Patterns and frequency: Evidence from the *Turkish National Corpus* (TNC)[*]

Mustafa Aksan & Yeşim Aksan

1. Introduction

The count of the frequencies of lexical items is a traditional undertaking. Previously, the primary motivation in these studies was mainly practical rather than theoretical in the sense that quantification information is expected to provide better description for individual items as well as for their combinations. Recently however, research on frequencies concluded that statistical regularities and distributional aspects of lexical structures have theoretical significance, bringing new insights into the role of lexis in grammar and in patterning.

Advances in corpus software development and corpus analytic tools provided additional empirical evidence for a renewed understanding of lexical structures. Concordance data have identified various patterns in ordinary language use, alongside formulaic expressions and various other forms of fixed expressions. In the general framework of British linguistic tradition (Stubbs, 1993, 2013), work on corpus data argued that lexical structures encode such properties that cannot be captured within the confines of individual word or lexeme. Sinclair (1998) thus proposes the term lexical item to account for recurrent and regular patterns that expand beyond size of a single item.

The present study will show data of frequent and recurrent patterns that are extracted from the Turkish National Corpus (TNC) (http://www.tnc.org.tr). The patterns that we will review here cover sequences of (i) lexical items (i.e. the *multiword units*), (ii) the regular frequent patterns formed by inflectional categories (i.e., the *multimorpheme* units), and (iii) patterns found among adjacent lexical item (i.e., *interlexical* units). In sum, the data here represent an initial typology of such structures and their observed frequencies.

This paper is organized as follows. The first part will review basics of fixed expressions and corpus-based analysis of frequent patterns. In the second part of the paper, the data of patterns of lexemes and morphemes will be given with their distributional frequencies. The frequencies of these recurrent patterns are indicative for a proper understanding Turkish lexicon in general.

2. Patterns, frequency and text organization

Bybee (2006) observes that high frequency words and expressions differ from low frequency words and expressions as having different set of properties. She further argues that emphasis on general patterns of language structures to derive abstract generalizations obscured the significant role of frequency in "producing a highly conventional set of general and specific structures that allow the expression of both conventional and novel ideas"(5).

A closer inspection of corpus data provided new evidence for the frequency and distribution of lexical items. It became clear that frequent patterns of use play a significant role in both production and comprehension of texts. Sinclair (1991:108) summarizes a basic conclusion from corpus analysis of text, “By far the majority of text is made of the occurrence of common words in common patterns, or in slight variants of these common patterns.” Sinclair (1991) also proposes two organizing principles that underscore the role of lexis in structuring and patterning of texts: Idiom Principle and Open Choice Principle. While the former principle holds that language users rely frequently on a vast stock of pre-fabricated lexico-grammatical patterns; the latter principle asserts that formal rules of grammar simply serve to combine these pre-formed patterns wherever the textual structure calls for such a choice.

3. Data and Method

3.1 The Corpus

Written component of the TNC is used to extract true rankings of the all types of n-grams in this study. The size of the TNC is 50,997,016 running words, representing a wide range of text categories spanning a period of 23 years (1990-2013). It consists of samples from textual data representing 9 different domains (98%) with 4978 documents and transcribed spoken data (2%) with 434 documents. The annotation system of the TNC covers over 90 inflectional morphemes, all of which are compatible with modern Turkish linguistics studies. Table (1) shows the distribution of texts in the written part of the TNC.

Table 1. Distribution of texts according to domains in TNC-written

Domain	No. of words	% of words
Imaginative: Prose	9.365.775	18.74 %
Informative: Natural and pure sciences	1.367.213	2.74 %
Informative: Applied science	3.464.557	6.93 %
Informative: Social science	7.151.622	14.31 %
Informative: World affairs	9.840.241	19.69%
Informative: Commerce and finance	4.513.233	9.03 %
Informative: Arts	3.659.025	7.32 %
Informative: Belief and thought	2.200.019	4.4 %
Informative: Leisure	8.421.603	16.85%
Total	49.983.288	100.00 %

3.2 Extraction of multiword, multimorpheme and interlexical units

To obtain multiword unit (MWU) candidates, first the TNC texts are optimized. After the optimization, the lower-cased, sentence-splitted, punctuation-delimited, ASCII-coded TNC texts are processed in Ngram Statistical Package (Text::NSP) tool (Pedersen et al., 2011) to generate rank order frequency lists of n-grams. Poisson-Stirling value as associative measures along with observed frequency are used for ranking and determining lexicalized MWUs. For simplicity in this study only observed frequencies of the data are given. Frequency cut-off for bi-grams 200 times; for tri-grams 100 times and for four-grams 5 times per million words is determined. As a second step, MWU candidates have been annotated by using the TNC-tagger which is based on words or lemmas in other words, free morphemes and available inflectional suffixes in the same word form. The tagging process is done by simply matching each word-form with the corresponding entry in the TNC-Natural Language Processing (NLP) dictionary. These entries include all information concerning the lemma, part-of-speech and inflectional suffixes that are observed in each word form of the given sequence. Finally, the grammatical sequences of these word forms are semi-automatically classified and validated by the researchers. The frequency of these grammatical sequences are also calculated as a final step and ranking them due to their observed frequency provided an overview of the constraints that are governing the MWU lexicalization in Turkish

As for multimorpheme unit extraction frequency count of sequences of nominal and verbal inflectional suffixes in Turkish are calculated. For achieving such a count morphological tagging of morphemes via the TNC-tagger is done. The annotation processes, including part-of-speech tagging, morphological tagging and lemmatization are done using a NLP dictionary based on the NooJ_TR module (Aksan & Mersinli, 2011). The unique, semi-automatic process of developing the NLP dictionary includes the automatic annotation of the type list with the NooJ_TR module and manual checking and revising the output and eliminating artificial/non-occurring ambiguities. After these stages, the entries of the NLP dictionary and actual running words of the corpus are matched via the PHP and MySQL-based interface of the TNC. However, we should note that consequences of the agglutinative nature of Turkish are reflected in various domains in the internal structure of words. While the order is mostly predictable and easy to process with clear-cut morpheme boundaries in many cases, the existence of a number of homographic morphemes, lemma + suffix combinations, suffix+suffix combinations as well as homographic lemmas present specific challenges for morphological tagging. The so-called challenges constitute the ambiguities that 15% of the TNC tokens contain.

Finally to extract interlexical units, again Text::NSP (Pedersen et al., 2011) is used and calculation based on observed frequencies of bi-grams (e.g., yap-*mak için*[1]) is achieved. The basic idea here is to obtain suffix-word pairs such as -*mAk için* 'in order to'. To do this, the first component of the unit is tagged by TNC-tagger schema (e.g., *güldürmek* → gül,VB+caus+nzmk) and the frequencies of all suffixes of the first unit, the recurrent suffixes in the construction and lexical patterns are all calculated (e.g. *caus+nzmk__için*). Then, the patterns cited are updated after calculation of observed frequencies of the closing suffix from the first unit and the pattern emerging with combination of the following lexeme (e.g., *nzmak__için*). Sequences of multiple affixes in interlexical units are further counted on the basis of patterns identified (e.g. *caus+nzmk__ için*).

4. Multiword units

In simple terms, a multiword unit (MWU) is "(…) the most frequent recurring lexical sequences in a register" (Biber, Conrad, Cortes, 2004: 371). Most of the MWUs do not form conventional forms like phrases or compounds and they may lack structural unity or semantic compositionality. In other words, fragments of lexical sequences or syntactically incomplete but meaningful strings are forming the MWUs (e.g. *süre sonra* 'after time', *başta olmak üzere* 'being in the first') though semantically full expressions are also automatically retrieved as lexical chunks (e.g. *ne de olsa* 'after all'). In the overall organization of a discourse, MWUs typically occur in between phrases or clauses, serving more like a bridging elements. Typically, larger MWUs incorporate smaller ones, i.e., a four-gram may have a tri-gram as its constitutive component, and similarly, a tri-gram may have a bi-gram in its constitution.

Research on Turkish MWUs can be classified as studies on NLP (e.g. Oflazer, Çetinoğlu & Say, 2004; Kumova-Metin & Karaoğlan, 2011) and on linguistic identifications and classification. Recently, to identify formal and functional properties of MWUs as well as to comment on methodological challenges in extracting them, corpus-driven studies have been carrid out. In this respect, Mersinli (2015) explores linguistic relevance of MWU ranking of 12 associative measures that Text::NSP contain on 10-million-word TNC Baby. Mersinli and Aksan (2016) discuss methodological considerations for clarifying appropriate processes for Turkish MWU extraction considering the agglutinative nature of Turkish. Durrant (2013) argues that frequent co-occurrence of elements attested at word level in English occurs at morphological level in Turkish, and thus psychological models of processing should include morphological patterns. Aksan and Aksan (2015 a,b) present the emerging formal category and internal structure of MWUs and their primary discourse functions on two domains of the TNC, namely imaginary and informative domains. These studies also demonstrate the register/genre specificity of MWUs identified for fiction and informative written text in Turkish. In a more recent study, Yıldız (2016) investigates the structural patterns and discourse functions of the most frequent 50 MWUs in the construction of academic texts in Turkish using a special corpus that has over 1,000,000 words that contain texts from 12 sub-disciplines belonging to the humanities and fundamental sciences.

The corpus-driven identification of the most frequent lexicalized MWUs across the written-TNC is given below in Tables 3 to 5 (see also http//: www.tncfrequency.org.tr).

Table 3. Top 5 bi-grams in the written-TNC

Rank	**Bi-gram**	**Freq.**	**Word class**
1	*ya da* 'or'	64871	conjunction
2	*bir şey* 'something'	36796	pronoun
3	*ortaya çıkmak* 'to show up'	23019	verb
4	*her şey* 'everything'	23013	pronoun
5	*hem de* 'besides'	22843	adverb

Table 4. Top 5 tri-grams in the written-TNC

Rank	Tri-gram	Freq.	Word class
1	*bir süre sonra* ‘after a while’	4419	adverb
2	*bir kez daha* ‘one more time’	4000	adverb
3	*ne var ki* ‘however’	3360	conjunction
4	*başka bir şey* ‘something else’	3293	pronoun
5	*ne yazık ki* ‘unfortunately’	3020	conjunction

Table 5. Top 5 four-grams in the written-TNC

Rank	Four-gram	Freq.	Word class
1	*kısa bir süre sonra* ‘ after a short while’	1057	adverb
2	*önemli bir rol oynamak* ‘ to play an important role’	459	verb
3	*her zaman olduğu gibi* ‘as usual’	424	conjunction
4	*önemli bir yer tutmak* ‘to keep a significant place’	371	verb
5	*temel hak ve özgürlükler* ‘basic rights and freedoms’	339	noun

A cursory analysis reveals that some of the tri-grams encompass bi-grams (e.g. *başka bir şey* ‘something else’ > *bir şey* ‘something’) and some of the four-grams encompass tri-gram (e.g. *kısa bir süre sonra* ‘after a short while’ > *bir süre sonra* ‘after a while’). The most frequent sequences consist of two-word and three-word units, while there are considerably fewer four-word sequences. Considering the grammatical categories of the MWUs, there is a variety in bi-grams and four-grams. Among the top five bi-grams and four-grams, nouns and verbs emerge as the distinct categories when compared to tri-grams which contains mostly adverbs and conjunctions. Overall, all the MWUs displayed are predominantly composed of function words, and thus there is a “‘world-out-there’ representation, dominated by impersonal constructions” (O’Keeffe et al., 2007:68). The functions of these MWUs are classified primarily under referential expressions and text organizers by following Biber, Conrad & Cortes, 2004; Carter & McCarthy, 2006. MWUs having the word class conjunction serve as text organizers, such as transitional signals (e.g. ya da ‘or’) by showing relationships between prior and coming discourse. The MWUs under the category of adverbs and pronouns are used as referential expressions to make direct reference to physical and abstract entities to identify the entity or to single out some particular aspects of the entity as important. For instance, *bir süre sonra* ‘immediately’ expresses time reference; *bir şey* ‘something’ indexes vague expression.

5. Multimorpheme units

Affixes play a significant role both in formation of new lexemes and in expression of various grammatical categories, yet there are only a very small number of studies on their various orderings or combinations. While one possible reason for this lack of interest relates to complicated nature of affix combinations that calls for different approaches simultaneously, the other is the lack of comprehensive corpus data that would present huge data of cited combinations a researcher cannot access individually. When actual data of use is observed, we find that only a very few of possible combinations are realized due to severe constraints that are at work.

Common tendency is to explain ordering of inflectional affixes by referring to formal constraints on each of the categories and consequently on their combinations, and for the ordering of derivational suffixes, it is the productivity and semantics of the categories are in question[2].

The first comprehensive study on frequencies of affixes and their combinations in Turkish is conducted by Pierce in early 1960s. Constructing very small-sized corpus of written and spoken Turkish, Pierce (1961) presents observed frequencies of Turkish derivational and inflectional suffixes, together with top 20 most frequent lexemes. Until the recent work in computational morphology[3] we find no other principled account of suffix frequencies in Turkish.

The observed frequencies of inflectional suffixes cited in the written-TNC are given below.

Table 6. Most frequent 1-morphgrams in the TNC

Rank	**1-morphgram**	**Frequency**	**Sample**
1	Bare	13306983	*bir*
2	Nom	8282759	*ev*
3	Acc	1942012	*onu*
4	p3s	1652940	*konusu*
5	Dat	1107401	*ortaya*

The bare in the list above refers to any uninflected token excluding the noun. Regardless of its particular grammatical function in a sentence, the uninflected noun is taken as representing the nominative. It is evident that proper annotation of categories requires syntactic and morphological parsers. Yet, the observed frequencies listed above provide a general frame for the simple quantities of such items in the language.

The inflectional suffixes in the verbal domain are ordered in terms of the slots available for each category. Göksel and Kerslake (2005) [4] provide two such constrained orderings for finite and nonfinite inflectional categories[5] in the verbal domain. In the following, the data of combinations with different number of morphgrams provide relevant cited orderings. Nominal inflectional categories are small in number compared to verbal affixes and as in the verbal domain, do not allow alternative ordering.

Table 7. Most frequent 2-morphgrams in the TNC

Rank	**2-morphgrams**	**Frequency**	**Sample**
1	p3s+loc	806453	*içinde*
2	pl+acc	725378	*evleri*
3	past+3s	539287	*dedi*
4	p2s+loc	536356	*evinde*
5	p3s+dat	517672	*yerine*

Other than (3) in the table above, all remaining 2-morphgrams represent nominal inflectional categories.

The grammatical requirement predicts that a simple inflected verb occurs minimally with a tense suffix and an agreement marker. The nominal categories may occur minimally with any of the nominal inflectional categories. Thus, any 3-morphgrams sequences in Turkish is expected to be an expansion of minimally inflected form combining with other categories from its domain. The expansion of the grammatically required minimal form with the verb is done by addition of a voice (5) category, a nominalizer (2, 3) or tense/aspect marker attached to copula (1) (the compound tense). The observed frequencies suggest that nominalizers dominate 3-morphgrams sequences.

Table 8. Most frequent 3-morphgrams in the TNC

Rank	3-morphgrams	Frequency	Sample
1	vi+past+3s	222758	*idi*
2	pcdk+p3s+acc	143388	*olduğunu*
3	pcdk+p2s+acc	141595	*gördüğünü*
4	p3s+loc+kia	117398	*arasındaki*
5	pasv+nzma+p3s	107434	*edilmesi*

The 4-morphgram sequences are derived by addition of a tense/aspect marker to a copula, motivated by a structural requirement. In (3), (4), and (5), we also find introduction of voice categories, here the passive, which is also the most productive among voice categories.

Table 9. Most frequent 4-morphgrams in the TNC

Rank	4-morphgrams	Frequency	Sample
1	imprf+vi+past+3s	102740	*ediyordu*
2	perf+vi+past+3s	85516	*olmuştu*
3	pasv+perf+cop+3s	60307	*edilmiştir*
4	pasv+cont+cop+3s	48364	*görülmektedir*
5	pasv+va1+aor+3s	34167	*edilebilir*

Further addition of voice affixes and nominalizers from nonfinite verb template expand other *n*-morphgrams that we will not present here given the limitations of space. As expected, the number of citations of morphgrams in the corpus decreases significantly as the number of affixes that enter into combinations increase. The list of top 5 most frequent 9-morpgrams (from a total of 31) gives a simple idea about concatenation of affixes in between from the simplest to the most complex.

Table 10. The most frequent 9-morphgrams in the TNC

Rank	9-morphgrams	Freq.	Sample
1	recp+caus+pasv+va1+neg+aor+vi+past+3s	5	*karşılaştırılamazdı*
2	caus+caus+pasv+va1+neg+aor+vi+past+3s	2	*çıkartılamazdı*
3	recp+pasv+va1+neg+nzma+p3s+vi+past+3s	2	*anlaşılamamasıydı*
4	caus+caus+pasv+neg+nzma+p3s+vi+past+3s	1	*çıkartılmamasıydı*
5	caus+caus+pasv+va1+dsub+aor+vi+past+3s	1	*çıkartılabilirdi*

In recurrent 9-morphgrams we find sequences of voice categories, occurring in their licenced sequence. The negative and nominalizers also expand the combination significantly. It seems that Turkish do not produce such word monsters that might have included more affixes than the combinations given above. In fact, the average length is relatively small to include about 2-3 affixes per word, excluding the derivational affixes.

6. Interlexical compound forms

The formulaicity of case marker and postposition sequences is observed in many contexts. Phrasal or clausal uses of such recurrent patterns headed by a postposition are commonly treated as compound forms in grammars. On the basis of their compositional semantics and textual functions, such patterns are discussed as adjectival or adverbial forms, most often combining with other affixes to form richer sequences (as in *-mamasına rağmen* 'despite not being […]'; *-dığından ötürü* 'it because of […]' among many others). Textual functions of these patterns are well-recognized and documented with extensive categorizations for their adverbial uses. Some of these patterns are analysed in papers specifically devoted for their semantic contributions to the texts.

Table 11 below list the observed frequencies of interlexical compound forms cited in the TNC. These patterns occur both in their possible attested simple forms and are expanded with further addition of affixes preceding the closing case marker.

Table 11. Top 5 interlexical patterns

Rank	**Lemma (2)**		**Suffix (1)**	**Citation**	**(1) + Lemma (2) Freq**
1	*için*	'for'	nzmk	et*mek_için*	40,872
2	*gibi*	'as, like'	pcdk+p3s	ol*duğu_gibi*	18,020
3	*sonra*	'after'	pcdk+abl	ol*duktan_sonra*	15,825
4	*üzere*	'about to'	nzmk	ol*mak_üzere*	15,338
5	*için*	'for'	gen	bun*un_için*	15,336

İçin 'for' with its different complements (1) and (5) outnumber the other interlexical compound forms. As discussed in grammars as well, *için* 'for' appears to be most productive postposition to head more varied types of complements. The observed frequencies indicate that the total of *için* 'for' citations are more that the total of other patterns among the top five. There are also compound forms with auxiliaries that recur in the texts. *Olmak* 'to be, to become' is very frequent in such compound forms, a buffer stem to carry inflectional affixes or tense/aspect as in *-mış olduk* 'we have been (…)'; frequencies of *olmak* 'to be, to become' forms above indicate its recurrent use as an auxiliary in nominalization.

The observed frequencies of lemmas in these interlexical compound forms are given below:

Table 12. Top 5 lemma in interlexical patterns

Lemma	**Freq**

1	*için*	'for'	136,796
2	*gibi*	'as, like'	56,494
3	*sonra*	'after'	38,679
4	*göre*	'as for'	37,850
5	*kadar*	'until'	35,912

Here again, we observe the dominance of *için* 'for' citations, accounting almost half of the top five lemmas in compound patterns. The high frequency of *için* 'for' forms is also evident in among the most common patterns, sampling its three different types complementation.

Table 13. Top 5 frequent patterns

	lemma		**lemma +suff+ lemma**	**freq**
1	için	'for'	onlar_*için*	8734
2	için	'for'	gerçekleştirmek_*için*	7973
3	doğru	'towards'	geriye_*doğru*	6583
4	için	'for'	yılı_*için*	6398
5	kadar	'until'	sonuna_*kadar*	6014

A relatively detailed distribution of *için* 'for' complements identifies various nominalizers in its complements. *Infinitive için* 'for' sequence constitutes the most recurrent pattern, followed by the other nominalizers, representing almost all available such forms in Turkish.

Table 14. *İçin* 'for' complements

	suffix	**anno.**	**citation**	**freq.**
1	-mAk	nzmk	et*mek_için*	58,535
2	-mA	nzma+p3s	ol*ması_için*	15,638
3	-dIk	pcdk+p3s	ol*duğu_için*	20,153
4	-An	pcan+pl	ol*anlar_için*	844
5	-AcAk	pcck+p3s	ol*acağı_için*	263
6	Diğer	p1p	ülke*miz_için*	41,363
	Total			136,796

When we look at the observed frequencies of infinitive complements of *için* 'for', we find that -*mAk* 'to infinitive' is commonly preceded by voice suffixes, modality markers and the negative. A 3-morphgram complement of *için* 'for' with –*mAk* 'to infinitive' also cites the same categories with voice affixes. It is interesting to note that while passive is the most frequent among voice categories in the corpus data, it is the least cited among –*mAk* 'to infinitive' complement of *için* 'for', and with no citation at all in 3-morphgrams complements below.

Table 15. *İçin* 'for' morphgrams with -*mAk*

n-morphgram	Suffix sequence	Citation	Freq.
1-morphgram	nzmk	et*mek_için*	40,872
2-morphgram	caus+nzmk	gerçekleş*tirmek_için*	7,973

	va1+nzmk	ed*ebilmek_için*	5,943
	neg+nzmk	ol*mamak_için*	1,724
	pasv+nzmk	koru*nmak_için*	667
	refl+nzmk	gör*ünmek_için*	278
	recp+nzmk	gör*üşmek_için*	255
	Total		16,840
3-morphgram	caus+va1+nzmk	sür*dürebilmek_için*	493
	caus+neg+nzmk	kaç*ırmamak_için*	100
	recp+caus+nzmk	ara*ştırmak_için*	87
	caus+caus+nzmk	çık*artmak_için*	84
	recp+neg+nzmk	karşıla*şmamak_için*	59
	Total		823

The patterns we have identified above suggest the prevalent nature of these specific patterning of affixes and lemmas interlexically. Here, we have illustrated a possible way of progressing toward analysing internal structure of such sequences concentrating on their observed frequencies.

7. Conclusion

In this paper we have presented our corpus-driven analysis of recurrent patterns in Turkish. We have identified citations of patterns with three types of units. The multiword units are extracted from the TNC by using specialized software. The recurrent sequences of morphemes are identified and extracted from corpus since they are partially annotated in the construction of the corpus. The frequently used interlexical units are also identified and their observed frequencies are calculated.

At this stage, we have illustrated a case of using corpus analytic tools to derive frequency information from a corpus as well as procedures of unit identification based on observed quantities of these possible units. The information on quantities of units as they are observed in data, may further be enriched by conducting detailed statistical analyses to calculate and predict the combination potentials of elements from slots available in the entire construction of the pattern.

The identification of emerging patterns and units on the basis of their frequency of use provides an input for a formal and functional investigation of such units. A detailed typology of these patterns with different units will provide a much more reliable picture of lexis in Turkish. Particular textual functions of patterns, defined as semantic sequences, will contribute our understanding of textual relations.

Abbreviations

3s	3rd sing.	imprf	imperfective	past	past
abl	ablative	kia	adjectival	pasv	passive
acc	accusative	loc	locative	pcan	adjectival
aor	aorist	neg	negative	pcdk	nominalizer
caus	causative	nom	nominative	perf	perfective
cont	continuous	nzma	nom. -mA	pl	number/person
cop	copula	nzmk	nom. -mAk	recp	reciprocal

dat	case-dative	p1p	poss. 1st pl.	refl	reflexive
dsub	desubjectivizer	p2s	poss. 2nd sing	va1	aux verb
gen	genetive	p3s	poss. 3rd sing	vi	buffer verb "i"

References

Aksan, M., & Aksan, Y. 2015a. Multi-word in imaginative and informative domains. In: Zeyrek, D., Sağın-Şimşek, Ç., Ataş, U. & Rehbein, J. (eds.) 2015. *Ankara papers in Turkish and Turkic linguistics*. Wiesbaden: Harrassowitz Verlag, 316-327.

Aksan, M., & Aksan, Y. 2015b. Multi-word expressions in genre specification. *Mersin University Journal of Linguistics and Literature,* 12, 1-42.

Aksan, M., & Mersinli, Ü. 2011. A corpus-based Nooj module for Turkish. *Proceedings of the Nooj 2010 international conference and workshop*. Komotini, 29-39.

Balthasar B., & Nichols, J. 2013. Inflectional Synthesis of the Verb. In: Dryer, Matthew S. & Haspelmath, M. (eds.) 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute.

Biber, D., Conrad, S. & Cortes, V. 2004. If you look at ... : Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25, 371–405.

Bybee, J. 2006. *Frequency of use and the organization of language.* Oxford: Oxford University Press.

Carter, R. A., & McCarthy, M. J. 2006. *Cambridge grammar of English*. Cambridge: Cambridge University Press.

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

Enç, M. 2004. Functional categories in Turkish. *Proceedings of WAFL 1, MIT working papers in linguistics*, 46, 208-225.

Göksel, A. 2001. The auxiliary verb *ol* at the morphology-syntax interface. In: Erguvanlı, E. (ed.) *The Verb in Turkish*. Amsterdam.: John Benjamin, 151-181.

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

Güngör,T. 2003. Lexical and morphological statistics for Turkish. *Proceedings of international XII. Turkish symposium on artificial intelligence and neural networks*.

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

Kumova-Metin, S. K., & Karaoğlan, B. 2011. Measuring collocation tendency of words. *Journal of Quantitative Linguistics* 18, 174-187.

Mersinli, Ü. 2015. Associative measures and multi-word extraction in Turkish. *Mersin University Journal of Linguistics and Literature* 12, 43-61.

Mersinli, Ü., & Aksan, Y. 2016. A methodology for multi-word unit extraction in Turkish. *Proceedings of the first international conference on Turkic computational linguistics,* 27-31.

O'Keeffe, A., McCarthy, M.J. & Carter, R.A. 2007. *From corpus to classroom*. Cambridge: Cambridge University Press.

Oflazer, K., Çetinoğlu, Ö., & Say, B. 2004. Integrating morphology with multi-word expression in Turkish. *Proceedings of the 2nd ACL workshop on multiword expressions: Integrating processing*, 64-71.

Pedersen, T., Banerjee, S., McInnes, B. T., Kohli, S., Joshi, M., & Liu, Y. 2011. The ngram statistics package (Text::NSP): A flexible tool for identifying ngrams, collocations, and word associations. *Proceedings of the workshop on multiword expressions: From parsing and generation to the real world*, 131–133.

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

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

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. 1998. The lexical item. In: Weigand, E. (ed.) 1998. *Contrastive lexical semantics*. Amsterdam: John Benjamins.

Stubbs, M. 1993. British tradition in text analysis: From Firth to Sinclair. In Baker, M., Francis, G., & Tognini-Bonelli, E. (eds.) 1993. *Text and technology: In honour of John Sinclair*. Amsterdam: John Benjamins, 1-33.

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

Turkish National Corpus. http://www.tnc.org.tr. 24.08.2016.

Turkish National Corpus-Word and Multi-word frequencies in Turkish. http://www.tudfrekans.org.tr. 14.11.2016.

Yıldız, İ. 2016. Multi-word units in Turkish scientific texts: A corpus-based genre analysis. Unpubished Ph.D. Dissertation, Mersin University.

[*] This study is supported by TÜBİTAK (grant no: 113K039).

[1] Since the annotations of the forms encode the meaning, we will not give their English equivalents.

[2] Sezer (2001) analyzes restrictions on combinations of various tense and agreement affixes with their particular head features. Enç (2004) follows the same idea only to argue the nature of some of these functional heads. Göksel (2001) also notes morphological constraints that license or block affix combinations. As for ordering of derivational suffixes, there exist only occasional references in a number of papers.

[3] Hankamer (1989) calculates frequencies of Turkish affixes from a data of newspaper articles. He concludes that average number of affixes per word is 3.06 and the ratio of words with five or more suffixes is 19.8. Güngör (2003) presents his count 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 per word as 2.4.

[4] Slots in the finite verbal inflection are as follows: (1) optative (2) modality, (3) tense-aspect, (4) copula attached categories, and (5) copula *-DIr*. The non-finite categories include (1) voice, (2) negative, (3) subordinators, and (4) agreement.

[5] Baltasar and Nichols (2013) discuss verbal synthesis in of languages with number of categories in their "verbal synthesis". Turkish ranks in middle among the list of languages with relatively less (0-1) or (12-13) more number of inflectional categories.