DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

TURKISH LANGUAGE CHARACTERISTICS AND AUTHOR IDENTIFICATION

by

Feriştah ÖRÜCÜ

July, 2009

İZMİR

TURKISH LANGUAGE CHARACTERISTICS AND AUTHOR IDENTIFICATION

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 Master of Science
in Computer Engineering, Computer Engineering Program

by
Feriştah ÖRÜCÜ

July, 2009
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"TURKISH LANGUAGE CHARACTERISTICS AND AUTHOR IDENTIFICATION"** completed by **FERİŞTAH ÖRÜCÜ** under supervision of **ASST. PROF. DR. GÖKHAN DALKILIÇ** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Gökhan DALKILIÇ

Supervisor

(Jury Member)

(Jury Member)

Prof.Dr. Cahit HELVACI
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to thank to my thesis advisor Assist. Prof. Dr. Gökhan Dalkılıç for his help, suggestions and guidance.

I also thank to all my colleagues to share their computers for my test studies and to my family and my sincere friends for their patience and support.

Feriştah ÖRÜCÜ

TURKISH LANGUAGE CHARACTERISTICS
AND AUTHOR IDENTIFICATION

ABSTRACT

Models of natural languages and language characteristics are widely used in many computer science applications such as data security, language identification, spell checking, data compression, authorship attribution and speech recognition. In the scope of this study, a large scale corpus is created and used to discover language characteristics of Turkish. Word and letter based analyses are made on this corpus to build a base for several NLP studies.

In the next step of the study, we used two different methods based on word n-grams to identify author of an anonymous text. For 16 authors, training and test set articles are collected, and mentioned two methods are applied on these article sets. Finally, obtained results from two methods are compared with each other and most successful method is determined.

Keywords : Turkish, Corpus, N-gram, Zipf's Law, Author Identification, Term Frequency, Inverse Document Frequency

TÜRK DİLİNİN KARAKTERİSTİKLERİ VE YAZAR TANIMA

ÖZ

Doğal dil modelleri ve dil karakteristikleri, bilgisayar bilimleri alanında veri güvenliği, dil teşhisi, imla denetimi, veri sıkıştırma, yazar tanıma ve ses tanıma gibi bir çok alanda sıklıkla kullanılmaktadır. Bu çalışma kapsamında, büyük ölçekli bir Türkçe külliyat oluşturularak, Türk diline ait karakteristiklerin keşfedilmesi amacı ile bir uygulama geliştirilmiştir. Çeşitli NLP çalışmalarına zemin hazırlamak amacıyla, külliyat üzerinde kelime ve harf bazlı bir çok analiz gerçekleştirilmiştir.

Çalışmanın bir sonraki adımında, yazarı bilinmeyen bir makalenin yazarını tahminlemek amacı ile, kelime n-gramları tabanlı iki farklı yöntem kullanılmıştır. 16 yazar için, çalışma ve test grubu makaleleri derlenmiş ve bahsi geçen iki yöntem bu makaleler üzerinde denenmiştir. Son olarak iki yöntemden elde edilen sonuçlar karşılaştırılarak, en verimli yöntem saptanmıştır.

Anahtar sözcükler : Türkçe, Külliyat, N-gram, Zipf's Kanunu, Yazar Tanıma, Terim Frekansı, Ters Döküman Frekansı

CONTENTS

Page

CHAPTER ONE

INTRODUCTION

The goal of this study is to obtain some statistical results about contemporary Turkish language and to determine important characteristics of Turkish by the analysis on a large scale Turkish text. Then we continue by comparing collected results with the results obtained from smaller corpora in previous studies or results obtained for different languages. Success and variation of the generated results are related with the amount of text used for analyzing. Therefore, 234,067 articles are collected to have sufficiently large text collection. Collected articles consist of articles of Akşam, Hürriyet, Milliyet, Radikal, Sabah, Tercüman, Vatan and Yeniasır newspapers.

1.1 Recent Studies

One of the first studies on corpus linguistics area is the study of Randolph Quirk 'Towards a description of English Usage' in 1960. Another important study was the publication by Henry Kucera and Nelson Francis of 'Computational Analysis of Present-Day American English' in 1967. This study was a work based on the analysis of the Brown Corpus, a carefully compiled selection of daily American English. A variety of computational analysis on compiled rich and assorted corpus, combining elements of linguistics, language teaching, psychology, statistics, and sociology was subjected by Kucera and Francis. Shortly thereafter, Houghton-Mifflin approached Kucera to supply a million words, three-line citation base for its new American Heritage Dictionary, the first dictionary to be compiled using corpus linguistics.

The Brown Corpus has also spawned a number of similarly structured corpora: the LOB Corpus (1960s British English), Kolhapur (Indian English), Wellington (New Zealand English), Australian Corpus of English (Australian English), the Frown Corpus

(early 1990s American English), and the FLOB Corpus (1990s British English). Other corpora represent many languages, varieties and modes.

Models of natural languages and language characteristics are widely used in many computer science applications such as data security (Stinson, 1995), (Seberry & JPieprzyk, 1988), language identification, correcting OCR (optical character recognition) text, spell checking (Teahan, 1998), data compression (Witten, Moffat & Bell, 1999), (Diri, 2000), authorship ascription (Gayde & Karslıgil, 2000), speech recognition (Santos and Alcaim, 2000), etc.

Previous studies in Turkish can be exemplified by Töreci (1975), Sezgin (1993), Koltuksuz (1995), Güngör (1995), Çiçekli & Temizsoy (1997), Oflazer (2000), Diri (2000), and Dalkılıç M.E. & Dalkılıç G. (2001).

1.2 Linguistic Features

In this part, definitions about linguistic features like Type/Token Ratio, Hapax Legomena Ratio, Index of Coincidence, Entropy, Redundancy and Unicity Distance will be explained. Type/Token Ratio is some kind of vocabulary diversity in language. Hapax Legomena Ratio is used to describe Lexical diversity. The Index of Coincidence for a text is the probability that two letters selected from it are identical. Entropy gives lower bound to the average number of bits per symbol needed to encode a message for a language. Redundancy is a measure for amount of constraint imposed on a text in the language and Unicity Distance is the minimum number of letters of encrypted text that have to be intercepted in order to render identification of the key. All these features change according to the language and the text. These features will be explained more detailed on the next parts of this chapter.

1.2.1 Type/Token Ratio (TTR)

Measurements of vocabulary diversity play an important role in language research and linguistic fields. The common measures used are based on the ratio of different words *(Types)* to the total number of words *(Tokens)*. This is known as the *Type-Token Ratio (TTR)* and can be calculated with the Formula 1.

$$TTR = \frac{Number\ of\ types\ \times 100}{Total\ number\ of\ tokens} \qquad (1)$$

If a text is 10,000 words long, it is said to have 10,000 "*Tokens*". But lots of these words will be repeated, and there may be only 5,000 "*Types*" means different words in the text. The ratio between types and tokens in this example would be 50%. But the type/token ratio (TTR) varies in accordance with the length of the text collection which is being studied. Larger samples give lower values for TTR. A 10,000 word text might have a TTR of 50%; a shorter one might reach 80%. Largest TTR means richer language usage.

1.2.2 Hapax Legomena Ratio (HR)

Hapaxes are words, which we used in the corpus only once. The Hapax Legomena Ratio (HR) is the ratio in percent between once-occurring types (*hapax legomena*) and the vocabulary size. This ratio is calculated by using Formula 2 given below.

$$HR = \frac{Number\ of\ once\ occuring\ types\ \times 100}{Total\ number\ of\ types} \qquad (2)$$

Type-token ratio (TTR) and *hapax legomena* ratio (HR) are used to describe Lexical diversity. These values can help notice differences of languages, or different authors of same language.

1.2.3 Index of Coincidence (IC)

IC was introduced by William Friedman in *The Index of Coincidence and its Applications in Cryptography* (Friedman, 1922). *Index of Coincidence (IC)* is a statistical measure of text which distinguishes encrypted text from plain text. The Formula 3 used to calculate IC:

$$IC = \sum_{i=1}^{N} \frac{(f_i \times (f_i - 1))}{N(N-1)} \qquad (3)$$

where f_i is the frequency of the i^{th} letter of the alphabet and N is the number of letters in alphabet.

1.2.4 Entropy (H)

In information theory (Shannon, 1948), the fundamental coding theorem states that the lower bound to the average number of bits per symbol needed to encode a message is given by its *entropy*.

The entropy is a statistical parameter which measures, in a certain sense, how much information is produced on the average for each letter of a text in the language. If the language is translated into binary digits (0 or 1) in the most efficient way, entropy H is the average number of binary digits required per letter of the original language.

Entropy values of n-gram series are calculated by using the Formula 4. "x" is every n-gram observed in the corpus, "p" is the probability of n-gram.

$$\mathrm{H}(X) = -\frac{1}{n} \sum_{x \in X} \mathrm{p(x)} \log_2 \mathrm{p(x)} \qquad (4)$$

Entropy is the lower bound to the number of bits per symbol required to encode a long string of text drawn from a language.

1.2.5 Redundancy (R)

The redundancy, measures the amount of constraint imposed on a text in the language due to its statistical structure. Number of characters in studied corpus, P is equal to 30 for Turkish (with space character). The *maximum redundancy* occurs when all the symbols have equal likelihood, and is equal to $\log_2 P = 4.91$ bits/letter.

Redundancy of an n-gram series is calculated by taking difference of its entropy from maximum redundancy value as shown in Formula 5.

$$\mathrm{R} = \log_2 P - H \qquad (5)$$

1.2.6 Unicity Distance (U)

In cryptology, substitution ciphers can be solved by exhaustively searching through the key space for the key that produces the decrypted text most closely resembling meaningful text. Instead, patterns and redundancy can be used to greatly narrow the search. As the amount of available cipher text increases, solving substitution ciphers becomes easier.

Unicity Distance is usually understood as the number of letters of encrypted text that have to be intercepted in order to render identification of the key and hence unique decryption possible. The *unicity distance*, defined as the entropy of the key space divided by per character redundancy, is a theoretical measure of the minimum amount of cipher text required by an adversary with unlimited computational resources. The expected unicity distance is accordingly Formula 6 given below:

$$\mathrm{U} = \frac{\mathrm{H(k)}}{\mathrm{R}} \qquad (6)$$

where U is the unicity distance, H(k) is the entropy of the key space and R is defined as the plaintext redundancy in bits per character.

In the next chapters, statistical analyses are given to make a base to future studies as author identification. Experimental results based on linguistic features which are collected from large text collection or corpus will be given. Following them, several letter and word based analyses and n-gram based analyses will be appended. Then, answers will be looked for if Turkish word and letter n-grams fit Zipf's Law.

CHAPTER TWO

GENERAL STATISTICS

2.1 General Statistics on Articles

Newspaper articles are used to obtain statistical results of contemporary Turkish. Some of these statistics are listed on Table 2.1.

The article collection consists of 234,067 articles and 109,300,288 words. Average word used per article is computed as 467. Number of total distinct words in the collection is observed as 1,173,041 (with affixes). Amount of distinct words observed per article was calculated as 330.

These analyses are made before construction of corpus, on article collection which includes punctuation marks and words which have characters like Q, X, W that are not belong to Turkish alphabet. Also article based analyses have to make before collecting all articles together.

Table 2.1 Some statistical results for Turkish article collection.

Total Article Count	234,067
Total Word Count	109,300,288
Word Count Per Article	466.962
Total Distinct Word Count	1,173,041
Count of Distinct Words Per Article	330.034
Type/Token Ratio	0.720
Count of Words Occurring Only Once (Hapax)	440,859
Count of Words Occurring Only Once Per Article	268.291
Hapax Legomena Ratio	0.812
Average Sentence Length	11.511
Average Word Length	6.159
Word Based Entropy	2.396

2.1.1 Punctuation Mark Frequencies in Turkish

Frequencies of some important punctuation marks are shown in Table 2.2. According to this table, for example, comma is used once per 15.912 words on average and exclamation mark is observed once in every 253.088 words.

Table 2.2 Frequencies of some major punctuation marks.

Punctuation Mark	Average Word Period
,	15.912
!	253.088
?	144.807
;	264.614
:	218.339

2.1.2 Type/Token Ratio (TTR) for Turkish Text

Value of Type/Token Ratio per article is calculated about 72% as seen from Table 2.1. In other words, 72 of every 100 words are different from each other. If TTR value is calculated on whole collection, TTR value decreases to a very low value like:

Total Distinct Word Count / Total Word Count = 1,173,041 / 109,300,288 ≅ 1.073%.

The fact under that is while the text collection is getting larger, instead of continuing to observe new words, some observed words are repeating.

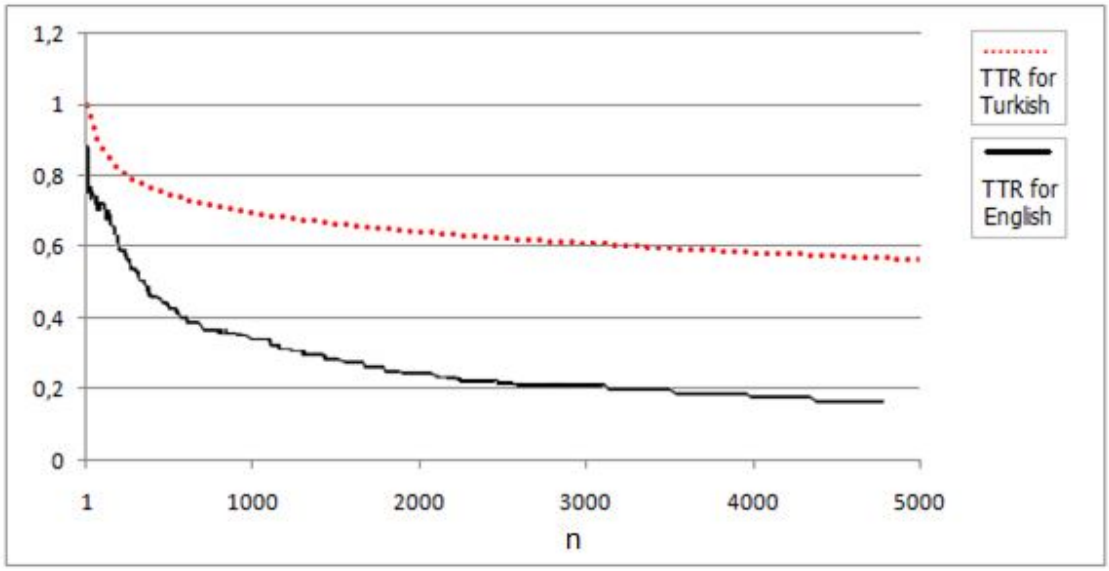


Figure 2.1 Type-Token Ratios for English and Turkish Corpus.

There exists a different strategy for computing TTR to prevent such low values for large texts. The standardized type/token ratio (STTR) is computed for every *n* (*n* = {1, 10, 20, 30, …, 5000} as can be seen Figure 2.1) words from each text file. In other words, if n is assumed as 1,000, the ratio is calculated for the first 1,000 running words, and then calculated afresh for the next 1,000, and so on to the end of corpus. A running average is computed, which means that an average type/token ratio based on consecutive 1,000-word chunks of text is computed.

Figure 2.1 shows relation between token count and TTR for an English text (Youmans, 1990). If same analysis is made on Turkish corpus, it can be seen that, TTR values is higher than English text. The fact under that is Turkish belongs to the group of *agglutinative languages* and Turkish morphology is quite complex, so words can be used with several affixes. Standardized TTR values for Turkish also can be seen on Figure 2.1.

2.1.3 Hapax Legomena Ratio for Turkish

Value of Hapax Legomena Ratio (HR) for whole text collection is calculated about 0.3758 from Table 2.1. In other words, 37.58 of each 100 words are used in collection, observed only once. When we look at the newspaper articles, average Hapax Legomena Ratio is calculated as 81.292% shown as below.

Average Hapax Legomena / Average Type Count = 268.291 / 330.034 ≅ 81.292%

Table 2.3 shows Hapax Legomena Ratios for English and German Texts (Schrader, 2006). Average hapax legomena value for Turkish newspaper articles is also given on this table.

Table 2.3 Hapax Legomena Ratios for English, German and Turkish Texts

Language	Tokens	Types	Hapax Legomena
English	29,077,024	101,967	39,200 (38.44%)
German	27,643,792	286,330	140,826 (49.18%)
Turkish	109,300,288	1,173,041	440,859 (37.58%)

2.2 Letter Based Analyses on Corpus

In this part of the study, one of the largest Turkish corpora was created by collecting a large amount of newspaper articles. This new corpus contains 105,863,484 words and 776,755,254 characters. Size of the corpus on disk is about 857 MB. It consists of 30 different characters; 29 characters of Turkish alphabet and the space character. All words containing Q, W, X characters which don't belong to Turkish alphabet are eliminated completely.

As collected texts contain newspaper articles instead of regular and errorless texts like stories and novels, the corpus is closed to contemporary Turkish language. Therefore the corpus has an extensive word variety. Several analyses based on letters and words were made on the corpus. In spite of working with such a large corpus have

many difficulties because of memory and time limitations. By using different algorithms like virtual corpus (Kit & Wilks, 1998) and partial corpus methods, difficulties were overcome and n-gram analysis were made (n = 1 to n = 100). Also 2-gram probability distribution table, entropy, redundancy, unicity distance values prepared for the corpus.

Turkish alphabet consists of 8 vowels (V) {A,E,I,İ,O,Ö,U,Ü} and 21 consonants (C) {B,C,Ç,D,F,G,Ğ,H,J,K,L,M,N,P,R,S,Ş,T,V,Y,Z}. In this study, also space character was used to separate words. Characters which are other than these 30 characters, like punctuation marks or letters of foreign languages are eliminated. Corpus contains only words which are formed by 29 Turkish capital letters and one space character between each sequential word.

Letter based analyses like Letter N-gram Distributions, Bigram[1] Distribution Table, Index of Coincidence, Entropy, Redundancy, Perplexity, Unicity Distance values for corpus, Most Common Letter N-grams and Letter Positions in Turkish are given in next parts of this section.

2.2.1 Letter N-gram Distributions

Table 2.4 shows maximum number of distinct n-grams that can be observed in corpus, the exact number of observed distinct n-grams and ratio between these two values. Maximum values are calculated as n^{th} power of alphabet's letter count (L^n). For example, as corpus contains 30 distinct characters, 30^2=900 different 2-grams can be observed. However, 899 different n-grams were observed in the corpus. The only missing 2-gram is "##" of course. "#" character is used instead of space character. While corpus has been created, just one space character is allowed to situate between two words. As a result, observation ratio of 2-gram letters is about 99.89%.

[1] Unigram (or monogram), bigram (or digram), trigram, tetragram, pentagram, hexagram, heptagram, octagram, nanogram, and decagram are used for respectively 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10-grams.

Table 2.4 Number of maximum and observed n-grams (*1≤n≤30)* for Turkish

	Maximum	Observed	Ratio %		Maximum	Observed	Ratio %
1-gram	30	30	100	**16-gram**	4.305E+23	415,591,550	-
2-gram	900	899	99.89	**17-gram**	1.291E+25	459,811,521	-
3-gram	27000	20,189	74.77	**18-gram**	3.874E+26	497,925,784	-
4-gram	8.100E+05	192,585	23.78	**19-gram**	1.162E+28	529,394,771	-
5-gram	2.430E+07	1,004,623	4.13	**20-gram**	3.487E+29	555,192,937	-
6-gram	7.290E+08	3,793,749	0.52	**21-gram**	1.046E+31	576,014,886	-
7-gram	2.187E+10	11,013,232	0.05036	**22-gram**	3.138E+32	595,068,519	-
8-gram	6.561E+11	25,460,011	0.00388	**23-gram**	9.414E+33	609,434,840	-
9-gram	1.968E+13	50,522,029	2.56 E-4	**24-gram**	2.824E+35	620,478,621	-
10-gram	5.905E+14	87,007,201	1.4 E-5	**25-gram**	8.473E+36	629,423,647	-
11-gram	1.771E+16	134,346,905	7.6 E-7	**26-gram**	2.542E+38	635,747,911	-
12-gram	5.314E+17	189,116,676	-[2]	**27-gram**	7.626E+39	640,750,275	-
13-gram	1.594E+19	248,904,914	-	**28-gram**	2.288E+41	644,599,440	-
14-gram	4.783E+20	308,424,787	-	**29-gram**	6.863E+42	647,193,362	-
15-gram	1.435E+22	364,355,219	-	**30-gram**	2.059E+44	649,634,588	-

While "*n*" in "n-gram" getting bigger, observation ratios are decreasing. When we look at the 8-grams, it can be seen that maximum value is 30^8=656,100,000,000, observed distinct 8-gram count 25,460,011 and the observation ratio is 0.0039. After 11-grams, observation ratios are too low to pay attention.

In contradiction to corpus collected from newspaper articles, observation ratios, calculated from corpora which are collection of stories, novel texts are a bit lower because of newspaper articles consist words just seen in speaking language. So, it is possible to see more varieties of n-gram combinations.

Observation ratios given on Table 2.4 are higher than the ratios calculated by using 11.5 MB corpus in the study of Dalkılıç M. E. & Dalkılıç G. (2001). Observation ratios are 95.11%, 42.13%, 8.45%, 1.10%, and 0.11% for 2-grams through 6-grams respectively in mentioned study. So, corpus size is an important factor on observation ratios.

[2] - shows discarded ratios

Table 2.5 Frequencies of Turkish bigrams per million letters

	#¹	A¹	B¹	C¹	Ç¹	D¹	E¹	F¹	G¹	Ğ¹	H¹	I¹	İ¹	J¹	K¹	L¹	M¹	N¹	O¹	Ö¹	P¹	R¹	S¹	Ş¹	T¹	U¹	Ü¹	V¹	Y¹	Z¹	∑
#²	0	16144	170	48	892	245	17886	430	131	63	373	10631	15290	51	9236	3233	4896	21268	793	6	1465	15448	771	2136	2731	5719	1506	152	802	3773	136289
A²	10012	291	4530	2493	1002	9513	69	1122	765	426	3985	8	176	67	7682	14805	8408	4013	22	*	1706	7171	4758	1489	5579	118	4	1859	8670	1666	102408
B²	16557	2088	35	14	111	14	527	5	2	5	47	103	856	*	31	119	132	185	131	28	5	384	17	186	144	135	53	7	96	25	22040
C²	1250	1579	6	13	*	1	1739	1	1	9	5	250	379	1	10	135	154	1660	314	15	1	330	44	4	31	197	243	58	21	89	8541
Ç²	3289	1090	*	*	2	*	903	5	*	2	105	45	1740	*	244	126	2	266	31	18	20	441	8	48	179	298	358	*	2	1	9224
D²	11698	3962	211	38	2	346	3251	4	14	94	45	361	1175	19	71	2366	767	6856	217	144	20	3108	10	6	27	309	169	96	658	742	36786
E²	4945	25	2487	2615	1358	11049	63	572	3520	474	1479	1	46	104	3103	12436	5915	5114	14	*	571	4588	3111	991	4797	18	17	4284	4639	1777	80113
F²	1750	896	*	*	*	1	515	39	*	1	2	98	355	*	18	12	46	94	115	20	1	63	43	10	46	58	74	3	53	1	4317
G²	7861	91	1	*	5	6	95	13	3	1	1	10	107	*	6	626	33	507	143	1	2	567	11	42	9	204	2	104	323	176	10952
Ğ²	2	1635	*	*	*	*	1807	*	*	*	*	1477	1542	*	*	*	*	*	866	231	*	*	*	*	*	910	230	*	*	*	8702
H²	5188	2064	1	157	5	5	304	3	10	*	20	8	569	*	13	29	178	18	48	8	51	119	71	18	106	212	22	7	37	10	9281
I²	413	5	306	720	1360	3604	4	334	247	2247	192	11	*	48	2053	4183	2258	5135	*	*	624	5138	4216	1721	3241	*	*	49	2240	996	41344
İ²	7644	436	7939	1087	1978	5988	36	597	2210	2832	1095	1	141	191	4901	6645	4016	5575	16	1	342	6817	4833	1415	4545	17	1	690	1458	1118	74565
J²	98	150	8	*	*	2	66	*	*	*	*	*	19	*	*	3	*	13	144	*	*	50	1	*	*	1	7	*	*	*	564
K²	10793	6925	2	28	24	11	5413	30	11	1	86	2618	3424	2	352	992	133	737	1955	158	123	2442	461	853	284	879	1213	32	142	46	40171
L²	811	6953	99	100	424	67	5333	230	34	525	173	3370	7687	31	3166	1733	1365	3667	5171	409	752	2750	639	1144	1415	2468	1125	387	2096	1133	55260
M²	4622	4704	40	11	227	5	3569	2	11	110	170	2164	3661	3	412	2683	97	1063	537	58	232	1773	236	968	1044	2031	1010	31	172	366	32011
N²	2333	14121	32	12	*	43	10002	8	6	13	61	9841	12581	*	107	54	44	162	3612	1774	51	292	49	6	31	4471	3213	6	468	14	63406
O²	6375	22	683	104	988	1145	64	260	145	4	182	14	42	23	2032	414	358	482	44	*	531	738	1930	35	894	45	2	39	5071	271	22938
Ö²	2465	2	459	3	144	493	*	15	1641	*	2	*	1	2	326	7	6	24	*	1	1	102	633	100	161	*	1	*	392	2	6984
P²	2320	1952	2	*	*	9	461	5	1	*	113	358	611	*	110	37	152	6	590	69	10	101	202	4	12	390	118	*	26	2	7662
R²	1291	16213	219	73	12	156	13552	135	249	565	125	3121	8404	*	536	1	64	453	6006	1532	450	93	82	23	523	2892	2694	468	349	8	60288
S²	8791	4499	5	3	19	30	3309	26	13	6	81	904	3040	1	707	358	376	966	415	226	148	1153	176	81	170	829	474	43	203	83	27134
Ş²	1769	3513	*	*	*	*	972	1	*	*	22	2183	2733	*	75	2	27	53	225	62	1	483	*	3	*	859	853	16	12	*	13866
T²	6327	3896	2	8	213	11	4873	245	3	*	210	300	1632	*	2084	582	19	820	418	190	361	1951	2381	1583	1083	830	405	1	65	10	30503
U²	1314	50	3967	773	94	2259	45	177	368	1070	358	2	5	12	1830	2309	1504	2794	33	*	117	2920	1344	454	1185	15	*	315	824	283	26419
Ü²	1491	1	768	222	360	1646	1	22	1557	229	283	*	1	6	1004	965	975	922	*	*	38	991	667	443	2117	*	2	39	1155	365	16270
V²	4759	1135	1	1	*	14	1634	*	1	1	34	40	243	*	14	24	2	17	87	85	*	113	20	78	112	108	148	45	43	4	8764
Y²	9003	5438	34	5	4	120	2808	30	2	*	22	1873	5089	2	46	361	29	434	822	886	38	88	421	24	29	1314	976	17	64	110	30089
Z²	1117	2528	31	10	*	1	813	3	6	23	11	1552	3018	*	3	17	55	102	168	1063	*	73	1	1	5	1092	1349	15	8	39	13109
∑	136289	102408	22040	8541	9224	36786	80113	4317	10952	8702	9281	41344	74565	564	40171	55260	32011	63406	22938	6984	7662	60288	27134	13866	30503	26419	16270	8764	30089	13109	1000000

2.2.2 Turkish Bigram Distribution

Table 2.5 shows frequencies of all observed 2-grams in the corpus. When this table is examined carefully, several important properties of Turkish can be determined. Columns present first character of 2-grams while rows present second characters. The numbers represented in the table are frequencies of 2-grams observed per million letters.

The value of the N^1-$\#^2$ couple (2-gram N#) 21,268 is the highest value in the table. As a result, it can be said, Turkish words are mostly ending with the letter N. If words ending with N are proportioned to all Turkish words, it can be seen 15.61% of all Turkish words end with the letter N. Likewise if we look at 17,886 times observed "E#" 2-grams, it can be seen that 13.12% of words end with the letter E, with the frequency of 16,144 "A#", 11.85% of words end with the letter A and with the frequency 15,448, 11.33% of words end with the letter R. 51.91% of all Turkish words are terminated by one of the these four letters.

When the bigrams which begin with space character are analyzed, it can be seen that 12.15% of words begin with "#B" bigram which has frequency 16,557; 8.58% of them begin with "#D" bigram which has frequency 11,698; 7.92% of them begin with "#K" diagram which has frequency 10,793; 7.35% of them begin with "#A" bigram which has frequency 10,012; 6.61% of them begin with "#Y" and 6.45 % of them begin with "#S" bigram. These six letters are stated as first character in 49.05% of all Turkish words.

When this table is examined carefully, although there is no word in Turkish beginning with the "Ğ" letter, frequency of "#Ğ" bigram is 2 per million. When the reason of this situation is researched, some usages listed below are explored;

- *" Erdoğan'ın başı **ğ**öğe mi ereeer... "*
- *" '**Ğ**' planımız var! "*
- *" Hem na **ğ**mağlup unvanı gitti, hem de şampiyonluk yolunda çok ama çok önemli 3 puanı Diyarbakır'da bıraktılar. "*

- *“ Bölge İdare Mahkemeleri’ne **ğ**önderme yapmayı ihmal etmedi. ”*

As can be seen from the examples, most important causes of “#Ğ” bigram are misspelling and using the letter “Ğ” by itself.

If the observation ratio of a bigram is less than 1 per million, it’s observation ratio is discarded and is shown with the “*” character.

The results in Table 2.5 are compared with the results in the similar table which are obtained by using only 11.5 MB corpus in the study of Dalkılıç M. E. & Dalkılıç G. (2001). Differentiation between frequencies of most commonly used 5 bigrams “N#”(21268 per 1,000,000), “E#”(17886), “#B”(16557), “AR”(16213), and “A#”(16144) in two tables are 0.0423%, 2.5926%, 10.4446%, 0.8864%, and 1.5550%. When we look at bigrams which have maximum differences, “KP”(110), “GS”(13), “MF”(46), “BY”(34), “DN”(43), “BN”(32), and “PN”(51) have differentiation rates as 2100.0%, 1200.0%, 1050.0%, 1033.3%, 975.0%, 966.7%, and 920.0%. According to these results, it can be said that, bigrams which have high frequencies have stable observation ratios independent from the size of corpus.

2.2.3 Index of Coincidence (IC)

For the corpus studied, N is equal to 30 (29 letters and space character). The *index of coincidence* for a text is the probability that two letters selected from it are identical. If such a text is generated randomly, the chance of pulling out an A is $^1/_{30}$. The probability of pulling out two As simultaneously is $(^1/_{30})*(^1/_{30})$. The chance of drawing any pair of letters is $30*(^1/_{30})*(^1/_{30}) = (^1/_{30}) = 0.0333$. So the IC of an evenly distributed set of corpus letters of a 30 letter alphabet is 0.0333.

Table 2.6 Frequency distribution for Turkish corpus characters.

Unigram	Ratio	Unigram	Ratio	Unigram	Ratio
#	13.629%	M	3.201%	G	1.095%
A	10.241%	T	3.050%	H	0.928%
E	8.011%	Y	3.009%	Ç	0.922%
İ	7.457%	S	2.713%	V	0.876%
N	6.341%	U	2.642%	Ğ	0.870%
R	6.029%	O	2.294%	C	0.854%
L	5.526%	B	2,204%	P	0.766%
I	4.134%	Ü	1.627%	Ö	0.698%
K	4.017%	Ş	1.387%	F	0.432%
D	3.679%	Z	1.311%	J	0.056%
				Total	**100**

When the Formula 3 is applied on values given in Table 2.6, IC value is calculated for Turkish as given below.

$IC = (R_{\#})^2 + (R_A)^2 + (R_E)^2 + \ldots + (R_J)^2$

$IC = (13.629)^2 + (10.241)^2 + (8.011)^2 + \ldots + (0.056)^2 = 0.063$

IC values of some other languages can be seen on Table 2.7 (Menezes, 1996).

Table 2.7 IC values of some languages.

Language	IC
French	0.0778
Spanish	0.0775
German	0.0762
Italian	0.0738
English	0.0667
Russian	0.0529
Turkish	0.0630

Cipher text encrypted with a substitution cipher would have an IC closer to 0.0333, since the frequencies would be closer to random. Turkish plaintext would have an IC closer to 0.063. This measure allows computers to score possible decryptions effectively. In cryptology, alphabet which is used for IC computation should not contain space character. In this case, IC value is 0.0596 for Turkish and 0.065 for English. These results are completely identical with the IC results obtained by using 11.5 MB sized smaller corpus in the study of Dalkılıç M. E. & Dalkılıç G. (2001).

2.2.4 Entropy (H)

The entropy of English text is between 1.0 and 1.5 bits per letter, or as low as 0.6 to 1.3 bits per letter, according to estimates by Shannon based on human experiments (Shannon, 1948). Previous studies were made with humans predicting text, found that the entropy of Turkish between 1.34 and 1.47 bit per letter, or as low as 0.56 to 0.62 (Dalkılıç M. E. & Dalkılıç G., 2001).

Computation on such a large scale corpus has many difficulties. Virtual corpus method (Kit & Wilks, 1998) assisted to overcome these difficulties. But after 6-grams, this method was not enough alone. Partial corpus method was used to compute n-gram entropy and frequency values. In partial corpus method, large scale corpus is separated into many equal sized small corpora and computations are made on these corpora. Finally, results are collected together on files by line by line iteration on partial results.

When calculated entropy values given on Table 2.8 are compared with the results calculated by Dalkılıç M. E. & Dalkılıç G. (2001) by using only 11.5 MB corpus, entropy values for first six n-gram groups are almost identical. This means corpus size is not important in these types of linguistic studies.

As can be seen in the Figure 2.2, entropy values form exponential distribution. Computed entropy value for 100-gram letters, is 0.29 which is dissimilar with the values predicted by Shannon tests for Turkish which is 0.56 to 0.62. For 100-grams and consequent n-gram series, entropy values are so close to normalized entropy value. Therefore, 0.29 is accepted as entropy of studied corpus.

According to Table 2.4 after 11-grams sample spaces for n-gram series are too low. For 100 grams sample space is equal to $\mathbf{776,750,007/30^{100}}$. If enough sample space for 100-grams was available, it can be possible to estimate entropy value for Turkish language. But it is theoretically impossible.

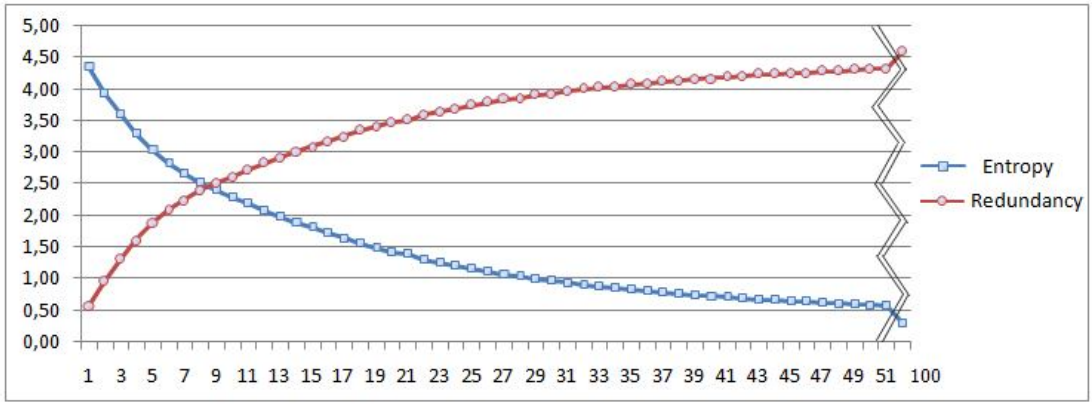


Figure 2.2 Entropy and Redundancy values for n-gram letters (1≤n≤100)

2.2.5 Redundancy (R)

Figure 2.2 shows redundancy values for Turkish letter n-grams calculated by using Formula 5. For example, redundancy of unigram letters in Turkish is

$$\mathrm{R} = \log_2 P - H = \log_2 30 - 4.35 = 0.56 \text{ bits.}$$

As seen from the figure the highest redundancy value is 4.62 which is for 100-grams.

2.2.6 Perplexity (PP)

The perplexity (PP) of a language is defined as entropy to the power of 2.

$$\boldsymbol{Perplexity = 2^H}$$

Perplexity is equal to $2^{4.3517} = 20.42$ for unigram letters and can be seen in Figure 2.3 for all n-gram groups. 20.42 goes down to 1.23 for 100-grams.

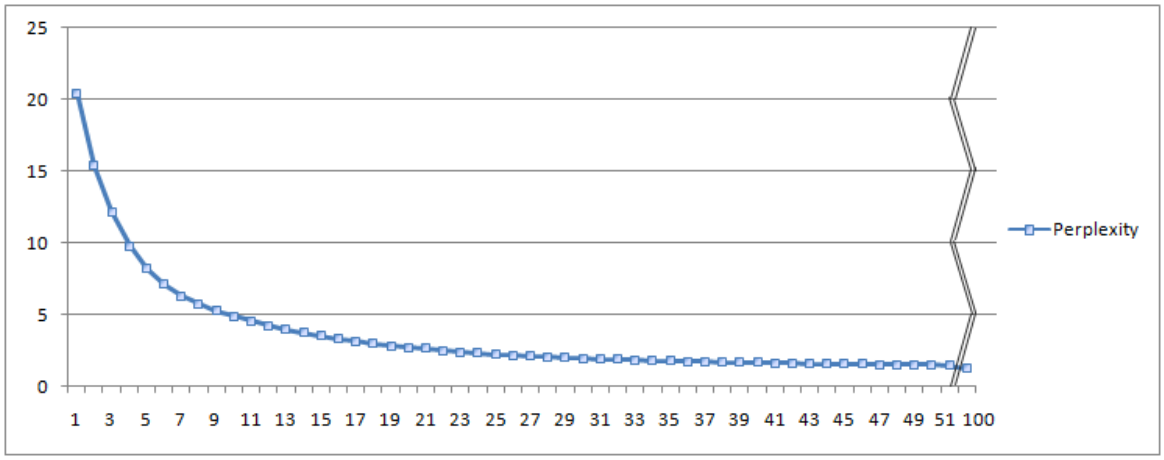


Figure 2.3 Perplexity values for n-gram letters (1≤n≤100)

Table 2.8 n^{th} order ($1 \leq n \leq 100$) Entropy, Redundancy, Unicity Distance and Perplexity values for corpus

	Entropy(bit/letter)	**Redundancy(bit/letter)**	**Unicity Distance**	**Perplexity**
1-gram	4.3517	0.56	192.92	20.42
2-gram	3.9411	0.97	111.17	15.36
3-gram	3.6034	1.31	82.43	12.15
4-gram	3.2923	1.62	66.58	9.80
5-gram	3.0342	1.88	57.42	8.19
6-gram	2.8277	2.08	51.73	7.10
7-gram	2.6611	2.25	47.89	6.33
8-gram	2.5215	2.39	45.09	5.74
9-gram	2.4021	2.51	42.95	5.29
10-gram	2.2919	2.62	41.14	4.90
11-gram	2.1880	2.72	39.57	4.56
12-gram	2.0881	2.82	38.17	4.25
13-gram	1.9928	2.92	36.92	3.98
14-gram	1.9004	3.01	35.79	3.73
15-gram	1.8113	3.10	34.76	3.51
16-gram	1.7264	3.18	33.83	3.31
17-gram	1.6457	3.26	33.00	3.13
18-gram	1.5697	3.34	32.25	2.97
19-gram	1.4983	3.41	31.57	2.83
20-gram	1.4316	3.48	30.97	2.70
21-gram	1.3693	3.54	30.42	2.58
22-gram	1.3122	3.60	29.94	2.48
23-gram	1.2586	3.65	29.50	2.39
24-gram	1.2085	3.70	29.10	2.31
25-gram	1.1619	3.75	28.74	2.24
26-gram	1.1183	3.79	28.41	2.17
27-gram	1.0777	3.83	28.11	2.11
28-gram	1.0398	3.87	27.83	2.06
29-gram	1.0043	3.91	27.58	2.01
30-gram	0.9712	3.94	27.35	1.96
31-gram	0.9402	3.97	27.13	1.92
32-gram	0.9110	4.00	26.93	1.88
33-gram	0.8835	4.03	26.75	1.84
34-gram	0.8576	4.05	26.58	1.81
35-gram	0.8332	4,08	26.42	1.78
36-gram	0.8101	4.10	26.27	1.75
37-gram	0.7883	4.12	26.13	1.73
38-gram	0.7676	4.14	26.00	1.70
39-gram	0.7479	4.16	25.88	1.68
40-gram	0.7293	4.18	25.76	1.66
41-gram	0.7115	4.20	25.65	1.64
42-gram	0.6946	4.22	25.55	1.62
43-gram	0.6785	4.23	25.45	1.60
44-gram	0.6631	4.25	25.36	1.58
45-gram	0.6484	4.26	25.27	1.57
46-gram	0.6343	4.28	25.19	1.55
47-gram	0.6208	4.29	25.11	1.54
48-gram	0.6079	4.30	25.04	1.52
49-gram	0.5955	4.31	24.96	1.51
50-gram	0.5836	4.33	24.90	1.50
51-gram	0.5721	4.34	24.83	1.49
100-gram	0.2919	4.62	23.33	1.23

2.2.7 Unicity Distance (U)

An alphabet of 32 characters can carry 5 bits of information per character (as $32 = 2^5$). In general the number of bits of information is log_2N, where N is the number of characters in the alphabet. So for English each character can convey $log_226 = 4.7$ bits of information.

However the average amount of actual information carried per character in meaningful English text is only about 1.5 bits per character. So the plain text redundancy is $R = 4.7 - 1.5 = 3.2$.

Basically the bigger unicity distance is the better. For a one time pad, given the unbounded entropy of the key space, we have $U = \infty$, which is consistent with the one-time pad being theoretically unbreakable.

For a simple substitution cipher, the number of possible keys is 26! = 4.0329 * 1026, the number of ways in which the alphabet can be permuted. Assuming all keys are equally likely, $H(k) = log_2(26!) = 88.4$ *bits*. For English text $R = 3.2$, thus $U = 88.4/3.2 = 28$. (Waters, 1976).

So given 28 characters of cipher text it should be theoretically possible to work out an English plaintext and hence the key.

If this study is made on Turkish corpus, each character can convey $log_230 = 4.91$ bits of information (N=30, 29 alphabet characters and space character). Average amount of actual information carried per character is only about 0.294 bits per character (computed for 100-gram letters). So redundancy value for studied corpus is $R = 4.91 – 0.294 = 4.616$ and unicity distance value of corpus is $U = log_2(30!)/4.616 = 23.33$. Redundancy values for all n-gram groups are given in Figure 2.4.

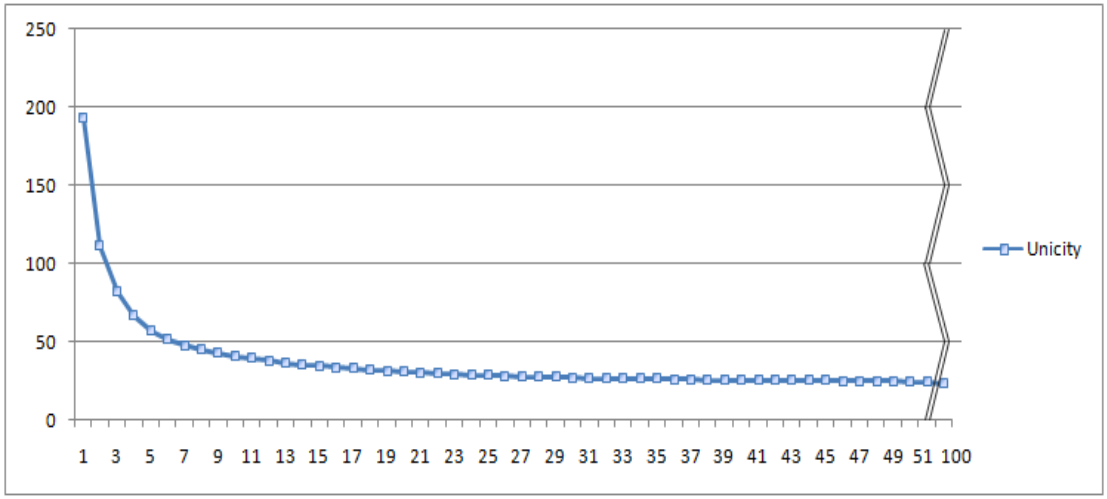


Figure 2.4 Unicity Distance values for n-gram letters (1≤n≤100)

Table 2.8 shows entropy, redundancy, unicity distance and perplexity values of Turkish corpus for n-gram groups (1≤n≤100). While variation between 1-grams' and 2-grams' entropy values is 0.4106, variation between 100-grams' and 101-grams' entropy decreasing to 0.0028. Since 100-grams, variances are becoming very low values and entropy values being stable. So entropy of 100-grams, 0.2919 ≈ 0.3, can be accepted as entropy of studied corpus. Same acceptance can be made for redundancy, unicity distance and perplexity values.

2.2.8 Most Common Letter N-grams

Table 2.9 shows most frequently used 30 letter n-grams of Turkish. Although n-gram analysis were made for 1-grams to 100-grams, as average word length is about 6.34 in Turkish shown on Table 2.15, to present meaningful values, only $1 \leq n \leq 7$ *n*-grams were illustrated.

As seen from Table 2.9, space character has the ratio of 13.629% which is the maximum according to all the letters. The most commonly used Turkish alphabet character is "A" with the 10.241% of ratio. The least commonly used Turkish letter is "J" with the ratio 0.056%. The most frequently used Turkish consonant letter is *N* with

the ratio of 6.341%. The 13 most frequently used characters together count 78.32 percent of letter occurrences.

Table 2.9 Most frequently used n-grams (*1≤n≤7)* for Turkish

1	**%**	**2**	**%**	**3**	**%**	**4**	**%**	**5**	**%**	**6**	**%**	**7**	**%**
#	13.629	N#	2.127	LAR	0.696	#BİR	0.427	#BİR#	0.3396	#İÇİN#	0.0799	TÜRKİYE	0.0624
A	10.241	E#	1.789	#Bİ	0.595	BİR#	0.351	LARIN	0.1749	LARIN#	0.0631	#TÜRKİY	0.0623
E	8.011	#B	1.656	LER	0.547	LARI	0.323	LERİN	0.1556	#TÜRKİ	0.0625	#KADAR#	0.0433
İ	7.457	AR	1.621	AN#	0.515	LERİ	0.289	INDA#	0.1259	TÜRKİY	0.0624	#OLDUĞU	0.0414
N	6.341	A#	1.614	İN#	0.487	#VE#	0.250	LARI#	0.1228	ÜRKİYE	0.0624	#OLARAK	0.0404
R	6.029	R#	1.545	İR#	0.480	YOR#	0.220	LERİ#	0.1106	LARINI	0.0613	OLARAK#	0.0403
L	5.526	İ#	1.529	EN#	0.469	ERİN	0.210	#İÇİN	0.1074	N#BİR#	0.0612	#DEĞİL#	0.0373
I	4.134	LA	1.481	ERİ	0.464	#BU#	0.207	İNDE#	0.1023	INDAN#	0.0585	LARINI#	0.0358
K	4.017	AN	1.412	DA#	0.463	INDA	0.206	İYOR#	0.0965	LERİNİ	0.0563	#SONRA#	0.0355
D	3.679	ER	1.355	#YA	0.456	LAR#	0.200	#TÜRK	0.0936	#DAHA#	0.0560	LERİNİ#	0.0327
M	3.201	İN	1.258	BİR	0.451	ARIN	0.197	İNİN#	0.0914	İ#BİR#	0.0527	ASINDA#	0.0297
T	3.050	LE	1.244	#DE	0.429	NDA#	0.184	N#BİR	0.0849	#GİBİ#	0.0524	LARINDA	0.0290
Y	3.009	#D	1.170	#KA	0.428	NİN#	0.163	NDAN#	0.0823	LERİN#	0.0514	#BÜYÜK#	0.0275
S	2.713	DE	1.105	ARI	0.427	İNDE	0.160	İÇİN#	0.0815	#DEĞİL	0.0500	ÜRKİYE#	0.0275
U	2.642	#K	1.079	DE#	0.420	İYOR	0.160	ININ#	0.0762	#KENDİ	0.0471	LERİNDE	0.0263
O	2.294	I#	1.063	YOR	0.364	DEN#	0.157	IYOR#	0.0744	#KADAR	0.0455	ARININ#	0.0250
B	2.204	#A	1.001	IN#	0.358	DAN#	0.156	#DEĞİ	0.0742	LARAK#	0.0442	#BAŞKAN	0.0241
Ü	1.627	EN	1.000	#BU	0.356	LER#	0.151	ARIN#	0.0711	KADAR#	0.0433	ERİNDE#	0.0237
Ş	1.387	IN	0.984	AR#	0.355	NIN#	0.145	#OLMA	0.0676	#SONRA	0.0433	ERİNİN#	0.0233
Z	1.311	DA	0.951	#VE	0.352	ERİ#	0.144	ARINI	0.0670	#BAŞKA	0.0430	YORLAR#	0.0227
G	1.095	K#	0.924	#OL	0.344	ARI#	0.143	ANLAR	0.0646	ASINDA	0.0429	#DEVLET	0.0226
H	0.928	#Y	0.900	#BA	0.335	#BAŞ	0.137	ERİNİ	0.0630	E#BİR#	0.0426	LARININ	0.0225
Ç	0.922	#S	0.879	ARA	0.322	İNİ#	0.136	#ÇOK#	0.0630	ERİNDE	0.0424	#İÇİNDE	0.0225
V	0.876	YA	0.867	NDA	0.309	#DE#	0.134	#OLDU	0.0627	OLDUĞU	0.0416	NLARIN#	0.0224
Ğ	0.870	MA	0.841	#GE	0.307	NLAR	0.134	TÜRKİ	0.0626	#OLDUĞ	0.0414	N#SONRA	0.0220
C	0.854	İR	0.840	ER#	0.287	#OLA	0.133	RKİYE	0.0625	İNDEN#	0.0407	K#İÇİN#	0.0217
P	0.766	Bİ	0.794	N#B	0.277	İNE#	0.132	NLARI	0.0624	YORUM#	0.0406	#GERÇEK	0.0217
Ö	0.698	#G	0.786	İNİ	0.270	INI#	0.131	ÜRKİY	0.0624	OLARAK	0.0404	RASINDA	0.0213
F	0.432	İL	0.769	#HA	0.263	#DA#	0.131	ARAK#	0.0622	#OLARA	0.0404	#GÖSTER	0.0213
J	0.056	KA	0.768	İLE	0.259	NDE#	0.122	ANIN#	0.0619	#KARŞI	0.0403	LERİNİN	0.0213
∑	**100**		**35.35**		**12.09**		**5.63**		**2.83**		**1.51**		**0.91**

Table 2.10 Letter positions in Turkish words which are 1 to 26 characters length.

	1	2	3	4	5	6	7	8	9	10	11	12	13
a	7,347	**21,340**	5,101	**16,210**	11,003	10,154	**14,163**	9,479	11,191	10,271	9,299	9,823	8,540
b	**12,148**	0,404	1,549	1,285	0,692	0,515	0,529	0,224	0,135	0,232	0,069	0,056	0,088
c	0,918	0,263	1,363	0,864	1,499	1,174	1,410	1,013	0,777	0,996	0,808	0,542	0,685
ç	2,413	1,574	1,590	0,759	0,314	0,487	0,182	0,201	0,297	0,048	0,088	0,031	0,030
d	8,583	1,349	3,664	4,130	2,71	4,875	3,351	5,064	4,686	3,556	5,340	4,168	4,627
e	3,629	17,904	3,104	11,990	9,197	8,381	11,900	8,339	10,517	9,512	7,785	9,843	9,107
f	1,284	0,180	0,716	0,333	0,705	0,217	0,120	0,234	0,026	0,148	0,013	0,007	0,008
g	5,768	0,060	0,958	1,292	0,096	0,165	0,204	0,061	0,030	0,036	0,011	0,020	0,008
ğ	0,001	0,584	1,994	0,144	1,235	1,509	1,175	2,153	1,491	1,674	1,880	1,310	1,551
h	3,806	0,358	1,760	0,429	0,644	0,196	0,069	0,241	0,063	0,014	0,010	0,009	0,027
ı	0,303	2,735	1,328	6,853	5,047	7,282	8,760	7,886	11,302	9,816	11,253	12,389	11,712
i	5,609	10,928	3,399	11,146	8,353	9,048	10,675	8,944	**13,028**	10,848	**13,703**	13,388	12,737
j	0,072	0,013	0,055	0,090	0,119	0,029	0,127	0,135	0,004	0,031	0,003	0,012	0,001
k	7,919	1,380	6,714	4,518	4,45	4,602	2,920	3,628	3,780	4,216	3,642	3,314	3,118
l	0,595	5,699	9,010	7,221	8,874	**10,619**	6,188	7,358	5,976	4,188	5,176	3,815	3,595
m	3,391	1,017	4,490	3,956	5,084	5,858	4,215	3,666	3,450	2,975	2,995	3,176	2,910
n	1,711	3,502	9,186	4,352	**11,37**	7,385	10,376	**12,619**	10,228	**18,436**	13,394	**17,836**	**20,540**
o	4,677	6,423	0,642	1,015	1,63	2,067	1,427	2,131	1,632	1,132	1,250	0,793	0,636
ö	1,809	3,048	0,017	0,132	0,132	0,041	0,086	0,015	0,015	0,025	0,004	0,003	0,004
p	1,703	0,175	2,450	0,486	1,006	0,256	0,270	0,214	0,142	0,181	0,068	0,056	0,061
r	0,947	2,953	**16,782**	3,973	7,965	6,126	7,057	11,431	9,165	10,825	11,619	9,604	10,644
s	6,450	1,680	3,347	2,103	3,039	3,724	1,920	2,925	1,800	1,645	2,259	1,023	1,271
ş	1,298	0,824	3,173	1,040	2,618	1,600	0,889	1,766	1,307	0,846	0,975	0,620	0,912
t	4,642	1,323	5,242	4,076	3,619	4,740	3,482	2,081	2,087	1,662	0,931	1,062	0,722
u	0,964	5,849	1,598	5,657	2,542	2,782	2,628	2,406	2,447	1,893	2,502	1,686	1,761
ü	1,094	5,219	0,360	3,309	1,413	1,111	1,220	0,523	0,864	0,365	0,433	0,243	0,181
v	3,492	0,430	1,823	0,463	0,384	0,159	0,181	0,057	0,036	0,022	0,021	0,009	0,008
y	6,606	1,658	4,614	1,727	3,472	3,873	3,128	3,603	1,877	2,087	2,019	1,119	1,117
z	0,820	1,124	3,968	0,449	0,789	1,026	1,346	1,601	1,647	2,321	2,454	4,047	3,400

	14	15	16	17	18	19	20	21	22	23	24	25	26
a	8,615	8,860	6,309	5,803	6,089	5,698	6,677	5,045	4,404	4,122	5,276	6,038	13,821
b	0,110	0,118	0,150	0,033	0,046	0,057	0,061	0,046	0,137	0,429	0,315	0,884	0,542
c	0,936	0,501	0,628	0,359	0,739	0,824	1,056	0,193	0,372	0,472	0,315	0,442	0,813
ç	0,037	0,024	0,027	0,030	0,039	0,062	0,065	0,055	0,078	0,043	0,079	0,147	-
d	5,547	3,468	3,895	3,893	3,903	5,911	3,932	2,950	3,230	3,092	3,543	2,651	5,962
e	8,665	10,971	7,748	9,412	9,300	8,944	12,100	10,467	8,710	8,373	8,740	12,224	**14,092**
f	0,005	0,006	0,005	0,010	0,009	0,012	-	0,009	0,059	-	0,079	-	-
g	0,007	0,010	0,013	0,017	0,017	0,030	0,042	0,083	0,098	0,301	0,236	0,147	0,542
ğ	1,423	2,227	2,009	1,234	1,076	0,838	1,507	1,728	3,425	0,644	0,630	0,295	1,626
h	0,008	0,010	0,008	0,016	0,021	0,034	0,046	0,037	0,861	0,215	0,630	0,589	0,813
ı	11,846	9,835	11,326	8,974	8,945	6,917	6,356	5,514	5,950	5,410	6,535	3,976	5,420
i	15,123	15,161	16,009	**21,699**	17,671	**22,094**	**18,762**	**23,240**	**23,899**	**27,265**	**20,945**	**26,804**	13,279
j	0,002	0,006	0,001	0,005	0,002	0,004	0,008	0,009	0,020	-	-	0,295	-
k	2,844	3,071	3,475	2,220	2,852	1,870	2,991	3,694	3,132	1,589	1,654	2,062	2,168
l	3,366	2,293	2,259	2,431	2,028	1,511	1,572	2,169	1,429	2,576	2,756	4,271	3,794
m	2,881	2,317	2,171	1,918	1,782	1,399	1,220	1,424	1,781	1,760	5,197	2,356	0,813
n	**17,376**	**20,590**	**21,838**	18,484	**25,131**	19,466	18,013	19,610	16,559	17,862	18,898	12,960	12,195
o	0,637	0,813	0,656	0,561	0,337	0,371	0,233	0,671	0,392	0,472	0,866	0,589	1,626
ö	0,003	0,003	0,004	0,005	0,011	0,016	0,008	0,055	0,059	0,086	0,158	0,147	-
p	0,036	0,042	0,031	0,015	0,033	0,030	0,054	0,046	0,078	-	0,158	0,147	0,542
r	10,284	9,085	10,285	9,777	9,030	11,054	12,261	11,496	13,310	12,538	9,370	10,604	11,924
s	1,110	1,352	1,890	2,023	1,078	1,414	0,776	2,132	0,998	1,331	1,260	1,915	1,084
ş	0,756	0,857	0,997	0,976	0,866	0,879	0,742	1,048	0,998	0,773	0,787	1,620	2,168
t	0,638	0,678	0,776	0,913	0,894	0,749	0,643	0,845	0,783	1,589	1,024	0,736	1,084
u	1,621	1,304	1,191	1,343	1,143	1,765	1,201	1,158	0,607	1,202	1,811	1,031	1,084
ü	0,125	0,128	0,067	0,108	0,046	0,057	0,080	0,083	0,215	0,301	0,472	0,442	0,813
v	0,006	0,007	0,010	0,011	0,019	0,027	0,031	0,028	0,020	0,086	0,315	0,295	0,271
y	1,088	0,875	0,949	0,636	0,665	0,522	0,623	0,616	0,529	1,417	0,394	1,326	1,084
z	4,908	5,389	5,274	7,095	6,230	7,446	8,941	5,551	7,869	6,054	7,559	5,007	2,439

2.2.9 Letter Positions in Turkish Words

Table 2.10 shows, presence ratios for all Turkish letters for each word position. Most common letter which is situated on a position is highlighted. As can be seen on the table "b" is the most common character that Turkish words begin with.

Table 2.11, was generated using ratio values shown in Table 2.10. When CV patterns of Turkish are examined, 5 of most common 10 CV patterns are matched with the CV sequence seen in Table 2.11. These patterns are CVCVC, CVC, CV, CVCV and VCVC.

Table 2.11 Most common letters for each positions and CV (*consonant-vowel*) forms.

1	2	3	4	5	6	7	8	9	10	11	12	13
B	A	R	A	N	L	A	N	İ	N	İ	N	N
C	V	C	V	C	C	V	C	V	C	V	C	C
14	15	16	17	18	19	20	21	22	23	24	25	26
N	N	N	İ	N	İ	İ	İ	İ	İ	İ	İ	E
C	C	C	V	C	V	V	V	V	V	V	V	V

2.3 Word Based Analysis on Corpus

Most Common Word N-Grams, Word Beginnings and Endings, Word Length Distributions, Sentence Length Distribution, Word CV Patterns are important word based analyses for learning characteristics of Turkish and determining differences from other languages. In this part of study these word based analyses will be explained.

2.3.1 Most Common Word N-Grams

Table 2.12 shows most frequently used word n-grams for Turkish. Most common words are "BİR", "VE", "BU", "DE", "DA" and these five words form 0.078 of all words. Top five most common 2-gram words are "YA DA", "BÖYLE BİR", "HEM DE", "BİR ŞEY", "NE KADAR" and their total ratio equal to 0.0033 of all 2-grams. First five most common 3-gram words are "NE#YAZIK#Kİ", "BİR#KEZ#DAHA",

"NE#VAR#Kİ", "ÇOK#ÖNEMLİ#BİR" and "BİR#SÜRE#SONRA". These five 3-grams form 0.00047 of whole 3-grams.

Table 2.12 Most frequently used word n-grams (*1≤n≤3)* in Turkish

Unigram	**%%%[3]**	**Bigram**	**%%%**	**Trigram**	**%%%**
BİR	249.140	YA#DA	9.930	NE#YAZIK#Kİ	1.348
VE	183.070	BÖYLE#BİR	5.950	BİR#KEZ#DAHA	1.343
BU	152.240	HEM#DE	5.930	NE#VAR#Kİ	0.685
DE	98.640	BİR#ŞEY	5.750	ÇOK#ÖNEMLİ#BİR	0.677
DA	96.420	NE#KADAR	5.170	BİR#SÜRE#SONRA	0.676
İÇİN	58.620	BİR#DE	4.510	MİLLİYET#COM#TR	0.634
ÇOK	46.210	BU#KADAR	4.220	BİR#AN#ÖNCE	0.630
NE	42.200	YENİ#BİR	3.900	NE#OLURSA#OLSUN	0.621
DAHA	41.110	VE#BU	3.770	HER#NE#KADAR	0.554
AMA	41.090	BÜYÜK#BİR	3.640	BAŞKA#BİR#ŞEY	0.530
GİBİ	38.420	EN#BÜYÜK	3.360	BİR#ŞEY#YOK	0.517
O	37.810	O#ZAMAN	3.290	BİR#YANDAN#DA	0.472
İLE	35.700	BU#KONUDA	3.290	AMA#YİNE#DE	0.424
EN	32.150	O#KADAR	3.210	BÖYLE#BİR#ŞEY	0.412
KADAR	31.780	ÖNEMLİ#BİR	3.210	BİR#SÜRE#ÖNCE	0.395
VAR	30.980	DAHA#DA	3.030	RECEP#TAYYİP#ERDOĞAN	0.383
OLARAK	29.590	BEN#DE	3.010	DAHA#ÖNCE#DE	0.378
Kİ	29.230	DE#BU	2.970	BAŞTA#OLMAK#ÜZERE	0.371
HER	28.320	BİR#BAŞKA	2.860	O#KADAR#ÇOK	0.357
DEĞİL	27.390	BAŞKA#BİR	2.800	HER#GEÇEN#GÜN	0.335
SONRA	26.040	BU#ARADA	2.770	HER#ŞEYDEN#ÖNCE	0.330
OLAN	23.990	GİBİ#BİR	2.750	YÖNETİM#KURULU#BAŞKANI	0.329
BÜYÜK	20.170	O#DA	2.620	KISA#BİR#SÜRE	0.318
TÜRKİYE	20.130	BU#NEDENLE	2.590	ÇOK#BÜYÜK#BİR	0.310
DİYE	19.780	DA#BU	2.590	İÇ#VE#DIŞ	0.303
İKİ	19.330	İÇİN#DE	2.570	BAŞBAKAN#RECEP#TAYYİP	0.294
YA	18.290	DAHA#ÇOK	2.480	O#ZAMAN#DA	0.278
YENİ	18.220	DAHA#FAZLA	2.440	AVRUPA#İNSAN#HAKLARI	0.273
İSE	17.610	ÇOK#DAHA	2.440	ÇOK#DAHA#FAZLA	0.271
YOK	17.300	AMA#BU	2.440	BU#NEDENLE#DE	0.269

2.3.2 Word Beginnings and Endings

Table 2.13 shows the first and last letter distributions of Turkish words. 12.148% of Turkish words begin with letter "B" while 15.605% of them end with letter "N". So, "B" is most frequently used letter which starts words and the letter "N" is most frequently observed letter which terminates words. 60,429% of all words begin with one of the letters "B", "D", "K", "A", "Y", "S", "G", or "İ" while 81,900% of them end with one of the letters "N", "E", "A", "R", "İ", "I", or "K".

[3] %%% means per 1,000,000.

Table 2.13 Probability distribution for word beginning and ending characters

Letter	First	Last	Letter	First	Last
A	7.346%	11.845%	M	3.391%	3.592%
B	**12.148%**	0.125%	N	1.712%	**15.605%**
C	0.918%	0.035%	O	4.677%	0.582%
Ç	2.413%	0.655%	Ö	1.809%	0.005%
D	8.584%	0.180%	P	1.703%	1.075%
E	3.629%	13.123%	R	0.947%	11.335%
F	1.284%	0.315%	S	6.450%	0.566%
G	5.768%	0.096%	Ş	1.298%	1.567%
Ğ	0.001%	0.046%	T	4.642%	2.004%
H	3.806%	0.274%	U	0.964%	4.197%
I	0.303%	7.800%	Ü	1.094%	1.105%
İ	5.609%	11.219%	V	3.492%	0.112%
J	0.072%	0.037%	Y	6.606%	0.588%
K	7.919%	6.777%	Z	0.820%	2.769%
L	0.595%	2.372%	**Total:**	**100%**	**100%**

2.3.3 Word Length Distributions

15.37% of Turkish letters consist of five letters. Most frequently seen example of such words is “KADAR”. Other most frequently observed word lengths are 6 (12.22%), 7 (11.67%) and 4 (10.06%) as shown on Table 2.14.

Word lengths which have observation ratios less than 0.0001% are discarded (words with length longer than 26 letters). Total ratio of discarded words is 0.00023% .

When ratios given on Table 2.14 compared with the results obtained in recent study of Dalkılıç M.E. & Dalkılıç G. (2001), although ranks of most frequently observed word lengths are similar, average word length is calculated as 6.13 in previous study.

Table 2.14 Word length distribution

Word Length	Ratio %	Word	Ratio %	Word	Ratio %
1	0.7524%	10	5.6765%	19	0.0285%
2	8.1462%	11	3.6321%	20	0.0144%
3	9.7397%	12	2.5118%	21	0.0055%
4	10.0644%	13	1.4059%	22	0.0026%
5	15.3717%	14	0.7993%	23	0.0010%
6	12.2296%	15	0.4412%	24	0.0006%
7	11.6791%	16	0.2302%	25	0.0003%
8	9.2705%	17	0.1109%	26	0.0001%
9	7.8276%	18	0.0579%	**Total:**	**100%**

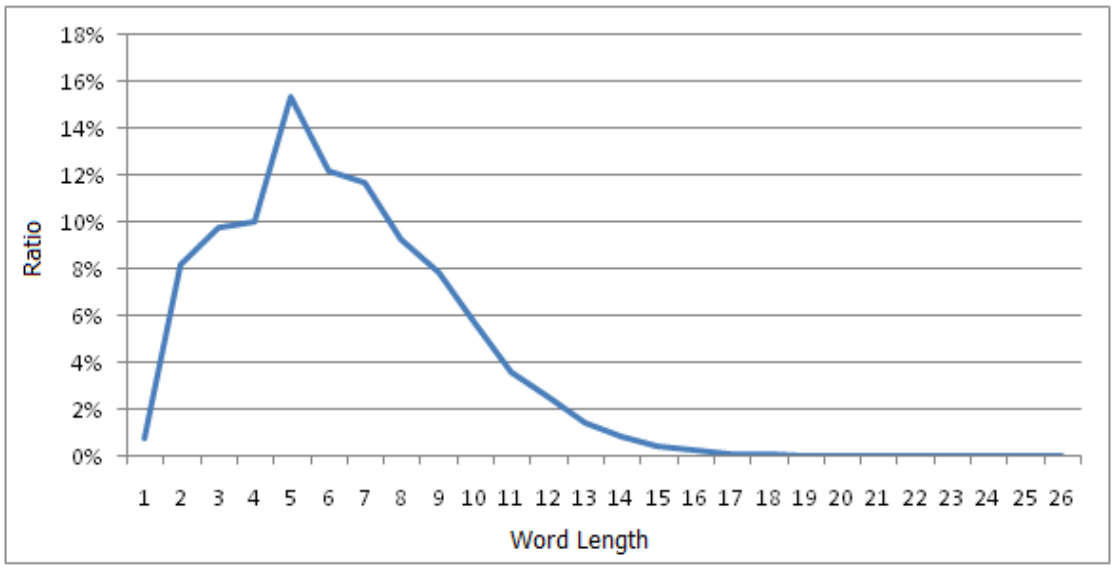


Figure 2.5 Word length distribution for Turkish

Word length distribution graph for Turkish is given in Figure 2.5. Average word length for Turkish is computed as 6.34 using the values on Table 2.14. Average word lengths for some European Languages are given at Table 2.15 (Hollink, Kamps, Monz, & de Rijke, 2004). Comparing with the given languages, Turkish and Finnish (which is an agglutinative language as Turkish) have longest word lengths.

Table 2.15 Average word lengths for some European Languages and Turkish

Dutch	English	Finnish	French	German	Italian	Spanish	Swedish	Turkish
5.4	5.8	7.3	4.8	5.8	5.1	5.1	5.4	6.34

2.3.4 Sentence Length Distribution

Most commonly observed sentences in Turkish are sentences which consist of 4, 5, 6, 7, 8, or 9 words as seen on Table 2.16. These sentences form 39.4% of all sentences in corpus. Sentence lengths which have observation ratio less than 0.00002% are discarded. Total ratio of discarded sentences is 0.00087%.

Table 2.16 Sentence length distribution

Sentence Length	Frequency	Ratio	Sentence Length	Frequency	Ratio
1	409,356	4.00%	21	142,638	1.40%
2	419,356	4.10%	22	122,533	1.20%
3	563,186	5.51%	23	107,487	1.05%
4	663,759	6.49%	24	91,316	0.89%
5	694,009	6.79%	25	77,704	0.76%
6	705,194	6.90%	26	66,895	0.65%
7	689,122	6.74%	27	57,262	0.56%
8	660,011	6.46%	28	49,688	0.49%
9	615,317	6.02%	29	42,286	0.41%
10	568,890	5.57%	30	36,236	0.35%
11	517,412	5.06%	31	31,067	0.30%
12	465,596	4.56%	32	26,772	0.26%
13	416,491	4.07%	33	22,596	0.22%
14	368,433	3.60%	34	19,624	0.19%
15	326,787	3.20%	35	16,962	0.17%
16	287,352	2.81%	36	14,361	0.14%
17	250,645	2.45%	37	12,216	0.12%
18	219,348	2.15%	38	10,912	0.11%
19	190,984	1.87%	39	9,187	0.09%
20	165,480	1.62%	40	8,131	0.08%

As seen on Figure 2.6, observation ratios decrease to very low values by the sentences have lengths 40. Average sentence length is calculated as 10.692 according to values on Table 2.16.

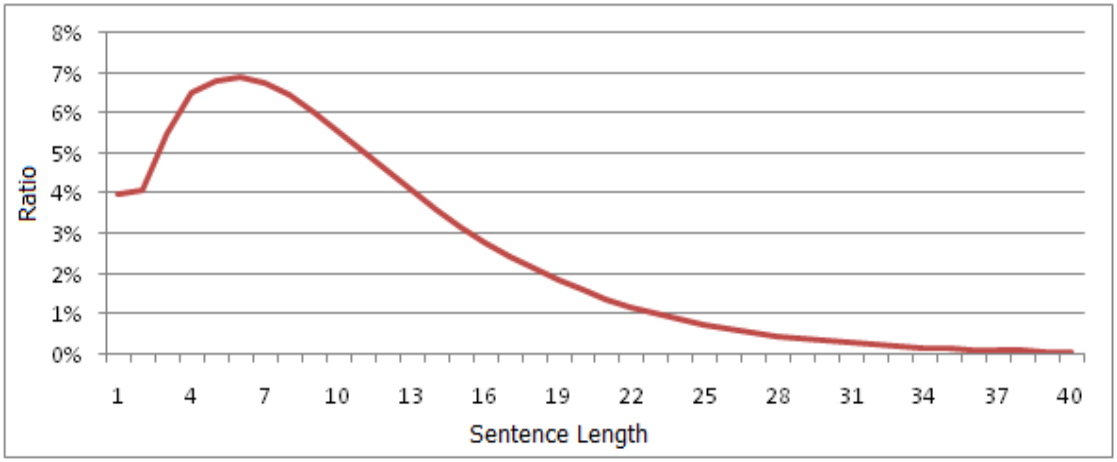


Figure 2.6 Sentence length distribution for Turkish

2.3.5 Word CV Patterns

Turkish alphabet consists of 8 vowels (V) {A,E,I,İ,O,Ö,U,Ü} and 21 consonants (C) {B,C,Ç,D,F,G,Ğ,H,J,K,L,M,N,P,R,S,Ş,T,V,Y,Z}. Corpus used in this study only contains these 8 vowels, 21 consonants and space character to separate words.

If all characters of the corpus are analyzed, results shown on Table 2.17 are obtained. According to these results consonants form 49.27% of whole corpus characters, vowels form 37.10% of them and space character forms 13.63% of all characters. If space character is omitted, consonants' ratio is 57.04% and vowels' ratio is 42.96% among all characters of corpus.

Table 2.17 Consonant, vowel, space character distributions

		Including Space Character	**Excluding Space Character**
	Total Occurrence	**%**	**%**
C	382,683,136	49.27	57.04
V	288,208,635	37.10	42.96
#	105,863,483	13.63	-

In this study, Turkish words analyzed with their CV forms. CV forms which have observation ratio higher than 0.2% and most frequently used word examples of these forms are listed on Table 2.18. CV forms listed Table 2.18 are 88.3% of all CV forms. 11.7% of them are omitted which have less than 0.2 observation ratio.

According to data on Table 2.18, most frequently observed CV form is CVCVC and most frequently used word in this form is "KADAR".

Top 12 of Turkish CV forms listed on Table 2.18 have same ranks with the Turkish CV forms which are obtained by using only 11.5 MB corpus in the study of Dalkılıç M. E. & Dalkılıç G. (2001).

Table 2.18 Observation ratios and sample words for most frequently seen 63 CV pattern in Turkish.

Pattern	%	Sample
CVCVC	7.73	KADAR
CVC	7.25	BİR
CV	6.92	VE
CVCV	4.90	DAHA
CVCCV	4.35	SONRA
CVCCVC	4.14	DEVLET
CVCVCV	3.49	SADECE
CVCVCVC	3.30	YENİDEN
CVCCVCV	2.99	TÜRKİYE
VCVC	2.52	İÇİN
CVCVCCV	2.15	ŞEKİLDE
CVCCVCVC	2.03	BAŞBAKAN
VCV	1.84	AMA
VCCVC	1.80	ANCAK
CVCVCCVC	1.67	DEĞİLDİR
VCCVCV	1.55	OLDUĞU
VCCV	1.54	ÖNCE
CVCVCVCV	1.50	BELEDİYE
VCVCVC	1.40	OLARAK
CVCVCVCVC	1.39	GEREKİYOR
CVCCVCVCV	1.25	TÜRKİYEDE
CVCVCCVCV	1.24	DEMOKRASİ
VC	1.08	EN
VCCVCVC	1.08	ERDOĞAN
CVCCVCVCVC	1.06	TÜRKİYENİN
CVCCVCCV	1.02	BİRLİKTE
VCVCV	1.01	ÜZERE
VCVCCV	0.89	İÇİNDE
CVCVCVCCV	0.87	KONUSUNDA
CVCVCCVCVC	0.81	DEMOKRATİK
CVCCVCCVC	0.77	GERÇEKTEN

Pattern	%	Sample
VCCVCVCV	0.70	OLDUĞUNU
CVCC	0.67	TÜRK
VCVCVCVC	0.58	EKONOMİK
CVCVCVCCVC	0.57	TARAFINDAN
V	0.57	O
VCCVCCVC	0.55	İSTANBUL
VCCVCCV	0.53	ASLINDA
CVCVCCVCVCV	0.52	GEREKTİĞİNİ
VCVCVCV	0.52	ÜZERİNE
CVCCVCCVCV	0.50	YARDIMCISI
VCCVCVCVC	0.49	İSTİYORUM
VCVCCVC	0.48	İLİŞKİN
CVCCVCVCCV	0.47	KARŞISINDA
CVCVCVCVCVC	0.43	GALATASARAY
CVCVCVCVCV	0.39	POLİTİKASI
VCVCCVCV	0.38	İÇİNDEKİ
VCC	0.38	İLK
CVCVCCVCCV	0.35	FENERBAHÇE
CVCVCCVCVCVC	0.35	DEMOKRASİNİN
VCVCVCCV	0.33	ARASINDA
CVCCVCVCVCV	0.29	KENDİLERİNE
CVCCVCVCCVC	0.29	BAŞBAKANLIK
VCCVCCVCV	0.29	İNSANLARI
CVCVCCVCCVC	0.27	GENELKURMAY
CVCCVCCVCVC	0.26	ŞİRKETLERİN
VCCVCVCCV	0.26	ÖNCELİKLE
VCVCVCCVC	0.24	AÇISINDAN
VCVCCVCVC	0.24	İLİŞKİLER
CVCCVCVCVCVC	0.23	CUMHURİYETİN
CVCVCVCCVCV	0.22	KONUSUNDAKİ
CVCVCVCVCCV	0.21	DOLAYISIYLA
CVCCVCCVCVCV	0.21	MİLLETVEKİLİ

Table 2.19 Observation ratios and sample words for most frequently seen 60 CV pattern in English.

Pattern	%	Sample	Pattern	%	Sample
VC	9.30	UP	VVC	0.53	OUR
CVC	8.48	FOR	CVCCVVC	0.52	REFRAIN
CCV	6.29	THE	**CVCCV**	0.47	KOLYA
CVCC	5.67	BOYS	**CVCCVCVC**	0.46	PANCAKES
CV	5.56	SO	CCVCVC	0.44	CHIMED
VCC	5.43	END	CCVCCVC	0.44	CLOTHES
CCVC	4.45	THAT	CCVVCC	0.43	THOUGH
CVCV	3.49	MORE	CVVCVC	0.39	VOICES
V	2.77	A	CVVCV	0.37	VOICE
CVVC	2.22	TOOK	CCVCCC	0.35	THIRTY
CCVCC	2.01	SHALL	**CVCVCCVC**	0.35	REMEMBER
CVCCVC	1.76	HURRAH	VCVVC	0.33	AGAIN
CVCVC	1.71	LIVES	VCCVCC	0.32	ALWAYS
CVVCC	1.54	TEARS	CVCVVC	0.31	BURIED
CC	1.42	MY	CVVCVCC	0.29	FEELING
CVCCC	1.33	FORTH	**CVCCVCV**	0.28	PICTURE
CCVCV	1.23	THERE	VCVCC	0.28	EVERY
CVCCVCC	1.18	TALKING	CVCCVCVCC	0.26	HAPPENING
C	1.08	S	CVVCCC	0.26	TAUGHT
CCVVC	1.03	CRIED	CVCVCVCC	0.26	HUMANITY
CVCVCC	0.96	FINISH	CVVCCVC	0.25	LAUGHED
VCVC	0.95	EVER	VCCVVC	0.25	AFRAID
CVV	0.94	SEE	CCVCCVCC	0.25	GLADNESS
VCCV	0.83	ONCE	CVCCCVC	0.24	PATCHED
VCV	0.81	ONE	CCVCVCC	0.24	FLOWERS
VCCVC	0.65	OTHER	CVVCCV	0.23	PEOPLE
CCVV	0.59	TRUE	**CVCCVCCVC**	0.23	KARTASHOV
CCC	0.58	WHY	CVCCCVCC	0.22	LANDLADY
CVCVCVC	0.54	FUNERAL	VCCC	0.21	ONLY
CVCVCV	0.53	BECOME	CVCCCV	0.21	LITTLE

20 of CV patterns (approximately 30%) are common for Turkish and English. These patterns are highlighted on both Table 2.18 and 2.19.

2.3.6 Zipf's Law

Zipf's law is an empirical law named after the Harvard linguist George Kingsley Zipf. It is based on the observation that the frequency of occurrence of some events is a function of its rank in the frequency table. This function can be expressed by the following equation:

$$f(k;s,N) = \frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)}$$

Where *N is* the number of elements, *k* is their rank and *s* is the value of the exponent characterizing the distribution. This equation states that the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word. Its graphical representation in a log-log scale is a straight line with a negative slope.

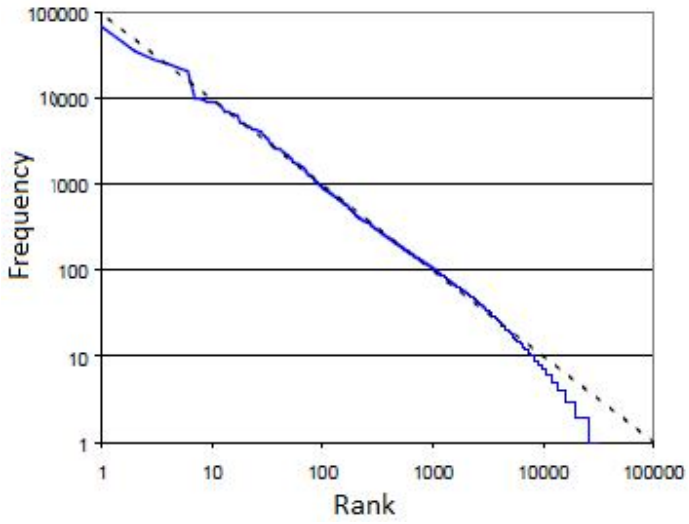


Figure 2.7 Zipf curve for the unigrams extracted from the 1 million words of the Brown corpus.

Word frequency and rank distribution graph for an English corpus, known as Brown Corpus, is given in Figure 2.7. (Ha, Garcia, Ming & Smith, 2002) The straight line shows Zipf's Law and the other dotted points are the actual values.

Table 2.20 Frequency-rank values of some sample word unigrams of Turkish

Word	Freq.	Rank	f*r	Word	Freq.	Rank	f*r
BİR	2,637,507	1	2,637,507	ULUS	4,417	3,000	13,251,000
VE	1,938,076	2	3,876,152	ARINÇ	3,299	4,000	13,196,000
BU	1,611,614	3	4,834,842	YERLEŞMİŞ	1,584	8,000	12,672,000
AMA	434,992	10	4,349,920	PARKI	1,236	10,000	12,360,000
DEĞİL	289,927	20	5,798,540	ÇOKÇA	556	20,000	11,120,000
YOK	183,166	30	5,494,980	ILGAZ	331	30,000	9,930,000
TÜRK	158,350	40	6,334,000	İÇİNDEDİRLER	225	40,000	9,000,000
ÖNEMLİ	136,448	50	6,822,400	ÜZECEK	164	50,000	8,200,000
OLDUĞUNU	121,355	60	7,281,300	KAÇINMALI	126	60,000	7,560,000
BÜTÜN	100,560	70	7,039,200	REPOYA	99	70,000	6,930,000
BİN	89,916	80	7,193,280	BOMBALANMIŞ	81	80,000	6,480,000
ORTAYA	85,094	90	7,658,460	DURDURMASINI	67	90,000	6,030,000
MİLYON	77,497	100	7,749,700	GİŞELERİ	57	100,000	5,700,000
ADAM	45,656	200	9,131,200	AYRILABİLİRDİ	29	150,000	4,350,000
DÖRT	33,235	300	9,970,500	DAYAMA	27	155,000	4,185,000
EDİLEN	25,860	400	10,344,000	DEVLETLERİNDEKİ	26	160,000	4,160,000
ALIYOR	22,476	500	11,238,000	CENTRUM	17	200,000	3,400,000
SIRADA	19,182	600	11,509,200	İMDB	12	250,000	3,000,000
PARTİSİ	16,642	700	11,649,400	DENKSİZCE	8	300,000	2,400,000
MUSUNUZ	15,017	800	12,013,600	HÜPÜRDETEN	5	400,000	2,000,000
ALANDA	13,518	900	12,166,200	CADDELERİMİZİN	3	500,000	1,500,000
BAŞ ARI	12,277	1,000	12,277,000	BULAŞIKÇIYA	2	600,000	1,200,000
İSTEDİĞİNİ	6,541	2,000	13,082,000	YAZAMAYACAKSAN	1	800,000	800,000

Zipf's law is useful as a rough description of the frequency distribution of words in human languages. Calculated frequency results of letter n-grams and word n-grams, as seen in Table 2.20, were exported to a Matlab application and then results were sorted by their frequencies in descending order, and finally used to form the Figure 2.8, 2.9 and 2.10. Table 2.21 shows point counts used to draw Figure 2.8, 2.9 and 2.10. For example, 72,131,395 points used to form word trigrams diagram. According to Figure 2.8, it can be said that, while 1, 2 and 3-grams fit Ziph's law, 4 and 5 grams deviate from Ziph's law. There is a clear deviation in graphs belong to $6 \leq n \leq 10$ interval.

There is a close similarity between Figure 2.8 and the monogram, bigram, trigram and tetra-gram rank-frequency graphs of TurCo, which is the corpus with a word count of 50,111,828. (Dalkılıç, G., & Çebi, Y., 2004).

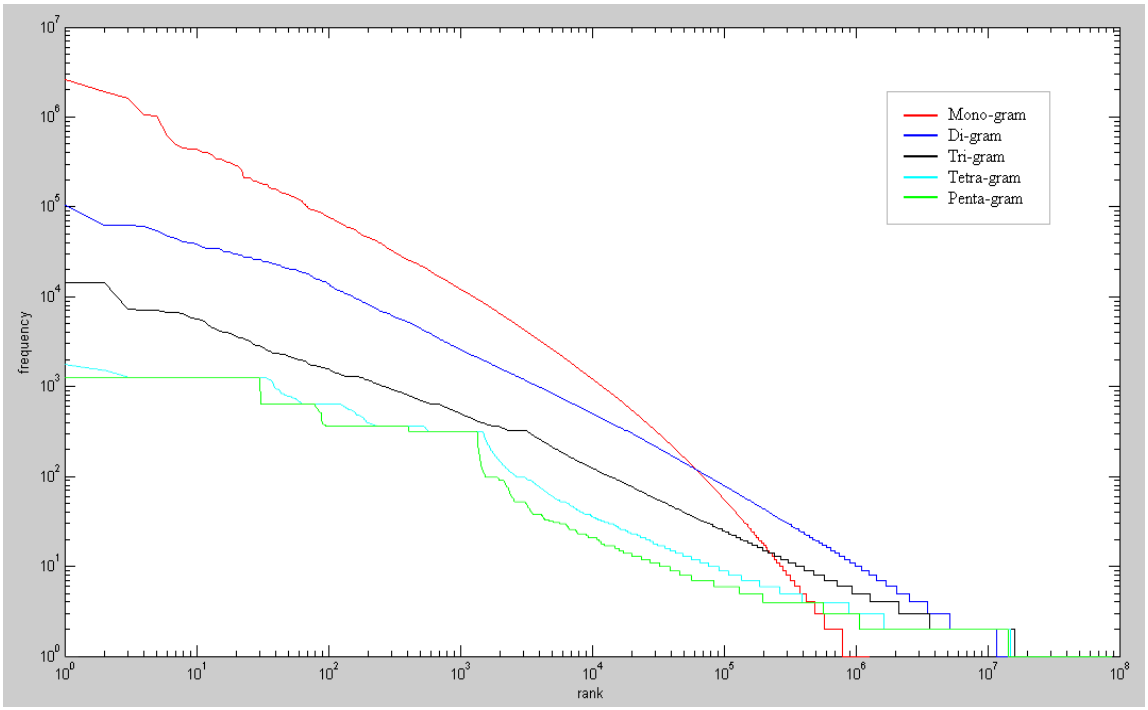


Figure 2.8 Frequency-rank data for word n-grams *(1≤n≤5)*.

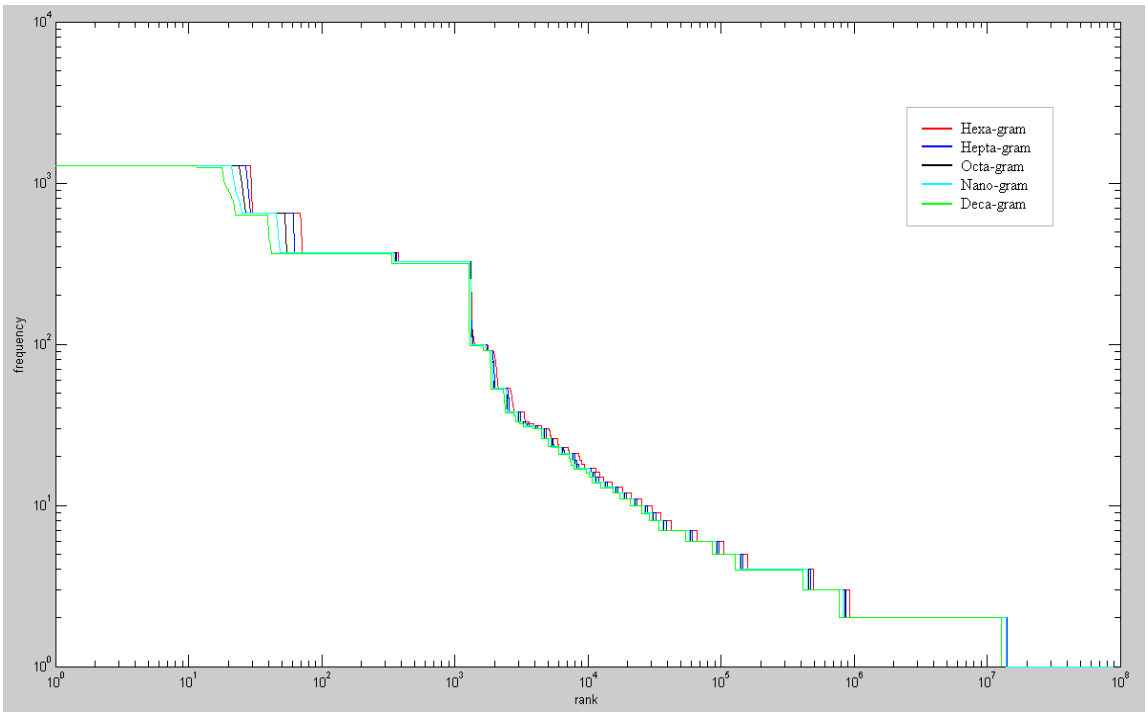


Figure 2.9 Frequency-rank data for word n-grams *(6≤n≤10)*.

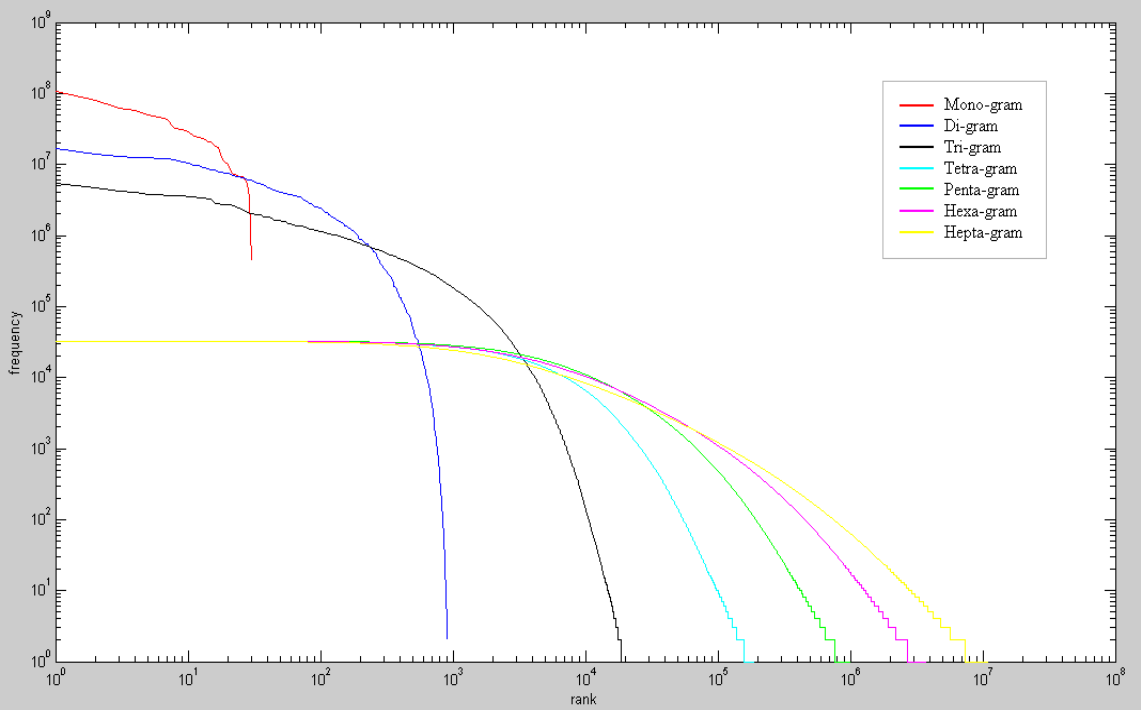


Figure 2.10 Frequency-rank data for letter n-grams *(1≤n≤7)*.

Table 2.21 Point counts used to construct Figure 2.8, 2.9 and 2.10.

	Word	Letter		Word	Letter
n-gram	**Point**	**Point**	**n-gram**	**Point**	**Point**
Monogram (n=1)	1,291,005	30	Hexagram (n=6)	89,193,263	3,793,750
Digram (n=2)	33,421,623	899	Heptagram (n=7)	89,456,732	11,013,233
Trigram (n=3)	72,131,395	20,190	Octagram (n=8)	89,613,648	-
Tetragram (n=4)	85,596,608	192,585	Nanogram (n=9)	89,732,140	-
Pentagram (n=5)	88,500,062	1,004,624	Decagram (n=10)	89,830,417	-

In conclusion, Zipf's Law provides a theoretical model that closely fits the data for word unigrams, bigrams and trigrams, but is seen to deviate for data associated with other word n-grams and letter n-grams. Although, there is a similarity between Zipf's Law's rank-frequency graph and the actual frequency-rank graph of some Turkish word n-grams, there is not any perfect match. Insufficiency of sample spaces of n-gram series after trigrams can be accepted cause of this situation. In these cases, other models may be more appropriate.

CHAPTER THREE

AUTHOR IDENTIFICATION

Natural Language Processing is a research area that is used for many different purposes and it becomes more popular continuously. Speech syntheses, speech recognition, machine translation, spelling correction and author identification are some of the applications of NLP.

Author identification is the task of identifying the author of a given text. Aim is to automatically determine the corresponding author of an anonymous text. It can be seen as a classification problem, where a set of documents with known authors are used for training. The main idea under computer-based author identification is to define an appropriate characterization of documents to determine the writing style of authors.

Related with innovations in computer science of identification technologies such as cryptographic signatures, intrusion detection systems, author identification have been used in areas such as intelligence, criminal law, civil law, and computer security, verifying the authorship of e-mails and newsgroup messages.

Some important techniques used for author identification are vocabulary richness and lexical repetition, word frequency distributions, syntactic analysis, word collocations, grammatical errors, and word, sentence, clause, paragraph lengths. Many studies combine features of different types using multivariate analysis techniques.

In the last 50 years there were many studies in the author identification area. Amongst the pioneers of authorship attribution are Morton (1965), who focused on sentence lengths, and Brainerd (1974), who focused on syllables per word. In 1984, Mosteller and Wallace took the Federalist Papers and determined a very credible

attribution of authorship on the basis of a range of discriminates and used Bayesian analysis. Burrows (1992) focused on common high-frequency words. Cavnar (1994) described an n-gram based approach to text categorization is tolerant of textual errors. Holmes (1994) used word counts and document length features, Twedie ve Baayen (1998) used ratio between different word count and total word count. F¨urnkranz (1998) described an algorithm for efficient generation and frequency-based pruning of 2-gram and 3-gram features. Brinegar (2000), who focused on word lengths and Stamatatos (2000) have applied Multiple Regression and Discriminant Analysis using 22 style markers.

Important studies in Turkish can be exemplified by Tan (2002) developed an algorithm by using 2-grams, Çatal (2003) developed a system named NECL by using n-grams, Diri and Amasyalı (2003) formed 22 style markers to determine author and type of a document, and in their another study (2006) they used 2 and 3-grams to determine author and type of a document and gender of author.

Recent studies based on n-grams, generally focused on letter n-grams. In this study, we used and compared two main method based on word n-grams and some style markers are formed to identify authors. Linguistic statistics are collected such as type/token ratio, hapax legomena ratio, average sentence length, average word length, word count per article, punctuation mark frequencies, entropy, and most frequently used word n-grams ($1 \leq n \leq 6$) for all authors. In the next parts of this chapter, details of the methods will be explained; collected statistics and obtained results will be given.

3.1 Preliminary Studies

At the beginning of the study, 16 authors are selected to work for author identification process. These authors write articles in different categories such as economy, education, politics, sports etc. These authors' articles had to be collected before starting to statistical studies about authors. To collect articles of several

newspapers and authors, different download programs were constructed. One of them can be seen in Figure 3.1.

Download program firstly takes a web page source code which contains web addresses for author's articles. These addresses are splitted and listed on "Article Links" section. Then, source codes of these links are downloaded, unnecessary content and tags are eliminated using code block seen on Table 3.1. Finally, all articles are saved in a folder with the name same as its author.

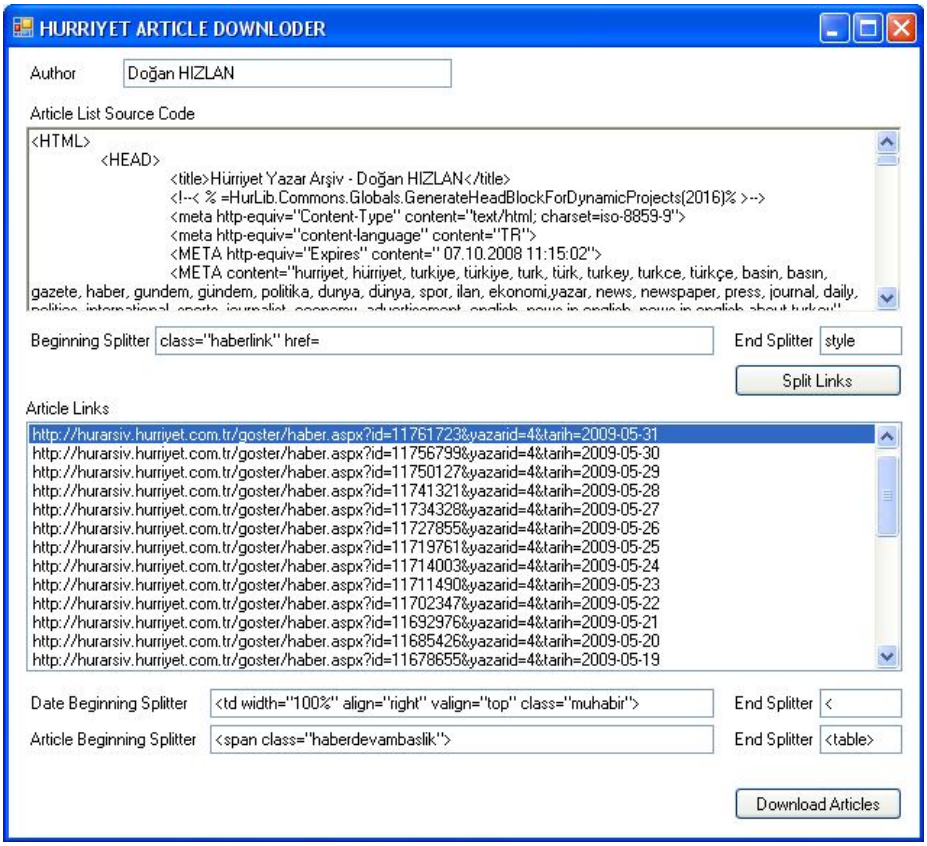


Figure 3.1 Article downloader for Hürriyet newspaper.

Table 3.1 Code block for eliminating HTML tags from web page's source code.

```
article = article.Replace("<BR>","\r\n");
article = article.Replace("<B>","");
article = article.Replace("</B>", "");
article = article.Replace("<P>", "");
article = article.Replace("</P>", "");
article = article.Replace("<br>", "\r\n");
article = article.Replace("<b>", "");
article = article.Replace("</b>", "");
article = article.Replace("<p>", "");
article = article.Replace("</p>", "");
article = article.Replace(" ", "");
```

After download process, total 33,666 articles are collected as training set. Distribution of articles between authors is given on Table 3.2.

Table 3.2 Authors and count of training set articles used for statistical analysis.

Author	Article Count
Abbas GÜÇLÜ	2,338
Bekir COŞKUN	1,884
Doğan HIZLAN	2,381
Ercan KUMCU	1,777
Ertuğrul ÖZKÖK	1,999
Güngör URAS	2,469
Güzin Abla	1,847
Hadi ULUENGİN	1,544
Hasan CEMAL	1,973
Hasan PULUR	2,522
Mehmet Ali BİRAND	1,671
Oktay EKŞİ	1,932
Sami KOHEN	2,016
Taha AKYOL	2,563
Yalçın BAYER	2,057
Yalçın DOĞAN	2,693

Some extra articles, which are not used in statistical analysis, are needed to use for testing prediction ratios in authorship attribution studies. Therefore test set articles that are seen on Table 3.3, are downloaded.

Table 3.3 Count of test set articles collected for all authors.

Author	Article Count
Abbas GÜÇLÜ	320
Bekir COŞKUN	259
Doğan HIZLAN	333
Ercan KUMCU	209
Ertuğrul ÖZKÖK	282
Güngör URAS	1026
Güzin ABLA	279
Hadi ULUENGİN	209
Hasan CEMAL	804
Hasan PULUR	847
Mehmet Ali BİRAND	210
Oktay EKŞİ	271
Sami KOHEN	656
Taha AKYOL	823
Yalçın BAYER	336
Yalçın DOĞAN	273

3.2 Author Based Statistical Results

After article collection statistical results about authors are calculated. These results help understanding characteristics of authors and gain affluence and efficiency to author identification processes. Average word count per article, distinct word count per article, type\token ratio, count of words occurring once per article, hapax legomena ratio, average sentence length, average word length, observation periods for some punctuation marks like “,”, “!”, “?”, “;”, “:”, entropy are the statistical results that are obtained. Results, obtained by analysis on articles of authors, can be seen on Table 3.4 and Table 3.5.

Table 3.4 Punctuation mark frequencies and entropy values for authors.

Author	,	!	?	;	:	Entropy
	Frequency					
Abbas GÜÇLÜ	16.4	336.98	151.1	881.51	230.35	2.38
Bekir COŞKUN	13.0	2,276.10	74.44	115.88	133.36	2.18
Doğan HIZLAN	11.0	1,251.87	258.7	366.03	150.71	2.37
Ercan KUMCU	19.1	3,502.42	531.4	5,053.83	1,896.7	2.33
Ertuğrul ÖZKÖK	18.0	18,048.7	121.3	1,858.95	189.33	2.43
Güngör URAS	17.7	910.94	189.9	2,191.62	391.22	2.34
Güzin Abla	14.6	595.42	119.7	233.46	247.48	2.27
Hadi ULUENGİN	11.5	104.99	204.6	144.50	942.59	2.49
Hasan CEMAL	17.7	141.98	85.73	1,944.88	99.66	2.44
Hasan PULUR	7.81	53.99	64.81	386.90	93.90	2.36
Mehmet Ali BİRANI	13.5	1,482.03	186.0	1,625.73	388.71	2.57
Oktay EKŞİ	23.5	1,035.28	135.9	712.01	208.07	2.37
Sami KOHEN	15.3	1,206.64	213.1	982.53	274.03	2.37
Taha AKYOL	13.8	74.03	157.3	347.95	98.23	2.36
Yalçın BAYER	19.1	747.36	121.5	145.91	261.71	2.71
Yalçın DOĞAN	9.60	81.43	163.0	23,340.3	190.52	2.36

According to Table 3.4, Abbas GÜÇLÜ uses comma once per 16.4 words, uses exclamation mark once per 336.98 words and so on. Word based entropy of Abbas GÜÇLÜ's texts are calculated as 2.38.

Table 3.5 Some statistical results for authors.

Author	Total Article Count	Total word count	Word per article	Total distinct word count	Distinct word per article	Type Token Ratio	Count of words occuring once	Count of words occuring once per article	Hapax Legomena Ratio	Average sentence length	Average word length
Abbas GÜÇLÜ	2,338	961,727	411.346	78,817	298.293	0.73	40,094	245.057	0,82	10.564	6.211
Bekir COŞKUN	1,884	439,287	233.167	63,032	179.220	0.77	33,613	149.751	0,84	8.781	6.139
Doğan HIZLAN	2,381	956,432	401.693	106,790	299.416	0.76	55,695	251.836	0,84	12.143	6.404
Ercan KUMCU	1,777	753,021	423.760	56,838	279.049	0.61	28,105	216.313	0,78	11.215	6.763
Ertuğrul ÖZKÖK	1,999	992,680	496.588	90,885	344.412	0.98	45,734	274.870	0,80	9.701	6.017
Güngör URAS	2,469	1,117,727	452.704	91,171	291.419	0.65	44,376	223.135	0,77	9.431	6.175
Güzin Abla	1,847	956,251	517.732	80,795	349.705	0.73	41,116	280.277	0,80	9.062	5.973
Hadi ULUENGİN	1,544	796,491	515.862	100,892	383.281	0.77	53,841	318.117	0,83	15.909	5.958
Hasan CEMAL	1,973	1,081,353	548.076	84,931	366.206	0.68	40,789	289.255	0,79	9.061	6.075
Hasan PULUR	2,522	965,713	382.916	105,419	280.094	0.73	53,899	229.621	0,82	11.660	6,030
Mehmet Ali BİRAND	1,671	1,123,376	672.278	96,309	461.723	0.69	49,969	372.419	0,81	10.110	6.350
Oktay EKŞİ	1,932	720,552	372.957	70,695	279.626	0.75	36,450	232.236	0,83	13.868	6.181
Sami KOHEN	2,016	936,355	464.462	57,265	314.793	0.68	26,444	255.310	0,81	14.906	6.113
Taha AKYOL	2,563	1,031,337	402.394	90,521	286.792	0.72	44,816	234.380	0,82	11.738	6.268
Yalçın BAYER	2,057	2,102,314	1,022.030	154,019	700.637	0.70	76,113	563.717	0,80	11.902	6.235
Yalçın DOĞAN	2,693	1,143,676	424.685	80,419	290.246	0.69	36,342	226.902	0,78	9.205	5.950

3.3 Word N-gram Computing For Authors

Before word n-gram computation for 16 authors, corpara were created for each individual author. Size of created corpora can be seen in Figure 3.2.

Name	Size
Abbas GÜÇLÜ.txt	7.679 KB
Bekir COŞKUN.txt	3.447 KB
Doğan HIZLAN.txt	7.831 KB
Ercan KUMCU.txt	6.439 KB
Ertuğrul ÖZKÖK.txt	7.705 KB
Güngör URAS.txt	8.809 KB
Güzin Abla.txt	7.365 KB
Hadi ULUENGİN.txt	6.102 KB
Hasan CEMAL.txt	8.453 KB
Hasan PULUR.txt	7.466 KB
Mehmet Ali BİRAND.txt	9.087 KB
Oktay EKŞİ.txt	5.727 KB
Sami KOHEN.txt	7.368 KB
Taha AKYOL.txt	8.315 KB
Yalçın BAYER.txt	16.790 KB
Yalçın DOĞAN.txt	8.740 KB

Figure 3.2 Corpus files and size for authors.

Using these corpora, most frequently used 500 n-grams were calculated for each n-gram groups ($1 \leq n \leq 6$) and stored in an SQL database table. Also top 500 Turkish n-grams, obtained from combination of corpora of authors, for each n value were added to this table. Design of database table can be seen in Figure 3.3. This table holds author name, n-gram string, n value which is the number in n-gram, frequency and probability of n-gram. Sample data in database table can be seen in Figure 3.4.

Table - dbo.AUTHORS_NGRAMS*

Column Name	Data Type	Allow Nulls
Author	nvarchar(50)	☑
Word	nvarchar(500)	☑
N	char(1)	☑
Count	int	☑
Probability	float	☑

Figure 3.3 Design of AUTHORS_NGRAMS table.

	Author	Word	N	Count	Probability
	Abbas GÜÇLÜ	BİR	1	21327	0,02217626
	Abbas GÜÇLÜ	MİLLİ EĞİTİM	2	1714	0,001782253
	Abbas GÜÇLÜ	MİLLİ EĞİTİM BAKANLIĞI	3	593	0,0006166138
	Abbas GÜÇLÜ	MİLLİ EĞİTİM BAKANLIĞI NIN	4	129	0,0001341369
	Abbas GÜÇLÜ	MİLLİ EĞİTİM BAKANI HÜSEYİN ÇELİK	5	73	7,590693E-05
	Abbas GÜÇLÜ	KONUDA OLDUĞU GİBİ BU KONUDA DA	6	31	3,223445E-05
	GENERAL	BİR	1	564826	0,02327376
	GENERAL	YA DA	2	22600	0,0009312373
	GENERAL	NE YAZIK Kİ	3	3665	0,000151017
	GENERAL	AVRUPA İNSAN HAKLARI MAHKEMESİ	4	763	3,143956E-05
	GENERAL	EKİBİ TARAFINDAN TERCÜME EDİLDİKTEN SONRA	5	726	2,991497E-05
	GENERAL	GEÇEN YILIN AYNI DÖNEMİNE GÖRE YÜZDE	6	70	2,884363E-06

Figure 3.4 Sample data contained by AUTHORS_NGRAMS table.

3.4 Author Identification Based on Author Specific N-gram Method

In this model when a sample article with an unknown author is handled, firstly word n-grams of this article are computed. If an n-gram of sample article is also an element of most frequently used Turkish n-grams, this n-gram is eliminated thus, n-grams which are specific to sample article are obtained. In the same way for all 16 authors, n-grams specific to each author are found. Finally, sample article's and authors' specific n-grams are compared and the author having more common n-grams with sample article is accepted as the author of the sample article.

There are three parameters used in this model. These are shown in Figure 3.5. ***A*** is the number of most frequently used author n-grams, ***S*** is the number most used n-grams belongs to sample article and ***G*** is the number of most frequently used general Turkish n-grams used in n-gram elimination and author specific n-gram determination studies.

Figure 3.5 Parameters in *Author Specific N-gram Method*

Output screen of the test program for 1-grams can be seen in Figure 3.6 and Figure 3.7. At the bottom of the page most frequently used ***G*** general Turkish n grams are listed. Above of them is the most frequently used ***S*** n-grams belong to sample article are listed. If an n-gram listed in this section is also a general n-gram then this n-gram is eliminated and shown as strikethrough. Otherwise an n-gram is specific to sample article and author; n-gram is shown with yellow background. At the top of the page most frequently used ***A*** n-grams of each 16 authors are listed. If an n-gram in this section is common with generally used n-grams, this n-gram is eliminated and is strikethrough, too. Author specific words have yellow background. If an author specific n-gram is common with n-gram specific to sample article then this n-gram is shown with red background.

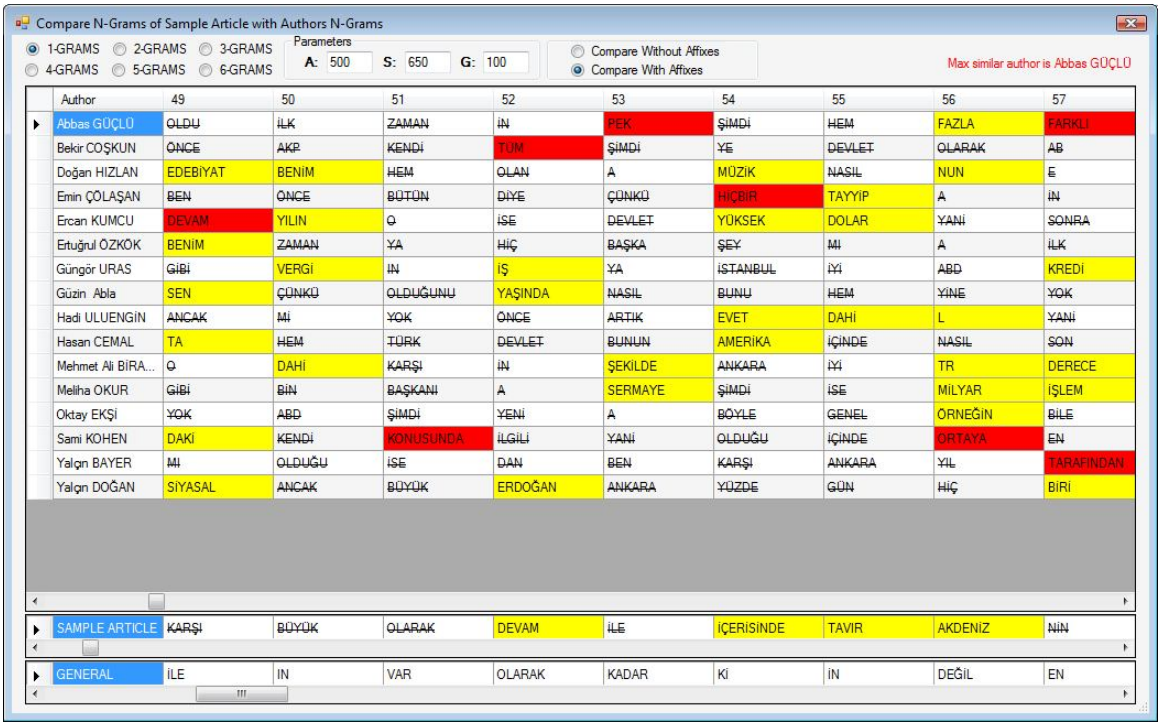


Figure 3.6 Sample output screen for "Author Specific N-gram Method"

As can be seen in Figure 3.7 results are stated at right side of the page. Abbas GÜÇLÜ who has 71 common n-grams with sample article and has a probability of 0.052 is accepted as author of sample article.

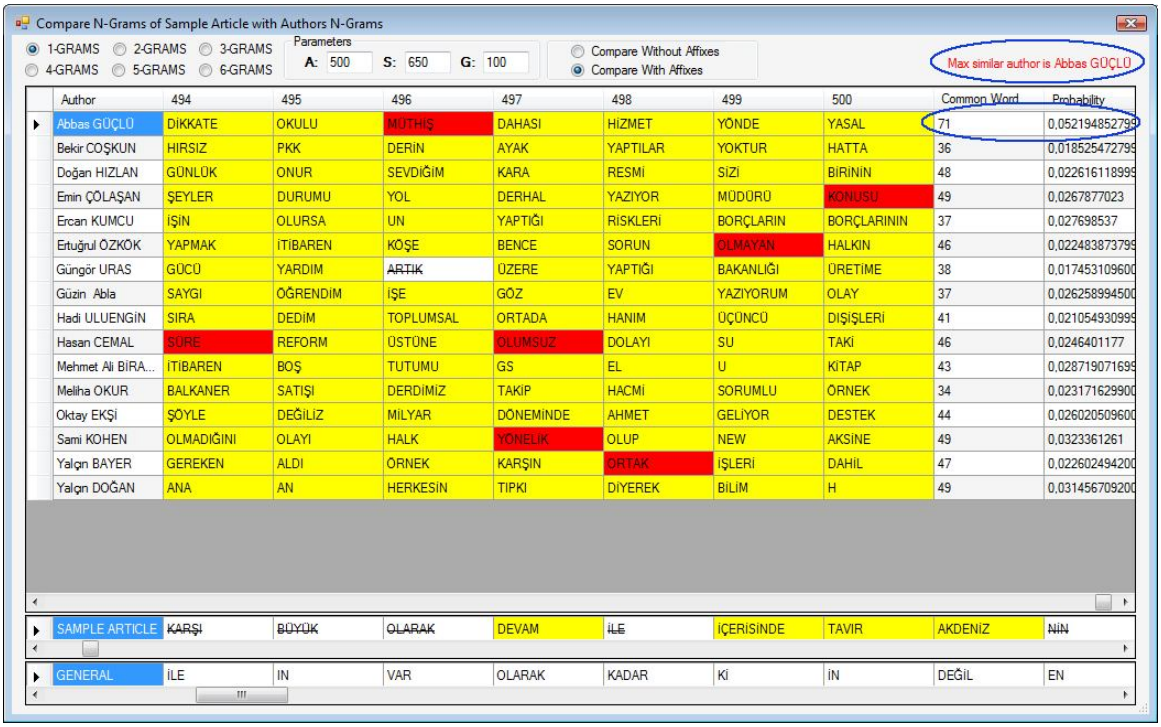


Figure 3.7 Sample result set for "Author Specific N-gram Method"

Program generates results for each n-gram group ($1 \leq n \leq 6$). So when a sample article is handled, 6 results are generated. To evaluate success of method, 100 articles are selected randomly for each author and method is applied to these 100 random articles. Results are stored in a text file with similar format shown in Figure 3.8. At the end of the file, success ratios are given.

For the sample given on Figure 3.8, for 95 of 100 random selected articles, 1-gram group makes true estimation and for 94 of 100 articles 2-grams make true estimation, so on. In 98 of 100 random articles, at least one of the n-gram groups makes true estimation. 6-grams' success ratio is only about 39%. For 61 articles, 6-grams generate wrong results or cannot generate any results. As can be seen in Figure 3.8, sometimes n-gram groups cannot make estimation because of no common n-grams existing between sample article and author n-grams. For 71 articles, all result generated groups make true estimations.

```
All Authors N-gram Count (A) = 500
Sample N-gram Count     (S) = 650
General N-gram Count    (G) = 100
--------------------------------------------------------------------------
1) D:\ Authors\Abbas GÜÇLÜ\14.5.2002.txt

1GRAM SIMILARITY : Abbas GÜÇLÜ -------> True
2GRAM SIMILARITY : Abbas GÜÇLÜ -------> True
3GRAM SIMILARITY : Yalçın BAYER -------> False
4GRAM SIMILARITY : Güzin  Abla -------> False
5GRAM SIMILARITY : Güzin  Abla -------> False
6GRAM SIMILARITY : Güzin  Abla -------> False
--------------------------------------------------------------------------
2) D:\ Authors\Abbas GÜÇLÜ\28.6.1999.txt

1GRAM SIMILARITY : Abbas GÜÇLÜ -------> True
2GRAM SIMILARITY : Abbas GÜÇLÜ -------> True
3GRAM SIMILARITY : Meliha OKUR -------> False
4GRAM SIMILARITY : Abbas GÜÇLÜ -------> True
5GRAM SIMILARITY : --------------------
6GRAM SIMILARITY : --------------------
--------------------------------------------------------------------------
.
.
.
--------------------------------------------------------------------------
100) D:\Authors\Abbas GÜÇLÜ\27.5.2005.txt

1GRAM SIMILARITY : Abbas GÜÇLÜ -------> True
2GRAM SIMILARITY : Abbas GÜÇLÜ -------> True
3GRAM SIMILARITY : Abbas GÜÇLÜ -------> True
4GRAM SIMILARITY : Oktay EKŞİ -------> False
5GRAM SIMILARITY : --------------------
6GRAM SIMILARITY : --------------------
--------------------------------------------------------------------------
Results

At least one true estimation ratio = 98 / 100
1-Grams true estimation ratio = 95 / 100
2-Grams true estimation ratio = 94 / 100
3-Grams true estimation ratio = 86 / 100
4-Grams true estimation ratio = 78 / 100
5-Grams true estimation ratio = 53 / 100
6-Grams true estimation ratio = 39 / 100
All true estimation success ratio = 71 / 100
```

Figure 3.8 Sample result file for randomly chosen 100 articles.

3.4.1 Experimental Results for Training and Test Sets

When "Author Specific N-gram Method" is applied to 100 randomly selected training set articles for each author, results seen on Table 3.6 are obtained. According to this table, 1-grams give best results for Abbas GÜÇLÜ and Sami KOHEN's articles. On the average, 1-grams have 90.06% success ratio, while 2-grams have 88.13%, 3-grams have 77.63%, 4-grams have 61.00%, 5-grams have 37.88% and 6-grams have 22.94% success ratio. With the ratio of 97.00%, at least one of n-grams gives correct result. Authors of articles, with the percentage 58.19%, are estimated truly for each n-gram group which could generate a result.

The results seen on Table 3.6 and 3.7 are obtained by comparing n-grams with their affixes. So, two n-grams are accepted as the same if two n-grams have same affixes.

If the same method is applied to 100 randomly selected test set articles (out of training set) for each author, results seen on Table 3.7 are obtained. In this case, 1-grams and 2-grams give best results for Hadi ULUENGİN's articles with a 100% success ratio. But, on the average 1-grams have 87.13% success ratio, 2-grams have 83.31%, 3-grams have 69.44%, 4-grams have 50.25%, 5-grams have 23.44% and 6-grams have 8.56% success ratio for all authors. At least one of the n-grams gives true result with the ratio 93.94%. Authors of articles, with the percentage 49.31%, are estimated truly for each n-gram method. When the training set and test set results are analyzed, it can be seen easily that, 1-grams are most successful n-gram group in "Author Specific N-gram Method". Also this model is more efficient on training set articles than test set articles.

Table 3.6 Author identification success ratios for *Author Specific N-gram Method* on training set articles

AUTHOR	1 GRAM	2 GRAM	3 GRAM	4 GRAM	5 GRAM	6 GRAM	AT LEAST 1	ALL
Abbas GÜÇLÜ	99	98	91	86	56	35	100	82
Bekir COŞKUN	84	82	61	42	22	13	98	51
Doğan HIZLAN	92	84	71	55	28	15	99	59
Ercan KUMCU	95	96	92	70	44	30	99	79
Ertuğrul ÖZKÖK	87	93	82	51	43	31	99	51
Güngör URAS	80	82	73	52	22	15	93	54
Güzin ABLA	98	98	92	81	34	17	98	87
Hadi ULUENGİN	97	97	91	70	33	7	100	74
Hasan CEMAL	80	83	72	63	50	29	96	48
Hasan PULUR	84	77	59	45	31	23	96	40
Mehmet Ali BİRAND	97	95	94	89	67	54	99	76
Oktay EKŞİ	89	81	73	60	50	28	97	48
Sami KOHEN	99	99	99	78	55	38	100	74
Taha AKYOL	83	83	62	49	33	19	96	42
Yalçın BAYER	90	85	73	49	32	11	84	41
Yalçın DOĞAN	87	77	57	36	6	2	98	25
AVERAGE %	90.06	88.13	77.63	61.00	37.88	22.94	97.00	58.19

Table 3.7 Author identification success ratios for *Author Specific N-gram Method* on test set articles

AUTHOR	1 GRAM	2 GRAM	3 GRAM	4 GRAM	5 GRAM	6 GRAM	AT LEAST 1	ALL
Abbas GÜÇLÜ	94	92	87	72	39	19	99	65
Bekir COŞKUN	83	71	52	22	7	1	92	38
Doğan HIZLAN	90	87	81	48	26	7	98	57
Ercan KUMCU	99	98	95	74	39	15	100	86
Ertuğrul ÖZKÖK	91	86	72	42	22	9	98	44
Güngör URAS	79	81	61	45	24	9	90	45
Güzin ABLA	99	99	96	84	41	20	99	81
Hadi ULUENGİN	100	100	96	64	16	3	100	72
Hasan CEMAL	75	79	56	45	17	6	88	34
Hasan PULUR	90	79	48	32	17	10	97	35
Mehmet Ali BİRAND	90	89	78	66	30	3	99	55
Oktay EKŞİ	91	80	74	68	41	17	98	57
Sami KOHEN	95	97	85	61	28	10	99	61
Taha AKYOL	47	41	25	20	10	5	56	10
Yalçın BAYER	86	82	64	38	14	2	94	32
Yalçın DOĞAN	85	72	41	23	4	1	96	17
AVERAGE %	87.13	83.31	69.44	50.25	23.44	8.56	93.94	49.31

3.4.2 Effects of Affixes on Author Specific N-gram Method

Results given in the previous section are obtained using words with affixes. In this case n-grams have same roots but different affixes are accepted as different n-grams. For matching of two n-grams with the same roots but different affixes, firstly roots of n-grams must be determined. Zemberek is used for root detection of words and n-grams with the same roots are accepted as common n-grams (Google 2008). Zemberek is an open source, platform independent, general purpose Natural Language Processing library and toolset designed for Turkic languages, especially Turkish.

Results calculated by n-grams without affixes can be seen on Table 3.8 for training set articles and on Table 3.9 for test set articles. In this model 1-grams' success ratios are decreasing to 55.94% on average, 2-grams' success ratios are decreasing to 69.44% and the 3-grams' success ratios are decreasing to 57.50% for training set. These values are 49.88%, 64.19% and 50.63% respectively for test set articles. In this case 2-grams and 3-grams are more effective than 1-grams.

Satisfying results are obtained for articles of Ercan KUMCU, Güzin ABLA, Sami KOHEN also in these method.

If the results of Table 3.8 and Table 3.9 are compared with Table 3.6 and 3.7, it can be said that, affixes are important elements for specifying characteristics of an author. Therefore, words will be used with their affixes in the next parts of this study.

Table 3.8 Author identification success ratios for *Author Specific N-gram Method* on training set articles without word affixes

AUTHOR	1 GRAM	2 GRAM	3 GRAM	4 GRAM	5 GRAM	6 GRAM	AT LEAST 1	ALL
Abbas GÜÇLÜ	77	69	73	77	57	38	98	37
Bekir COŞKUN	54	52	47	39	21	13	88	14
Doğan HIZLAN	64	72	56	43	27	13	93	16
Ercan KUMCU	95	92	88	70	48	37	98	61
Ertuğrul ÖZKÖK	58	67	54	43	38	30	92	18
Güngör URAS	48	60	53	45	23	18	83	20
Güzin ABLA	99	90	84	74	42	22	99	65
Hadi ULUENGİN	20	73	50	57	32	10	93	3
Hasan CEMAL	13	69	47	47	37	28	90	9
Hasan PULUR	46	76	37	36	29	23	93	6
Mehmet Ali BİRAND	72	74	77	83	77	55	98	45
Oktay EKŞİ	45	48	43	48	44	29	80	11
Sami KOHEN	94	94	88	70	59	38	100	45
Taha AKYOL	29	76	51	43	37	30	92	11
Yalçın BAYER	61	59	46	35	23	16	83	11
Yalçın DOĞAN	20	40	26	21	7	3	1	69
AVERAGE %	55.94	69.44	57.50	51.94	37.56	25.19	86.31	27.56

Table 3.9 Author identification success ratios for *Author Specific N-gram Method* on training set articles without word affixes

AUTHOR	1 GRAM	2 GRAM	3 GRAM	4 GRAM	5 GRAM	6 GRAM	AT LEAST 1	ALL
Abbas GÜÇLÜ	70	73	64	60	41	28	92	23
Bekir COŞKUN	50	48	45	29	11	2	84	12
Doğan HIZLAN	66	76	61	45	22	9	94	17
Ercan KUMCU	100	98	83	75	59	25	100	72
Ertuğrul ÖZKÖK	46	55	40	34	25	12	81	9
Güngör URAS	43	54	42	30	12	5	76	7
Güzin ABLA	100	99	95	84	54	26	100	70
Hadi ULUENGİN	25	84	47	53	23	13	94	3
Hasan CEMAL	8	59	42	43	25	10	75	5
Hasan PULUR	43	78	31	17	5	11	91	1
Mehmet Ali BİRAND	33	46	55	52	31	10	85	11
Oktay EKŞİ	54	39	46	52	43	23	83	17
Sami KOHEN	92	90	68	51	31	14	98	16
Taha AKYOL	11	41	36	20	11	11	64	3
Yalçın BAYER	50	56	41	37	18	7	78	10
Yalçın DOĞAN	7	31	14	19	5	1	54	0
AVERAGE %	49.88	64.19	50.63	43.81	26.00	12.94	84.31	17.25

3.5 Author Identification based on Support Vector Machine Method

This method is used for information retrieval by Baeza-Yates & Ribeiro-Neto (1999). SVM Method is used to identify author of an anonymous article. Firstly, n-grams of sample text are computed, n-grams which are common with general Turkish most frequently used n-grams are eliminated like previous method thus, sample article's specific n-grams are determined. Then for each remaining n-gram of sample article, *term frequency matrix* similar to Table 3.10 is prepared for each author. In the sample below, only 18 n-grams are selected. But the study made on all specific n-grams of sample article. If an n-gram is also an element of any 16 authors' top 500 n-grams, its frequency called as *term frequency* is given; otherwise 0 is assigned to related cells. For example, n-gram "MİLLİ" is one of the 500 most frequently used n-grams of Abbas GÜÇLÜ and its term frequency is 1868.

Table 3.10 Term frequency matrix for the sample article

N-Gram\Author	Sample Text	Abbas GÜÇLÜ	Bekir COŞKUN	Doğan HIZLAN	Emin ÇOLAŞAN	Ercan KUMCU	Ertuğrul ÖZKÖK	Güngör URAS	Güzin Abla	Hadi ULUENGİN	Hasan CEMAL	Mehmet Ali BİRAND	Meliha OKUR	Oktay EKŞİ	Sami KOHEN	Yalçın BAYER	Yalçın DOĞAN
KANUNU	1	0	0	0	0	0	0	317	0	0	0	0	126	0	0	0	0
MİLLİ	3	1868	128	277	489	1003	446	1269	0	178	433	316	0	533	0	1428	442
OLDU	2	1445	442	372	893	1046	823	1307	982	496	937	1031	214	450	787	1605	0
KAFALAR	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DÜN	2	656	369	319	1197	0	1082	316	0	486	854	0	178	660	805	974	1034
DEĞİŞİKLİĞİ	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DÜZENLEME	1	0	0	0	0	0	0	0	0	0	0	0	131	0	0	0	0
EDİLDİĞİ	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KOMİSYONDA	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
İZLEMİŞTİK	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HÜKÜMETİ	1	0	0	0	0	0	0	0	0	0	475	302	0	0	360	0	343
ARASINDAKİ	1	281	0	423	0	268	0	321	0	0	282	366	0	0	504	0	353
GERGİNLİK	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DEĞİŞİRDİ	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BEKLEYEN	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HÜKÜMETE	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SONUÇLARI	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BAHARA	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

After construction of frequency matrices, some mathematical operations like calculating weights for *tf* as term frequency, *idf* as inverse document frequency values of the text matrix and taking normalization with the cosine similarity formula are done.

In this method, $tf_{t,d}$ is the frequency of an n-gram, where t is a term which is a specific n-gram of sample article and d is accepted as a document which is formed by top 500 n-grams of authors. The weight of the term frequency is calculated like the following formula;

$$if\ tf_{t,d} > 0\ \ tf = 1 + \log_{10}(tf_{t,d}), \qquad else\ \ tf = 0.$$

The aim of the weighted term frequency is to put numbers in smaller values. After calculation of weights of term frequencies given on Table 3.10, Table 3.11 is obtained.

Table 3.11 Weights of the term frequencies for the sample article

N-Gram\Author	Sample Text	Abbas GÜÇLÜ	Bekir COŞKUN	Doğan HIZLAN	Emin ÇOLAŞAN	Ercan KUMCU	Ertuğrul ÖZKÖK	Güngör URAS	Güzin Abla	Hadi ULUENGİN	Hasan CEMAL	Mehmet Ali BİRAND	Meliha OKUR	Oktay EKŞİ	Sami KOHEN	Yalçın BAYER	Yalçın DOĞAN
KANUNU	1.000	0	0	0	0	0	0	3.501	0	0	0	0	3.100	0	0	0	0
MİLLİ	3.000	4.271	3.107	3.442	3.689	4.001	3.649	4.103	0	3.250	3.636	3.500	0	3.727	0	4.155	3.645
OLDU	2.000	4.160	3.645	3.571	3.951	4.020	3.915	4.116	3.992	3.695	3.972	4.013	3.330	3.653	3.896	4.205	0
KAFALAR	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DÜN	2.000	3.817	3.567	3.504	4.078	0	4.034	3.500	0	3.687	3.931	0	3.250	3.820	3.906	3.989	4.015
DEĞİŞİKLİĞİ	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DÜZENLEME	1.000	0	0	0	0	0	0	0	0	0	0	0	3.117	0	0	0	0
EDİLDİĞİ	2.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KOMİSYONDA	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
İZLEMİŞTİK	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HÜKÜMETİ	1.000	0	0	0	0	0	0	0	0	0	3.677	3.480	0	0	3.556	0	3.535
ARASINDAKİ	1.000	3.449	0	3.626	0	3.428	0	3.507	0	0	3.450	3.563	0	0	3.702	0	3.548
GERGİNLİK	3.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DEĞİŞİRDİ	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BEKLEYEN	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HÜKÜMETE	4.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SONUÇLARI	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BAHARA	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Another weighting operation used in this method is taking document frequencies. The inverse document frequency is calculated with the following formula;

$$idf_t = \log_{10}(N/df_t)$$

where df_t is the document frequency of term t. In this study df_t means the number of authors whose most frequently used 500 n-grams set contains n-gram t and N is the total number of authors in collection. Calculated inverse document frequency values of Table 3.11 can be seen on Table 3.12.

Table 3.12 Inverse document frequency values for the sample article

N-GRAM	IDF
KANUNU	1.2304
MİLLİ	1.2515
OLDU	1.2520
KAFALAR	0
DÜN	1.2515
DEĞİŞİKLİĞİ	0
DÜZENLEME	1.2041
EDİLDİĞİ	0
KOMİSYONDA	0
İZLEMİŞTİK	0
HÜKÜMETİ	1.2430
ARASINDAKİ	1.2492
GERGİNLİK	0
DEĞİŞİRDİ	0
BEKLEYEN	0
HÜKÜMETE	0
SONUÇLARI	0
BAHARA	0

Weight value $tf_{t,d} \times idf_t$ is defined as product of term frequency weight and inverse document frequency weight. It gives the weight of term t in document d and it is calculated by the following formula;

$$W_{t,d} = (1 + \log_{10} tf_{t,d}) \times \log_{10}(N/df_t) = tf_{t,d} \times idf_t$$

After calculation of weight values ($W_{t,d}$) of example, Table 3.13 is obtained.

Table 3.13 Weight values calculated for the sample article.

N-Gram\Author	Sample Text	Abbas GÜÇLÜ	Bekir COŞKUN	Doğan HIZLAN	Emin ÇOLAŞAN	Ercan KUMCU	Ertuğrul ÖZKÖK	Güngör URAS	Güzin Abla	Hadi ULUENGİN	Hasan CEMAL	Mehmet Ali BİRAND	Meliha OKUR	Oktay EKŞİ	Sami KOHEN	Yalçın BAYER	Yalçın DOĞAN
KANUNU	1.230	0	0	0	0	0	0	4.308	0	0	0	0	3.815	0	0	0	0
MİLLİ	3.755	5.346	3.889	4.308	4.617	5.008	4.567	5.136	0	4.068	4.551	4.380	0	4.664	0	5.200	4.562
OLDU	2.504	5.208	4.564	4.470	4.947	5.033	4.902	5.154	4.998	4.627	4.973	5.025	4.170	4.574	4.878	5.265	0
KAFALAR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DÜN	2.503	4.777	4.464	4.385	5.104	0	5.049	4.380	0	4.614	4.920	0	4.068	4.780	4.888	4.992	5.024
DEĞİŞİKLİĞİ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DÜZENLEME	1.204	0	0	0	0	0	0	0	0	0	0	0	3.754	0	0	0	0
EDİLDİĞİ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KOMİSYONDA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
İZLEMİŞTİK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HÜKÜMETİ	1.243	0	0	0	0	0	0	0	0	0	4.570	4.326	0	0	4.421	0	4.395
ARASINDAKİ	1.249	4.308	0	4.530	0	4.282	0	4.380	0	0	4.310	4.451	0	0	4.625	0	4.432
GERGİNLİK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DEĞİŞİRDİ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BEKLEYEN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HÜKÜMETE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SONUÇLARI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BAHARA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Finally, a normalization operation must be implemented over the matrix values on Table 3.13, because these numeric values are independent from each other and they must be accumulated between 0 and 1. For the normalization operation, the following cosine similarity formula is used;

$$N_{W_{t,d}} = W_{t,d} \times \frac{1}{\sqrt{{W_1}^2 + {W_2}^2 + {W_3}^2 + \cdots + {W_m}^2}}$$

where m is the number of n-grams and for each author $N_{W_{t,d}}$ (normalized weight value) is calculated. After calculation of cosine normalization Table 3.14 is constructed.

Table 3.14 Normalized weight values calculated for sample article.

N-Gram\Author	Sample Text	Abbas GÜÇLÜ	Bekir COŞKUN	Doğan HIZLAN	Emin ÇOLAŞAN	Ercan KUMCU	Ertuğrul ÖZKÖK	Güngör URAS	Güzin Abla	Hadi ULUENGİN	Hasan CEMAL	Mehmet Ali BİRAND	Meliha OKUR	Oktay EKŞİ	Sami KOHEN	Yalçın BAYER	Yalçın DOĞAN
KANUNU	0.041	0	0	0	0	0	0	0.136	0	0	0	0	0.134	0	0	0	0
MİLLİ	0.125	0.121	0.137	0.147	0.132	0.171	0.137	0.162	0	0.129	0.126	0.118	0	0.124	0	0.134	0.122
OLDU	0.083	0.118	0.160	0.153	0.141	0.172	0.147	0.162	0.153	0.147	0.138	0.136	0.146	0.121	0.141	0.136	0
KAFALAR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DÜN	0.083	0.108	0.157	0.150	0.146	0	0.152	0.138	0	0.146	0.136	0	0.142	0.127	0.141	0.129	0.134
DEĞİŞİKLİĞİ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DÜZENLEME	0.040	0	0	0	0	0	0	0	0	0	0	0	0.131	0	0	0	0
EDİLDİĞİ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KOMİSYONDA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
İZLEMİŞTİK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HÜKÜMETİ	0.041	0	0	0	0	0	0	0	0	0	0.127	0.117	0	0	0.128	0	0.117
ARASINDAKİ	0.042	0.097	0	0.155	0	0.146	0	0.138	0	0	0.119	0.120	0	0	0.133	0	0.118
GERGİNLİK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DEĞİŞİRDİ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BEKLEYEN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HÜKÜMETE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SONUÇLARI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BAHARA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

After normalization operation, by using Euclidean distance formula;

$$\mathrm{Sim}(X,Y) = \sum (X_i \times Y_i)$$

where, X and Y are the authors, X_i and Y_i are weight values of related n-grams. Similarities between the sample article and other authors' profiles are calculated and the similarity matrix is created as in the Table 3.15.

Table 3.15 Similarity matrix created for sample article.

Authors	Sample Text	Abbas GÜÇLÜ	Bekir COŞKUN	Doğan HIZLAN	Emin ÇOLAŞA	Ercan KUMCU	Ertuğrul ÖZKÖK	Güngör URAS	Güzin Abla	Hadi ULUENG	Hasan CEMAL	Mehmet Ali BİRAND	Meliha OKUR	Oktay EKŞİ	Sami KOHEN	Yalçın BAYER	Yalçın DOĞAN
Sample Text	1.000	0.711	0.404	0.426	0.522	0.430	0.501	0.440	0.374	0.429	0.527	0.517	0.451	0.596	0.477	0.525	0.496
Abbas GÜÇLÜ	0.711	1.000	0.621	0.648	0.738	0.660	0.701	0.612	0.628	0.687	0.715	0.742	0.624	0.770	0.659	0.754	0.758
Bekir COŞKUN	0.404	0.621	1.000	0.622	0.662	0.481	0.716	0.494	0.574	0.757	0.624	0.637	0.514	0.707	0.583	0.620	0.680
Doğan HIZLAN	0.426	0.648	0.622	1.000	0.714	0.594	0.752	0.612	0.667	0.714	0.723	0.718	0.604	0.664	0.566	0.635	0.698
Emin ÇOLAŞAN	0.522	0.738	0.662	0.714	1.000	0.573	0.759	0.662	0.600	0.686	0.753	0.679	0.662	0.749	0.611	0.765	0.764
Ercan KUMCU	0.430	0.660	0.481	0.594	0.573	1.000	0.601	0.692	0.630	0.546	0.640	0.662	0.567	0.561	0.671	0.645	0.633
Ertuğrul ÖZKÖK	0.501	0.701	0.716	0.752	0.759	0.601	1.000	0.636	0.652	0.802	0.811	0.735	0.653	0.731	0.649	0.744	0.745
Güngör URAS	0.440	0.612	0.494	0.612	0.662	0.692	0.636	1.000	0.595	0.620	0.697	0.665	0.715	0.606	0.585	0.689	0.662
Güzin Abla	0.374	0.628	0.574	0.667	0.600	0.630	0.652	0.595	1.000	0.701	0.701	0.697	0.588	0.644	0.625	0.634	0.692
Hadi ULUENGİN	0.429	0.687	0.757	0.714	0.686	0.546	0.802	0.620	0.701	1.000	0.743	0.714	0.635	0.729	0.683	0.674	0.776
Hasan CEMAL	0.527	0.715	0.624	0.723	0.753	0.640	0.811	0.697	0.701	0.743	1.000	0.802	0.723	0.714	0.705	0.769	0.825
Mehmet Ali BİRA...	0.517	0.742	0.637	0.718	0.679	0.662	0.735	0.665	0.697	0.714	0.802	1.000	0.666	0.691	0.718	0.703	0.757
Meliha OKUR	0.451	0.624	0.514	0.604	0.662	0.567	0.653	0.715	0.588	0.635	0.723	0.666	1.000	0.595	0.575	0.668	0.672
Oktay EKŞİ	0.596	0.770	0.707	0.664	0.749	0.561	0.731	0.606	0.644	0.729	0.714	0.691	0.595	1.000	0.689	0.824	0.778
Sami KOHEN	0.477	0.659	0.583	0.566	0.611	0.671	0.649	0.585	0.625	0.683	0.705	0.718	0.575	0.689	1.000	0.699	0.728
Yalçın BAYER	0.525	0.754	0.620	0.635	0.765	0.645	0.744	0.689	0.634	0.674	0.769	0.703	0.668	0.824	0.699	1.000	0.778
Yalçın DOĞAN	0.496	0.758	0.680	0.698	0.764	0.633	0.745	0.662	0.692	0.776	0.825	0.757	0.672	0.778	0.728	0.778	1.000

According to the results shown on Table 3.15, similarities of authors with themselves are equal to 1. Most similar author profile with the sample article is Abbas GÜÇLÜ with the similarity value 0.711.

Table 3.16 Author identification success ratios for SVM method on training set articles

AUTHOR	1 GRAM	2 GRAM	3 GRAM	4 GRAM	5 GRAM	6 GRAM	AT LEAST 1	ALL
Abbas GÜÇLÜ	94	95	86	81	57	35	99	70
Bekir COŞKUN	65	74	58	43	22	13	99	36
Doğan HIZLAN	88	86	72	64	31	19	100	49
Ercan KUMCU	89	87	85	71	44	31	96	68
Ertuğrul ÖZKÖK	59	87	73	57	43	34	98	36
Güngör URAS	71	81	68	54	22	15	93	47
Güzin ABLA	95	97	86	75	40	26	99	73
Hadi ULUENGİN	98	96	92	67	29	6	100	67
Hasan CEMAL	73	78	75	60	50	32	96	40
Hasan PULUR	69	72	53	51	32	23	93	34
Mehmet Ali BİRAND	90	93	88	88	67	55	98	68
Oktay EKŞİ	69	82	69	55	45	27	94	32
Sami KOHEN	94	99	95	64	55	35	100	56
Taha AKYOL	76	77	66	52	32	19	95	36
Yalçın BAYER	85	86	76	56	32	18	94	37
Yalçın DOĞAN	67	68	57	33	4	2	93	14
AVERAGE %	80.13	84.88	74.94	60.69	37.81	24.38	96.69	47.69

Table 3.17 Author identification success ratios for SVM method on test set articles

AUTHOR	1 GRAM	2 GRAM	3 GRAM	4 GRAM	5 GRAM	6 GRAM	AT LEAST 1	ALL
Abbas GÜÇLÜ	85	87	80	70	43	27	97	50
Bekir COŞKUN	65	64	49	26	8	1	86	31
Doğan HIZLAN	89	88	76	58	29	9	99	55
Ercan KUMCU	92	95	91	74	39	16	99	74
Ertuğrul ÖZKÖK	69	84	65	44	24	11	98	29
Güngör URAS	71	70	61	42	21	12	90	30
Güzin ABLA	98	98	94	88	54	36	99	80
Hadi ULUENGİN	94	99	94	61	16	4	100	64
Hasan CEMAL	69	69	55	43	16	6	90	29
Hasan PULUR	66	67	48	29	16	8	89	20
Mehmet Ali BİRAND	76	80	76	57	31	6	98	37
Oktay EKŞİ	75	74	65	64	43	21	93	40
Sami KOHEN	87	90	78	55	27	11	97	34
Taha AKYOL	48	42	21	18	10	7	57	11
Yalçın BAYER	76	70	64	38	10	3	90	24
Yalçın DOĞAN	56	62	48	20	4	1	88	5
AVERAGE %	76.00	77.44	66.56	49.19	24.44	11.19	91.88	38.31

3.5.1 Experimental Results for Training and Test Sets

After applying SVM method on training set articles, as can be seen on Table 3.16, most successful results obtained for Sami KOHEN by 2-grams with 99% success ratio and for Hadi ULUENGİN by 1-grams with the ratio of 98%. On the average, SVM method reaches 80.13% success ratio by 1-grams, 84.88% success ratio by 2-grams, 74.94% for 3-grams, 60.69% for 4-grams, 37.81% for 5-grams and 24.38% success ratio for 6-grams. At least one of the six n-gram groups gives true result for 96.69% of articles while 47.69% of them are predicted truly by all n-gram groups that generated any results whether true or false.

On the other hand, according to values on Table 3.17, SVM method gives best results for Hadi ULUENGİN by 2-grams with the success ratio 99% and for Güzin ABLA by 1-grams with the success ratio 98% when applied to test set articles. Average success ratios are 76% by 1-grams, 77.44% by 2-grams, 66.56% by 3-grams, 49.19% by 4-grams, 24.44% by 5-grams, and 11.19% by 6-grams. Authors of 91.88% of articles are predicted correctly by at least one of the six n-gram groups and authors of 38.31% of articles are identified by all n-gram groups that generated any results.

In SVM method, 2-grams are more effective than 1-grams on both training and test set articles. When looked at average success ratio comparisons on Table 3.18, it can be said that, author specific n-gram method is more successful than SVM method for both training and test set articles.

Table 3.18 Comparison table for author identification success ratios

		1 GRAM	2 GRAM	3 GRAM	4 GRAM	5 GRAM	6 GRAM	AT LEAST 1	ALL
Author Specific N-gram Method	**Training Set**	90.06	88.13	77.63	61.00	37.88	22.94	97.00	58.19
	Test Set	87.13	83.31	69.44	50.25	23.44	8.56	93.94	49.31
SVM Method	**Training Set**	80.13	84.88	74.94	60.69	37.81	24.38	96.69	47.69
	Test Set	76.00	77.44	66.56	49.19	24.44	11.19	91.88	38.31

Used methods and obtained success ratios of previous studies can be exemplified by Kjell (1994) performed experiments with neural networks and Bayesian classifiers in Authorship attribution area. Testing was performed using 30 samples of text from two authors. Each sample was 6,000 letters long and obtained about 80-90% success.

The usefulness of function words is examined by Shlomo & Levitan (2000). The authors conducted experiments with support vector machine classifiers in twenty novels of eight authors and they obtained success rates above 90%. They concluded that, using function words is a valid and good approach in authorship attribution.

According to researchers in 2001, Stamatatos, Fakotakis & Kokkinakis have measured a success rate of %65 and %72 in their study for authorship recognition, which is an implementation of Multiple Regression and Discriminant Analysis on 30 texts of 10 authors.

Also in 2003, Diederich and his collaborators conducted experiments with support vector classifiers and detected author with %60-80 success rates with different parameters.

Diri & Amasyalı (2003), figured out 22 of style markers and by considering them as having equal weights a success rate of %67 has been measured on an author group consists of 18 different authors. Results with the artificial neural networks have %60 of success rate using MLP and %72 of success rate using Radial Base Function. In the second phase 11 of style markers among the 22 style marker has been selected as equal weights and the success rate improved to %78. But the MLP success was %60 and Radial Base Function success was %61. In the third phase the style markers SM3, SM13, SM17 and SM21 has been taken with different weights and they have measured a success rate of %84. In their other study, Amasyali & Diri (2006) have handled the text as a whole and they have extracted the character bi-grams and the tri-grams and obtained a success rate of 83% for 18 different authors, with 35 different texts written by each author.

In the study of Bozkurt, Bağlıoğlu & Uyar (2007), gaussian classifiers on the stylometry feature set also worked well obtaining around % 60 success rates. Support vector machine classifier is also seen as a very good classifier for authorship attribution obtaining a success rate around %95 on bag of words feature set. The number of authors is 18 and the experiments are done according to data which have 500 articles from 18 different writers.

CHAPTER FOUR

IMPLEMENTATION

In the scope of this study, a program was developed to process articles and to obtain necessary results. This program was developed by using Visual Studio .NET technology and Visual C# programming language. Lots of functionality is needed for analysis collected by this program.

Main form of the program can be seen on Figure 4.1. Some important functionalities such as corpus creation and correction, n-gram counting, statistics, author identification and article searching, can be reached from main form. Shown as Figure 4.1, on corpus creation page, user can determine path of collected text and corpus file will be created. User can create a corpus by filtering unwanted characters from text collection.

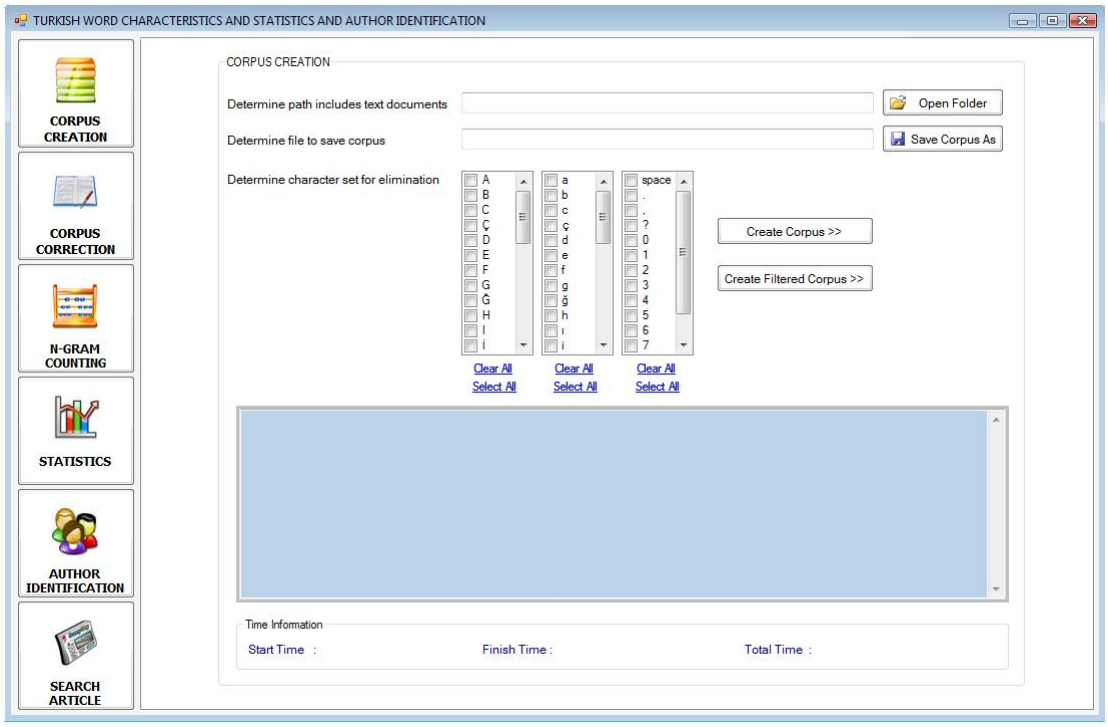


Figure 4.1 Corpus creation page

Mistaken words in corpus can be eliminated or corrected by using corpus correction page shown in Figure 4.2. These words arise from web pages where the articles are downloaded from. By correcting mistaken words only once, other instanses of them in the corpus will be corrected automaticaly. In the same way, by eliminating a mistaken word only once, other instances of it will be eliminated too.

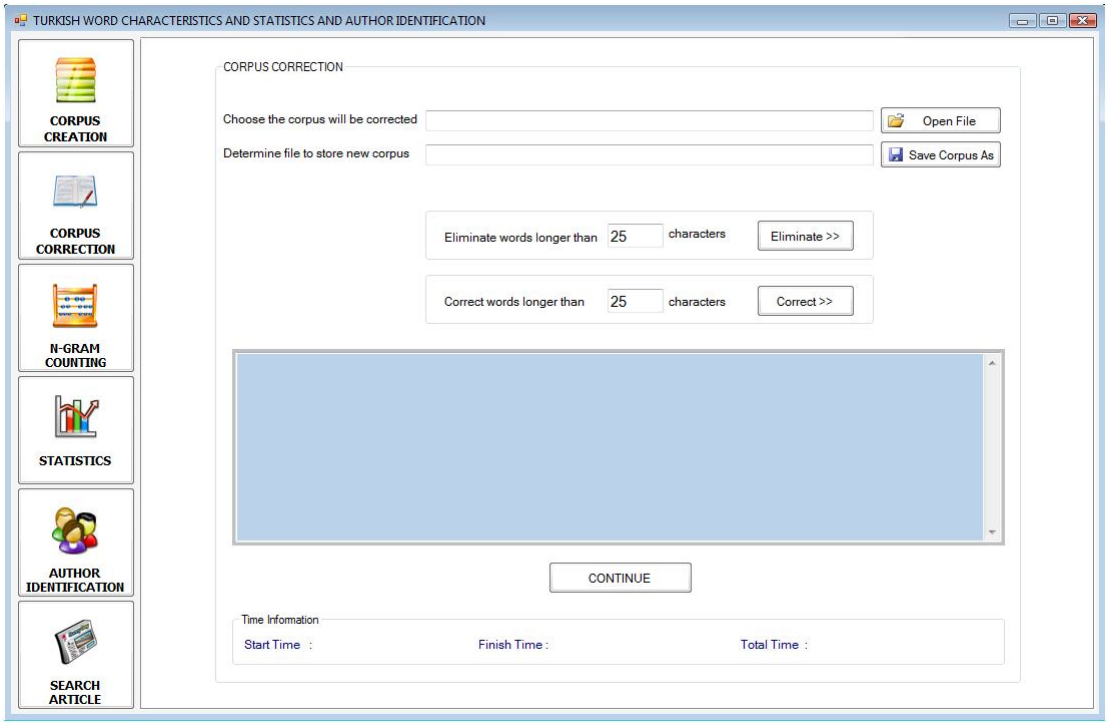


Figure 4.2 Corpus correction page

N-gram counting page shown on Figure 4.3 is the page which user can select an existing corpus file that contains corpus character alphabet and a file to collect counting results. Results are stored in a text file with descending order according to the frequencies as given in Table 4.1.

Table 4.1 N-gram counting result file format

N-gram : frequency (probability)
BİR : 97671 (0,02437534) VE : 68217 (0,01702463) BU : 57516 (0,01435403) DE : 43178 (0,01077575) İÇİN : 22810 (0,005692596) :

By using page given in Figure 4.3, both letter and word based n-gram counting can be done. Before starting n-gram counting, value of n and the minimum occurrence value as a threshold frequency must be determined.

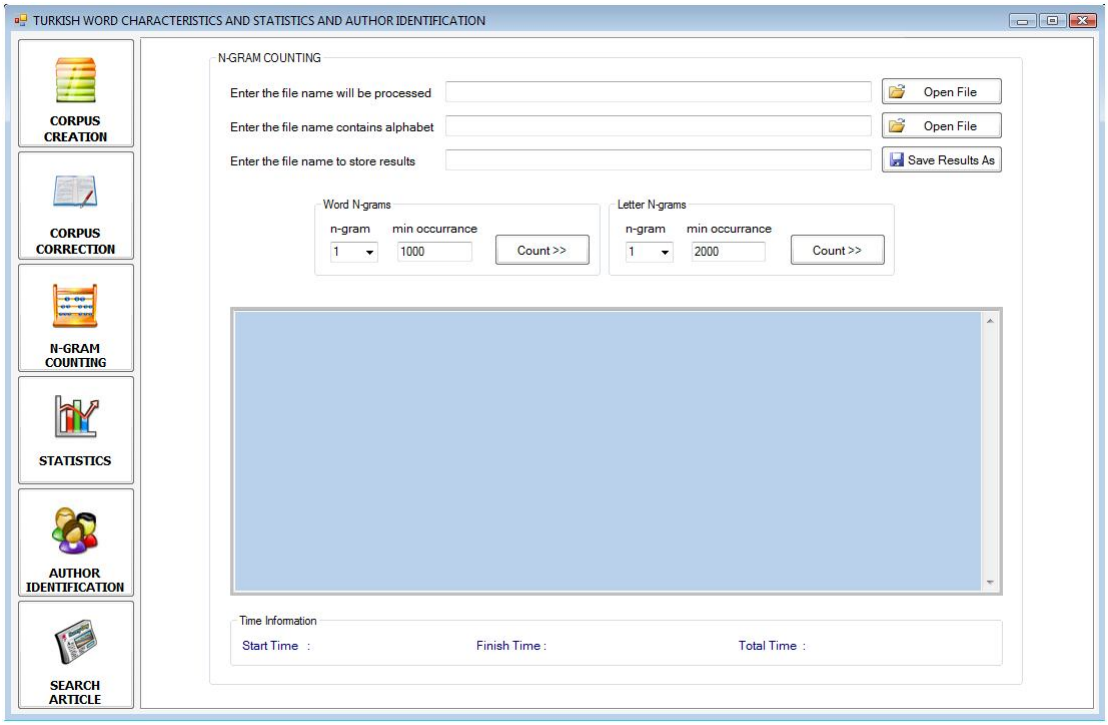


Figure 4.3 N-gram counting page

By using statistics page shown on Figure 4.4, a user can list most or least used, letter or word based n-grams of Turkish. Also, user can determine number of n-grams that will be listed. Lists are prepared by descending order according to the n-gram frequencies.

Author identification page shown on Figure 4.5, contains functionalities for identifiying author of an anonymous article. Two methods, used for author identification which were explained in section 3.4 and 3.5, studied in this page. Methods can be applied on both an anonymous article or randomly selected 100 articles. Parameters (A, S, G) can be set and n-gram comparision type (with affixes or without affixes) can be selected in this form.

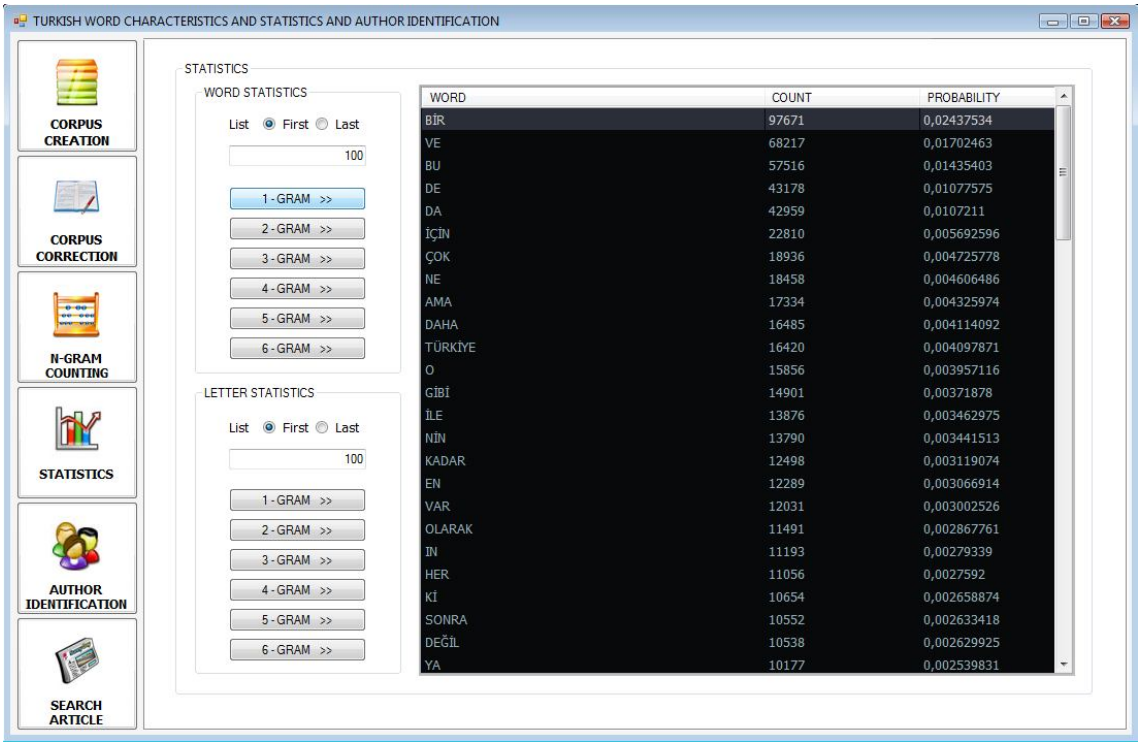


Figure 4.4 Statistics page.

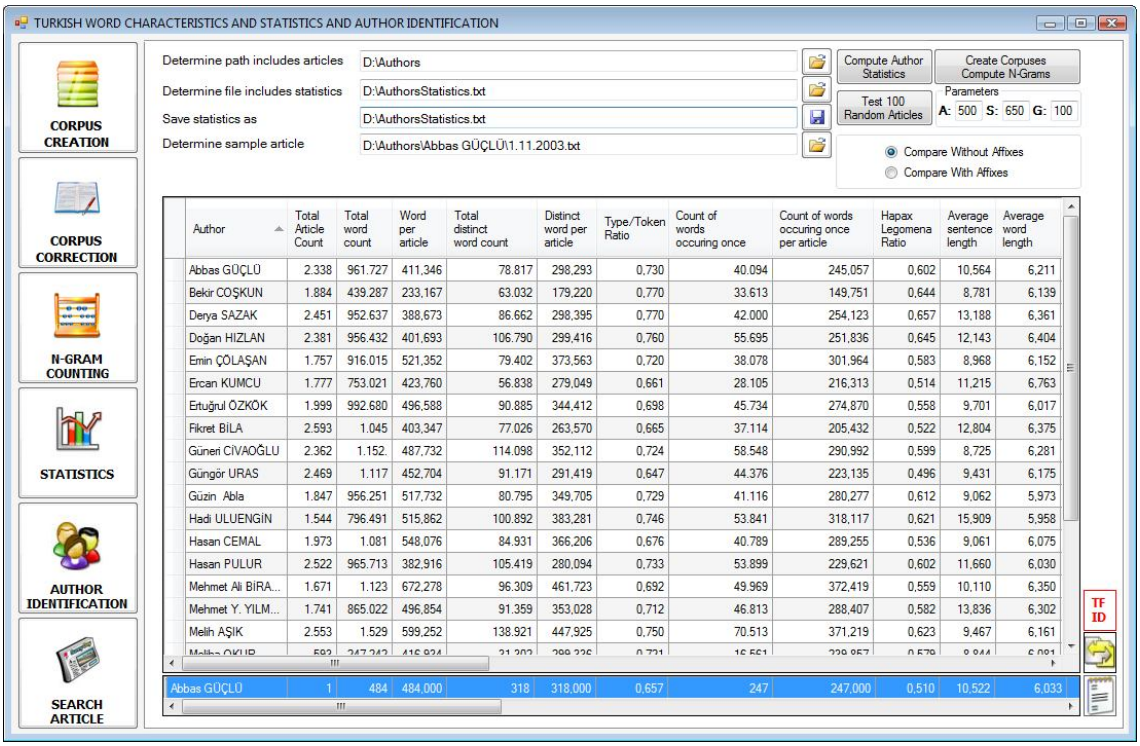
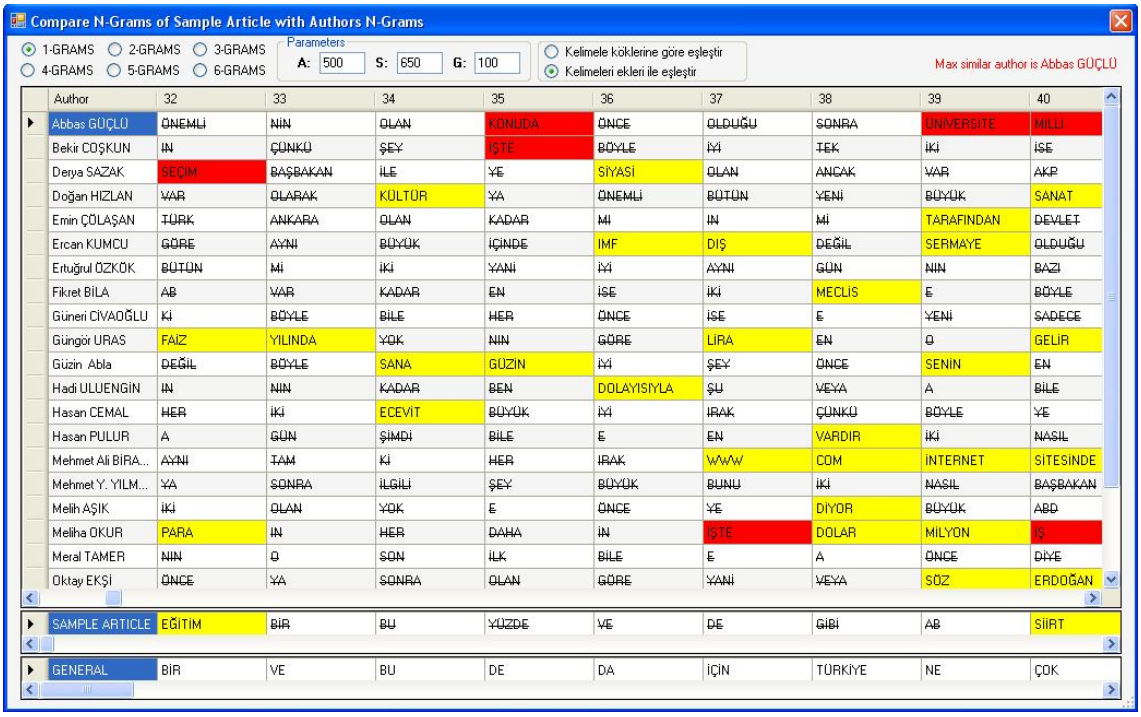


Figure 4.5 Author identification page.

The page developed for author specific n-grams method can be seen on Figure 4.6. N-gram group, parameters and n-gram comparison type (with affixes, without affixes) can be set in this page too. Finally, an author has more common n-grams with the anonymous text, is selected as author of the anonymous text.

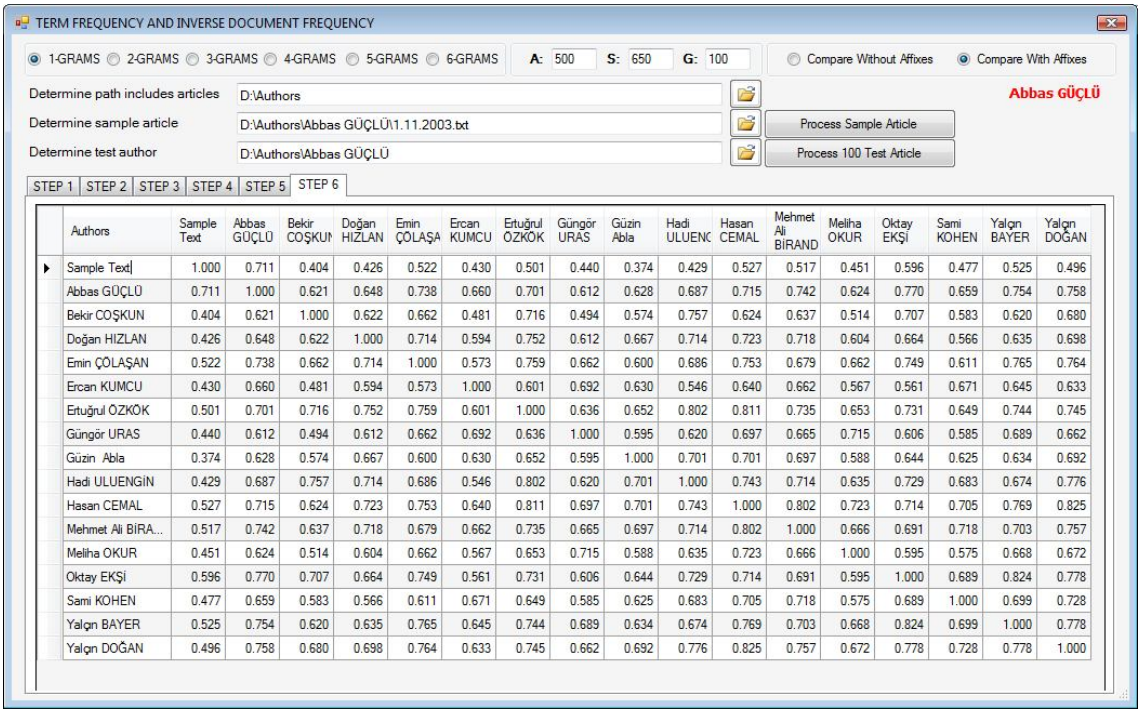
Yellow backgrounded cells are sample article or authors' specific n-grams which are not elements of generally used Turkish n-grams. While strikethrough cells are common elements with generally used Turkish n-grams. Red backgrounded cells are the common n-grams with an author's n-grams and sample article.



Author	32	33	34	35	36	37	38	39	40
Abbas GÜÇLÜ	ÖNEMLİ	NİN	OLAN	KONUDA	ÖNCE	OLDUĞU	SONRA	ÜNİVERSİTE	MİLLİ
Bekir COŞKUN	İN	ÇÜNKÜ	ŞEY	İŞTE	BÖYLE	İYİ	TEK	İKİ	İSE
Derya SAZAK	SEÇİM	BAŞBAKAN	İLE	YE	SİYASİ	OLAN	ANCAK	VAR	AKP
Doğan HIZLAN	VAR	OLARAK	KÜLTÜR	YA	ÖNEMLİ	BÜTÜN	YENİ	BÜYÜK	SANAT
Emin ÇÖLAŞAN	TÜRK	ANKARA	OLAN	KADAR	Mİ	İN	Mİ	TARAFINDAN	DEVLET
Ercan KUMCU	GÖRE	AYNI	BÜYÜK	İÇİNDE	İMF	DIŞ	DEĞİL	SERMAYE	OLDUĞU
Ertuğrul ÖZKÖK	BÜTÜN	Mİ	İKİ	YANİ	İYİ	AYNI	GÜN	NİN	BAZI
Fikret BİLA	AB	VAR	KADAR	EN	İSE	İKİ	MECLİS	E	BÖYLE
Güneri CİVAOĞLU	Kİ	BÖYLE	BİLE	HER	ÖNCE	İSE	E	YENİ	SADECE
Güngör URAS	FAİZ	YILINDA	YOK	NİN	GÖRE	LİRA	EN	O	GELİR
Güzin Abla	DEĞİL	BÖYLE	SANA	GÜZİN	İYİ	ŞEY	ÖNCE	SENİN	EN
Hadi ULUENGİN	İN	NİN	KADAR	BEN	DOLAYISIYLA	ŞU	VEYA	A	BİLE
Hasan CEMAL	HER	İKİ	ECEVİT	BÜYÜK	İYİ	IRAK	ÇÜNKÜ	BÖYLE	YE
Hasan PULUR	A	GÜN	ŞİMDİ	BİLE	E	EN	VARDIR	İKİ	NASIL
Mehmet Ali BİRA...	AYNI	TAM	Kİ	HER	IRAK	WWW	COM	İNTERNET	SİTESİNDE
Mehmet Y. YILM...	YA	SONRA	İLGİLİ	ŞEY	BÜYÜK	BUNU	İKİ	NASIL	BAŞBAKAN
Melih AŞIK	İKİ	OLAN	YOK	E	ÖNCE	YE	DİYOR	BÜYÜK	ABD
Meliha OKUR	PARA	İN	HER	DAHA	İN	İŞTE	DOLAR	MİLYON	İŞ
Meral TAMER	NİN	O	SON	İLK	BİLE	E	A	ÖNCE	DİYE
Oktay EKŞİ	ÖNCE	YA	SONRA	OLAN	GÖRE	YANİ	VEYA	SÖZ	ERDOĞAN
SAMPLE ARTICLE	EĞİTİM	BİR	BU	YÜZDE	VE	DE	GİBİ	AB	SİİRT
GENERAL	BİR	VE	BU	DE	DA	İÇİN	TÜRKİYE	NE	ÇOK

Figure 4.6 Author prediction page for author specific n-gram method.

On the Figure 4.7, author prediction page for SVM method is given. N-gram group, parameters and n-gram comparison type (with affixes, without affixes) can also be set in this page. SVM method can be applied on an anonymous article or randomly selected 100 anonymous articles. All steps from frequency table to similarity table are given in this page clearly. All author profiles give similarity value equal to 1 when compared with themselves. For other profiles, similarity values are between 0 and 1.



Authors	Sample Text	Abbas GÜÇLÜ	Bekir COŞKUN	Doğan HIZLAN	Emin ÇOLAŞA	Ercan KUMCU	Ertuğrul ÖZKÖK	Güngör URAS	Güzin Abla	Hadi ULUENG	Hasan CEMAL	Mehmet Ali BİRAND	Meliha OKUR	Oktay EKŞİ	Sami KOHEN	Yalçın BAYER	Yalçın DOĞAN
Sample Text	1.000	0.711	0.404	0.426	0.522	0.430	0.501	0.440	0.374	0.429	0.527	0.517	0.451	0.596	0.477	0.525	0.496
Abbas GÜÇLÜ	0.711	1.000	0.621	0.648	0.738	0.660	0.701	0.612	0.628	0.687	0.715	0.742	0.624	0.770	0.659	0.754	0.758
Bekir COŞKUN	0.404	0.621	1.000	0.622	0.662	0.481	0.716	0.494	0.574	0.757	0.624	0.637	0.514	0.707	0.583	0.620	0.680
Doğan HIZLAN	0.426	0.648	0.622	1.000	0.714	0.594	0.752	0.612	0.667	0.714	0.723	0.718	0.604	0.664	0.566	0.635	0.698
Emin ÇOLAŞAN	0.522	0.738	0.662	0.714	1.000	0.573	0.759	0.662	0.600	0.686	0.753	0.679	0.662	0.749	0.611	0.765	0.764
Ercan KUMCU	0.430	0.660	0.481	0.594	0.573	1.000	0.601	0.692	0.630	0.546	0.640	0.662	0.567	0.561	0.671	0.645	0.633
Ertuğrul ÖZKÖK	0.501	0.701	0.716	0.752	0.759	0.601	1.000	0.636	0.652	0.802	0.811	0.735	0.653	0.731	0.649	0.744	0.745
Güngör URAS	0.440	0.612	0.494	0.612	0.662	0.692	0.636	1.000	0.595	0.620	0.697	0.665	0.715	0.606	0.585	0.689	0.662
Güzin Abla	0.374	0.628	0.574	0.667	0.600	0.630	0.652	0.595	1.000	0.701	0.701	0.697	0.588	0.644	0.625	0.634	0.692
Hadi ULUENGİN	0.429	0.687	0.757	0.714	0.686	0.546	0.802	0.620	0.701	1.000	0.743	0.714	0.635	0.729	0.683	0.674	0.776
Hasan CEMAL	0.527	0.715	0.624	0.723	0.753	0.640	0.811	0.697	0.701	0.743	1.000	0.802	0.723	0.714	0.705	0.769	0.825
Mehmet Ali BİRA...	0.517	0.742	0.637	0.718	0.679	0.662	0.735	0.665	0.697	0.714	0.802	1.000	0.666	0.691	0.718	0.703	0.757
Meliha OKUR	0.451	0.624	0.514	0.604	0.662	0.567	0.653	0.715	0.588	0.635	0.723	0.666	1.000	0.595	0.575	0.668	0.672
Oktay EKŞİ	0.596	0.770	0.707	0.664	0.749	0.561	0.731	0.606	0.644	0.729	0.714	0.691	0.595	1.000	0.689	0.824	0.778
Sami KOHEN	0.477	0.659	0.583	0.566	0.611	0.671	0.649	0.585	0.625	0.683	0.705	0.718	0.575	0.689	1.000	0.699	0.728
Yalçın BAYER	0.525	0.754	0.620	0.635	0.765	0.645	0.744	0.689	0.634	0.674	0.769	0.703	0.668	0.824	0.699	1.000	0.778
Yalçın DOĞAN	0.496	0.758	0.680	0.698	0.764	0.633	0.745	0.662	0.692	0.776	0.825	0.757	0.672	0.778	0.728	0.778	1.000

Figure 4.7 Author prediction page for SVM method.

Search article page shown on Figure 4.8, is used to search and display articles according to date or author. After selection of an article, picture of the author appears on the top right of the screen and text of the article states on the center area. User can change font, text color, and background color by using the windows shown on Figure 4.9 and can select one of the themes shown on Figure 4.10.

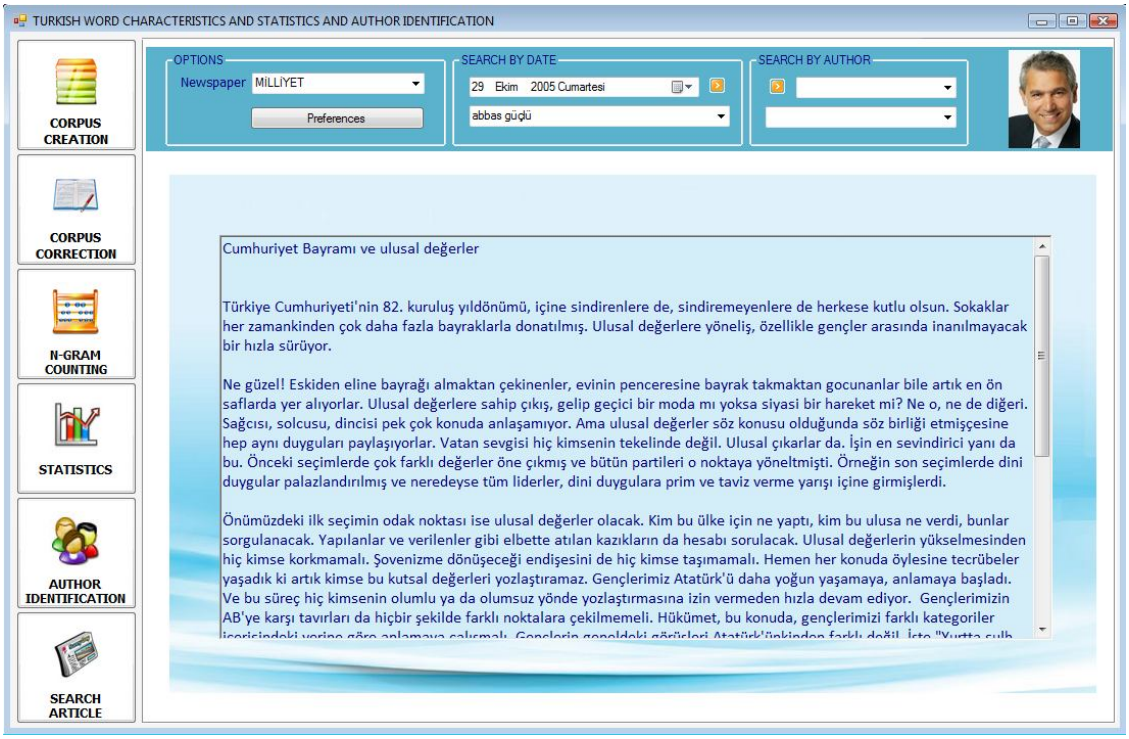


Figure 4.8 Article searching page.

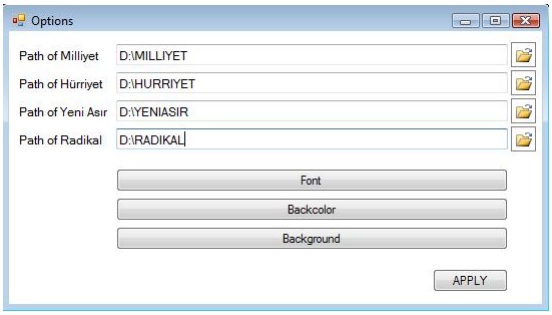


Figure 4.9 Options for “Search Article” page.

Figure 4.10 Background options.

CHAPTER FIVE

CONCLUSION & FUTURE WORK

In the scope of this thesis, firstly some linguistic studies are made to determine important characteristics of Turkish by using a large scale Turkish corpus. Studied linguistic features consist of Type/Token Ratio, Hapax Legomena Ratio, Index of Coincidence, Entropy, Redundancy and Unicity Distance. Also some other features of Turkish like most common letter and word n-grams, letter position distributions on Turkish words, word and sentence length distributions and most commonly observed CV patterns, are collected and analysis to see if Turkish word and letter n-grams fits Zipf's Law are made.

Type/Token Ratio per article is calculated about 72% and Hapax Legomena Ratio is calculated as 81.292% for Turkish. Index of coincidence value for Turkish is calculated as 0.063. While 100-grams' entropy value 0.29 is accepted as entropy of Turkish, highest observed redundancy value is 4.62 which is for 100-grams.

Turkish words mostly end with the letter N (15.61%), after that E, A, and R are observed as terminated letters of words. 12.15% of words begin with B and D, K, A, Y, S are other frequently observed letters which initiate words. While B (12.15%) is the most commonly used letter as first character of words, A is the most frequently observed letter in the second (21.34%) and fourth (16.21%) place in a word. For third place R (16.782) and for fifth place N (11.370) is being used commonly.

BİR, VE, BU, DE, and DA are top 5 frequently used words of Turkish. "YA DA", "BÖYLE BİR", "HEM DE", "BİR ŞEY" and "NE KADAR" are most commonly used bigrams while "NE YAZIK Kİ", "BİR KEZ DAHA", "NE VAR Kİ", "ÇOK ÖNEMLİ BİR" and "BİR SÜRE SONRA" are top 5 trigrams.

Average word length is computed as 6.34 and average sentence length is calculated as 10.69 for Turkish.

Most commonly observed CV forms are CVCVC, CVC, CV, CVCV, CVCCV, CVCCVC, CVCVCV, CVCVCVC, CVCCVCV and VCVC. 20 of top 60 CV patterns of Turkish and English (approximately 30%) are common.

While word 1, 2 and 3-grams fit Zipf's law, word 4 and 5-grams deviate from Zipf's law. There is a clear deviation for word n-grams in $6 \leq n \leq 10$ interval and for all letter n-grams.

On the next part of the study, two methods, *Author Identification Based on Author Specific N-gram Method* and *Author Identification based Support Vector Machine (SVM) Method* are applied on training and test sets of 16 authors. The first method gives more successful results than the second method. First method reaches success ratios as 90% for training sets and 87% for test sets with 1-grams while second method has success ratios as 85% for training sets and 77% for test sets with 2-grams.

Obtained statistics can be used for many computer science areas such as data security, language identification, spell checking, data compression and speech recognition. Also, more successful results can be obtained in author identification studies by adding new features to existing features and combining results obtained by several methods.

REFERENCES

Amasyalı M.F., & Diri B. (2006). *Automatic Written Turkish Text Categorization in Terms of Author, Genre and Gender*. 11th International Conference on Applications of Natural Language to Information Systems, Austria.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman, Boston.

Bozkurt, İ.N., Bağlıoğlu, O., & Uyar, E. (2007). *Authorship Attribution: Performance of various features and classification methods*. Computer and information sciences, pp:158-162

Brainerd B. (1974). *Weighting Evidence in Language and Literature: A Statistical Approach*. University of Toronto Press

Burrows J.F. (1992). *Not unless you ask nicely: the interpretative nexus between analysis and information*. Literary Linguist Computing (7), pp.91-109

Cavnar W.B. (1994). *Using an n-gram-based Document Representation with a Vector Processing Retrieval Model*. In Proceedings of the Third Text Retrieval Conference (TREC-3)

Çatal, Ç., Erbakırcı, K., & Erenler, Y. (2003). *Computer-based Authorship Attribution for Turkish Documents*. Turkish Symposium on Artificial Intelligence and Neural Networks

Çiçekli, İ., & Temizsoy, M. (1997). *Automatic Creation of a Morphological Processor in Logic Programming Environment.* Proc. of 5th Intl. Conf. on Practical Applications of Prolog, pp. 95-106.

Dalkılıç, M. E., & Dalkılıç, G. (2001). *Some Measurable Language Characteristics of Printed Turkish*. International Symposium on Computer and Information Sciences (ISCIS) XVI, 5-7 November, Antalya

Dalkılıç, G., & Çebi, Y. (2004). *Zipf's Law and Mandelbrot's Constants for Turkish Language Using Turkish Corpus*, Lecture Notes in Computer Science Volume 3621, pp. 273-282.

Diederich, J., Kindermenn, J., Leopold, E., & Pass, G. (2003). *Authorship attribution with Support Vector Machines*. Applied Intelligence. pp.109-123.

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.

Diri B., & Amasyalı, M. F. (2003). *Automatic Author Detection for Turkish Texts, Artificial Neural Networks and Neural Information Processing*. Pp. 138-141

Friedman, W. (1922). *The Index of Coincidence and Its Applications in Cryptography*. Publication No. 22. Geneva IL: Riverbank Publications.

F¨urnkranz J. (1998). *A Study using n-gram Features for Text Categorization*. Austrian Research Institute for Artifical Intelligence

Gayde, Ş., & Karslıgil, M. Y. (2000). *A Natural Language Processing Tool for the Analysis of Turkish Texts and a Fuzzy-based Statistical Approach for Author Recognition*. ISCIS XV, pp. 1-12.

Google 2008. *Zemberek*. Retrieved Nov 05, 2008, http://code.google.com/p/ zemberek/

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

Ha, L. Q., Garcia, E. I. S., Ming, Ji, & Smith, F. J. (2002). *Extension of Zipf's law to words and phrases*. In: Proceedings of COLING 2002, pp 315-320.

Hollink, V., Kamps, J., Monz, C., & de Rijke, M. (2004). *Monolingual Document Retrieval for European Languages*. Information Retrieval, Volume 7, Numbers 1-2, pp. 33-52

Holmes D.I. (1994). *Authorship Attribution*. Computers and the Humanities, 28:87-106

Kjell, B. (1994). *Authorship Attribution of Text Samples using Neural Networks and Bayesian Classifiers.* IEEE International Conference on 2-5 Oct 1994, Volume: 2, pp. 1660-1664.

Kit, C., & Wilks, Y. (1998). *The Virtual Approach to Deriving Ngram Statistics from Large Scale Corpora*. International Conference on Chinese Information Processing Conference, Beijing, China, pp. 223—229

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

Menezes, A., van Oorschot, P. & Vanstone, S. (1996). *Handbook of Applied Cryptography*. CRC Press

Morton A.Q. (1965). *The Authorship of Greek Prose Journal of the Royal Statistical Society*, Series A, 128:169-233

Mosteller F., Wallace D.L. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Reading, MA:Addison-Wesley

Oflazer, K. (2000). *Developing a morphological analyzer for Turkish*. NATO ASI on Language Engineering for Lesser-studied Languages, Ankara, Turkey.

Santos, S., & Alcaim, A. (2000). *Reduced Sets of Subword Units for Continuous Speech Recognition of Portuguese*. Electronic Letters, vol. 36 (6), pp. 586-588.

Schrader, B. (2006). *Non-probabilistic alignment of rare German and English nominal expressions.* LREC-2006: Fifth International Conference on Language Resources and Evaluation. pp.1274-1277

Seberry, J., & Pieprzyk, J. (1988). *Cryptography: an Introduction to Computer Securit.* Pren-Hall.

Sezgin, F. (1993). *Dil ve Edebiyatta İstatistik ve Bilgisayar Uygulamaları.* Dergah Yayınları, İstanbul

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

Shlomo, A., & Levitan, S.(2000). *Measuring the usefullness of function words for Authorship Attribution*. Proceedings of ACH/ALLC Conference 2005 in Victoria, BC, Canada.

Stamatatos E., Fakotakis N., & Kokkinakis G. (2000). *Automatic Text Categorization in Terms of Genre and Author.* Computational Linguistics, pp.471-495

Stamatatos, E, Fakotakis, N., & Kokkinakis, G. (2001). *Computer- Based Authorship Attribution without lexical measures*. Computers and Hummanities, pp.193-214.

Stinson, D.R. (1995). *Cryptography Theory and Practice*. CRC Press

Tan C. M., Wang Y. F., Lee C. D. (2002). *The Use of Bi-grams to Enhance Journal Information Processing and Management*. Vol:30 No:4 pp.529-546

Teahan, W.J. (1998). *Modeling English Text*. Ph.D. Dissertation, The Univ. of Waikato, New Zeland.

Töreci, E. (1975). *Statistical Investigations on the Turkish Language Using Digital Computers.* Master Thesis, Middle East Technical University, Ankara, Turkey.

Tweedie F., & Baayen, H. (1998). *How Variable may a Constant be Measures of Lexical Richness in Perspective*. Computers and the Humanities, 32(5):323-352

Wikipedia 2009. *Corpus Linguistics*. Retrieved Apr 27, 2009, http://en.wikipedia.org/wiki/Corpus_linguistics

Wikipedia 2009. *Unicity distance*. Retrieved May 9, 2009, http://en.wikipedia.org/wiki/Unicity_distance

Waters, R. C. (1976). *Cryptology and Data Communications*. Massachusetts Institute of Technology Artificial Intelligence Laboratory Cryptology and Data Communications, WP 136.

Witten, I. H., Moffat, A., & Bell, T. C. (1999). *Managing Gigabytes: Compressing and Indexing Documents and Images*. 2nd ed., Morgan and Kaufmann Publishers.

Youmans, G. (1990). *Measuring Lexical Style and Competence: The Type-Token Vocabulary Curve.* Department of English University of Missouri-Columbia Columbia, MO 65211