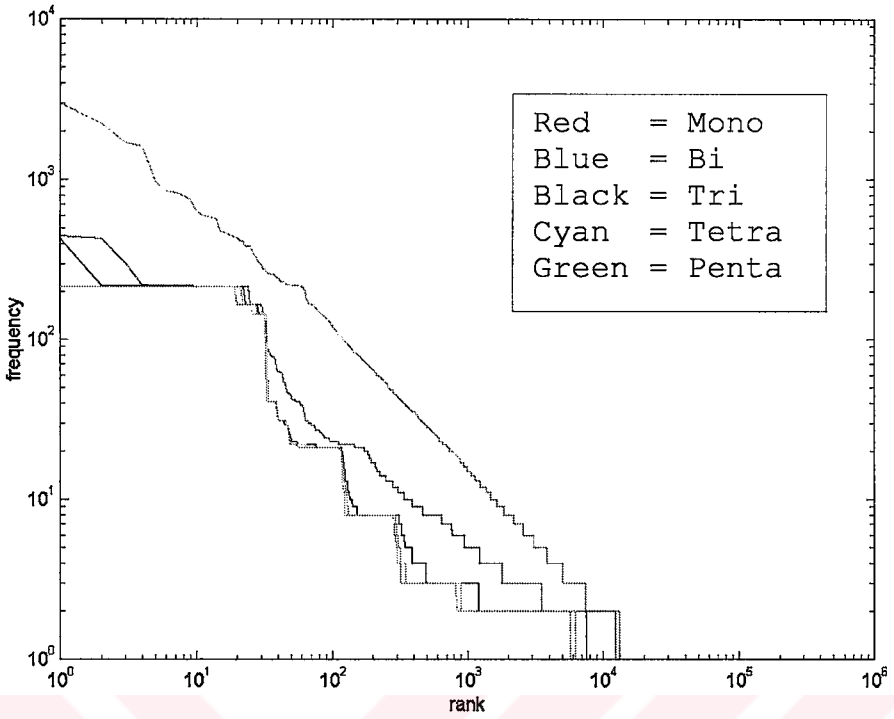
B.3. Devlet İstatistik Enstitüsü



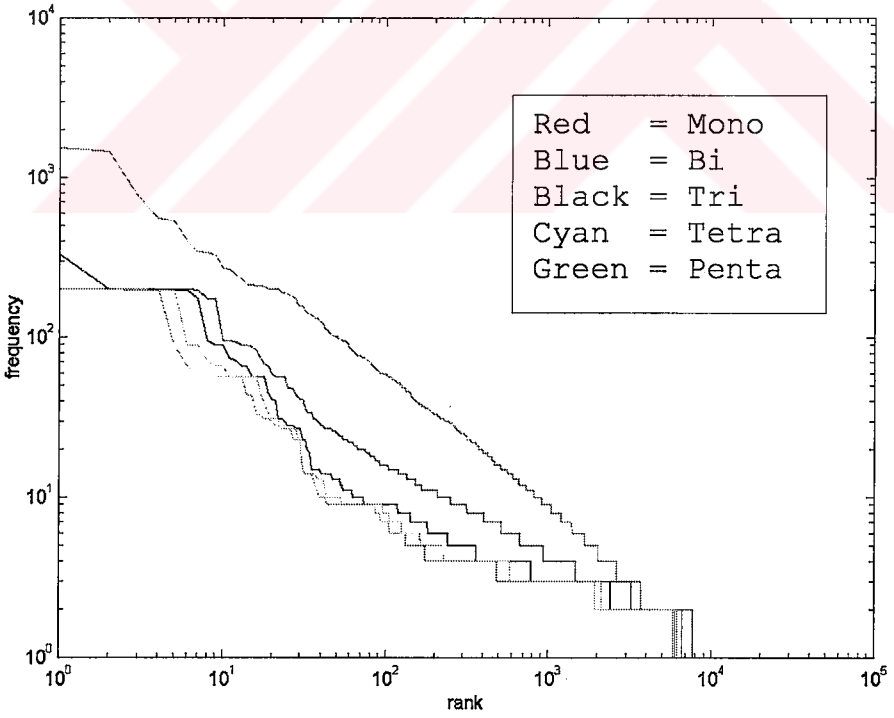
B.4. Hürriyet



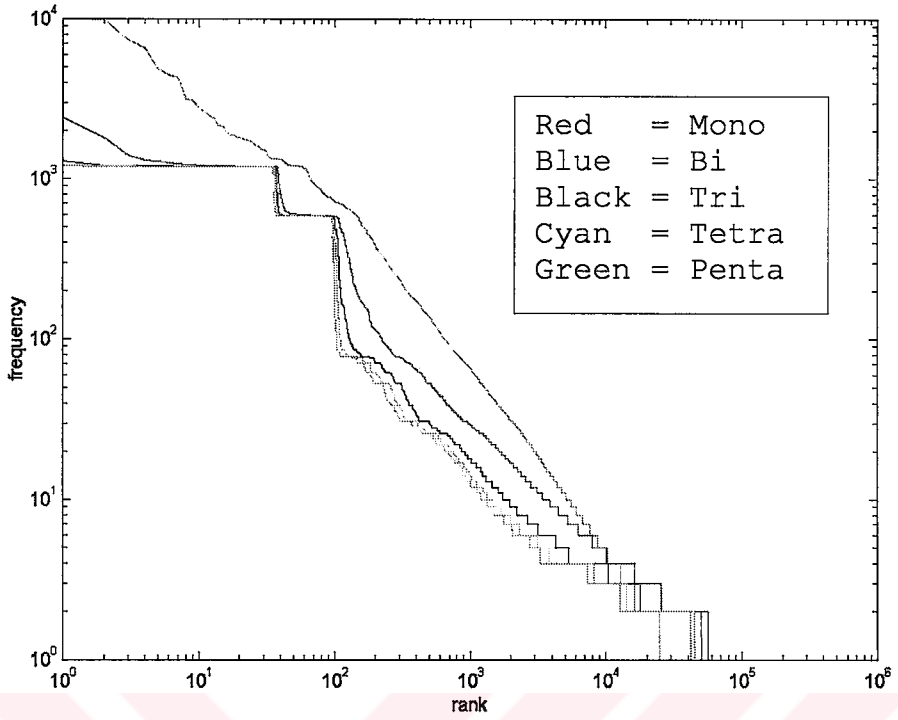
B.5. Lazland



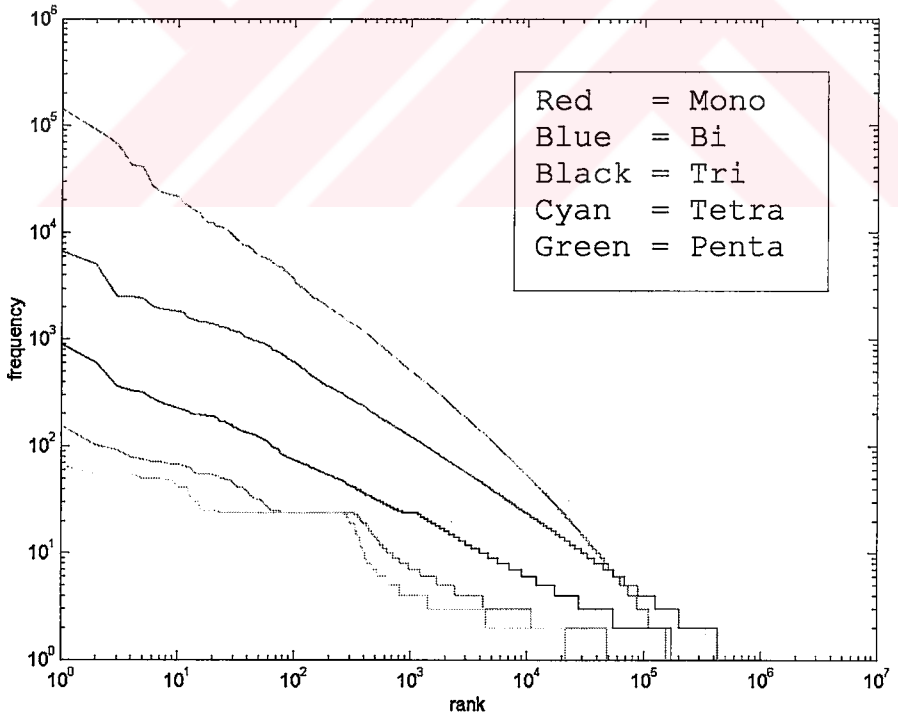
B.6. Pankitap



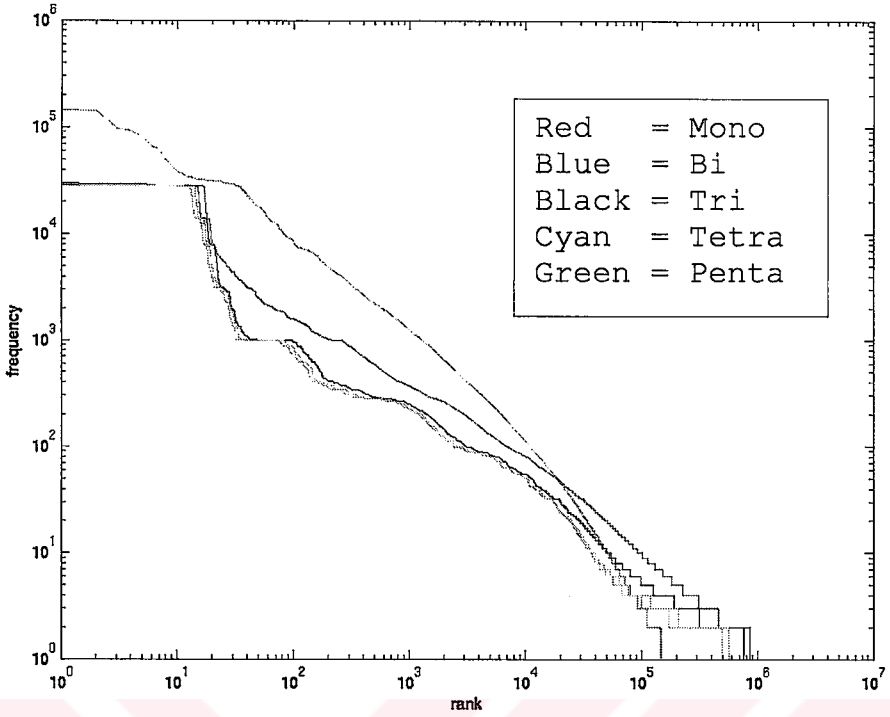
B.7. PCMagazin



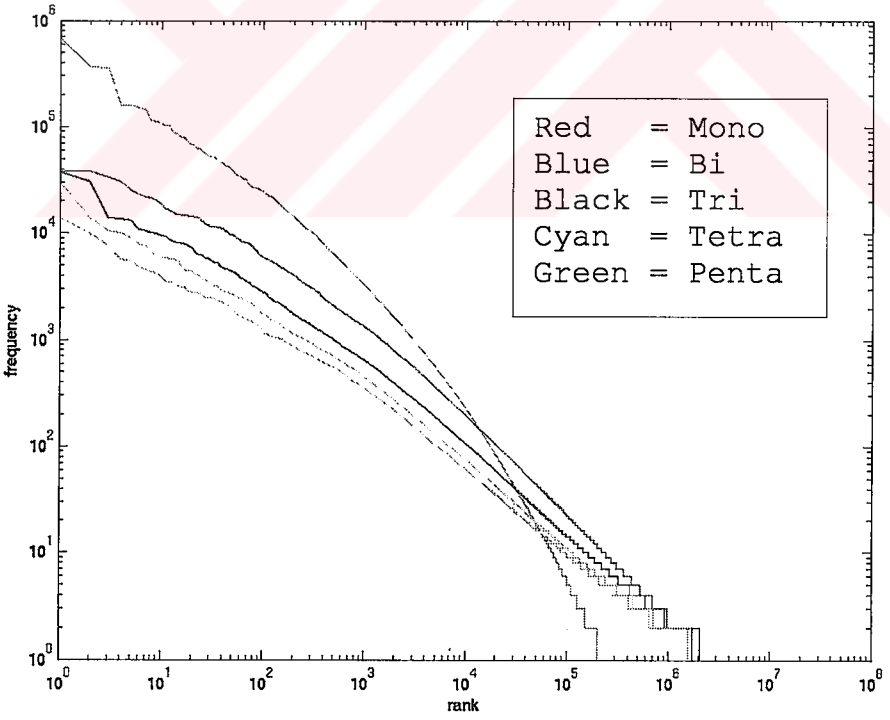
B.8. Novels & Stories



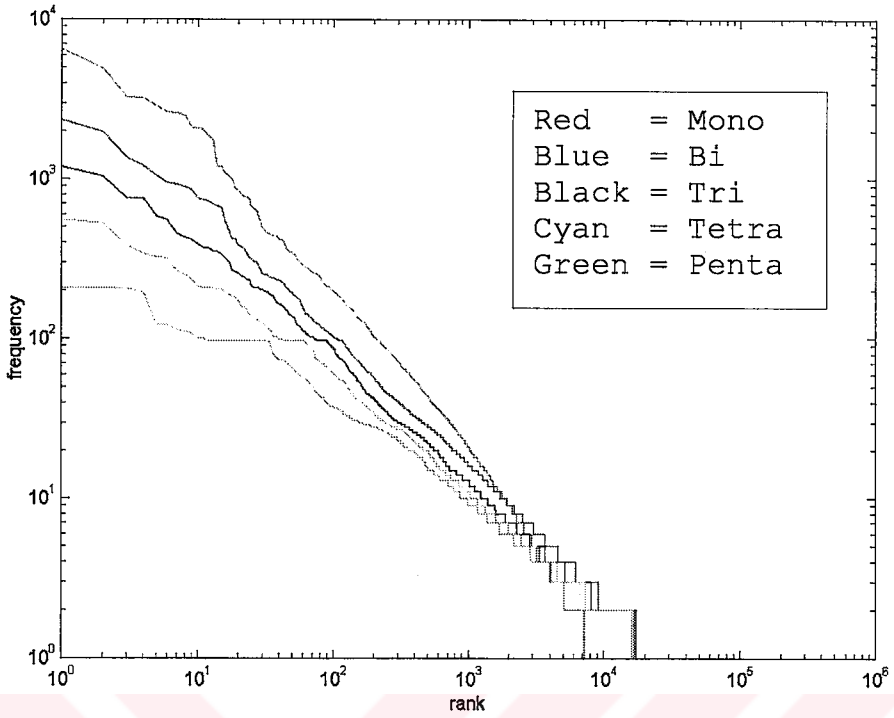
B.9. Star Gazete



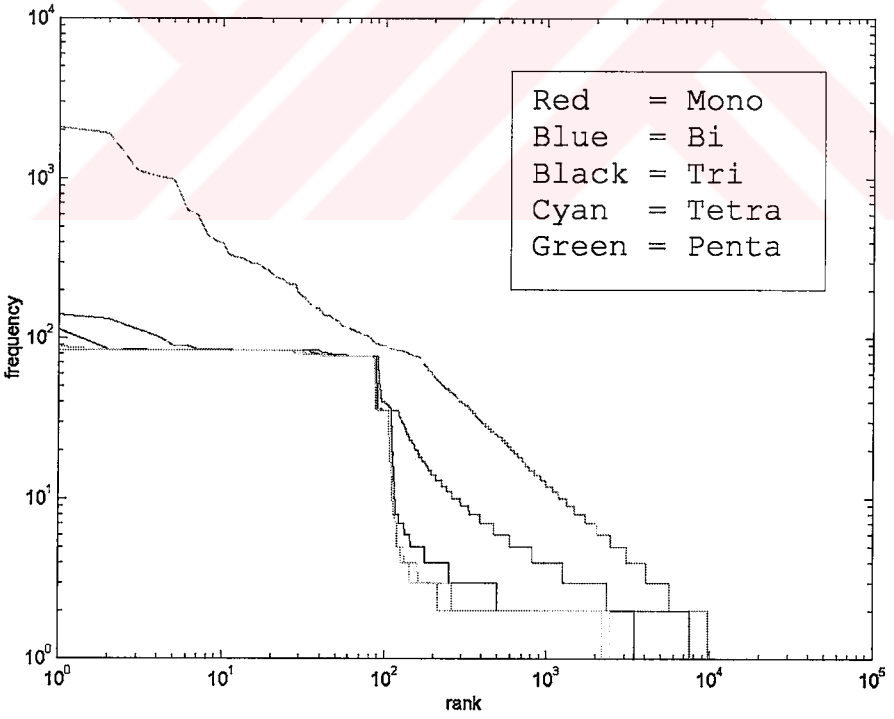
B.10. TBMM



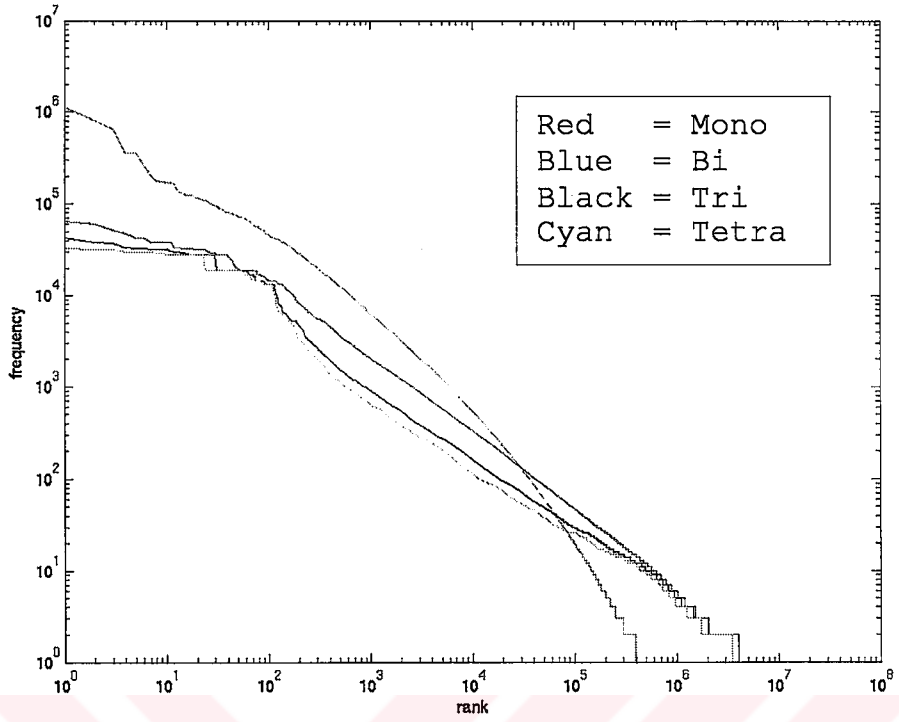
B.11. Ulusal Program



B.12. Yeni Asır



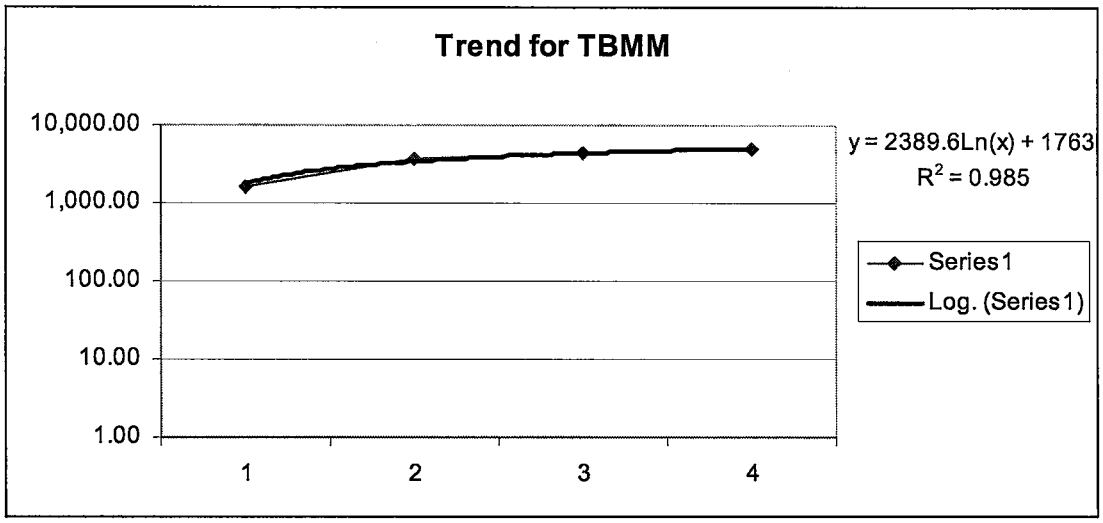
B.13. TurCo



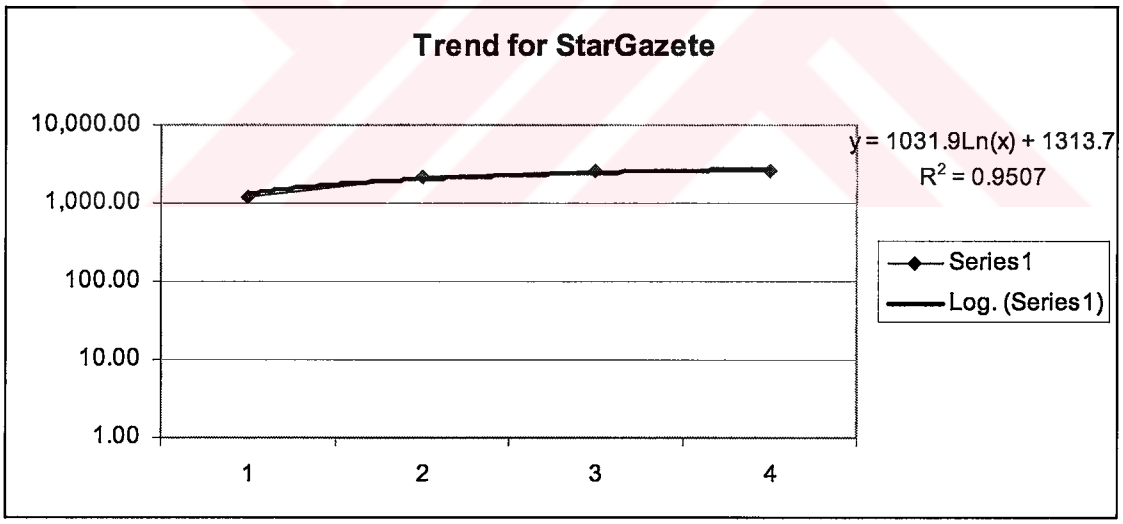
Renkli çıktı olmayacak ise
Beklil verilmeli.

C. NODW Increase Trend Between N-grams

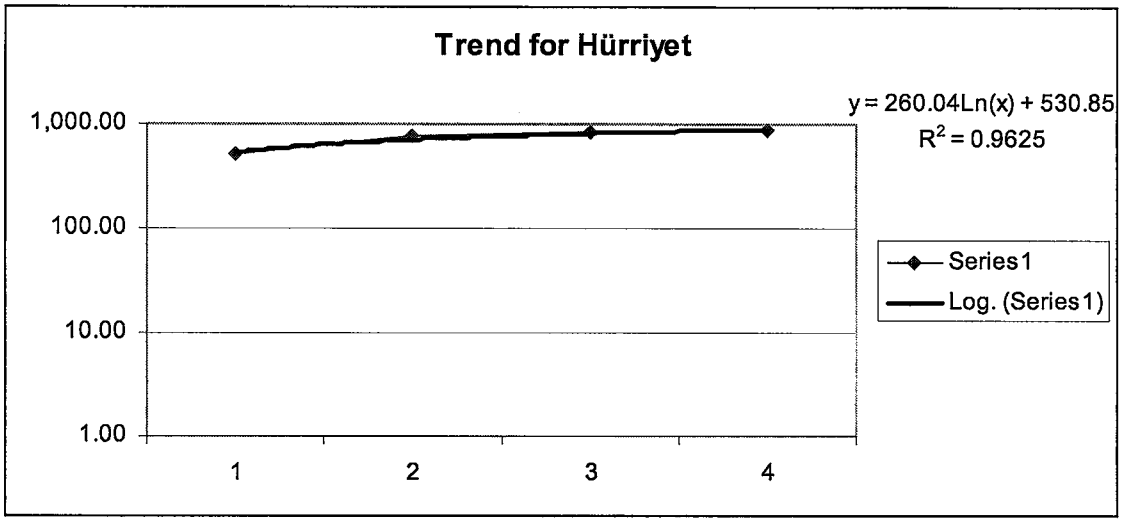
C.1. TBMM



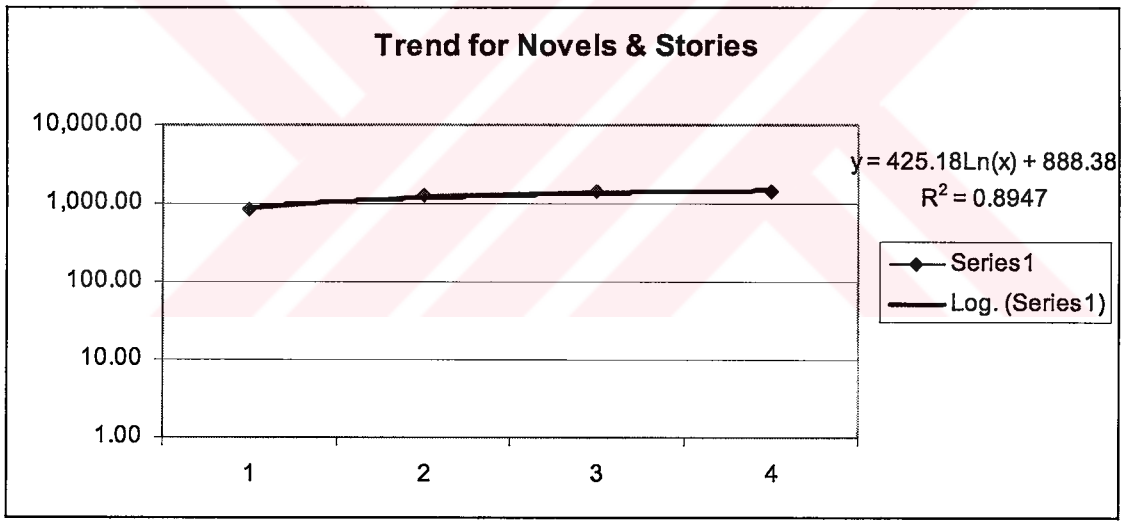
C.2. StarGazete



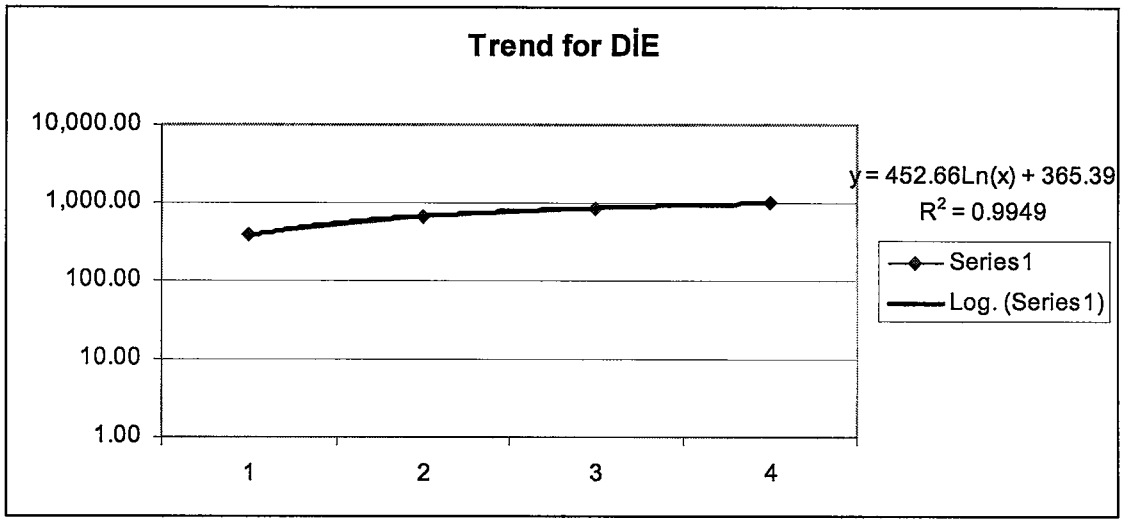
C.3. Hürriyet



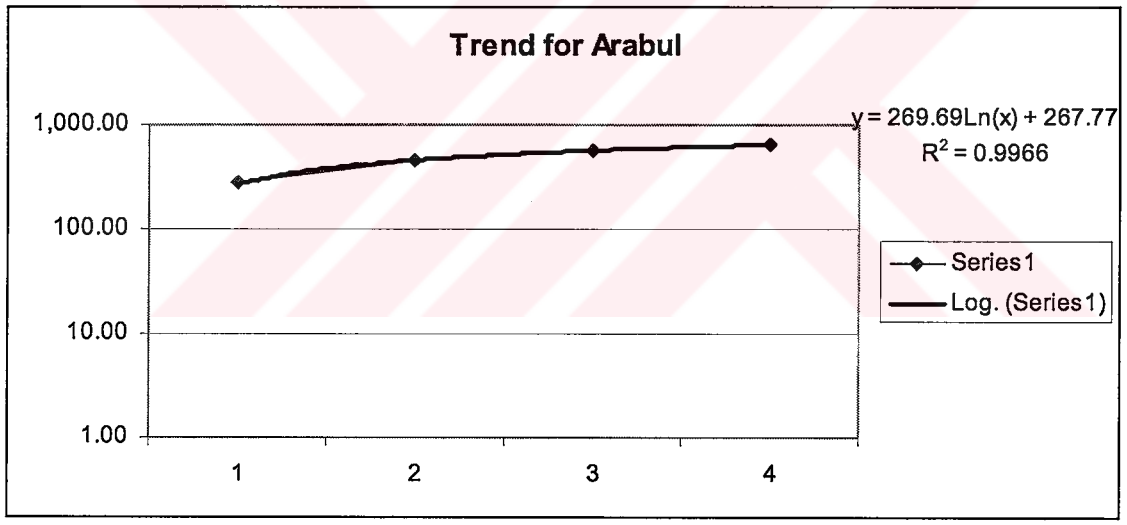
C.4. Novels & Stories



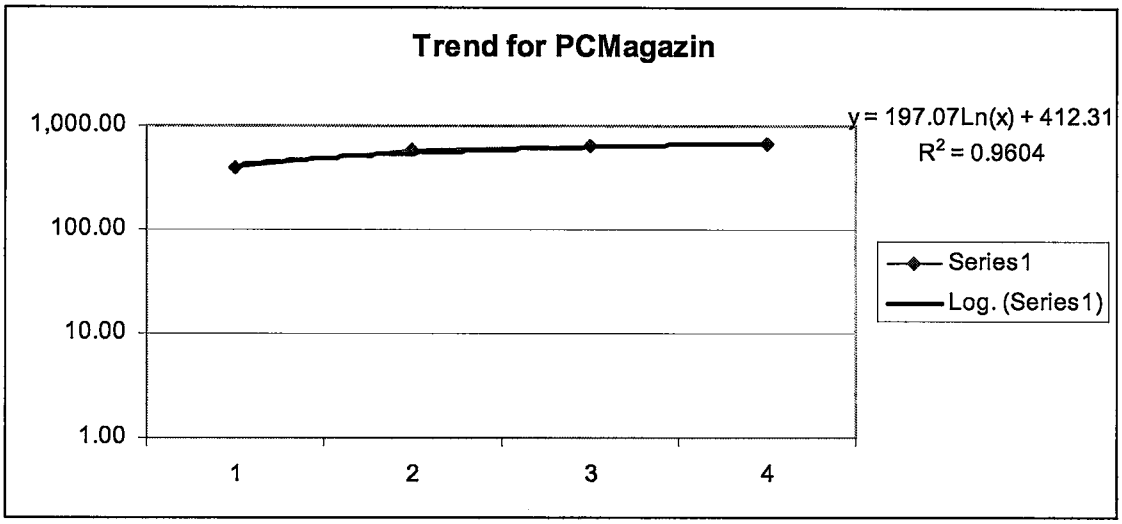
C.5. DİE



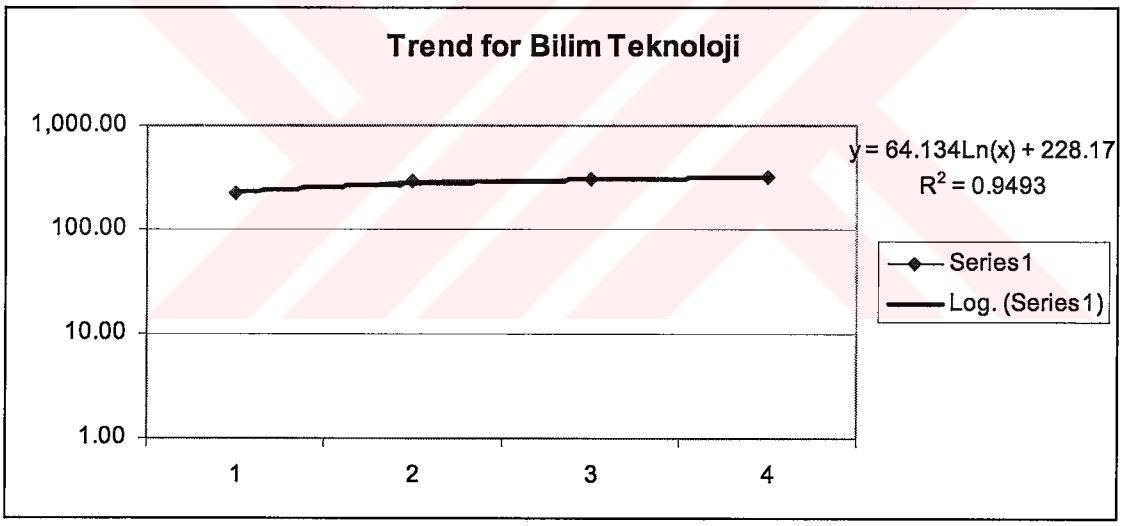
C.6. Arabul



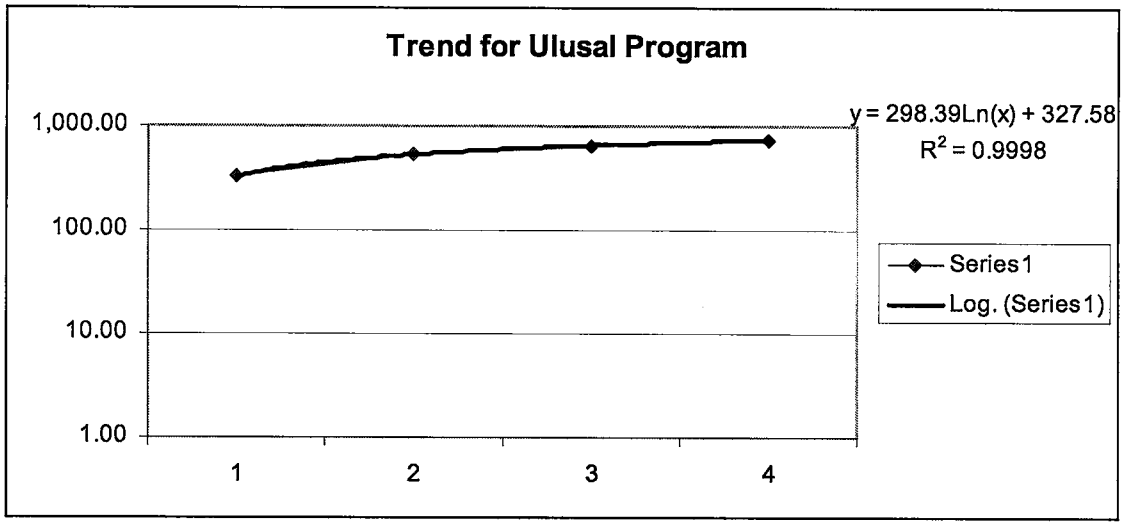
C.7. PCMagazin



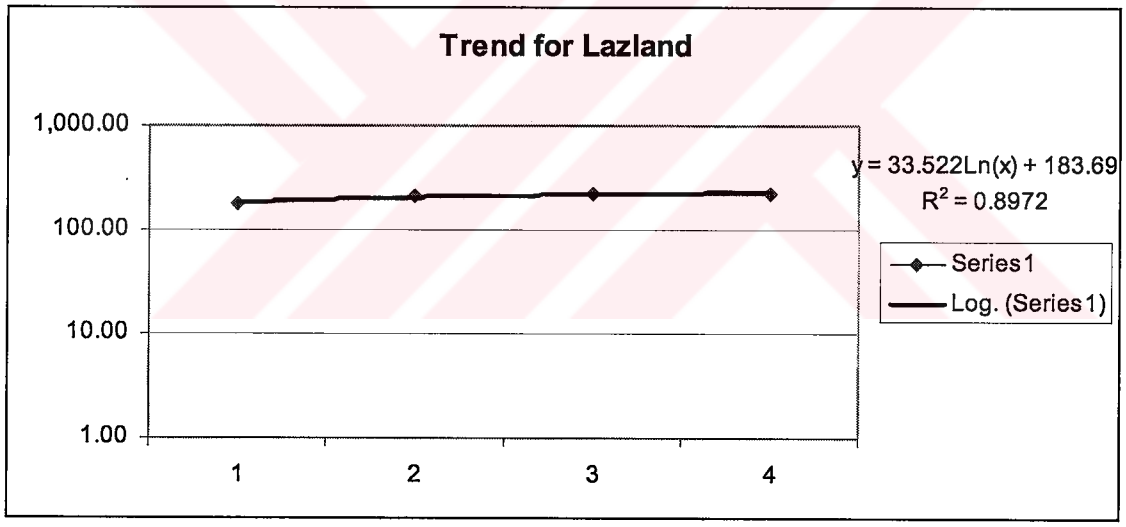
C.8. Bilim Teknoloji



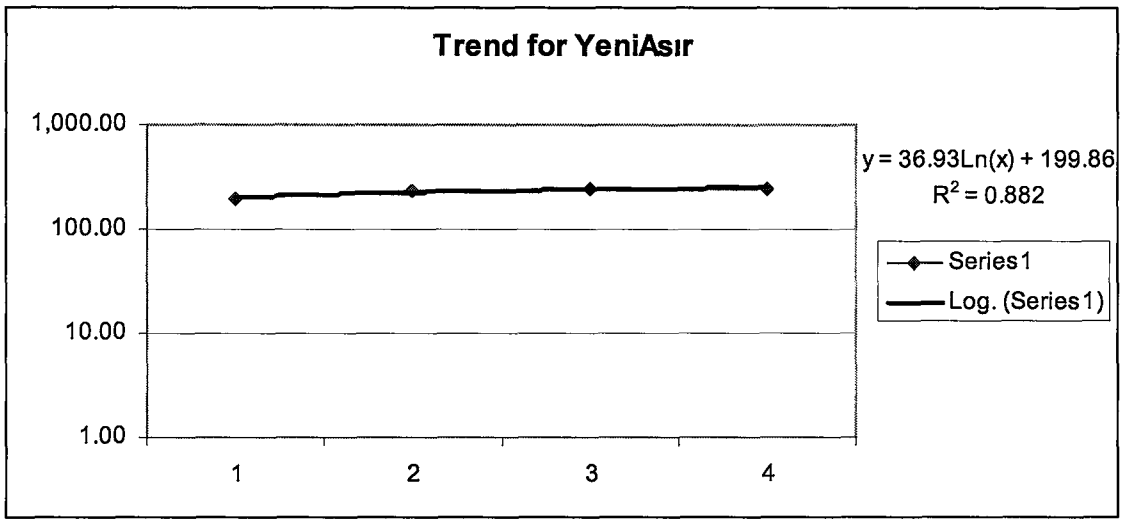
C.9. Ulusal Program



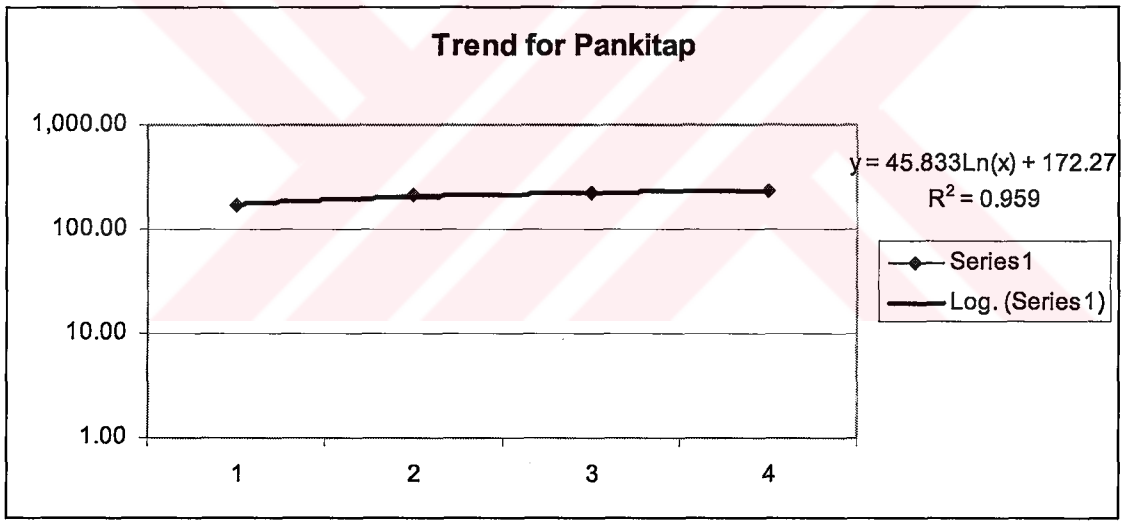
C.10. Lazland



C.11. YeniAsır



C.12. PanKitap



D. The Contents of the Novels & Stories

Author	Name of the Novel or Story	Size (Bytes)
Fahir Armaoğlu	20'inci Yüzyıl Siyasi Tarihi	1,789,935
Orhan Hançerlioğlu	Düşünce Tarihi	1,278,301
Haz. Bekir Onur	Çocuk Ve Ergen Gelişimi	847,028
Emre Kongar	Devrim Tarihi Ve Toplumbilim Açısından Atatürk	771,856
Gabriel Garcia Marquez	Yüzyıllık Yalnızlık	756,736
Stefan Zweig	Kendi Hayatının Şiirini Yazanlar	729,928
Murathan Mungan	Üç Aynalı Kırk Oda	721,717
P.Nikitin	Ekonomi Politik	716,305
Ahmet Hamdi Tanpınar	Huzur	712,727
Eleanor H. Porter	Pollyanna	632,607
Cengiz Aytmatov	Dişi Kurdun Rüyaları	618,686
Fakir Baykurt	Tırpan	598,925
Anonymous	Irvin D. Yalom	592,402
Yaşar Kemal	İnce Memed 1	589,135
İnci Aral	Erkek Ölü Kuşlar	573,938
Bekir Onur	Gelişim Psikolojisi Yetişkinlik Yaşlılık Ölüm	520,620
Sabahattin Ali	Bütün Öyküleri – II	509,423
İpek Ongun	Ve Gençler - Lütfen Beni Anla	507,717
Milan Kundera	Varolmanın Dayanılmaz Hafifliği	474,342
Doğan Cüceloğlu	İnsan İnsana	470,555
Yasar Kemal	Binboğalar Efsanesi	462,907
Selim Doğanay	Gözün Anatomisi Ve Görme Fizyolojisi	459,401
Sabahattin Ali	Değirmen,Kağnı,Ses	448,939
Bertrand Russell	Sorgulayan Denemeler	401,928
Yakup Kadri Karaosmanoğlu	Kiralık Konak	385,415
M. Sunullah Arısoy	Halk Şiiri Antolojisi	378,058
Yakup Kadri Karaosmanoğlu	Yaban	370,111
Montaigne	Denemeler	366,500
Georges Politzer	Felsefenin Başlangıç İlkeleri	356,535
Emile Zola	Apartman	347,185
Jean Paul Sartre	Duvar	335,843
Muzaffer İzgü	Zıkkımın Kökü	335,112
Nur Serter	Dinde Siyasal İslam Tekeli	306,110
Volney	Yıkıntılar	293,453
Erdal Öz	Gülünün Solduğu Akşam	293,328
John Steinbech	Yukarı Mahalle	275,534
Erdal Atabek	Kırmızı Işıkta Yürümek	272,033
Ferenc Molnar	Pal Sokağı Çocukları	269,720
İlhan Başgöz	Yunus Emre	263,107
Cengiz Aytmatov	Beyaz Gemi	258,254
Nadir Nadi	27 Mayıs'tan 12 Mart'a	255,556
Alphonse Daudet	Değirmenimden Mektuplar	248,638
Anonymous	Talât Paşa'nın Hatıraları	245,941
Nihat Behram	Darağacında Üç Fidan	243,574
Turgenyev	Rudin	243,500

D. The Contents of the Novels & Stories (Cont'd)

Author	Name of the Novel or Story	Size (Bytes)
John Grew	Atatürk Ve İnönü	239,036
Alan Durning	Tüketim Toplumu Ve Dünyanın Geleceği	238,783
Leyla Navaro	Beni Duyuyor Musun	238,260
Anonymous	Düşünüyorum Öyleyse Vurun	238,133
Berthe Georges Gaulis	Kurtuluş Savaşı Sırasında Türk Milliyetçiliği	226,982
Aziz Nesin	Şimdiki Çocuklar Harika	225,466
Theodor Storm	Kır Atlı	217,103
Nurer Uğurlu	Avrupa İle Asya Arasındaki Adamgazi Mustafa Kemal	216,832
Fiyodor Dostoyevski	Yeraltından Notlar	214,414
Jose Mauro De Vasconcelos	Şeker Portakalı	211,966
Cengiz Aytmatov	Toprak	209,433
Alan Lightman	Yıldızların Zamanı	209,431
Ron Coleman&Giles Barrie	Yöneticinin Klavuzu	206,448
Mahmut Adem	Atatürkçü Düşünce Işığında Eğitim Politikamız	203,413
Enver Ziya Karal	Tanzimat-I Hayriye Devri	202,301
Mahlon B. Hoagland	Hayatın Kökleri	199,826
Anonymous	Jules Amcam - Seçme Öyküler	196,903
Puşkin Dubrovski	Maça Kızı	196,135
Reşat Nuri Güntekin	Yaprak Dökümü	196,131
Oya Baydar	Elveda Alyoşa	195,260
Orhan Kemal	Cemile	194,826
O'henry	Nasıl Sevdi	192,575
Baki Öz	Atatürk'ün Anadolu'ya Gönderiliş Olayının İçyüzü	189,294
Dostoyevski	Beyaz Geceler-Uysal Kız	187,630
Anton Çehov	Korkunç Bir Gece	185,686
Sabahattin Eyüboğlu	Köy Enstitüleri Üzerine	185,328
Imre Madach	İnsanın Trajedisi	184,385
Actonio Krögerbu	Alacakaranlıkta	184,000
Voltaire	Candıde Ya Da İyimserlik	181,830
Bozkurt Güvenç	Kültürün Abc'si	180,108
Henrik Ibsen	Yaban Ördeği	173,669
E.T.A. Hoffmann	Uğursuz Miras	172,363
Kalman Mikszath	Konuşan Kaftan	171,803
Cengiz Aytmatov	Cemile	165,684
Pierre Loti	Doğudaki Hayalet	164,560
Anonymous	Mukaddes Ankara'dan-Mektuplar	155,411
Simone De Beauvoir	Sessiz Bir Ölüm	150,168
Aristoteles	Atinalıların Devleti	144,105
Anonymous	Satranç Üzerine	141,003
Ostrovski	Bu Hesapta Yoktu	140,457
Monumentum Ancyranum	Ankara Anıtı	140,394
Nurer Uğurlu	30 Ağustos Hatıraları	140,269
Tarık Z. Tunaya	Hürriyet'in İlanı	130,732
Turgenyev	Bozkırda	130,411
Platon	Devlet_Adamı	129,736
Maksim Gorki	Bozkırda	123,620

D. The Contents of the Novels & Stories (Cont'd)

Author	Name of the Novel or Story	Size (Bytes)
Firenc Herczeg	Bizans	122,884
H. De Balzac	Top Oynayan Kedi Mağazası	122,242
H. De Balzac	Bilinmeyen Başyapıt - Kırmızı Han	119,505
Peter Schlemihl	Adelbert Von Chamisso	118,449
Denis Diderot	Aykırı Düşünceler	117,897
A. Şemsutdinov	Kurtuluş Savaşı Yıllarında Türkiye - Sovyetler Birliği İlişkileri	117,285
Platon	Mektuplar	116,440
Muzaffer Ramazanoğlu	Gılgamış Destanı	115,456
Oscar Wilde	Lady Windermere'in Yelpazesi	112,564
Carlo Goldoni	Yazlık Dönüşü	108,432
Rıfat Miser	Toplum Kalkınması	108,356
Nurer Uğurlu	Kemalizm Sonrasında Türk Kadını	104,875
Beaumarchais	Sevil Berberi	99,616
Dostoyevski	Başkasının Karısı - Namuslu Hırsız	99,462
Francis Bacon	Yeni Atlantis	98,024
Sami Selçuk	Demokrasi manifestosu	91,481
Anonymous	Johann Wolfgang Goethe	85,781
Chateaubriand	Son İbni Sirac'ın Serüvenleri	82,132
Hamdi Tanses	Beste Ve Güfteleriyle Halk Türküleri	78,346
Anonymous	Satranç Ve Tarihçesi	74,353
Lessing	Yahudiler	74,037
Luigi Pirandello	Üç Kısa Oyun	71,082
Ömer Hayyam	Rubailer	69,908
Halil Köseler	Görme Özürlülerle İlgili Özel Eğitim Sorunları Ve Çözüm Yolları	66,382
Puşkin	Bakır Atlı	63,927
Anonymous	Vatandaşlık Ve Güçlendirme	41,591
Ahmet Gülüm	Dikkat Yazılı Var	38,187
	TOTAL	33,570,562

E. Turkish Alphabet

E.1. Lowercase Letters

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
a	b	c	ç	d	e	f	g	ğ	h	ı	i	j	k	l	m	n	o
19	20	21	22	23	24	25	26	27	28	29							
ö	p	r	s	ş	t	u	ü	v	y	z							

Consonants: {b, c, ç, d, f, g, ğ, h, j, k, l, m, n, p, r, s, ş, t, v, y, z}

Vowels: {a, e, ı, i, o, ö, u, ü}

E.2. Uppercase Letters

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	B	C	Ç	D	E	F	G	Ğ	H	I	İ	J	K	L	M	N	O
19	20	21	22	23	24	25	26	27	28	29							
Ö	P	R	S	Ş	T	U	Ü	V	Y	Z							

Consonants: {B, C, Ç, D, F, G, Ğ, H, J, K, L, M, N, P, R, S, Ş, T, V, Y, Z}

Vowels: {A, E, I, İ, O, Ö, U, Ü}