Mersin Üniversitesi Dil ve Edebiyat Dergisi, MEUDED, 2016; 13 (2), 1-33.

FORMULAICITY IN TURKISH: EVIDENCE FROM THE TURKISH NATIONAL CORPUS[1]

Selma Ayşe ÖZEL[2], Yasin BEKTAŞ[3], Hakan YILMAZER[4]

Çukurova University

Abstract: Formulaic sequences are the most frequently occurred forms in a language. Identification of formulaic sequences in language is useful for a wide range of areas including linguistics, second language learning, natural language processing, etc. To identify formulaic sequences in a language, the most preferred method is to use a corpus, which may be formed from written texts or tape-recorded conversations in the language, and count the frequencies of sequences in the corpus. Then, most frequently occurring sequences are examined to find formulas. Numerous studies have been made to identify formulas for several languages like English. There exists only few studies about formulaicity in Turkish and most of these studies focus on identifying formulas in the forms of multi word units. Turkish, however, is an agglutinating language having a rich and complex morphology, therefore formulaic sequences in affixation should be discovered. Only very limited

[1] This study was supported by TÜBİTAK (Grant no:113K039). We express our gratitudes to TÜBİTAK.

[2] Çukurova University, Department of Computer Engineering, Adana, Türkiye, saozel@cu.edu.tr

[3] Çukurova University, Department of Electrical and Electronics Engineering, Adana, Türkiye, ybektas79@gmail.com

[4] Çukurova University, Department of Computer Engineering, Adana, Türkiye, yilmazerhakan@gmail.com

Makale gönderim tarihi: 7 Mart 2016; Kabul tarihi: 30 Mayıs 2016

studies about formulaicity in affixation of Turkish exist in the literature. In this study, we try to discover formulaic sequences in affixation of Turkish by counting frequent suffix n-grams in written and spoken Turkish by using the Turkish National Corpus, which is a balanced, large scale, and general-purpose corpus for contemporary Turkish. We list the most frequent suffix combinations not only for verbs but also for all lexical categories like noun, adjective, verb, and adverb for both written and spoken corpora from Turkish National Corpus, and discuss similarities and differences in affixation in written and spoken usage of Turkish. We observe that, we prefer shorter suffix sequences in spoken Turkish than in written Turkish, and as the length of the suffix n-grams increase, we use different formulaic sequences in written and spoken Turkish.

Keywords: *Frequent suffix n-grams, written Turkish, spoken Turkish, Turkish National Corpus*

TÜRKÇE'DE KALIP ANLATIMLAR: TÜRKÇE ULUSAL DERLEMİ'NDEN GÖRÜNÜMLER

Öz: Kalıp anlatımlar yada sabit ifade dizileri (formüller) bir dilde en sık gözlenen biçimlerden oluşur. Dildeki formüllerin belirlenmesi; dilbilimi, yabancı dil öğrenimi, doğal dil işleme gibi pek çok alan için faydalıdır. Bir dildeki sabit ifade dizilerini belirleyebilmek için en çok tercih edilen yöntem bir derlem kullanmak ve derlemdeki dizilerin sayısını belirlemektir. Türkçe'deki formüller ile ilgili az sayıda çalışma bulunmaktadır. Bu çalışmada, Tükçe'de eklerde yer alan formül dizilerini Türkçe Ulusal Derlemi'ni kullanarak, yazılı ve sözlü Türkçe'de en sık görülen n'li biçimbirim dizilerinin sayısal dağılımını ortaya çıkarmaya çalışmaktayız. Tüm sözcük kategorileri için en sık ek kombinasyonları listelenmektedir. Sözlü Türkçe'de yazılı Türkçe'ye göre daha kısa ek dizilerinin tercih edildiği, n'li biçimbirim dizilerinin uzunlukları arttıkça yazılı ve sözlü Türkçe'de farklı formül dizilerinin kullanıldığı görülmektedir.

Anahtar sözcükler: *Sık n'li biçimbirim dizileri, yazılı Türkçe, sözlü Türkçe, Türkçe Ulusal Derlemi*

1. INTRODUCTION

A considerable proportion of natural languages contain formulaic sequences that are predictable, fixed, or semi-fixed chunks (Wray,

2002). Formulaic sequence is defined as "a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray & Perkins, 2000).

As the definition of formulaic sequences implies, identification of formulaic sequences in a language is useful for a wide range of areas including linguistics, second language learning, natural language processing, spell checking, POS tagging, and the like (e.g., Durrant, 2013; Güngör, 2003; Wray, 2008). Most of the studies about formulaic sequences identification belong to English language (Wray & Perkins, 2000; Wray, 2002; Wray, 2008; Biber, 2009; Biber, Conrad & Cortes, 2004; Durrant & Mathews-Aydınlı, 2011, among many others) in which formulaic sequences are in the form of multi words or word n-grams.

One of the earliest studies for identifying formulaic sequences in Turkish language belongs to Tannen and Öztek (1977) who lists formulas used by native Turkish speakers for certain contexts and compares these formulas with the ones used by native Greek speakers. As shown by Tannen and Öztek (1977), formulas can change when the language and culture of the speaker change and therefore formulaic sequences should be identified for different languages and contexts separately.

Identifying formulaic sequences for a language is not an easy task, however, according to Durrant & Mathews-Aydınlı (2011), formulas can be considered as "the most frequent recurrent forms in a relevant corpus" (Durrant & Mathews-Aydınlı, 2011). Therefore, it is possible to define formulas by counting high frequency linguistic combinations from a corpus (Biber, 2009; Biber, Conrad & Cortes, 2004; Durrant, 2013; Simpson-Vlach & Ellis, 2010, among many others). Doğançay (1990) lists formulaic expressions and routines in Turkish that are used by Turkish native speakers by using a dataset which is collected by tape-recordings of naturally occurring conversations. Formulaic sequences are identified by analyzing the conversations and listing the most frequent sequences, which consist of lexical n-grams, in the conversations (Doğançay, 1990).

Dalkılıç & Çebi (2004) identify most frequent lexical n-grams in Turkish by using a corpus which is collected from 12 different websites that include websites of newspapers, magazines, bookstores, etc., and has about 50 million words. Statistical analyzes for one, two, three, four and five-grams are made and it is observed that the most frequent n-grams are one and two-grams, the frequency of larger grams decreases exponentially as n increases which indicates that a word cannot be used in every position in a sentence of a natural language (Dalkılıç & Çebi, 2004).

When studies to identify formulaic sequences for Turkish language are analyzed, we have observed that the aim of these studies is to find formulaic sequences in the form of lexical n-grams. However, Turkish is an agglutinating language which has a rich and complex morphology. New words can be derived by attaching derivational suffixes to a root. Therefore, it is possible to find formulaic sequences within single word units.

Formualicity in affixation for Turkish has not been studied widely. Güngör (2003) gives some statistics about affixation like minimum, maximum, and average suffix length; minimum, maximum, and average number of suffixes in a word for a corpus compiled from several newspapers, periodicals, and a few novels. Durrant (2013) provides more detailed analysis for 20 most frequent verbs including syntagmatic association between inflectional suffixes, fixed sequences of suffixes, and associations between particular lexical and grammatical forms for a corpus collected from 7 online newspapers in Turkey. Apart from the existing studies, in this study, our aim is to show formulaicity in affixation for both written and spoken Turkish by using the Turkish National Corpus (TNC), which is a balanced, large scale, and general-purpose corpus for contemporary Turkish (Aksan et al., 2012). We list the most frequent suffix combinations not only for verbs but also for all lexical categories such as, noun, adjective and adverb for both written and spoken parts of the TNC; and then we discuss similarities and differences in the affixation in written and spoken usage of Turkish. We also compare suffix combinations on the basis of lexical categories.

2. TURKISH NATIONAL CORPUS (TNC)

TNC (Aksan et al., 2012) is a balanced, large scale, and general-purpose corpus for contemporary Turkish. It has approximately 51 million words and follows the framework of British National Corpus. To develop TNC open-source software are used for specific tasks; as an example PHP is used for graphical user interface implementation, MySQL database management system is used for storing text contents, Perl scripting language is employed for index construction and some statistical computations. TNC is a free resource for non-commercial use. To use the TNC, a user needs to have a username and password which are provided by the system administrator, then the user can ask queries about usage of words and suffixes over the corpus and the results as well as statistical analysis of the results are also given as the output.

The corpus covers Turkish documents from a period of 20 years (1990 - 2009). Balance of the corpus is achieved through a wide range of text categories it covers so that 98% of the TNC consists of written text and the remaining 2% consists of transcribed spoken data. Number of words in the TNC is computed as 50,086,419 for written corpus, and 998,383 for spoken corpus. Number of distinct words is found as 1,316,462 for written corpus, and 114,044 for spoken corpus. Number of word stems is 71,437 for written corpus, 13,429 for spoken corpus. Domains of the documents that form the corpus and their ratios are presented in Table 1. Numbers of documents for each genre are given in Table 2.

Table 1. Domain of the documents in TNC

Domain	Ratio
1. World Affairs	20.05 %
2. Imaginative	19.22 %
3. Leisure	14.96 %
4. Social Science	14.55 %
5. Commerce and Finance	9.21 %
6. Art	7.50 %
7. Applied Science	7.19%
8. Belief and Thought	4.31 %
9. Natural Science	2.96 %
TOTAL	100 %

Table 2. Number of documents in TNC

	Genre	# of documents
	Unprinted written text	471
	Scientific prose	2142
	Unscientific prose	774
Written Part	Other printed written text	520
	Newspaper	469
	Fiction & poem	678
	TOTAL # of documents	5058
Spoken Part	Conversation & other	457

3. SUFFIX N-GRAMS

In this study, first of all, we list the most frequent suffix n-grams by counting their frequencies in both written and spoken parts of the TNC. After that, we compute number of suffix n-grams for each lexical category like noun, adjective, verb, adverb, etc. Then, we discuss similarities and differences in affixation in written and spoken usage of Turkish.

An *n*-gram is a contiguous sequence of *n* items from a given sequence of text or speech. As an example, suffix n-grams for the word "*almıştı*" are listed in Table 3.

Example: *almıştı* *al* VB perf+vi+past+3s

Table 3. Suffix n-grams for word "*almıştı*"

1-grams	2-grams	3-grams	4-grams
perf	perf+vi	perf+vi+past	perf+vi+past+3s
vi	vi+past	vi+past+3s	
past	past+3s		
3s			

As shown in Table 3, the word "*almıştı*" has four suffixes and each suffix forms a 1-gram, therefore from the word, we can find four 1-grams. Any adjacent two suffixes form a 2-gram therefore there are three 2-grams. Consecutive three suffixes are called as 3-gram, as shown in the table we have two 3-grams, and finally as we have only four suffixes in the word "*almıştı*" we have only one 4-gram, and it is not possible to have longer suffix n-grams from the word. For all

distinct words extracted from the TNC, we list all n-grams, and count their frequencies in the corpus.

First of all, we compute all 1-grams in the TNC. As shown in Table 4, in the written part of TNC we have 86 different 1-grams; in the spoken part of the TNC we observe 77 different 1-grams. When we compare these two values, we can conclude that in spoken Turkish, we do not use some suffixes which are only used in written Turkish. In Table 4, numbers of the expected and the observed suffix n-grams are also presented. We compute the expected value for an n-gram, where n is greater than 1, by computing the number of all possible distinct n-gram sequences from the observed 1-gram values in the written and spoken parts of the TNC. In the observed column of the table, numbers of distinct n-grams that are observed in the corpus are presented. As shown in the table, from 86 different suffixes, it is possible to generate 86*86=7396 distinct 2-grams, if repetitions of the suffixes are allowed, however, we only observe 983 different suffix 2-grams in the written part of the corpus. This shows that suffixes are not attached arbitrarily; there is a formula to combine them and form a Turkish word. In short, the combination of suffixes follows the rules of morphological suffix ordering in Turkish. According to Table 4, for n values which are greater than or equal to 5, as n increases, the number of unique suffix n-grams decreases sharply. When written and spoken parts of the corpus are compared, we observe that in spoken Turkish, we prefer shorter suffix n-grams. In written part, we observe up to 9-grams and 4 and 5-grams have the highest number of distinct instances. In the spoken part on the other hand, the longest n-gram observed has length 8, and 3 and 4-grams have the highest number of distinct instances. The length of suffixes reflects the differences in writing and speking. This is because of the interactive, interpersonal and spontenous nature of spoken language which generally necessitates shorter forms to convey the message.

Table 4. Number of expected and observed suffix N-grams in the TNC

N	# of Distinct N-grams in Written Part		# of Distinct N-grams in Spoken Part	
	Expected	Observed	Expected	Observed
1	86	86	77	77
2	7,396	983	5,929	574
3	636,056	3240	456,533	1456
4	54,700,816	5837	35,153,041	1743
5	4,704,270,176	5887	2,706,784,157	1046

N	# of Distinct N-grams in Written Part		# of Distinct N-grams in Spoken Part	
	Expected	Observed	Expected	Observed
6	$404.6 * 10^9$	3227	$208.4 * 10^9$	347
7	$34.8 * 10^{12}$	1030	$16.048 * 10^{12}$	70
8	$2.99 * 10^{15}$	169	$1.235 * 10^{15}$	7
9	$2.57 * 10^{17}$	13	$9.51 * 10^{16}$	0

In Table 5, the most frequently observed 1-grams from the written and spoken parts of the TNC are listed. The rows in the table are sorted in descending order according to the observed frequency of the suffix in the written and spoken parts of the TNC. According to Table 5, the most frequent suffixes are the nominal suffix showing possession"p3s" and the third person agreement suffix "3s" for the written and spoken parts respectively. Only the top 12 most frequent suffixes are included in the table. In the Written column, suffixes are listed with respect to their observed frequencies in descending order for written Turkish. "Rank in S" column shows the rank in spoken Turkish of the frequent suffix of written Turkish. As an example "p3s" is the most frequent suffix in written Turkish, while it is the second most frequent suffix in the spoken part of the TNC. Similarly, most frequent suffixes for spoken Turkish are listed with respect to their observed frequencies in descending order under the Spoken column. "Rank in W" column under the Spoken column shows the rank of the frequent suffix of spoken Turkish in the written part of the corpus. The most frequent suffix for spoken Turkish is "3s" and its rank in written Turkish is 3 so it is the third most frequent suffix for written Turkish. As shown in the table, most of the top 12 suffixes are common both in the written and spoken parts, however their rank may change.

Table 5. Top 12 most frequent 1-grams for written & spoken Turkish

	Written			Spoken		
Row	Rank in S	Suffix	Frequency (%)	Suffix	Frequency (%)	Rank in W
1	2	p3s	6.13	3s	5.28	3
2	3	acc	4.70	p3s	4.88	1
3	1	3s	4.36	acc	4.19	2
4	7	pl	3.63	loc	3.26	5
5	4	loc	3.40	past	3.24	10
6	9	p2s	3.21	dat	2.94	7
7	6	dat	3.02	pl	2.88	4

	Written			Spoken		
Row	Rank in S	Suffix	Frequency (%)	Suffix	Frequency (%)	Rank in W
8	10	gen	2.67	imprf	2.65	20
9	19	pasv	2.06	p2s	2.34	6
10	5	past	2.01	gen	2.23	8
11	17	nzma	1.76	neg	1.87	17
12	24	p3p	1.67	1s	1.84	25

Similarly, Table 6 presents the ranks of the most frequent twelve 2-grams for both written and spoken parts of the TNC. In Table 6, it is also observed that most of the top 12 frequent suffix 2-grams are common in written and spoken Turkish, however their ranks are different. Examples for the most frequent 2-grams from the corpus are given below. For the written Turkish, examples for top three most frequent suffix 2-grams are as follows:

aldı	al	VB	past+3s
sırasında	sırada	PP	p3s+loc
yorumlar	yorumla	VB	aor+3s

For the spoken Turkish, examples for top three most frequent suffix 2-grams are as follows:

aldı	al	VB	past+3s
alıyor	al	VB	imprf+3s
gidiyorlardı	git	VB	imprf+3p+vi+past

Top twelve most frequent suffix 3-grams for written and spoken parts of the TNC are presented in Table 7. The observed results for 3-grams are also similar to 1 and 2-grams. Examples for frequent 3-grams from the corpus are given below. For written Turkish, examples for top three most frequent suffix 3-grams are as follows:

açıktı	açık	AJ	vi+past+3s
alınması	al	VB	pasv+nzma+p3s
aldığını	al	VB	pcdk+p3s+acc

For spoken Turkish, examples for top three most frequent suffix 3-grams are as follows:

aldı	al	AJ	vi+past+3s
alsa	al	AJ	vi+avsa+3s
alırsak	al	VB	aor+vi+avsa+1p

Table 6. Top 12 most frequent 2-grams for written & spoken Turkish

	Written			Spoken		
Row	Rank in S	Suffix	Frequency (%)	Suffix	Frequency (%)	Rank in W
1	1	past+3s	1.41	past+3s	1.69	1
2	5	p3s+loc	1.07	imprf+3s	1.16	16
3	4	aor+3s	0.90	vi+past	0.91	4
4	3	vi+past	0.88	aor+3s	0.88	3
5	7	p3s+acc	0.83	p3s+loc	0.82	2
6	13	pcdk+p3s	0.79	past+1s	0.81	27
7	9	pl+acc	0.77	p3s+acc	0.61	5
8	12	p2s+loc	0.69	perf+3s	0.60	16
9	14	p3s+dat	0.64	pl+acc	0.59	7
10	18	nzma+p3s	0.59	imprf+1s	0.58	54
11	15	p2s+acc	0.58	neg+imp2	0.53	17
12	26	pl+gen	0.51	p2s+loc	0.51	8

Table 7. Top 12 most frequent 3-grams for written & spoken Turkish

		Written		Spoken		
Row	Rank in S	Suffix	Freq. %	Suffix	Freq. %	Rank in W
1	1	vi+past+3s	0.68	vi+past+3s	0.586	1
2	8	pasv+nzma+p3s	0.24	vi+avsa+3s	0.231	9
3	11	pcdk+p3s+acc	0.20	aor+vi+avsa	0.169	20
4	12	pcdk+p2s+acc	0.20	vi+past+1s	0.159	17
5	5	imprf+vi+past	0.19	imprf+vi+past	0.155	5
6	10	perf+cop+3s	0.19	neg+aor+3s	0.144	13
7	27	cont+cop+3s	0.18	neg+imprf+3s	0.143	39
8	15	perf+vi+past	0.17	pasv+nzma+p3s	0.138	2
9	2	vi+avsa+3s	0.16	neg+imprf+1s	0.129	60
10	30	p3s+loc+kia	0.13	perf+cop+3s	0.124	6
11	29	pasv+pcdk+p3s	0.12	pcdk+p3s+acc	0.117	3
12	23	pasv+perf+3s	0.12	pcdk+p2s+acc	0.115	4

Table 8 and 9 present and give examples for the most frequently observed suffixes in 1, 2, and 3-grams for written and spoken Turkish, respectively.

Table 8. Most frequently observed suffixes (in 1, 2, 3-grams) in the written Turkish

TAG	Morpheme	Function	As in
pasv	l/n	Voice	salıver**il**ecek, izle**n**di
nzma	mA	Nominalizer	yüz**me**
p3p	lArI	Possessive	(onların) saç**ları**
p3s	I	Possessive	(onun) laf**ı**
pl	lAr	number/person	okul**lar**
gen	In	case-genitive	defter**in** (rengi)
cont	mAktA	TAM_continuous	git**mekte**dir
loc	DA	case-locative	defter**de**
pcdk	DIk	Nominalizer	git**tik**lerinden

Table 9. Most frequently observed suffixes (in 1, 2, 3-grams) in the spoken Turkish

TAG	Morpheme	Function	As in
imprf	yor	TAM_imperfective	gidi**yor**
neg	mA	Negative	git**me**dik
1s	(I)m	Person	geldi**m**, gidiyor**um**
past	DI	TAM_past / perfective	git**ti**
perf	mIş	TAM_referential/perfective	git**miş**
3s	Ø	Person	geliyor
imp2	Ø,sAnA AsIn, gıl	Imperative	gel, gel**sene**, gel**esin**,
aor	Ø, r, z	TAM_aorist	acımayız, uyu**r**sun, uyuma**z**
avsa	sA, A	Adverbial	git**se**, gid**e**ydi

Table 10. Top 12 most frequent 4-grams for written & spoken Turkish

	Written			Spoken		
	Rank in S	Suffix	Freq. %	Suffix	Freq. %	Rank in W
1	1	imprf+vi+past+3s	0.16	imprf+vi+past+3s	0.082	1
2	3	perf+vi+past+3s	0.13	aor+vi+avsa+3s	0.080	7
3	6	pasv+perf+cop+3s	0.07	perf+vi+past+3s	0.052	2
4	4	aor+vi+past+3s	0.06	aor+vi+past+3s	0.043	4
5	20	pasv+cont+cop+3s	0.06	imprf+vi+past+1s	0.041	9
6	7	caus+pasv+nzma+p3s	0.05	pasv+perf+cop+3s	0.036	3
7	2	aor+vi+avsa+3s	0.05	caus+pasv+nzma+p3s	0.031	6
8	18	pasv+val+aor+3s	0.04	perf+vi+past+1s	0.031	13
9	5	imprf+vi+past+1s	0.03	imprf+vi+perf+3s	0.030	45
10	13	val+neg+aor+3s	0.03	aor+vi+avsa+2p	0.027	55
11	25	neg+imprf+vi+past	0.02	aor+vi+past+1s	0.026	31
12	38	neg+pcdk+p3s+acc	0.02	aor+vi+avsa+2s	0.026	80

Table 10 lists most frequent suffix 4-grams for the written and spoken Turkish. Examples for the most frequent 4-grams can be given as follows:

For written Turkish:

araştırıyordu	ara	VB	recp+caus+imprf+vi+past+3s
almıştı	al	VB	perf+vi+past+3s
alınmıştır	al	VB	pasv+perf+cop+3s

For spoken Turkish:

alıyordu	al	VB	imprf+vi+past+3s
alırsa	al	VB	aor+vi+avsa+3s
almıştı	al	VB	perf+vi+past+3s

When the most frequently occurred 4-grams are analyzed, the most frequently used suffix in the 4-grams are copula “cop”, negative “neg” and accusative “acc” for the written part; and second person singular “2s” for the spoken part. Although “neg” is frequently used in the 4-grams of the written part, it is not frequent in the spoken part.

Table 11. Most frequently observed suffixes in 4-grams of written & spoken Turkish

TAG	Morpheme	Function	As in
cop	DIr	Copula	gitmekte**dir**
neg	mA	Negative	git**me**dik
acc	I	case-accusative	defter**i**
2s	sIn, In, n	Person	gel**sin**

From the most frequent 4-grams, it is also observed that majority of the 4-grams end with the same suffix. According to Figure 1, about 68% of the 4-grams in the written part end with "3s", however 50% of the 4-grams in the spoken part end with "3s". Percentage of third person singular suffix for both written and spoken parts are the same, however, the second person singular suffix "2p" is used only at the end of the 4-grams obtained from the spoken part. This is an expected finding since the speaker addresses the hearer in speech.

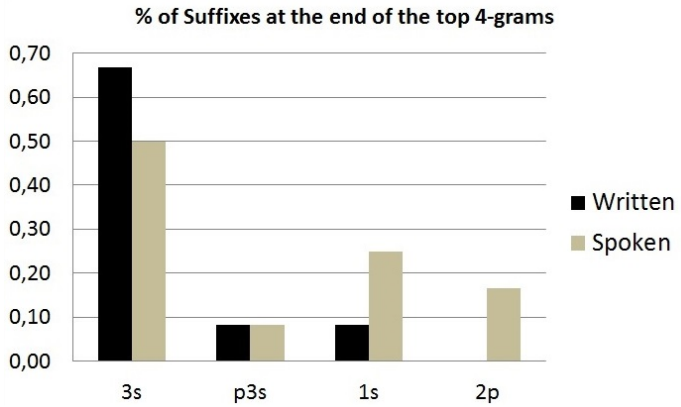

Figure 1. Percentage of suffixes observed at the end of the most frequent 4-grams

Most frequent suffix 5-grams for written and spoken parts of the TNC are presented in Table 12. As shown in the table, 5-grams occur very infrequently in spoken part as compared to the written part. Examples for frequent 5-grams are given below.

For written Turkish, three most frequent suffix 5-grams are:

alınmıştı	al	VB	pasv+perf+vi+past+3s
almıyordu	al	VB	neg+imprf+vi+past+3s
alınıyordu	al	VB	pasv+imprf+vi+past+3s

For spoken Turkish, three most frequent suffix 5-grams are:

almazsa	al	VB	neg+aor+vi+avsa+3s
almıyordu	al	VB	neg+imprf+vi+past+3s
alınmıştı	al	VB	pasv+perf+vi+past+3s

Table 12. Top 12 most frequent 5-grams for written & spoken Turkish

	Written			Spoken		
Row	Rank in S	Suffix	Freq %	Suffix	Freq %	Rank in W
1	3	pasv+perf+vi+past+3s	0.02	neg+aor+vi+avsa+3s	0.0145	11
2	2	neg+imprf+vi+past+3s	0.02	neg+imprf+vi+past+3s	0.0081	2
3	7	pasv+imprf+vi+past+3s	0.02	pasv+perf+vi+past+3s	0.0071	1
4	14	neg+perf+vi+past+3s	0.01	neg+imprf+vi+past+1s	0.0064	20
5	8	neg+aor+vi+past+3s	0.01	pasv+val+neg+aor+3s	0.0059	6
6	5	pasv+val+neg+aor+3s	0.01	pasv+aor+vi+avsa+3s	0.0054	8
7	10	caus+pasv+perf+cop+3s	0.01	pasv+imprf+vi+past+3s	0.0054	3
8	6	pasv+aor+vi+avsa+3s	0.01	neg+aor+vi+past+3s	0.0053	5
9	28	caus+imprf+vi+past+3s	0.01	p3s+loc+vi+past+3s	0.0052	13
10	30	caus+perf+vi+past+3s	0.01	caus+pasv+perf+cop+3s	0.0049	7
11	1	neg+aor+vi+avsa+3s	0.01	neg+aor+vi+past+1s	0.0043	34
12	18	val+aor+vi+past+3s	0.01	p2s+loc+vi+past+3s	0.0042	16

In Figure 2, a comparison of frequent 5-grams for written and spoken Turkish is given. As shown in Figure 2, all suffix 5-grams in written Turkish end with suffix “3s” whereas more than 80% of the suffix 5-grams in spoken Turkish end with “3s”, and the remaining suffix 5-grams end with “1s”. Another observation is that 67% of the 5-grams in written Turkish and 50% of the 5-grams in spoken Turkish end with “vi+past+3s”. 17% of the 5-grams in spoken Turkish end with “vi+past+1s”.

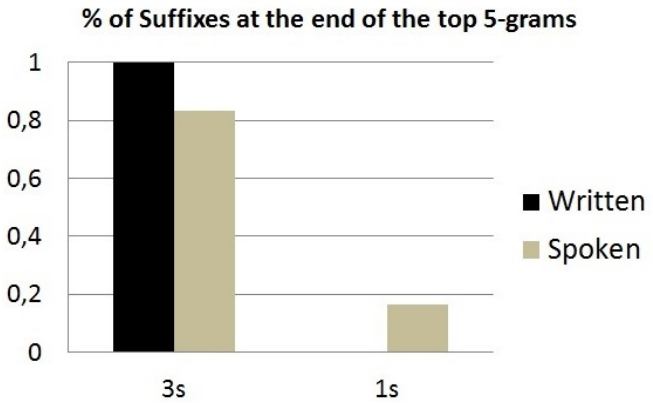

Figure 2. Percentage of suffixes observed at the end of the most frequent 5-grams

In Table 13, most frequent suffix 6-grams are listed for written and spoken Turkish. As we can easily guess, frequencies of suffix 6-grams decraeses for the spoken Turkish when compared to the written Turkish. Also, as n-increases frequencies of suffix n-grams also decrease sharply both for written and spoken Turkish. Examples for three most frequent suffix 6-grams are given as below:

For written Turkish:

alamıyordu	al	VB va1+neg+imprf+vi+past+3s
alamazdı	al	VB va1+neg+aor+vi+past+3s
alamamıştı	al	VB va1+neg+perf+vi+past+3s

For spoken Turkish :

alamazsa	al	VB va1+neg+aor+vi+avsa+3s
alamazdı	al	VB va1+neg+aor+vi+past+3s
alamıyordum	al	VB va1+neg+imprf+vi+past+1s

Table 13. Top 12 most frequent 6-grams for written & spoken Turkish

	Written			Spoken		
	Rank in S	Suffix	Freq %	Suffix	Freq %	Rank in W
1	4	val+neg+imprf+vi+past+3s	0.0041	val+neg+aor+vi+avsa+3s	0.00148	21
2	2	val+neg+aor+vi+past+3s	0.0039	val+neg+aor+vi+past+3s	0.00141	2
3	14	val+neg+perf+vi+past+3s	0.0023	val+neg+imprf+vi+past+1s	0.00141	5
4	16	caus+pasv+perf+vi+past+3s	0.0021	val+neg+imprf+vi+past+3s	0.00128	1
5	3	val+neg+imprf+vi+past+1s	0.0020	pasv+val+neg+perf+cop+3s	0.00094	6
6	5	pasv+val+neg+perf+cop+3s	0.0015	val+neg+aor+vi+past+1s	0.00094	13
7	23	pasv+val+aor+vi+past+3s	0.0015	pasv+val+neg+aor+vi+past	0.00081	17
8	10	pasv+neg+imprf+vi+past+3s	0.0014	pasv+neg+aor+vi+avsa+3s	0.00074	12
9	28	pasv+val+neg+pcck+p3s+acc	0.0014	caus+pasv+nzma+p3s+cop+3s	0.00067	14
10	32	pasv+val+neg+pcck+p2s+acc	0.0014	pasv+neg+imprf+vi+past+3s	0.00067	8
11	19	pasv+neg+aor+vi+past+3s	0.0014	val+neg+imprf+vi+avsa+3s	0.00060	31
12	8	pasv+neg+aor+vi+avsa+3s	0.0013	val+neg+aor+vi+past+1p	0.00054	33

In Figure 3, a comparison of frequent 6-grams for written and spoken Turkish is given. As shown in Figure 3, about 75% of suffix 6-grams in written Turkish end with suffix "3s" whereas more than 67% of the suffix 6-grams in spoken Turkish end with "3s", and the remaining suffix 5-grams end with "1s", "past", and "1p". None of the frequent suffix 6-grams in spoken Turkish end with "acc". Also 50% of the suffix 6-grams in written Turkish and 25% of the suffix 6-grams in spoken Turkish end with "vi+past+3s"; and 17% of the suffix 6-grams in spoken Turkish end with "vi+past+1s" as in suffix 5-grams.

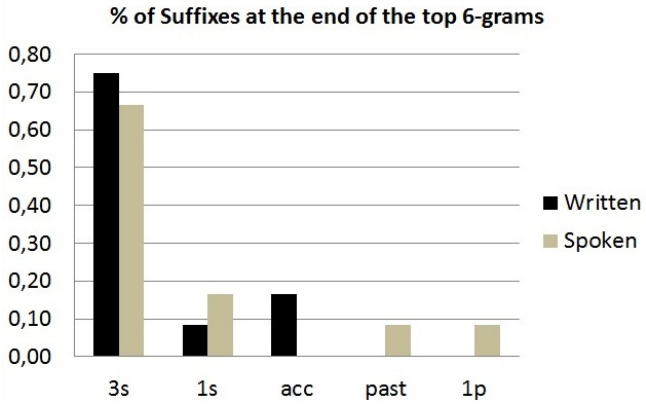

Figure 3. Percentage of suffixes observed at the end of the most frequent 6-grams

In Table 14, twelve most frequent suffix 7-grams are listed for written and spoken Turkish. As shown in the table, 50% of the most frequent suffix 7-grams observed in written Turkish is not used in spoken Turkish. There is only one suffix 7-gram "caus+va1+neg+past+vi+past+3s" which is used in spoken Turkish but not observed in written Turkish. Examples for frequent 7-grams are as below:

For written Turkish :

alınamazdı	al	VB	pasv+va1+neg+aor+vi+past+3s
alınamıyordu	al	VB	pasv+va1+neg+imprf+vi+past+3s
alıştıramıyordu	alış	VB	caus+va1+neg+imprf+vi+past+3s

Table 14. Top 12 most frequent 7-grams for written & spoken Turkish

	Written			Spoken		
	Rank in S	Suffix	Freq %	Suffix	Freq %	Rank in W
1	1	pasv+val+neg+aor+vi+past+3s	0.0010	pasv+val+neg+aor+vi+past+3s	0.00067	1
2	2	pasv+val+neg+imprf+vi+past+3s	0.0004	pasv+val+neg+imprf+vi+past+3s	0.00047	2
3	n/a	caus+val+neg+imprf+vi+past+3s	0.0003	pasv+val+neg+perf+vi+past+3s	0.00040	4
4	3	pasv+val+neg+perf+vi+past+3s	0.0003	caus+val+neg+aor+vi+avsa+1p	0.00020	55
5	8	pasv+val+neg+aor+vi+avsa+3s	0.0003	caus+pasv+val+aor+vi+avsa+3s	0.00020	20
6	56	caus+pasv+val+neg+perf+cop+3s	0.0002	recp+caus+pasv+val+neg+futr+3s	0.00013	25
7	n/a	caus+val+neg+aor+vi+past+3s	0.0002	pasv+val+neg+nzma+p3s+cop+3s	0.00013	18
8	n/a	pasv+val+neg+imprf+vi+avsa+3s	0.0002	pasv+val+neg+aor+vi+avsa+3s	0.00013	5
9	n/a	pasv+pasv+neg+aor+vi+past+3s	0.0002	caus+val+neg+imprf+vi+past+1s	0.00013	12
10	n/a	caus+val+neg+perf+vi+past+3s	0.0002	caus+pasv+nzma+p3s+vi+past+3s	0.00013	15
11	n/a	caus+pasv+val+aor+vi+past+3s	0.0002	caus+pasv+val+avsa+vi+past+3s	0.00013	90
12	9	caus+val+neg+imprf+vi+past+1s	0.0002	caus+val+neg+past+vi+past+3s	0.00007	n/a

For spoken Turkish :

alınamazdı	al	VB	pasv+va1+neg+aor+vi+past+3s
yapılamıyordu	yap	VB	pasv+va1+neg+imprf+vi+past+3s
bulunamamıştı	bul	VB	pasv+va1+neg+perf+vi+past+3s

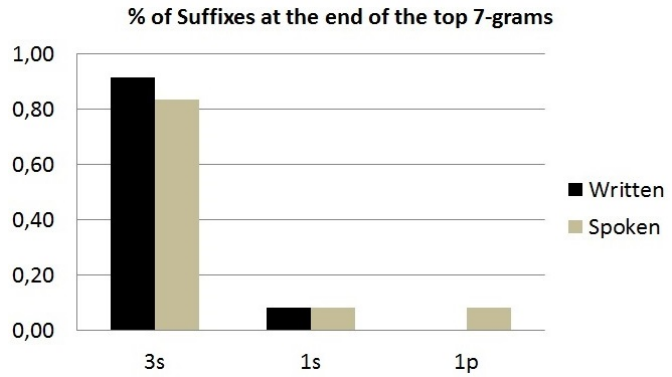

Figure 4. Percentage of suffixes observed at the end of the most frequent 7-grams

According to Figure 4, more than 90% of suffix 7-grams in written Turkish end with suffix "3s" whereas more than 82% of the suffix 7-grams in spoken Turkish end with "3s", and the remaining suffix 7-grams end with "1s" in written Turkish, "1s" and "1p" in spoken Turkish. None of the frequent suffix 7-grams in written Turkish end with "1p". Also 67% of the suffix 7-grams in written Turkish and 50% of the suffix 7-grams in spoken Turkish end with "vi+past+3s"; 50% of the suffix 7-grams in written Turkish end with "aor+vi+past+3s". "neg+past+vi+past+1s" only occurs in suffix 7-grams in spoken Turkish.

Table 15 lists the most frequent suffix 8-grams in written and spoken Turkish. As shown in Table 14, there are only seven suffix 8-grams for all spoken part of TNC, and two of these 8-grams only occur in the spoken part. For the written part, only one of the 8-grams occurs in the spoken Turkish, other most frequent eleven suffix 8-grams only observed in written Turkish. The frequencies of 8-grams are also very low with respect to shorter n-grams.

Table 15. Top 12 most frequent 8-grams for written & spoken Turkish

	Written			Spoken		
	Rank in S	Suffix	Freq %	Suffix	Freq %	Rank in W
1	n/a	caus+pasv+va1+neg+aor+vi+past+3s	5.6E-05	recp+caus+nzma+pl+loc+vi+avsa+3s	6.7E-05	n/a
2	n/a	caus+pasv+va1+neg+aor+vi+avsa+3s	4.9E-05	recp+pasv+va1+neg+aor+vi+avsa+3s	6.7E-05	47
3	n/a	caus+pasv+va1+neg+imprf+vi+past+3s	4.3E-05	recp+caus+pasv+neg+imprf+vi+past+3s	6.7E-05	27
4	n/a	caus+pasv+va1+neg+perf+vi+past+3s	3.9E-05	recp+pasv+va1+neg+imprf+vi+past+3s	6.7E-05	8
5	n/a	caus+pasv+va1+neg+imprf+vi+avsa+3s	1.9E-05	caus+caus+va1+neg+futr+vi+perf+1p	6.7E-05	n/a
6	n/a	pasv+va1+neg+nzma+p3s+vi+past+3s	1.9E-05	caus+caus+va1+neg+imprf+vi+past+1s	6.7E-05	16
7	n/a	recp+pasv+va1+neg+perf+vi+past+3s	1.6E-05	caus+caus+va1+neg+aor+3s+vi+avsa	6.7E-05	n/a
8	4	recp+pasv+va1+neg+imprf+vi+past+3s	1.4E-05			
9	n/a	recp+pasv+va1+neg+aor+vi+past+3s	1.3E-05			
10	n/a	pasv+pasv+va1+neg+aor+vi+past+3s	1.3E-05			
11	n/a	caus+pasv+va1+neg+nzma+p3s+cop+3s	1.3E-05			
12	n/a	caus+pasv+va1+neg+futr+vi+past+3s	9.4E-06			

Examples for frequent suffix 8-grams are listed below:

For written Turkish:

anlatılamazdı	anla	VB	caus+pasv+va1+neg+aor+vi+past+3s
anlatılamazsa	anla	VB	caus+pasv+va1+neg+aor+vi+avsa+3s
çıkartılamıyordu	çık	VB	caus+caus+pasv+va1+neg+imprf+vi +past+3s

For spoken Turkish:

karşılaştırmalardaysa	karşıla	VB recp+caus+nzma+pl+loc+vi +avsa+3s
uyuşulamazsa	uy	VB recp+pasv+va1+neg+aor+vi +avsa+3s
görüştürülmüyordu	gör	VB recp+caus+pasv+neg+imprf+vi +past+3s

In Figure 5 and 6, suffixes observed at the end and at the beginning of the most frequent suffix 8-grams are displayed.

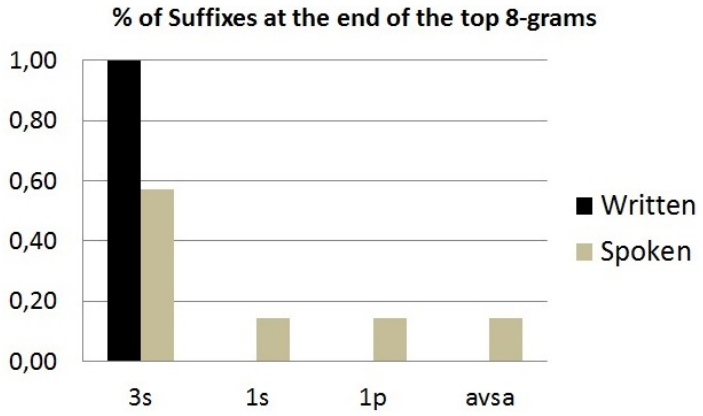

Figure 5. Percentage of suffixes observed at the end of the most frequent 8-grams

As shown in Figure 5, all suffix 8-grams end with “3s” in written Turkish, in spoken Turkish, 8-grams end with “3s”, “1s”, “1p”, and “avsa”. Also, 75% of the 8-grams ends with “vi+past+3s” in written Turkish. 8-grams start with “caus” and “recp” in written Turkish, they start with “caus”, “recp”, and “pasv” in spoken Turkish. When 8-grams in Table 15 are examined, all of them include “pasv+val+neg” in the written Turkish; however in the spoken Turkish suffix “neg” has different suffix combinations.

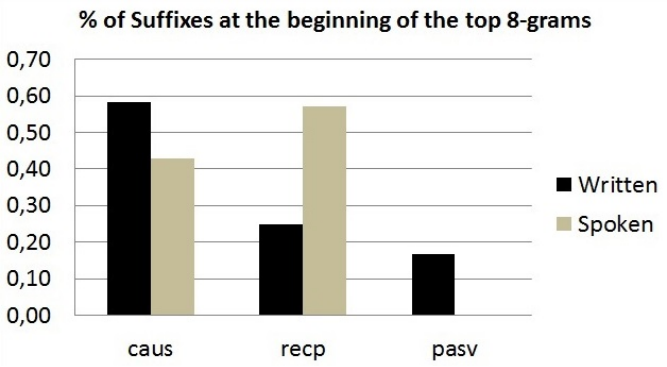

Figure 6. Percentage of suffixes observed at the beginning of the most frequent 8-grams

Table 16. Top 12 most frequent 9-grams for written Turkish

Written Turkish		
Rank	Suffix	Freq. %
1	recp+caus+pasv+va1+neg+aor+vi+past+3s	5.2E-06
2	recp+pasv+va1+neg+nzma+p3s+vi+past+3s	2.1E-06
3	caus+caus+pasv+va1+neg+aor+vi+past+3s	2.1E-06
4	caus+caus+pasv+neg+nzma+p3s+vi+past+3s	1.0E-06
5	caus+caus+pasv+va1+neg+imprf+vi+past+3s	1.0E-06
6	recp+caus+pasv+va2+neg+perf+vi+past+3s	1.0E-06
7	recp+caus+pasv+va1+neg+nzma+p3s+cop+3s	1.0E-06
8	caus+caus+va1+va1+neg+aor+vi+perf+3s	1.0E-06
9	pasv+va1+neg+pcan+pl+abl+vi+past+3s	1.0E-06
10	caus+caus+pasv+va1+neg+imprf+vi+avsa+3s	1.0E-06
11	caus+caus+pasv+va1+dsub+aor+vi+past+3s	1.0E-06
12	caus+pasv+va1+neg+pcck+p3s+vi+past+3s	1.0E-06

Table 16 lists the most frequent twelve suffix 9-grams in written Turkish. We do not observe any suffix 9-gram for the spoken Turkish. Examples for frequent 9-grams are given below:

karşılaştırılamazdı	karşıla	VB	recp+caus+pasv+va1+neg+aor+vi +past+3s
anlaşılamamasıydı	anla	VB	recp+pasv+va1+neg+nzma+p3s +vi+past+3s
çıkartılamazdı	çık	VB	caus+caus+pasv+va1+neg+aor +vi+past+3s

When frequent suffix 9-grams from written Turkish are examined, we observe that all 9-grams end with "3s"; 58.3% of the 9-grams begin with "caus"; 33.3 % of the 9-grams begin with "recp"; 8.4% of the 9-grams begin with "pasv"; and 9-grams occur very infrequently in the corpus.

3.1. SUMMARY

Table 17 lists the percent frequencies of all n-grams that are observed in the written and spoken part of the TNC. As an example 57.27% of all n-grams observed in the written Turkish are 1-grams; 26.57% of them are 2-grams. As the number of n increases, occurrence of an n-gram decreases rapidly for both written and spoken Turkish. 1-grams in spoken Turkish are more frequent than 1-grams in written Turkish. As shown in Table 17, we prefer shorter n-grams in spoken Turkish than written Turkish.

Table 17. Comparison of frequency of N-grams for all N values

N	Written Turkish (%)	Spoken Turkish (%)
1	57.27	62.15
2	26.57	27.62
3	8.05	7.69
4	2.50	2.09
5	5.51	0.39
6	0.08	0.05
7	0.01	0.007
8	0.0006	0.00047
9	0.00002	0

Table 18. Comparison of top-12 N-grams for written & spoken Turkish

N-gram	Both %	Written only %	Spoken only %
1	75 %	-	-
2	58.3 %	-	-
3	58.3 %	-	-
4	58.3 %	-	-
5	66.7 %	-	-
6	50 %	-	-
7	41.7 %	50%	8.3 %
8	8.3 %	91.7 %	43 %
9	-	100 %	-

In Table 18, we compare the most frequent twelve n-grams in written and spoken parts of TNC. According to Table 18, 75% of the top twelve 1-grams are common in both written and spoken parts, the remaining 25% are also observed in both parts but their ranks are lower. Up to 7-grams, all frequent shorter n-grams are observed both in written and spoken Turkish. However, some of the 7 and longer n-grams occur only in written or in spoken Turkish. As an example only 8.3% of the most frequent twelve 8-grams occurred in written corpus also observed in spoken corpus; the remaining 91.7% are specific to written part. 57% of the 8-grams occurred in spoken part also occur in the written part; however 43% of them are specific to spoken Turkish.

3.2. LEXICAL CATEGORIES AND DISTINCT SUFFIX N-GRAMS

Figures 7 – 12 show numbers of distinct 1, 2, 3, 4, 5, and 6-grams respectively, for all lexical categories in written and spoken Turkish.

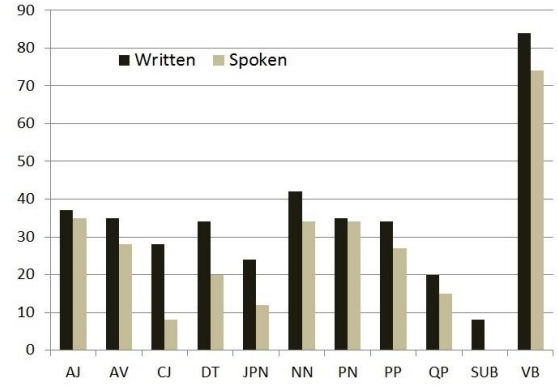

Figure 7. Number of distinct 1-grams for lexical categories

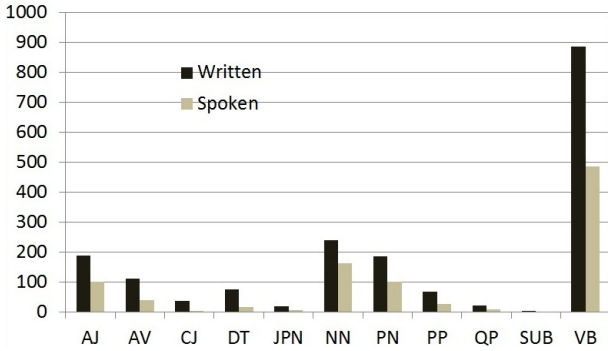

Figure 8. Number of distinct 2-grams for lexical categories

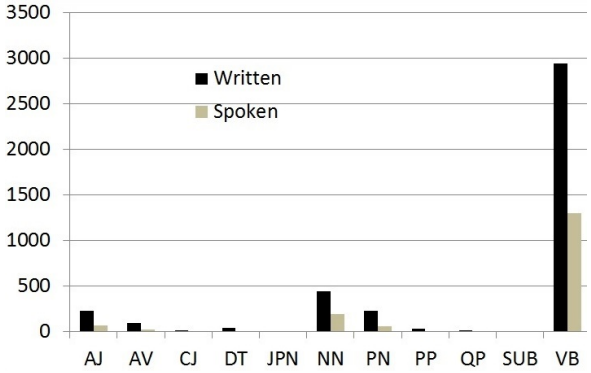

Figure 9. Number of distinct 3-grams for lexical categories

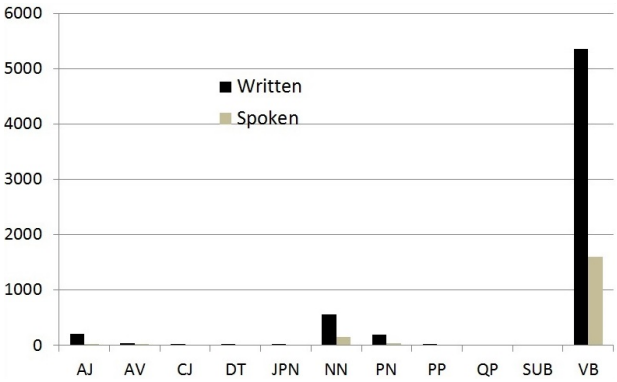

Figure 10. Number of distinct 4-grams for lexical categories

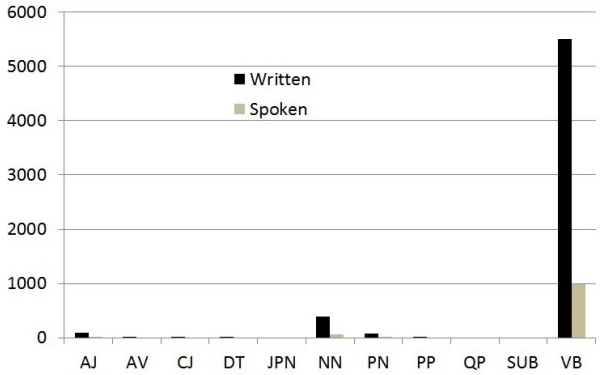

Figure 11. Number of distinct 5-grams for lexical categories

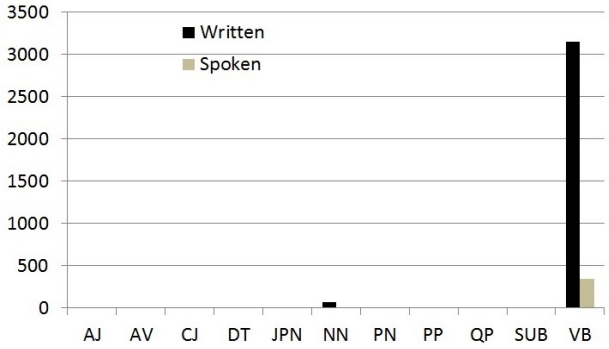

Figure 12. Number of distinct 6-grams for lexical categories

Table 19 shows number of distinct 7, 8, and 9-grams for all lexical categories in written and spoken Turkish. As shown in Figures 7 – 12 and Table 19, as n increases suffix n-grams are attached to VB lexical category. AJ, NN and PN are the other three lexical categories which can take n-grams up to 8-grams.

Table 19. Number of distinct 7, 8, 9-grams for lexical categories

	7-gram		8-gram		9-gram	
Lex. Categ.	Written	Spoken	Written	Spoken	Written	Spoken
AJ	2	-	1	-	-	-
AV	-	-	-	-	-	-
CJ	-	-	-	-	-	-
DT	-	-	-	-	-	-
JPN	-	-	-	-	-	-
NN	5	-	2	-	-	-
PN	1	-	-	-	-	-
PP	-	-	-	-	-	-
QP	-	-	-	-	-	-
SUB	-	-	-	-	-	-
VB	1022	347	166	7	13	-

Table 20. Percentage distribution of N-grams for each lexical category in written Turkish

N	AJ	AV	CJ	DT	JPN	NN	PN	PP	QP	SUB	VB
1	2.1	0.26	0.0037	0.019	0.0327	44.9	2.5	0.18	0.14	0.0001	50.17
2	1.54	0.15	0.0039	0.017	0.0045	33.9	1.2	0.21	0.13	0.0001	62.8
3	1.29	0.067	3.28E-03	0.014	5.57E-03	10.2	0.4	0.12	0.15	1.36E-04	87.7
4	0.29	7.5E-03	7.9E-05	2.1E-03	5.16E-04	4.6	0.14	1.6E-03	0	0	95
5	7.9E-03	2.4E-04	6.6E-06	3.3E-05	0	0.17	1.97E-03	1.97E-05	0	0	99.8
6	5.8E-03	0	0	0	0	2.6	0.012	0	0	0	97.4
7	0.022	0	0	0	0	0.065	0.01	0	0	0	99.9
8	0.18	0	0	0	0	0.35	0	0	0	0	99.5
9	0	0	0	0	0	0	0	0	0	0	100

Table 21. Percentage distribution of N-grams for each lexical category in spoken Turkish

N	AJ	AV	CJ	DT	JPN	NN	PN	PP	QP	SUB	VB
1	2.04	0.29	0.013	0.021	0.027	37.5	4.8	0.12	0.36	0	54.7
2	1.92	0.19	0.005	0.021	0.003	27.4	1.6	0.14	0.26	0	68.4
3	2.29	0.13	0.007	0.015	0.004	11.7	0.5	0.08	0.35	0	84.9
4	0.26	0.01	0	0	0	5.7	0.2	0	0	0	93.8
5	0.14	0	0	0	0	6.2	0.09	0	0	0	93.6
6	0	0	0	0	0	0	0	0	0	0	100
7	0	0	0	0	0	0	0	0	0	0	100
8	0	0	0	0	0	0	0	0	0	0	100

Tables 20 and 21 show the percentage distribution of n-grams for all lexical categories in written and spoken parts of TNC. According to Table 20, 50.17% of all 1-grams are used with VB, 44.9% of them are attached to NN, and the remaining 1-grams are used with other lexical categories. This distribution is also similar for spoken Turkish. As n increases, suffix n-grams are attached to VB lexical category for both in written and spoken Turkish.

4. CONCLUSIONS

In this study, we document suffix n-grams and conduct a quantitative analysis on suffixes that are used in all lexical categories of the written and spoken parts of the TNC. For the analysis, we generate all suffix n-grams that are observed in the TNC, and count their frequencies to identify the formulaic sequences in affixation of written and spoken Turkish. As a result of this analysis, we observed that maximum number of affixation is equal to 9 for written Turkish, and 8 for spoken Turkish. Maximum number of distinct suffix n-grams is observed from suffix 5-grams for written Turkish, and suffix 4-grams for spoken Turkish. When frequencies of all suffix n-grams are counted, as it is expected 1 and 2-grams are the most frequent n-grams, and as n increases, observed frequency of n-grams decreases sharply. The ratio of frequencies of suffix 1 and 2-grams to all n-grams are higher in spoken Turkish than in written Turkish. This ratio is similar for 3 and 4-grams in spoken and written Turkish, however, the frequency ratio for suffix 5, 6, 7, 8, and 9-grams are higher in written Turkish than in spoken Turkish. This can be emerged from the structurally complex and elaborate nature of writing which necessitates higher number of suffix n-grams to express the content of the message. Also writing permits a wide range of linguistic expressions which lead to the use of various longer n-grams. On the other hand, “speech is highly constrained in its typical linguistic characteristics” (Biber & Conrad, 2009, p. 261) so we can conclude that shorter affixations are preferred in spoken Turkish. When the content of the suffix n-grams is analyzed, we also found out that 4 and longer n-grams end with 3s, p3s, 1s, 2p, 1p which are all person agreement suffixes. Among the person suffixes, 1p and 2p are used in spoken Turkish, whereas 3s is more frequently used in written Turkish due to the differences in purpose, interactiveness and author involment of written and spoken register (Biber & Conrad, 2009). 8 and 9-grams

start with causative, recproical, passive which are all voice suffixes and they are the suffixes which considerably increase the number of suffix n-grams. However, it is found out that passive is used only in written Turkish. As also Paltridge (2006) maintains in informal speech the occurrence of passive construction is hardly observed. Non-attribution of agency provided by passive construction is typical for written languge. Up to 7-grams, all most frequent first twelve n-grams are common in both written and spoken Turkish, only their rankings are different. When longer suffix n-grams are compared, we found that the observed 7, 8, and 9-grams are different in written and spoken Turkish which have their own specific affixations. For further research the corpus-driven and comprehensive findings of this study can be used to describe the differences and similarities in written and spoken registers of Turkish.

REFERENCES

Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T., Mersinli, Ü., Demirhan, U. U., Yılmazer, H., Kurtoğlu, Ö., Atasoy, G., Öz, S., & Yıldız, İ. (2012). Construction of the Turkish National Corpus (TNC). In N. Calzolari, K. Choukri, T. Declerck et al. (Eds.), *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)* (pp. 3223-3227). İstanbul, Turkey: LREC 2012.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing, *International Journal of Corpus Linguistics*, 14, 275–311.

Biber, D., & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Cortes, V. (2004) If You Look at ... : Lexical Bundles in University Teaching and Textbooks, *Applied Linguistics*, 25 (3), 371–405.

Doğançay, S. (1990). Your eye is sparkling: Formulaic expressions and routines in Turkish, *Penn Working Papers in Educational Linguistics*, 6 (2), 49-64.

Dalkılıç, G., & Çebi, Y. (2004). Word statistics of Turkish language on a large scale text corpus – TurCo, *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04)*.

Durrant, P. (2013). Formulaicity in an agglutinating language: the case of Turkish, *Corpus Linguistics and Linguistic Theory*, 2013, 9 (1): 1-38.

Durrant, P, & Mathews-Aydınlı, J. (2011). A function-first approach to identifying formulaic language in academic writing, *Journal of English for Specific Purposes*, 30(1), 58–72.

Güngör, T. (2003). Lexical and morphological statistics for Turkish. Retrieved from https://www.cmpe.boun.edu.tr/~gungort/papers/Lexical%20and%20Morphological%20Statistics%20for%20Turkish.doc

Paltridge, B. (2006). *Discourse Analysis*. London: Continuum.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formula list: New methods in phraseology research, *Applied Linguistics*, 31 (4), 487-512.

Tannen, D., & Öztek, P. C. (1977). Health to our mouths: Formulaic expressions in Turkish and Greek author(s), *Proceedings of the 3rd Annual Meeting of the Berkeley Linguistic Society* (pp. 516-534).

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2008). *Formulaic language: pushing the boundaries*. Oxford: Oxford University Press.

Wray, A., & Perkins, M. R. (2000). The functions of a formulaic language: an integrated model. *Language & Communication*, 20, 1-28.

MORPHOLOGICAL ABBREVIATIONS

Abb.	Morpheme	Function	Example
1p	(I)k, (I)z	person	geldi**k**, gelmiş**iz**
1s	(I)m	person	geldi**m**, gidiyor**um**
2s	sIn, In, n	person	gel**sin**
3p	lAr	person	gidenleriydi**ler**, gitti**ler**
3s	Ø	person	geliyor
abl	DAn	case-ablative	defter**den**
acc	I	case-accusative	defter**i**
aor	r, z	TAM_aorist	uyu**r**sun, uyuma**z**
avca	cA	adverbial	çocuk**ça**, doğru**ca**
avip	Ip	adverbial	gelmey**ip**
avken	ken	adverbial	gider**ken**, gider**kene**
avmdn	mAdAn	adverbial	gel**meden** önce
avnce	IncA	adverbial	yaz**ınca**
avrek	ArAk	adverbial	yaz**arak**
avsa	sA, A	adverbial	git**se**, gid**e**ydi
c1s	Im	person_copula	nöbetçiy**im**
caus	t, Dır	voice	uyu**t**tu, yap**tır**dı
cont	mAktA	TAM_continuous	git**mekte**dir
cop	DIr	copula	gitmekte**dir**
dat	A	case-dative	defter**e**
futr	AcAk	TAM_future	gid**ecek**, gid**ecek**lerden
gen	In	case-genitive	defter**in** rengi
imp2	Ø, sAnA	imperative	gel, gel**sene**
imp3	sIn	imperative	gel**sin**
imp5	sAnIzA, In, InIz	imperative	gel**senize**, gid**in**, gid**iniz**
imprf	yor	TAM_imperfective	gidi**yor**
ins	ile	case-instrumental	defter**le**
kia	ki	adjectival	masada**ki**
loc	DA	case-locative	defter**de**
neg	mA	negative	git**me**dik
nom	Ø	case-nominative	masa
nzma	mA	nominalizer	yüz**me**
nzmk	mAk	nominalizer	uyu**mak**
p1p	mIz	possessive	andı**mız**

p1s	m	possessive	arı**m**
p2p	nIz	possessive	duaları**nız**
p2s	n	possessive	saçı**n**, başı**n**
p3p	lArI	possessive	onların saç**ları**
p3s	I	possessive	onun laf**ı**
past	DI	TAM_past / perfective	git**ti**
pasv	l/n	voice	salıver**il**ecek, izle**n**di
pcan	An	adjectival	gid**en**ler
pcck	AcAk	nominalizer	gid**eceğ**inden
pcdk	DIk	nominalizer	git**tik**lerinden
perf	mIş	TAM_evidentiality/perfective	git**miş**
pl	lAr	number/person	okul**lar**
recp	(I)ş	voice	döv**üş**tüler
va1	A, Abil	auxiliary verb	gel**e**mez, gel**ebil**ir
va2	ver	auxiliary verb	yapı**ver**di
Vi	i	Verb	gitti**y**se