Mersin Üniversitesi Dil ve Edebiyat Dergisi, MEUDED, 2016; 13 (2), 53-70.

VERB SYNTHESIS AND FREQUENCY

Mustafa AKSAN[1], Devrim ALICI[2], Umut Ufuk DEMİRHAN[3]

Mersin University

Abstract: Successive affixation in agglutinative languages derives complex structures. This study introduces frequency information of affix sequences in verbal domain from a corpus data. Recurent patterns of "morphgrams" formed by combinations of voice suffixes from non-finite template with other verbal inflections from finite template are extracted from the corpus. Starting with the simple two-morphgram patterns to the most complex nine-morphgrams, affix sequences are cited in the corpora. Samples indicate that Turkish do not derive monsterous words but rather limits the number of affixes that may be attacehed to a verb root or stem. Various statistical calculations also indicated the significance of grammatical patterns of affixes. The method and findings of the study have implications for morphological processing in agglutinative languages.

Keywords: *Frequency, affix order, verbal synthesis, Turkish National Corpus*

[1] Mersin University, Faculty of Science and Letters, Department of English Language and Literature, mustaksan@gmail.com

2 Mersin University, Faculty of Education, Department of Educational Sciences, devrimo.alici@gmail.com

[3] Mersin University, Faculty of Science and Letters, Department of English Language and Literature, umutufuk@gmail.com

Makale gönderim tarihi: 15 Ocak 2016; Kabul tarihi: 3 Haziran 2016

EYLEMCİL BİREŞİM VE SIKLIK

Öz: Sondan eklemeli dillerde, birbiri ardına eklenebilen çok sayıda biçimbirim son derece karmaşık diziler oluşturmaktadır. Derlem-çıkışlı bu çalışma derlem verisinde saptanan biçimbirim dizilerinin sıklık temelli sayısal bilgisini betimlemektedir. Dengeli ve temsil yeterliği olan Türkçe Ulusal Derlemi verisinde açıklaması yapılmış metinler taranarak eylemler üzerine eklenebilen çekim ve türetim ulamlarının sık ve birlikte kullanımı ile oluşan biçimbirim dizileri saptanmış ve bunların sayısal dağılımları hesaplanmıştır. Yalın iki biçimbirimli dizilerden karmaşık dokuz biçimbirimli dizilere kadar oluşan gözlenen sıklık değerleri yüksek diziler içyapılarını oluşturan ögeler ve sayısal dağılımları ile sıralanmıştır. Saptanan ve sıklık verisi hesaplanan dizilerin oluşturdukları sıralamanın istatistiksel değerlendirmesi ayrıca yapılmıştır. Çalışma daha sonra yapılacak derlem-çıkışlı betimlemeler için yöntem ve sayısal verinin anlamlıklarının hesaplanmasında yol gösterici olacaktır.

Anahtar sözcükler: *Sıklık, ek dizilimi, eylemcil bireşim, Türkçe Ulusal Derlemi*

1. INTRODUCTION

In their investigation on inflectional categories from a typological perspective, Bickel & Nichols (2013) define a synthetic construction as follows: "Grammatical categories like tense, voice, or agreement can be expressed either by individual words or by affixes attached to some other word (or the stem of a word). If a word combines with affixes, the resulting construction is said to be synthetic; if not, it is said to be analytic". Within categories that may be found in languages, they list the "prime" inflectional categories as agreement, tense/aspect/mood, evidentials, status, polarity, illocution, and voice. The measure they propose to account for typological status of a language is category-per-word ("cpw" value"). The cpw value of a maximally inflected verb in English would be 2 as it expresses agreement and tense.[4] The typological investigation concludes that Vietnamese is identified as a language with 0 cpw, lacking any inflectional category in a verb, and on the other extreme, there is Koasati with a cpw value of 13 categories that may be found in an inflected verb. In the categorization

[4] Göksel (1998) observes that in Turkish an inflected verb is marked minimally for tense and agreement.

of languages based on their cpw values, Turkish is listed among 31 others with 6-7 values within a total of 145 languages investigated. Bickel & Nichols (2013) note that universally, the most common type of languages are those with 4-8 cpw values.

This paper aims to address issues of (i) the order of inflectional affixes that are concatenated on a verb root, and (ii) observed frequencies of these verbal inflection affixes, particularly, voice categories. Studies on affixation of inflectional categories have uncovered grammatical aspects of forms and their orderings. We believe that corpus data in morphological analyses provide better understanding of affixes and their concatenation via quantificational information (Stubbs, 2013). Furthermore, such analyses will provide a new outlook to "lexical item" (Sinclair, 1998) as they are formed in agglutinative languages.

2. METHODOLOGY

2.1. THE CORPUS

Observed frequencies of affixes are extracted from the written part of the Turkish National Corpus (TNC). The size of TNC is 50,997,016 running words, incorporating a wide range of text categories covering a period of 23 years (1990-2013). TNC consists of samples from textual data (98%) and transcribed spoken data (2%).

To achieve representativeness and balance, the samples in the corpus are distributed for each text domain, time, and medium (Aksan et al., 2012). Table 1 and 2 show the distribution of texts in the written part of the TNC across domain and medium, respectively.

Table 1. The distribution of texts according to domains in the TNC

Domain	No. of words	% of words
Imaginative: Prose	9,365,775	18,74 %
Informative: Natural and pure sciences	1,367,213	274 %
Informative: Applied science	3,464,557	6,93 %
Informative: Social science	7,151,622	1431 %
Informative: World affairs	9,840,241	19,69 %
Informative: Commerce and finance	4,513,233	9,03 %
Informative: Arts	3,659,025	7,32 %

Domain	No. of words	% of words
Informative: Belief and thought	2,200,019	4,4 %
Informative: Leisure	8,421,603	16,85 %
Total	49,983,288	100,00 %

Table 2. The distribution of texts across mediums in the TNC

Medium	No. of words	% of words
Unspecified	10.541	0.02 %
Book	31.456.426	62.93 %
Periodical	15.968.240	31.95 %
Miscellaneous: published	958.999	1.92 %
Miscellaneous: unpublished	1.589.082	3.18 %
Total	49.983.288	100.00 %

The written texts are selected in accordance with the criteria of text domain, medium, and time. Here, domain refers to selection of texts according to imaginative and informative types. In the imaginative domain, texts are representatives of fiction; the informative domain is represented by texts from the social sciences, arts, commerce-finance, belief-thought, world affairs, applied sciences, natural-pure sciences, and leisure. The criterion of medium concerns text production where texts are collected to represent the written medium are selected from books, periodicals, published or unpublished documents and texts written-to-be-spoken such as news broadcasts and screenplays, among others.

2.2. ANNOTATION AND DATA PROCESSING

The texts in the corpus are analyzed and tagged by the TNC-tagger to calculate observed frequencies of verbal suffixes in Turkish. The lists are based on lemmas and morphological tags. An NLP dictionany is created by NooJ_TR module to annotate part-of-speech, and the output of the module covers morphological tagging and lemmatization information of the words in TNC (Aksan & Mersinli, 2011). The graph-based finite-state transducer of NooJ_TR module annotated the morphemes adopting a root-driven, non-stochastic rule-based apporoach. After the semi-automatic processing, the output is checked manually to eliminate artificial/non-occurring ambiguities. Following identification of non-canonical spellings and revision of tagged items, the NLP dictionary and its entries are matched via the PHP and MySQL-based interface of the corpus (Aksan et al., 2016).

Other than very few occasions, the order of affixes in Turkish is rigid, allowing little or no alternations. Thus, the order is predictable and morpheme boundaries are clear-cut due to conditioned phonological shape of allomorphs. However, the problem for any form of processing of morpheme sequences is the existence of a number of homograhic morphemes or homographic sequences derived in lemma+suffix combinations, suffix+suffix combinations and also of homographic lemmas. The end result is a total of ambiguous tags that relate about 15% of the TNC tokens.

2.3. CORPUS-DRIVEN PATTERN ANALYSIS

By now, it is customary to distinguish between pre-and post-corpus studies on patterns and within corpus linguistic approaches, to distinguish between corpus-based and corpus-driven studies. Gray & Biber (2015, p. 126) summarize major differences in corpus studies on phraseology as follows:

Table 3. Design parameters of corpus-based and corpus-driven phraseology

A. Research goals	B. Nature of multi-word units
Scope and methodological approach	*Idiomatic status*
1. explore the use of pre-selected lexical expressions (corpus-based approach) vs.	1. fixed idiomatic expressions vs.
2. identify and describe the full set of multi- word sequences in a corpus (corpus- driven approach)	2. non-idiomatic sequences that are very frequent
Role of register	*Length*
3. comparisons of phraseological patterns across registers vs.	3. relatively short combinations: 2–3 words vs.
4. focus on patterns in a single register vs.	4. extended multi-word sequences: 3+ words
5. focus on general corpora with no consideration of register	
Discourse function	*Continuous/discontinuous*
6. consideration of discourse functions vs.	5. continuous (uninterrupted) sequences vs.
7. no consideration of discourse functions	6. discontinuous sequences with variable “slots”

For the purposes of this study, the corpus-driven investigation of multiword patterns and multimorpheme sequences are not different (Durrant, 2013). Adopting a corpus-driven approach to recurrent patterns, this study will confine itself to the identification of these frequent units, following a similar research goal noted in (2) above. Furthermore, given the limits of space here, the role of register is also discarded and as in (5), by focusing on general corpora. Finally, functions of such units in a discourse (7) will also be left aside other than occasional references.

3. VERBAL INFLECTION IN TURKISH

Works on inflectional suffixes in Turkish distinguish finite and non-finite inflectional categories (Sezer, 2001; Enç, 2004). In the verbal domain, inflectional affixes are analyzed into two major templates (Göksel & Kerslake, 2005).[5] For each affix, there is a specific slot in both templates which ultimately dictates their grammatical orderings.

Table 4. Finite verb template

(1)-(y)A	(2) -(y)Abil -(y)Iver -(y)Agel -(y)Akal -(y)Adur	(3) -DI -mIş -(A/I)r/-z -(y)AcAK -(I)yor -mAlI -mAktA -(y)A	(4) -(y)DI -(y)mIş -(y)sA	(5)-DIr

Agreement is the final category that is marked in the order and there are four different groups or "paradigms". Selection of a particular agreement marker from any of these four paradigms is determined in the order by the affix representing tense/aspect/modality position preceding.[6]

[5] Here, the term "template" is used for easy of reference in discussion with no theoretical significance.

[6] From natural language processing perspective, Hakkani-Tür, Oflazer & Tür (2002) develop a morphological disambiguation procedure for Turkish. They argue for identifying "inflection groups" to resolve such ambiguities.

Table 5. Non-finite verb template

Root	Voice	Neg	Subordinator	Agr
V	Causative	-mA	-DIK	Agr
	Passive		-AcAk	
	Reflexive		-Iş	
	Reciprocal			

The template above gives the relative positions of voice categories that will be calculated in this study. In Turkish, all of the four categories *passive*, *causative*, *reciprocal* and *reflexive*, immediately follow the root. The productivity of each of the voice categories and frequency of their combination can be followed from the tables given below.

Pierce (1961) is the earliest study on frequencies of Turkish suffixes. He compiles a very small-sized corpus of written and spoken Turkish, and quantifies raw frequencies of Turkish derivational and inflectional suffixes. The count of Hankamer (1989) and, within natural language processing frame, the work of Güngör (2003) are the studies on the quantificational distribution of affixes in Turkish following the early work of Pierce.[7] Hence, due to lack of a large-scale representative corpus, the frequencies of affixes and their combinations are yet to be calculated for their implications on the structure of language.

4. VOICE SUFFIXES AND FREQUENCY

The list of five most frequent inflectional suffixes (nominal and verbal) in Turkish extracted from the written component of TNC are given below:

Table 6. Most frequent inflectional suffixes in TNC

Rank	Suffix	%	Frequency
1	bare	16,45	13,027,015
2	nominative	10,25	8,120,248
3	possesive	7,41	5,869,510
4	accusative	5,69	4,503,459
5	person (3s)	5,27	4,176,400

[7] Hankamer (1989) gives results of his count of Turkish affixes. He extracts the affixes and their quantities from small corpus of newspaper articles. The average number of affixes per word is 3.06 and proportions of words with five or more suffixes is 19.8 in his findings. Güngör's (2003) counts are from a 2,200,000-word corpus of newspapers and periodicals. He finds the maximum number of suffixes in a sequence as 8 and the average number of suffix length as 2.4.

The top ranking "bare" in the list refers to uninflected tokens that are not tagged as noun. The uninflected noun is identified as nominative, irrespective of its grammatical role. While this may sound counterintuitive; however, the list and the observed frequencies provide a general idea about the distribution of inflectional categories.

Table 7. Frequencies of voice categories

Rank	2-morphgrams	%	Frequency
11	passive (pasv)	2,50	1,976,830
21	causative (caus)	1,35	1,071,278
41	reciprocal (recp)	0,33	262,302
46	reflexive (refl)	0,14	108,156

The passive is the most frequent and the productive category in naturally occuring language data. It is confined with the least number of constraints and may attach to transitive and intransitive verbs. Causative ranks the second in the list with a frequency score almost half of the passive. The lesser productivity of reciprocal and the reflexive are due to their semantics since these two categories are confined to a small number of roots that are compatible in meaning. The non-finite template above (cf. Table 5) asserts that a voice category is followed by negative, subordinator and agreement marker. The derived verb formed by attachment of a voice affix may also receive affixes from the finite template.

The most frequently recurring inflectional affix combinations are given below. In the list of top five frequent affix combinations, we do not find voice categories.

Table 8. Top five 2-morphgrams

Rank	2-morphgrams	Frequency	Sample
1	past+3s	1,354,449	ara*dı*
2	p3s+loc	1,021,648	şahs*ında*
3	aor+3s	864,146	al*ır*
4	vi+past	838,367	ev*di*
5	p3s+acc	797,890	iç*ini*

The voice categories and their combinations in 2-morphgrams are given in the following table:

Table 9. Rank frequencies of 2-morphgrams including voice categories

Rank	2-morphgrams	Frequency	Sample
22	pasv+pcan	313,078	kır*ılan*
25	pasv+nzma	296,006	lisanslan*ma*
34	caus+pasv	220,228	yap*tırılacak*
48	caus+nzma	157,896	hikayeleş*tirme*
84	recp+caus	68,683	tut*uştur*uruz
134	recp+pasv	27,889	bak*ışıl*ır
157	pasv+pasv	19,251	bekle*nil*ir
173	refl+nzma	16,391	öv*ünme*
179	refl+neg	14,217	öv*ünme*

Combinations of passive with nominalizers *-An* and *-mA* take the first top two ranks in the list. Causative follows the passive in the list, as may be expected from their relative frequency scores as individual categories. Reciprocal combinations are cited with relatively less frequency scores and when they combine, they combine with other voice categories. Reflexive produces the least of frequency count given the severe constraints on its semantics.

While there are only a small number of citations of voice affixes in the list of productive 2-morphgrams, we observe an exponential increase in voice categories in 3-morphgrams. Passive is the only category that enters the list of top five in 3-morphgrams. The most frequent passive citations are with nominalizers from position 2 in non-finite template, followed by the same nominal agreement marker.

Table 10. Most frequent 3-morphgrams in the TNC

Rank	3-morphgrams	Frequency	Sample
1	vi+past+3s	646,729	adam*dı*
2	pasv+nzma+p3s	227,646	haşla*nması*
3	pcdk+p3s+acc	194,116	bil*diğini*
4	pcdk+p2s+acc	192,053	uyu*duğunu*
5	imprf+vi+past	186,393	acı*yordu*

Table 11. 3-morphgrams with voice in the TNC

Rank	3-morphgrams	Frequency	Sample
2	pasv+nzma+p3s	227,646	aç*ılması*
12	pasv+perf+3s	113,122	beğen*ilmiş*
29	caus+pasv+nzma	55,745	art*tırılma*
45	caus+past+3s	39,557	acı*ttı*
57	recp+caus+nzma	28,229	böl*üştürme*

The expansion of 3-morphgrams into 4-morphgrams is mainly due to attachment of additional voice categories to the sequence of affixes. In the list of top ten most frequent combinations, we find the first occurrence of a voice combination. Here again, passive outnumbers the other voice affixes and in all three citations of passive, it is followed by position 2 and position 3 suffixes, which themselves are followed by position 4 suffixes and then by agreement.

Table 12. Most frequent 4-morphgrams in the TNC

Rank	4-morphgrams	Frequency	Sample
1	imprf+vi+past+3s	149,758	al*ıyordu*
2	perf+vi+past+3s	127,325	coş*muştu*
3	pasv+perf+cop+3s	71,031	önle*nmiştir*
4	aor+vi+past+3s	58,818	dur*urdu*
5	pasv+cont+cop+3s	55,648	aç*ılmaktadır*
6	caus+pasv+nzma+p3s	50,398	bekle*tilmesi*
7	aor+vi+avsa+3s	46,222	iyileş*irse*
8	pasv+va1+aor+3s	40,355	ata*nabilir*
9	imprf+vi+past+1s	25,903	an*ıyordum*
10	va1+neg+aor+3s	25,437	tak*amaz*

We may argue from naturally occurring language data that an increase in the number of affixes in the verbal domain is mostly due to addition of a voice affix.

Table 13. Most frequent 4-morphgrams in the TNC

Rank	4-morphgrams	Frequency	Sample
3	pasv+perf+cop+3s	71,031	betimle*nmiştir*
5	pasv+cont+cop+3s	55,648	ara*nmaktadır*
6	caus+pasv+nzma+p3s	50,398	alış*tırılması*
15	pasv+nzma+p3s+acc	23,260	as*ılmasını*
18	pasv+pcdk+p2s+acc	21,975	yor*ulmanı*
20	pasv+neg+aor+3s	21,182	sez*ilmez*

Rank	4-morphgrams	Frequency	Sample
28	caus+pasv+perf+3s	16,530	oyna*tılmış*
35	recp+caus+neg+imp2	13,672	gör*üştürme*
41	caus+cont+cop+3s	11,554	bak*tırmaktadır*
161	refl+pasv+neg+aor	2,713	yet*inilmez*
162	refl+imprf+vi+past	2,707	mırılda*nıyordu*

The above list of 4-morphgrams suggests that when a voice suffix combines with a nominalizer, it is commonly followed by an agreement marker from the same template which itself precedes the case marker. Since case marking of nominalized clauses in Turkish is not different from ordinary nominals, finding case-marked nominalizations is expected.

We find more number of voice affixes and their combinations in the list of 5-morphgrams. 8 out of 10 morphgrams in this list includes voice affixes. The recurrent sequences include voice selecting position 3 suffix to follow which in turn followed by position 4 suffix attached to the copula. The sequence ends with the same agreement marker in all top 10 most frequent 5-morphgrams. In the previous patterns of affixes, we have cited nominalizers entering into sequences; however, there occur no nominalizers from non-finite template in five-morphgrams.

Table 14. Most frequent 5-morphgrams in the TNC

Rank	5-morphgrams	Frequency	Sample
1	pasv+perf+vi+past+3s	19,049	aktar*ılmıştı*
2	neg+imprf+vi+past+3s	17,498	ayır*mıyordu*
3	pasv+imprf+vi+past+3s	16,130	se*çiliyordu*
4	neg+perf+vi+past+3s	12,025	sek*memişti*
5	neg+aor+vi+past+3s	10,784	arzula*mazdı*
6	pasv+va1+neg+aor+3s	10,238	gözle*nemez*
7	caus+pasv+perf+cop+3s	9,783	uza*tılmıştır*
8	pasv+aor+vi+avsa+3s	9,322	çak*ılırsa*
9	caus+imprf+vi+past+3s	8,832	kokla*tıyordu*
10	caus+perf+vi+past+3s	7,881	morar*tmıştı*

Table 15. Most frequent 6-morphgrams in the TNC

Rank	6-morphgrams	Frequency	Sample
1	va1+neg+imprf+vi+past+3s	3,910	uç*amıyordu*
2	va1+neg+aor+vi+past+3s	3,768	çevir*emezdi*
3	va1+neg+perf+vi+past+3s	2,206	kaç*amamıştı*
4	caus+pasv+perf+vi+past+3s	1,965	soğu*tulmuştu*
5	va1+neg+imprf+vi+past+1s	1,879	tut*amıyordum*
6	pasv+va1+neg+perf+cop+3s	1,471	sök*ülememiştir*
7	pasv+va1+aor+vi+past+3s	1,400	bul*unabilirdi*
8	pasv+neg+imprf+vi+past+3s	1,351	at*ılmıyordu*
9	pasv+va1+neg+pcck+p3s+acc	1,349	yaz*ılamayacağını*
10	pasv+va1+neg+pcck+p2s+acc	1,318	zorlan*amayacağını*

In 6-morphgrams, as opposed to previous recurrent patterns, the increase of affixes in the sequence is not caused by addition of another voice affix. On the contrary, the number of voice affixes in the sequence decreases. Reciprocal and reflexive are not even cited among the top 10 of morphgrams. The source of increase is mainly due to modality suffixes from position 1 and position 2 and eight citations of negative in the sequences. The nominalizer *-AcAk* appears for the first time in a frequent affix sequence.

In the expansion of recurrent verbal affix sequences with seven, eight and nine suffixes, it is almost always the existence of voice categories or their combinations that produce these complex morpheme bundles.

Table 16. Most frequent 7-morphgrams in the TNC

	7-morphgrams	Freq.	Sample
1	pasv+va1+neg+aor+vi+past+3s	934	bil*inemezdi*
2	pasv+va1+neg+imprf+vi+past+3s	394	belirlen*emiyordu*
3	caus+va1+neg+imprf+vi+past+3s	323	oyna*tamıyordu*
4	pasv+va1+neg+perf+vi+past+3s	313	sağla*namamıştı*
5	pasv+va1+neg+aor+vi+avsa+3s	276	çiz*ilemezse*
6	caus+pasv+va1+neg+perf+cop+3s	239	cay*dırılamamıştır*
7	caus+va1+neg+aor+vi+past+3s	196	otur*tamazdı*
8	pasv+va1+neg+imprf+vi+avsa+3s	188	iç*ilemiyorsa*
9	pasv+pasv+neg+aor+vi+past+3s	186	de*nilmezdi*
10	caus+va1+neg+perf+vi+past+3s	181	öl*dürememişti*

The permanent category in all of the above 7-morphgrams is the negative. The template position of the negative imposes a grammatical

requirement that it is be followed by position 3 and 4 suffixes. High frequency of 3rd person marking also contributes to the formulaicity of morphgrams listed above.

Table 17. Most frequent 8-morphgrams in the TNC

	8-morphgrams	Freq	Sample
1	caus+pasv+va1+neg+aor+vi+past+3s	54	erit*ilemezdi*
2	caus+pasv+va1+neg+aor+vi+avsa+3s	47	tutt*urulamazsa*
3	caus+pasv+va1+neg+imprf+vi+past+3s	41	bin*dirilemiyordu*
4	caus+pasv+va1+neg+perf+vi+past+3s	37	uza*tılamamıştı*
5	caus+pasv+va1+neg+imprf+vi+avsa+3s	18	sön*dürülemiyorsa*
6	pasv+va1+neg+nzma+p3s+vi+past+3s	18	al*ınamamasıyla*
7	recp+pasv+va1+neg+perf+vi+past+3s	15	gör*üşülememişti*
8	recp+pasv+va1+neg+imprf+vi+past+3s	13	kayna*şılamıyordu*
9	recp+pasv+va1+neg+aor+vi+past+3s	12	payla*şılamazdı*
10	pasv+pasv+va1+neg+aor+vi+past+3s	12	de*nilemezdi*

The 8-morphgram sequences are expanded with additional voice categories. These additional voice affixations also produce triples of voice categories as in recp+caus+pasv, ranking top in the list of most productive 9-morphgrams list. As expected, passive combinations outnumber combinations of other voice combinations.

The non-voice 8-morphgrams are all incorporate nominalization affixes and the negative. In all of these sequences, the modality affix from position 1 of the finite template is followed by negative marker (obligatory in case of this particular modality suffix). Nominal agreement markers in the pattern are followed by copula which functions as a buffer to carry position 4 suffixes from the finite template.

Table 18. Non-voice 8-morphgrams in the TNC

	Non-voice 8-morphgrams	Freq.	Sample
1	va1+neg+pcan+pl+abl+vi+past+3s	5	tut*amayanlardandı*
2	va1+neg+pcdk+pl+p1p+vi+past+3s	3	yaşa*yamadıklarımızdı*
3	va1+neg+nzma+p3p+abl+vi+past+3s	2	bak*amamalarındandı*
4	va1+neg+nzma+p3s+abl+vi+past+3s	2	kurtul*amamasındandı*
5	va1+neg+pcdk+pl+p1s+vi+past+3p	2	yap*amadıklarımdı*

There are only 31 citations of morpheme bundles with 9 affixes in the corpus. Dominated by successive voice affixes, the most reccurent *caus-caus-pasv* sequences are attached to the same verb root, *çık* 'to go out'. The negative and the modality suffix from position 1 in the finite template are also frequent in 9-morphgrams.

Table 19. The most frequent 9-morphgrams in the TNC

	9-morphgrams	Freq	Sample
1	recp+caus+pasv+va1+neg+aor+vi+past+3s	5	karşıla*ştırılamazdı*
2	recp+pasv+va1+neg+nzma+p3s+vi+past+3s	2	anla*şılamamasıydı*
3	caus+caus+pasv+va1+neg+aor+vi+past+3s	2	çık*artılamazdı*
4	caus+caus+pasv+neg+nzma+p3s+vi+past+3s	1	çık*artılmamasıydı*
5	caus+caus+pasv+va1+neg+imprf+vi+past+3s	1	çık*artılamıyordu*
6	recp+caus+pasv+va2+neg+perf+vi+past+3s	1	geç*iştirilivermemişti*
7	recp+caus+pasv+va1+neg+nzma+p3s+cop+3s	1	ayr*ıştırılamamasıdır*
8	caus+caus+va1+va1+neg+aor+vi+perf+3s	1	düş*ürtebilemezmiş*
9	pasv+va1+neg+pcan+pl+abl+vi+past+3s	1	daya*namayanlardandı*
10	caus+caus+pasv+va1+neg+imprf+vi+avsa+3s	1	çık*artılamıyorsa*

The role of voice categories in the expansion of affix sequences in the verbal domain is clearly expressed in their distribution across n-morphgrams. In the table below, we summarize the number of citations of voice affixes in top 10 list of *n*-morphgrams:

Table 20. Voice affixes in morphgrams

2-morphgrams	0	5-morphgrams	7	8-morphgrams	10
3-morphgrams	1	6-morphgrams	6	9-morphgrams	10
4-morphgrams	4	7-morphgrams	10		

Göksel (1993) gives a list of possible voice combinations in Turkish with extensive discussions on each combination as well grammatical contraints that regulate such forms. Below is the list of grammatical combinations of voice categories.

1. V-REC-CAUS-PASS
2. V-REC-CAUS
3. V-REC-PASS
4. V-CAUS-PASS
5. V-CAUS-CAUS
6. V-REF-PASS
7. V-PASS-PASS

We note that all these permissible combinations are also cited in the corpus, of couse with different frequencies. Furthermore, corpus data reveals no counter-examples to the expected sequences.

5. STATISTICAL TESTS

From a quantificational perspective, we have checked the statistical significance of *voice+voice* combinations. In order to determine combination of the two variables - the first and the second slot entries from the template- are statistically significant or not, we applied chi-square test to categorical variables.

Table 21. Cross tabulation of voice suffixes

1st slot * 2nd slot Cross tabulation						
			2nd slot			
			caus	pasv	recp	Total
1st slot	caus	Observed freq.	17399	220228	0	237627
		Expected freq.	56415,0	179755,1	1456,9	237627,0
	pasv	Observed freq.	0	19251	2223	21474
		Expected freq.	5098,1	16244,2	131,7	21474,0
	recp	Observed freq.	68683	27889	0	96572
		Expected freq.	22927,2	73052,8	592,1	96572,0
	refl	Observed freq.	0	6915	0	6915
		Expected freq.	1641,7	5230,9	42,4	6915,0
Total		Observed freq.	86082	274283	2223	362588
		Expected freq.	86082,0	274283,0	2223,0	362588,0

Table 21 indicates that since zero cells (0%) have the expected frequency count less than 5, the chi-square result given on table 22 can be interpreted properly. We must note that we have excluded the values of reflexive from the chi-square analysis since it does not occur in the second slot.

Table 22. Chi-square test

Chi-Square Tests	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	198483,646[a]	6	,000
Likelihood Ratio	168477,606	6	,000
Linear-by-Linear Association	108046,078	1	,000
N of Valid Cases	362588		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 42,40.

Table 22 displays a statistically significant relation between the first and the second suffixes observed on verbs (χ^2=198483,646; p<0,05).

The results of the 1st chi-square test is as follows:

- 1st slot is filled by **causative** → χ^2=37552,63
- $\frac{(G-B)^2}{B} = \frac{(17399-56415)^2}{56415} + \frac{(220228-179755,1)^2}{179755,1} + \frac{(0-1456,9)^2}{1456,9}$
- 1st slot is filled by **passive** → χ^2=38875,66
- 1st slot is filled by **reciprocal** → χ^2=119829,1
- 1st slot is filled by **reflexive** → χ^2=2226,274

In the above list of conclusions, we can see the results of the first chi-square analysis. The passive suffix has the highest chi-square value. In the second stage of the analysis, the passive is excluded, and the chi-square test is applied again. The results of the second analysis show that the causative suffix has the highest chi-square value. For the third analysis, excluding the causative, the chi-square test is implemented once again. This time the first slot is filled by reciprocal and reflexive.

The results of the chi-square tests are again significant for both of these variables. Taken together, the results of chi-square tests suggest that the first and the second slot entries observed in verbal inflections cited in written part of the TNC are in statistically significant relationship.

The results of the 2nd and the 3rd chi-square tests are as follows:

- (χ^2=150518,933; p<0,05)→ The chi-square result of the 2nd analysis.
- 1st slot is filled by **causative** →χ^2= 40415,9
- 1st slot is filled by **reciprocal***:* χ^2= 107769
- 1st slot is filled by **reflexive***:* χ^2= 2334,048
- (χ^2=14623,349; p<0,05) → The chi-square result of the 3rd analysis.

6. CONCLUSION

Corpus data in morphological analysis provides information for researchers from varied contexts of language use across different domains, time period and medium that are inaccessible otherwise. A corpus-based study reveals quantificational aspects of language structure that help determine fundamental properties of units and patterns.

It has been argued that frequently occurring multi-word units in other languages correspond to multi-morpheme units in agglutinative languages. The current study presents data that yield support for the arguments to analyze multimorpheme units as patterns of lexical items. When we analyze the recurrent patterns of voice categories in Turkish, we observe that the emerging patterns follow the principles of combination that have been proposed on these sturctures previously. The wealth of corpus data makes it possible to advance a new approach to these frequent patterns as semantic sequences as well as their previously unnoticed discourse functions.

ACKNOWLEDGEMENTS

This study is supported by a research grant from the Scientific and Technological Research Council of Turkey (TÜBİTAK, Grant No: 113K039).

REFERENCES

Aksan, M., & Mersinli, Ü. (2011). A corpus-based Nooj module for Turkish. In Z. Gavriilidou et al. (Eds.), *Proceedings of the Nooj 2010 International Conference and Workshop* (pp. 29-39). Komotini, Greece: Democritus University of Thrace.

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.

Aksan, Y., Aksan, M., Özel, S. A., Yılmazer, H., Demirhan, U. U., Mersinli, Ü., Bektaş, Y., & Altunay, S. (2016). Web tabanlı Türkçe Ulusal Derlemi (TUD). In M. Akgül, U. Çağlayan, E. Derman & A. Özgit (Eds.), *Proceedings of the 16th Academic Computing Conference* (pp. 723-730). İstanbul: Gamze Yayıncılık.

Bickel B., & Nichols, J. (2013). *Inflectional synthesis of the verb*. Retrieved from http://wals.info/feature/22A#2/26.7/151.9

Durrant, P. (2013). Formulaicity in an agglutinating language. *Corpus Linguistics and Linguistic Theory*, 9, 1–38.

Enç, M. (2004). Functional categories in Turkish. In A. Csirmaz, Y. Lee and A. Walter (Eds.), *Proceedings of WAFL 1* (pp. 208-225). Cambridge: MIT Press.

Göksel, A. (1993). *Levels of representation and argument structure in Turkish* (Unpublished doctoral thesis). University of London.

Göksel, A. (1998). Word Length. In G. Booij, A. Ralli and S. Scalise (Eds.), *Proceedings of the First Mediterranean Morphology Meeting (*pp. 190-200). Patras: University of Patras.

Göksel, A., & Kerslake, C. (2005). *Turkish: A comprehensive grammar*. London: Routledge.

Gray, B., & Biber, D. (2015). Phraseology. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics (*pp. 125-145). Cambridge: Cambridge University Press.

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

Hakkani-Tür, D. Z., Oflazer, K., & Tür, G. (2002). Statistical morphological disambiguation for agglutinative languages. *Computers and Humanities*, 36, 381-410.

Hankamer, J. (1989). Morphological parsing and the lexicon. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 392-408). Cambridge: MIT Press.

Pierce, J. E. (1961). A frequency count of Turkish affixes. *Anthropological Linguistics*, 3, 31-42.

Sezer, E. (2001). Finite inflection in Turkish. In E. Erguvanlı (Ed.), *The verb in Turkish* (pp. 1-45). Amsterdam: John Benjamins.

Sinclair, J. (1998). The lexical item. In E. Weigand (Ed.), *Contrastive lexical semantics* (pp. 1-24). Amsterdam: John Benjamins.

Stubbs, M. (2013). Sequence and order: The neo-Firthian tradition of corpus semantics. In H. Hasselgard, J. Ebeling & S. O. Ebeling (Eds.), *Corpus perspectives on patterns and lexis* (pp. 13-33). Amsterdam: John Benjanims.