Mersin Üniversitesi Dil ve Edebiyat Dergisi, MEUDED, 2016; 13 (2), 71-108.

COLLIGATIONAL PATTERNS OF TURKISH MULTI-WORD UNITS[1]

Yeşim AKSAN[2], Ümit MERSİNLİ[3], Serap ALTUNAY[4]

Mersin University

Abstract: In multi-word unit (MWU) extraction studies, most of the challenges for rich morphology languages like Turkish can be overcome by the study of how colligational filtering works in our minds, along with how statistical and collocational sorting affects the process. Based on the assumption that lexicalization of any given collocation as a MWU also requires compatibility to some lexical or morphosyntactic constraints, this study will present the morphosyntactic tendencies observed in colligational patterns of Turkish MWUs and discuss their implications on language-specific MWU filtering processes. The aim of the study is to discuss if in Turkish, associative strength is enough for a collocation to be lexicalized as a MWU or not. Another purpose of the study is to show some morphosyntactic and lexical constraints that may validate collocations to be lexical multi-word units in Turkish. The paper will also underscore the methodological perspectives of MWU identification valid for rich-morphology languages. To achieve these goals, we first extracted MWU candidates -trigrams-

[1] This study was supported by TÜBİTAK (Grant no:113K039).

[2] Mersin University, Faculty of Science and Letters, Department of English Language and Literature, yesim.aksan@gmail.com

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

[4] Mersin University, Faculty of Science and Letters, Department of English Language and Literature, serap.altunay88@hotmail.com

Makale gönderim tarihi: 01 Mart 2016; Kabul tarihi: 10 Haziran 2016

from a 10-million-word sub-corpus of Turkish National Corpus (TNC) by using Text-NSP (Banerjee & Pederson, 2011). After that, the 3-grams were annotated by using the NLP dictionary of TNC-tagger, and classified according to their colligational patterns and lexical categories of the MWU. Most frequently observed colligational patterns are argued to be morphosyntactic tendencies governing MWU lexicalization in Turkish. In this respect, the study aims to contribute to the understudied area of formulaic language in Turkish.

Keywords: *Multi-word unit, colligational pattern, lexical frame, corpus-driven, Turkish National Corpus*

TÜRKÇEDE ÇOK SÖZCÜKLÜ BİRİMLERİN İŞLEV DİZİSİ ÖRÜNTÜLERİ

Öz: Çok sözcüklü birim (ÇSB) çıkarımı çalışmalarında, Türkçe gibi zengin biçimbilime sahip dillerde karşılaşılan pek çok güçlük, bu süreci etkileyen istatistik sıralamanın yanında, işlevsel ayıklamanın, zihnimizde nasıl işlediği üzerine çalışarak aşılabilir. Herhangi bir sözcük dizisinin ÇSB olarak sözlükselleşmesi için, bazı sözlüksel ve biçimsözdizimsel kısıtlamalara da uygun olması gerekeceği varsayımından hareketle, bu çalışma, Türkçe'de işlevsel örüntülerde gözlenen biçimsözdizimsel eğilimlere ve bu eğilimlere dayalı olarak, Türkçe'de ÇSB ayıklama sürecine ilişkin çıkarımlara değinecektir. Çalışmanın amacı, Türkçe'de bir sözcük dizisinin, ÇSB olarak sözlükselleşmesi için, içerdiği sözcükler arasındaki ilinti gücünün yeterli olmadığını göstermek ve bu sözcük dizilerinin kabul edilebilir ÇSB'ler olarak sözlükçemizde yer alması için gerekli olan biçimsözdizimsel ve sözlüksel kısıtlamaları tartışmaktır. Çalışma bu yönüyle, zengin biçimbilimli dillere özel bir ÇSB çıkarım yöntemiyle ilgili de bir bakış açısı sunmayı amaçlamaktadır. Belirtilen amaçlar doğrultusunda, öncelikle, Text-NSP (Banerjee & Pedersen, 2011) kullanılarak, Türkçe Ulusal Derlemi'nin 10 milyon sözcüklük bir alt-derleminden ÇSB adayları -üçlü diziler- çekilmiştir. Sonrasında, bu üçlü sözcük dizileri TUD-işaretleyicinin içerdiği Doğal Dil İşleme (DDİ) sözlüğü yardımıyla işaretlenmiş ve içerdikleri işlev dizileri ve sözcük türlerine göre sıralanmıştır. Sonuç olarak, bu en sık gözlenen işlev dizilerinin, Türkçe'de çok sözcüklü birimlerin sözlükçeye yerleşmesinde etken olan biçimsözdizimsel eğilimler olduğu savlanmıştır. Bu yönüyle çalışma, Türkçe'de fazlaca çalışılmayan kalıp dil kullanımı (İng. formulaic language) konusuna katkı sunmayı hedeflemektedir.

Anahtar sözcükler: *Çok sözcüklü birim, işlevsel örüntü, sözcük çerçevesi, derlem-çıkışlı, Türkçe Ulusal Derlemi*

1. INTRODUCTION

The frequently used word combinations or recurrent combination of two or more lexical items has aroused interest of the researchers and language teachers over the past three decades. There are numerous studies on linguistic analysis of phraseology, to determine different types of formulaic multi-word sequences and to describe how these sequences are used in everyday discourse (see Weinert 1995; Ellis 1996; Howarth 1996; Wray & Perkins 2000; Wray, 2000 for the reviews). Fixedness, formulaicity or the term of collocation in language is not a newly discovered phenomenon and related citations can be even dated back to 1920s. For English tradition, Jespersen (1924), Palmer (1932) and Firth (1951) can be named as the pioneers of formulaic language or phraseology for their views on the word combinations (for previous theoretical studies in this field see Pawley & Syder, 1983; Sinclair, 1991; Lewis, 1993; Weinert 1995; Howarth, 1998; Wray & Perkins 2000).

Wray (2002, p. 9) states that formulaic sequence is "a sequence, continuous or discontinuous, of words or other 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." Furthermore, she underscores that the use of prefabricated sequences in language is an underestimated part of our lexicon. Which is worse, formulaicity in agglutinative languages is an even more understudied subject, just because of the technical or computational difficulties in identifying the operational units. Unlike English - where the space character is a powerful operational delimiter for lemmas, stems or words - Turkish has its own morphosyntactic challenges, which even allows a sentence to be represented in a single word - as in *gitmişlermiş* 'they are said to have gone'.

This study aims to present colligational patterns occurring mostly in the formulaic Turkish, or in multi-word units (MWUs). We follow the basic principles of copus-driven methodology. We, also, take a frequency-driven approach to determine multi-word units in the present study. Multi-word sequences have been analyzed under a variety of labels and definitions. The frequently used terms are followings: *chunks, n-grams, prefabricated routines, multi-word units, multi-word*

expressions, lexical bundles, lexical phrase, formulaic expressions, formulaic sequences, clusters, fixed expressions, formulas, idioms. For the purpose of this study, we use 'n-grams' for any word sequence that are frequently observed and the terms 'multi-word unit' is utilized for valid, lexicalized word sequences that are stored as a single unit in the lexicon.

Section 2 reviews the corpus studies done on multi-word sequences. In section 3 data and methodology of the study is presented. Section 4 deals with the colligational rankings of tri-grams with reference to their internal structures by exploring what type of colligational patterning they involve. In section 5 the most frequent colligational patterns of 3-grams are analyzed to identify how morphosyntacic structure plays a role on the emergence of lexical frames as continuous (uninterrupted) and discontinuous sequences of multi-words.

2. CORPUS APPROACHES TO THE STUDIES OF MULTI-WORD SEQUENCES

In the late 1990s due to advancement in computers and their use in the analysis of language corpora, multi-word sequences have been studied empirically. For doing such an empirical research, Weinert (1995, p. 182) identifies two basic issues: (i) the best way to define and identify fixed multi-word units, and (ii) analysis of the discourse functions that these multiword units perform. Still these issues are considered to be the motivating force of the current studies. Although these empirical studies (e.g. Renouf & Sinclair, 1991; Nattinger & DeCarrico, 1992; Altenberg 1998; Aijmer, 1996; Granger, 1998; Moon, 1998; Partington, 1998; Hunston & Francis, 1999; Schmitt, 2004) highlight the significance of multi-word units, they differ in terms of "the research goals, the criteria used to identify multi-word units, the formal characteristics of multi-word units studied, the text samples drawn on, whether or not register comparison are made" (Biber, Conrad & Cortes, 2004, p. 372). These methodological parameters constitute the basis of corpus-based and corpus-driven studies done in the field of phraseology as summarized well enough in the table below by Gray & Biber (2013, p. 126).

Table 1. Design parameters of corpus-based and corpus-driven studies of 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"

The following brief review of the studies use corpus approaches to MWUs refers to the considerations summarized in Table (1).

2.1. CORPUS-BASED AND CORPUS-DRIVEN STUDIES OF MWUS

Intuitive approach in the analysis of formulaic language has a long tradition "with researchers making up lists of fixed expressions that they perceived as occurring frequently in the language" (Cortes, 2013, p. 34). For example, Pawley & Syder (1983) emphasize the importance of prefabricated language by making a long list of short and long expressions "which these authors perceived as frequent formulaic expressions in that geographical register" (Cortes, 2013, p.34). Using the frequency-based tradition, some studies have surveyed the literature on the occurrences of formulaic expressions and checked their frequency list in a corpus (Nattinger & DeCarrico, 1992). Yet, there have been few corpus-based studies to explore the use of specific multi-word units identified by earlier researches "mostly because

corpus linguists have not been convinced of the validity of the phrase lists proposed on an intuitive basis" (Gray & Biber, 2015, pp. 127-128).

On the contrary, there has been plenty of research applying some form of corpus-driven methodology. Here, corpus itself is analyzed inductively by utilizing software that automatically identifies multi-word expressions across the corpus texts or its relevant sub-corpora. Salem (1987) is one of the first studies in using corpus-driven approach to identify recurrent lexical phrases in French government documents. Altenberg (1998) was considered also to be the first study that examines recurrent phrases in spoken English on the basis of London-Lund Corpus. Using the data in the London-Lund Corpus, Eeg-Olofsson & Altenberg (1994) also conducted corpus-driven research to analyze discontinuous sequences for the first time. In this innovative study, they explore new computational and statistical techniques to analyze word combinations in the corpus. Butler (1997) adopts a similar approach to investigate 28 discontinuous frames in a corpus of Spanish texts. Around the same time, Biber et al. (1999) documented the most common lexical bundles in spoken and written registers. This study was distinctive in terms of adopting a register perspective, analyzing a large corpus consists of 5 million words for each register, using a frequency-based approach in the identification of multi-word units and focusing on longer multi-word units such as 4, 5, and 6-word sequences. Biber et al. (1999)'s analytic framework has lead to other register specific research. Biber, Conrad & Cortes (2004) compared the distribution, formal and functional characteristics of lexical bundle in four registers: conversation, university classroom teaching, university textbooks, and published academic articles; Partington & Morley (2004) examine the use of multi-words in spoken political discourse; Carter & McCarthy (2006) examine and list the functions of clusters in spoken and written discourse; Biber & Barbieri (2007) identify and describe the use of lexical bundles in written course syllabi and spoken advising sessions; Csomay (2013) focuses on the distribution of types of lexical bundles in spoken lectures; Hyland (2008) and Cortes (2013) report the discourse functions of multi-word units in written academic registers making comparisons across academic disciplines; Jablonkai (2010) studies the function of lexical bundles in English EU documents.

What is striking is that there has been a particular interest in multi-word units in academic register. The application of the findings of corpus research in the field of teaching and learning can be seen in these studies. Cortes (2004), for example, compared the use of lexical bundles by university students and published research articles in the field of history and biology. Comparing the use of multi-words by native-English and non-native English students' writings is the topic of several studies (e.g. Chen & Baker, 2010; Adel & Erman, 2012; Staples et al., 2013) whose purpose is to explore the patterns of language development in the use of these units. A part from comparing students' writing, Pan, Reppen & Biber (2016) compared the use of lexical bundles by L1-English versus L2-English academic professionals. They investigate the structural and functional types of lexical bundles utilized by L1 English and L1 Chinese professionals writing for English medium Telecommunications journals.

Most studies above have focused on continuous sequence of MWUs. However, researchers have investigated fixed discontinuous sequences of words which is defined as "recurrent word forming a "frame" for variable slots (e.g. too__to___)" (Gray & Biber, 2015, p. 132). Among several reseaches, Renouf & Sinclair (1991) were the first corpus-based study to analyse variable fillers in discontionus multi-words, referred to as "collactional framework". They determined seven specific collactional framework and find out the most common fillers in each frame. In the same vein, Marco (2000) found out that specific genres, which is medical journal articles, "attract particular types of frameworks, and shows that such frameworks can be related to the types of meaning that are important to the register involved" (Vincent, 2013, p.45). Stubbs (2007) proposed the term "phrase-frame (p-frame)" to investigate which one of the item is free to vary in a lexical phrase. Focusing on such p-frames helps us capture the greater variation in phraseology and also reveals which types of frames are commonly found in a particular register. Biber (2009), for example, compared conversation and academic writing and maintained that the most frequent 4-words academic p-frames consist of discontinuous sequences composed of closed class items with an internal slot, such as the * of the. Gray & Biber (2013) extended Biber's (2009) study by applying a corpus-driven approach to identify discontinuous frames. Römer (2010) also followed the same methodological approach to investigate frequent discontinuous sequeces in a corpus of academic book reviews.

2.2. CORPUS-BASED AND CORPUS-DRIVEN STUDIES OF TURKISH MWUS

Research on Turkish MWUs can be classified under two disciplines: studies on natural language processing (NLP) (e.g. Oflazer, Çetinoğlu & Say, 2004; Kumova-Metin & Karaoğlan, 2011) and the ones conducted in linguistics which aim to identify and describe multi-word expressions with their discourse functions. To identify formal and functional properties of MWUs as well as to comment on methodological challenges in extracting them, corpus-based and corpus-driven and hybrid studies have been carrid out lately. In this respect, Mersinli (2015) explores linguistic relevance of MWU ranking of 12 associative measures that Text::NSP contain on 10-million-word Baby Turkish Nationl Corpus (TNC). Mersinli and Aksan (2016) discuss methodological considerations to clarify appropriate processes for Turkish MWU extraction considering the agglutinative nature of Turkish by using corpus-driven methodology. Durrant (2013) following a hybrid approach, combining corpus-based and corpus-driven methodologies, 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. Again utilizing a hybrid approach, Aksan and Aksan (2015a,b) present, for the first time in Turkish, the emerging formal categories and internal structure of MWUs and their primary discourse functions adopting the framework of Biber, Conrad & Cortes 2004 on two domains of the TNC namely imaginary and informative texts. They focus on 2-grams and 3-grams in both continuous and discontinuous sequences. These studies also demonstrate the register/genre specificity of multi-words identified for fiction and informative written text in Turkish. In a more recent study Yıldız (2016) investigates the structural pattern and discourse functions of the most frequent 50 3-grams in the construction of academic texts as a register 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. This study follows the framework set by Hyland (2008) to investigate the discourse functions in academic register in Turkish. Once again, a hybrid approach is adopted in the analysis.

3. DATA AND METHODOLOGY

3.1. THE CORPUS

This study has used the data coming from 10-million-word sub-corpus of the TNC (Aksan et al., 2012; Aksan et al., 2016), namely TNC-Baby which is constructed following the design principles of the TNC. In this sense, it is a small size general corpus of contemporary Turkish. The 50 million-words size of the TNC is reduced to 10-million-words by preserving the quantificational distribution of the texts. The distribution of number of words in the corpus is determined proportionally for each text domain, time, and medium of text following the model of TNC. The whole corpus is sentence-splitted whose sentence boundary detection was automatically made by the software GENIA Sentence Splitter (GeniaSS) (Kim et al., 2003) and checked manually through the lines involving two or more combined sentences (Demirhan, 2013). Thus, sentence boundary detection made us observe the phraseology emerged in the combination of lexical units to form a cluster or candidates for multi-word expressions. Representativeness and balance of the sub-corpus is ensured by including a wide range of texts through equally sized samples. Overall, TNC-Baby contains samples from 1.413 different (1.055 written, 358 spoken) written and spoken texts. Detailed distribution of the content of TNC-Baby is seen in Tables 2 to 4 below.

Table 2. Domain-based distribution

Domain	Percentage	Total number of words
1. Imaginative Prose	19%	1.901.174
2. Informative Texts	81%	7.956.406

Table 3. Distribution of informative texts according to the domains

Domain	Percentage	Number of words
1. Informative: Natural and pure sciences	5,03%	400.207
2. Informative: Applied science	10,21%	812.349
3. Informative: Social science	20,08%	1.597.646
4. Informative: World affairs	22,57%	1.795.761
5. Informative: Arts	8,78%	698.572
6. Informative: Belief and thought	5,00%	397.820
7. Informative: Leisure	18,29%	1.455.226
8. Informative: Commerce and finance	10,04%	798.823

Table 4. Distribution of the texts according to the media

Media	Percentage	Total number of words
1. Books	46,1	3.667.944
2. Periodicals	37,1	2.951.859
2.1. Journals	14,9	1.185.466
2.2. Newspapers	11,1	883.176
2.3. Magazines	11,1	883.217
3. Other published written material	6,09	484.550
4. Unpublished written material	2,5	198.912
5. Spoken texts	8,21	653.228

3.2. EXTRACTION OF MWUS

As the case for most of the NLP studies, MWU extraction also relies on rule-based and statistical methods. For agglutinative languages, it is considered as a must to use hybrid strategies since word-forms are rarely core lexical units as in English and can sometimes form full sentences as in *gidecekler* "they will go".

Another justification for a hybrid methodology is that MWU formation cannot be explained solely by associative strengths of the given candidates. Numerous statistical formulas are implemented in the literature to reach a more accurate sorting of n-grams or MWU candidates but the case is still problematic especially in non-English languages. The problem here is that, languages like Turkish do not operate on word-forms but rather on lemmas and mostly inflectional suffixations, which makes the space character, that most of the statistical studies are based on as a delimiter, irrelevant and unreliable. Thus, in this paper, a morphosyntactic filtering is argued to accompany the frequency effect of overtly used word-form combinations, in other words, n-grams or MWU candidates.

The dual nature of the lexicalization of MWUs, which is the underlying assumption for this paper is given in Figure 1.

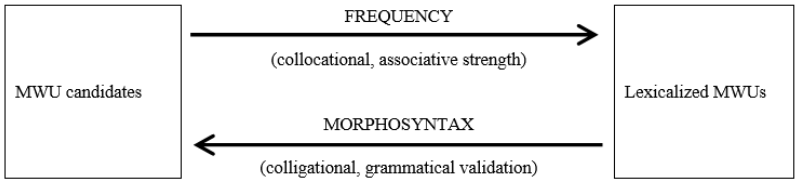

Figure 1. Dual nature of MWU formation

In this respect, the extraction process of collocations (MWU candidates) in Turkish is based on word-forms, i.e. any inflected or bare form of free morphemes delimited by a space character in written texts. The tool used for this first step is Text-NSP (Banerjee & Pedersen, 2011), which also provides frequency info of the extracted MWU candidates.

However, the second step, in which we have tagged these candidates by using TNC-tagger, is based on words/lemmas, i.e. 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 NLP dictionary of TNC. These entries include all information regarding the lemma, the part of speech (POS) and the inflectional suffixes that are observed in each word-form of the given collocation.

Finally, the colligations (grammatical patterns) of these word-forms, i.e. the morphosyntactic information for the collocations, are semi-automatically classified and validated by the researchers. The frequency of these colligations is also calculated in this final step. A sorting of these colligations according to their observational frequency provide an overview of the constraints that are governing the MWU lexicalization in Turkish.

An example collocation and colligation extracted through this process is given in (1).

(1)	Collocation:	*kısa*	*bir*	*süre*
		short	a	time
		"for a short time"		
	Colligation:	AJ,bare	DT,bare	NN,bare

4. COLLIGATIONAL RANKINGS OF 3-GRAMS

In this part of the study, the colligational patterns of 3-grams extracted from the TNC will be discussed under 3 titles. First group of colligations are full grammatical patterns of 3-grams which include lexical categories and inflectional suffixes. Second group excludes lexical categories and focuses on the inflectional morphology of each word-form in a given 3-gram. Turkish has two inflectional paradigms, i.e. verbal and nominal, and nominal inflection can occur on any non-verb stem. This phenomenon requires a separate analysis of inflectional suffixes excluding part-of-speech data. Third title in our analysis includes only lexical categories which may provide insight on overall lexical tendencies of MWU-formation. Finally, a sub-section is devoted to lemmas for a semantic discussion of 3-grams in Turkish.

4.1. CONTINUOUS AND DISCONTINOUS COLLIGATIONAL STRINGS

When all grammatical information and their sequences are ranked by their observed frequencies, the first observation to mention is that MWUs in Turkish are mostly composed of non-inflected word-forms or have empty morphemes such as the nominal case suffix. This tendency implies that MWU formation, which is in the blurry area between lexicon and grammar, mostly relies on lexical relations but not on grammatical operations. Table 5 lists the most frequent 10 continous and discontinuous sequences of colligations and the slight difference in number of occurrences of each item, implies that the target lexical category may be more important in a lexical analysis. Just in the same manner as derivational suffixes, we may classify these colligations as noun-forming colligations or adverb-forming colligations in future studies.

Table 5. Most frequent continuous and discontinuous sequences of colligations

	Colligation	Turkish	English	Freq.
1	AV,bare__AJ,bare__DT,bare	*çok önemli bir*	a very important	5292
2	AJ,bare__DT,bare__NN,nom	*kısa bir süre*	a short time	4809
3	NN,nom__CJ,bare__NN,nom	*radyo ve televizyon*	radio and television	4660
4	DT,bare__NN,nom__AV,bare	*bir süre sonra*	after a while	4525
5	AJ,bare__CJ,bare__AJ,bare	*ekonomik ve*	economic	3193

	Colligation	Turkish	English	Freq.
		sosyal	and social	
6	CJ,bare__AV,bare__AV,bare	*ama yine de*	but still	2680
7	NN,nom__NN,nom__CJ,bare	*ne var ki*	however, yet	2390
8	AJ,bare__DT,bare__NN,loc	*etkin bir şekilde*	efficiently	2216
9	AV,bare__DT,bare__NN,nom	*böyle bir şey*	such a thing	2204
10	CJ,bare__AJ,bare__DT,bare	*ile ilgili bir*	a … related to	1892

4.2. INFLECTIONAL SEQUENCES

An analysis of inflectional sequences supports the argument that, in order to be lexicalized, a MWU candidate should include as few inflectional suffixes as possible. If any inflection cannot be avoided, this would mostly be compounder -sI or case markers as seen in Table 6.

Table 6. Most frequent suffixes

	Inflections	Turkish	English	Freq.
1	bare_bare_bare	*çok önemli bir*	a very important	95565
2	bare_bare_loc	*etkin bir şekilde*	in an efficient way	6231
3	bare_bare_comp	*büyük millet meclisi*	grand assembly	5567
4	bare_bare_avrek	*bir araç olarak*	as a means	3055
5	bare_abl_bare	*bir yandan da*	on the other hand	2732
6	bare_dat_bare	*o güne kadar*	till that time	2556
7	bare_loc_bare	*bu konuda da*	in this respect	2380
8	bare_bare_ins	*başka bir deyişle*	in other words	2250
9	bare_bare_comp	*genel başkan yardımcısı*	vice chairman	1944
10	bare_comp_bare	*iş doyumu ve*	job satisfaction and	1672

4.3. LEXICAL CATEGORIES

The combinations of lexical categories provide valuable data on how nominals are the dominant POS for both the words internal constitute of a multi-word and also the target function of the given MWU. In other words, noun is the basic category both as a source and also as target POS in MWU-formation. The only verbs among the top 10 POS sequences are category changing inflection of light verb *-ol* ‘be’ in Turkish, which is a verb mostly serves as buffer lemma in certain inflections.

Table 7. Most frequent POS sequences

	Inflections	Turkish	English	Freq.
1	AJ_DT_NN	*kısa bir süre*	a short time	9174
2	NN_NN_NN	*büyükşehir belediye başkanı*	metropolitan mayor	8083
3	NN_CJ_NN	*radyo ve televizyon*	radio and television	6946
4	DT_NN_AV	*bir süre sonra*	after a ... time	6483
5	DT_NN_NN	*bir şey yok*	there's nothing ...	6352
6	DT_NN_VB	*bir araç olarak*	as a means	5637
7	AV_AJ_DT	*çok önemli bir*	a very important ...	5398
8	NN_NN_VB	*söz konusu olan*	the given ...	5361
9	AJ_NN_NN	*büyük millet meclisi*	grand assembly	4005
10	NN_NN_CJ	*ne var ki*	however	3648

4.4. LEMMAS

Most frequent lemmas observed in MWU-formation in Turkish are general nouns, light verbs, auxiliary verbs, adjectives, time adverbials and first person pronouns as can be seen in Table 8. This ranking is also compatible with overall rankings of postpositions in Turkish which supports the argument that the overall frequency of a given word lemma, possibly includes its uses in bigger multi-words. This argument also leads to another one that can be formulated as; word and multi-word frequency information cannot be studied in isolation and are strongly related to each other.

Table 8. Most frequent lemmas

	NOUN	ADJ.	VERB	ADV.	CONJ.	DET.	PRON.	POSP.
1	şey *thing*	önemli *important*	ol *be, become*	da, de *too, also*	ve *and*	bir *a/an*	bu *this*	için *for*
2	ne *what*	büyük *big*	et *do, act*	daha *more*	ya *or*	her *every*	o *that*	kadar *as ... as*
3	zaman *time*	iyi *good*	al *take*	çok *very*	ki *that -sub*	başka *other*	ben *I*	gibi *similar to*
4	yıl *year*	az *little, less*	yap *make*	en *most*	ile *with, by*	diğer *other*	şu *that*	üzere *in order to*
5	konu *subject*	aynı *same*	çık *leave*	sonra *then*	ama *but*	ilk *first*	sen *you*	göre *according to*
6	var *there's*	son *last*	gel *come*	önce *before*	hem *but also*	tam *exactly*	biz *we*	yana *aside from*
7	gün *day*	yeni *new*	gerek *be necessary*	böyle *as this*	diye *that -sub*	tek *the only*	biri *someone*	bağlı *related to*
8	yer *place*	genel *general*	geliş *develop*	hiç *no more*	veya *or*	bütün *all*	kim *who*	yüzden *because of*
9	süre *time*	ilgili *related*	gör *see*	hemen *immediately*	ancak *but*	hiçbir *no ...*	çoğu *most*	bile *even if*
10	ara *place*	kısa *short*	i *verbal p.*	herhangi *any*	yani *in short*	birkaç *some*	hepsi *all*	rağmen *although*

As a summary of the discussions above, we can conclude that; MWUs and their frequencies in Turkish;

1. mostly include bare forms rather than inflected.
2. are mostly observed in 3-grams, as closed projections.
3. rarely occur in the verbal paradigm.
4. are mostly noun phrases.
5. necessitate a re-examination of word frequencies.

5. STRUCTURAL TYPOLOGY OF MWUS IN TURKISH

In this part of the study the colligational patterns (morphosyntactic internal organization) of the MWUs is analyzed to identify how morphosyntacic structure plays role on the emergence of continuous sequences (full lexicalized MWUs) and discontinuous sequences (incomplete fragments). Based on this observation tendencies in morphosyntactic uses that govern MWUs formation is highlighted. Steps to be followed is summarized as such. First, we classified 4000 candidates of MWUs into two categories on the basis of their structural unity. These are (1) multi-word units with complete structures or continuous sequences (e.g. *etkin bir şekilde* 'in an efficient way'); (2) MWUs with incomplete structures or discontinuous sequences (e.g. *çok önemli* bir 'a very important'). Then, structural typology of 3-grams referring to their recurrent grammatical categories proposed by Aksan & Aksan (2015) is used. We analyzed word-based colligational rank frequency data employing the structural description of tri-grams to identify some of the outstanding morphosyntactic uses and the associated lexical frames emerge across the word-based colligations. We use the term 'frame' in a general sense not in a rigorous and statistical sense as in Biber (2009) and Gray & Biber (2013). By lexical frame we simply refer to our initial observations on multi-word formulaic sequences (e.g. *ne olursa olsun* 'in any case'), formulaic frames with variable slots (e.g. *ne kadar* * 'how *' as in *ne kadar güzel* 'how beautiful) along with fixed discontinuous sequences (e.g. *için * bir* 'for * a' as in *için önemli bir* 'for an important'). Note that asterisk is used to represent variable slots.

5.1. A STRUCTURAL TYPOLOGY OF TURKISH TRI-GRAMS

Aksan & Aksan (2015, pp. 7-10) is the first study in Turkish that define types of structures realized in 3-grams across the corpora by following the framework of Biber, Conrad & Cortes' (2004) classification. They propose 8 classes as, noun phrases (NPs) and noun phrase (NP) fragments; postpositional phrases (PPs); degree expressions; conjunctive patterns; *Ne* 'wh'-patterns; modality patterns; copular/existential construction and quotatives as shown in Table 9.

Table 9. Structural Types of tri-grams

TYPE I : NPs and NP-fragments

I.1 Indefinite NP fragments : degree+adjective+indefinite article
daha büyük bir 'something much bigger'
I.2 Indefinite NPs: adjective/demonstrative+indefinite article+*(some)thing*
kötü/öyle bir şey 'something bad/like that'
I.3 Indefinite NPs : Adjective+InArt+Head Noun
kısa/uzun bir süre 'for a short/long time'

TYPE II. Postpositonal Phrases

II.1 PPs with Indefinite NP complements: InArt+Noun+Postposition
bir süre/önce/sonra/için 'before/after/for a while'
bir an için/önce 'just for a moment / immediately'

II.2 PPs with oblique NP complements: demonstrative/quantifier+Noun+ postposition
*her şey**den** önce* 'first of all'
*o gün**den** sonra* 'ever after'
*o gün**e** kadar* 'until that day'
*her zaman**ki** gibi* 'as usual'
*başta ol**mak** üzere* 'as the first'

II.3 Postposition without complement combining following items: Postposition+participle/quantifier
için gerekli olan 'required by X'
için ne kadar 'how much for X'

TYPE III. Degree expressions

III.1 Adverbial *hiç* 'never, ever, no/any' patterns: ADV+Dem/InArt+N/P

hiç bir zaman	'never'
hiç bu kadar	'never that much'
hiç mi hiç	'not in the least'
bir daha hiç	'never again'
daha önce hiç	'never before'

III.2 Adverbial *çok* 'very' and *daha* 'more' patterns: ADV+ADV+ADJ

çok daha fazla	'much more'
çok daha iyi	'much better'
hem de çok	'even more'
o kadar çok	'that much'
bir kere/kez daha	'one more time'

TYPE IV. Conjunctive patterns

IV.1 Conjunctive *ve* 'and' patterns: CONJ+fragment from second conjunct

ve bir daha	'and once more'
ve bu arada	'and meanwhile"
ve bu nedenle	'and for this reason'
ve sonra da	'and after'

IV.2 Disjunctive *ya da* 'or' patterns: Disjunctive+demonstrative/determiner

ya da başka	'or another'
ya da bir	'or a/one'
ya da böyle	'or thus/in this manner'
ya da bu	'or this'
ya da daha	'or more'

IV.3 Additive *da* patterns: Adverbials+*da* 'additive'

bu kez de	'and this time'
bu nedenle de	'and for this reason'
daha önce/sonra da	'and even before/after'
diğer yandan da	'and on the other hand'
bir yandan da	'and on the other hand'

IV.4 Disjunctive *ama* 'but, however' patterns: Disjunctive+adverbials

ama bir türlü	'but in no way'
ama bu kez	'but this time'
ama gene de	'but still/yet/nevertheless'

TYPE V. *Ne*-patterns (wh-patterns): *ne*+conditional/adverbial/PRT

ne de olsa	'after all'
ne olursa olsun	'in any case'
ne kadar çok	'the more'
her ne kadar	'although'
ne var ki	'however'
ne yazık ki	'unfortunately'

TYPE VI. Modality patterns: modal adverb+particle+(demonstrative)

belki de bu	'maybe/perhaps this'
belki de en	'maybe/perhaps the most'
kim bilir belki	'who knows maybe/perhaps'

TYPE VII. Copular/existential constructions

VII.1 Linking: bir *(some)thing*+negative/become
bir şey değil/ol-du 'it is not something; something happened'
VII.2 Existential constructions: bir *(some)thing*+var/yok

bir şey vardı/yoktu	'there was something/nothing'

TYPE VIII. Quotatives

dedi kendi kendine	'said to her/himself'
dedim kendi kendime	'said to myself'
diye geçirdi içinden	's/he thought'
diye bir şey	'something called'

According to this classification, most of the MWUs are NPs or NP fragments, as similar in English. The listed types are almost exclusively NPs, yet more categories are identified to underscore NPs special role in the text due to their respective frequencies in the text. "For example, degree expressions and quantifiers as well as demonstratives are in fact NP elements. Similarly, those that combine with conjunctions are also part of the following NP or NP fragments" (Aksan & Aksan, 2015, p. 7). Furthermore, the above classification shows that tri-grams with a verbal element, excluding light verb constructions are quite rare in Turkish when compared to English. "This is probably due to the nature of functional categories in Turkish: those that would appear with verb are generally bound affixes rather than free words in their written forms fragments" (Aksan & Aksan, 2015, p. 10). All forms of tri-grams are composed either entirely or partially with function words. Those that are not function words,

undergo semantic bleaching and form non-compositional formulaic expressions. NPs and PPs are the most common in Turkish as it appears to be the case in English as well.

5.2. STRUCTURAL TYPOLOGY AND MWU CATEGORIES

When a general observation on the formation of continuous (uninterrupted) and discontinuous sequences of MWUs and structural typology are made, we first focus on less frequent structural types of 5 and 7; then move on the frequent ones especially NPs, PPs and conjunctives. We should note that in the data under examination rank and frequency figures refer to the total occurrences of the tagged grammatical sequence of a MWU and its ranking. The samples fall under this sequence having its own frequency figures. For instance, word-based colligation DT,bare_NN,nom_NN,nom sequence ranks 16 with a frequency of 1525; the sample *bir şey yok* 'there is nothing/no problem' occurs <270> times across the 10-million-word corpus of TNC-Baby. While discussing the sequences the most recurrent samples are primarily chosen.

5.2.1. COPULAR/EXISTENTIAL STRUCTURES AND NE-PATTERNS IN MWUS FORMATION

Copular/existential constructions are subsumed under two categories: Linking predicates (e.g. *değil* 'not', *ol-* 'to become' and existential constructions formed by *var* and *yok*. MWUs with these structures are not many in number and they form relatively fixed sequences which usually act as clause fragments in the texts as seen in the examples below.

(2) rank.16 - DT,bare_NN,nom_NN,nom – <freq. 1525>
bir şey yok 'there is nothing/no problem' <270>
bir şey var 'there is something that' <182>
bir şey değil 'lit. it is not something (important), not at all' <168>

(3) Bulgaristan çok ucuz bir ülke, ama alacak **bir şey yok** ülkede.
"Bulgaria is a cheap country, but there is nothing to buy."

(4) Bu, ayıp veya utanılacak **bir şey değil** ama hayatın zor bir gerçeği.
"This isn't something shameful or embarrassing, but a difficult reality of life."

(5) rank.26 - AV,bare_AJ,bare_NN,nom – <freq.946>
hiç önemli değil 'it doesn't have any importance' <26>
çok önemli değil 'it is not that much important <12>

(6) Fazla müzik aleti çalmak **hiç önemli değil**.
"To play lots of musical instruments is not so important."

(7) rank.119 - DT,bare_NN,nom_VB,past+3s – <freq.284>
bir şey oldu 'something happened' <67>

(8) O sırada hiç beklenmedik **bir şey oldu**.
"Meanwhile, something really unexpected happened."

From the corpus citations MWUs containing *var* is the most predominant among other predicates. Based on this property a lexical frame can be proposed as ***bir * var***, in which the attested content words occurring this frame are *ilişki* 'relation', *sorun* 'problem', *fark* 'difference', *yer* 'place', *nokta* 'point', iş 'job'', *konu* 'topic', *yol* 'way'. Note that other than copular/existential structures DT,bare_NN,nom_NN,nom and AV,bare_AJ,bare_NN,nom sequences give rise to MWs that can be classified under different structural typology. For instance, NPs with continuous MWUs such as, *bir bardak su* 'a glass of water' <27> or discontinuous MWU such as, *bir ilke imza* '(lit) a first signature; lead the way' <13>. We are not dealing with these structures in this section. Considering the DT,bare_NN, nomVB,past+3s sequence, it constructs lexical phrases primarily with *ol*-'to be; to become' as a predicate. However, we observe the use of different verbs (e.g. *yak*-'to light up', *sus*-'to keep quiet') and light verbs (*yap*-'to do; to make', *gel*-'to come; to happen by') other than *ol*-'to be; to become'. With these sequences, the only MWUs with verbs in Turkish emerge, such as *bir sigara yaktı* <41> 's/he lit a cigarette', *bir*

süre sustu <22>'s/he kept silence for a while', *bir şey geldi* <14>'something has come'.

Multi-words occur with the structure Ne-patterns (wh-patterns): *ne*+conditional/adverbial/PRT are the continuous sequences and usually function as the descriptive part of NPs (e.g. 9 below), or they are used as conjunctives, adverbials or clause fragments (e.g. 11) in a discourse.

(9) rank.54 - NN,nom_PP,bare_AJ,bare –< freq.545>
ne kadar güzel 'how beautiful' <79>
ne kadar önemli 'how important'<70>
ne kadar iyi 'how good' <53>
ne denli önemli 'how important'<19>

(10) Hala düşünebilmek ve soru sormak **ne kadar güzel**.
"How beautiful it is to be still able to think and ask questions."

(11) rank.64 - NN,nom VB,aor+vi+avsa+3s_VB,imp3 – <freq.453>
ne olursa olsun 'whatever the consequences are' <437>
ne yaparsa yapsın 'whatever he does' <16>

In NN,nom_PP,bare_AJ,bare sequence, out of 20 occurrences 18 of them are identified as in ***ne kadar** ** lexical frame which involves the following descriptor and classifier adjectives (Biber et al., 1999)5 as content words: *güzel* 'beautiful', *önemli* 'important', *iyi* 'good', *büyük* 'big', *zor* 'difficult', *doğru* 'right', *farklı* 'different', *yakın* 'close', *küçük* 'small', *uzak* 'far', *etkili* 'efficiant, *yanlış* 'wrong', *güçlü* 'strong', *ciddi* 'serious', *uzun* 'long', *kötü* 'bad', *başarılı* 'successful', mutlu 'happy'. The sequence of NN, nom_VB, aor+vi+avsa+3s_VB, imp3 leads to the formation of a fixed expression *ne olursa olsun* 'whatever the consequences are' with the frequency of 437 which

[5] Biber et al. (1999) define the semantic grouping of adjectives as such: "Descriptors are prototypical adjectives denoting such features as color, size, weight, chronology and age, emotion, and a wide range of other characteristics. (…) Classifiers can be grouped into subclasses, including relational, affiliative, and a miscellaneous topical class" (p. 509).

outnumbers other MWUs fall under the same sequence.

5.2.2. NPs, POSTPOSITIONAL PHRASES AND CONJUNCTIVE PATTERNS IN MWUS FORMATION

In this part of the paper, MWUs categorized under the structural typology of 1, 2, 4 are examined. Their role to produce continuous (uninterrupted, lexicalized) and discontinuous (incomplete) MWUs are discussed referring to the emerged lexical frames and word class of such units.

5.2.2.1. CONTINUOUS (UNINTERRUPTED) SEQUENCES

Multi-words with indefinite NPs

(12) rank.2 - AJ+bare_DT+bare_NN+nom - <freq.4809>
kısa bir süre 'a short time' <425>
önemli bir rol 'an important role'<142>

(13) Bernard Brodie, resmi stratejilerin oluşturulmasında da **kısa bir süre** görev almıştır.
"Bernard Brodie, has also been on duty for establishing official strategies."

(14) rank.9 - AV,bare_DT,bare_NN,nom- <freq.2204>
böyle bir şey 'such a thing' <299>
hiç bir şey 'nothing'<109>

(15) rank.19 - DT,bare_DT,bare_NN,nom – <freq.1418>
başka bir şey 'another thing' <628>
başka bir ifade 'another expression' <53>

Indefinite NP constructions above constitute complete MWUs which serve as NPs, manner and temporal adverbials mainly. We observe that among the MWUs formed with AJ+bare_DT+bare_NN+nom colligation the following determiner+noun combinations are forming the basis of lexical frames such as, **** bir süre***, **** bir zaman***, **** bir şey***

and ***önemli bir** ** which give rise to the productive and recurrent use of MWUs. The most common citations are, *kısa bir süre* ‘ a short period’<425>, *uzun bir süre* ‘a long period’<137>, *belli bir süre* ‘a definite period’ <53>, *bellirli bir süre* ‘a specific period’<48>; *kısa bir zaman* ‘ a short time’<49>, *belli bir zaman* ‘a definite time’ <48>, *uzun bir zaman* ‘a long time’<45>; *fazla bir şey* ‘something more’<96>, *yeni bir şey* ‘something new’<73>, *kötü bir şey* ‘something bad’<73>, *iyi bir şey* ‘something good’<54>; *önemli bir rol* ‘an important role’<142>, *önemli bir yer* ‘an important place’<138>, *önemli bir şey* ‘an important thing’ <65>, *önemli bir sorun* ‘an important problem’<51>, *önemli bir nokta* ‘an important point’<50>, *önemli bir adım* ‘an important step’ <42>. We should note that the same colligational string forms NPs with a wide variety of adjectives and nouns reflecting the subject matter of the corpus texts. Some of the examples contain *anlamlı bir ilişki* ‘a significant relation’ <61>, *son bir kez* ‘finally’<56>, *önemli bir adım* ‘an important step’ <42> and *yeni bir dünya* ‘a new world’<40>. By employing the colligational sequence AV,bare_DT,bare_NN,nom MWUs are produced functioning as indefinite determiner (*böyle bir şey* ‘such a thing’), pronouns (*hiç bir şey* ‘nothing’) and temporal adverbials (*hemen her zaman* ‘almost always’). Finally, DT,bare_DT,bare_NN,nom string displays a very interesting patterning regarding the syntagmatic association of the words. Seemingly synonymous two words (i.e. *başka* ‘different, other’ and *diğer* ‘other’) build the lexical frames as such ***başka bir ****, ***diğer bir ****, ***bir başka ****, ***bir diğer **** and the pattering with ***tek bir **** and ***bir tek ****. Following examples illustrate this case along with the preference of one order to another with reference to frequency of occurrence of the multi-words: *başka bir şey* ‘another thing’ <628>, *bir başka şey* ‘the other thing’<21>, *diğer bir husus* ‘another topic’<19>, *bir diğer husus* ‘another topic’<11>, *bir tek şey* ‘only thing’ <16> and *tek bir şey* ‘one thing’<15>.

Multi-words with locative marked NPs

(16) rank.8 – AJ,bare_DT,bare_NN,loc – <freq.2216>
etkin bir şekilde ‘in an efficient way’< 113>
açık bir şekilde ‘apparently’<103>
hızlı bir şekilde ‘in a fast way’<102>
etkin bir biçimde ‘in an efficient manner’<62>

kısa bir sürede 'in a short time' <52>

(17) Olayı etkin bir şekilde izleyecek kimse de yoktu.
"There was no one to follow the event in an efficient way."

What is striking with locative marked NPs is that they serve as manner and temporal adverbials and systematically appear in the form of lexical frames as **** bir şekilde*** and **** bir biçimde***. Out of 81 multi-words the ones that contain the lexical item *şekilde* 'in the way' are used 53 times and those that are formed with *biçimde* 'in the manner' are used 22 times.

Multi-words with instrumental marked NPs

(18) rank.22 - DT,bare_DT,bare_NN,ins – <freq. 1248>
başka bir deyişle 'in other words' <336>
bir başka deyişle 'in another words' <307>
diğer bir deyişle 'to put it differently' <226>

(19) rank.61 - AJ,bare_DT,bare_NN,ins – <freq.489>
büyük bir olasılıkla 'most likely'<83>
büyük bir ihtimalle 'probably'<49>
büyük bir dikkatle 'with great attention' <26>

Multi-words involving instrumental marked NPs either serve as conjunctions or as non-compositional formulaic units or as manner adverbials. Sequences having the colligational combination of (18) follows the similar manner in the production of MWs in terms of ordering the items as in (15). Out of 10 multi-words 7 of them involve the lexical frame ***başka bir **** and ***bir başka ****. In (19) the lexical frame ***büyük bir **** leads the list by forming specified morphosyntactic units. Out of 23 occurrences 10 of them encompass ***büyük bir **** frame.

Multi-words with VB, avrek

(20) rank.17 - DT,bare_NN,nom_VB,avrek – <freq.1516>
bir araç olarak 'being as a tool'<86>

bir sorun olarak 'being as a problem'<69>
bir varlık olarak 'being as an entity'<63>

(21) rank.123 - AV,bare_AJ,bare_VB,avrek – <freq.270>
daha ayrıntılı olarak 'being more detailed'<28>
en son olarak 'being the last'<20>

(22) Devlet ve halk arasında uzaklığın kaldırılması bir sorun olarak aydınların gündemine gelmiştir.
"That the distance between people and the state should be shortened, has been added to the agenda of intellectuals."

Multi-words serving as adverbials and including the unit VB,avrek lead to the formation of following lexical frames; ***bir * olarak***, ***daha * olarak***, ***en * olarak*** and ***çok * olarak***. Among them the most common frame in DT,bare_NN,nom_VB,avrek string is ***bir * olarak*** (66 out of 66 occurrences) in which a range of nouns, mostly topical or related to the subject matter of the corpus texts, occur in the missing slot of this frame.

Multi-words with definite NPs or NP fragments

Other colligational patterns producing complete sequences of multi-words are definite NPs (e.g. 23) and some definite NP fragments (e.g. 24) acting as modifiers of nouns as given below. In (23) the colligation string also encompasses a lexical frame of ***en önemli ** with a variety of nouns that complete the missing slot of the frame.

(23) rank.26 - AV,bare_AJ,bare_NN,nom – <freq. 946>
en önemli nokta 'the most important point' <34>
en önemli sorun 'the most important problem'<29>

(24) rank.5 - AJ,bare_CJ,bare_AJ,bare – <freq. 3193>
ekonomik ve sosyal 'economical and social'<209>
sosyal ve kültürel 'social and cultural'<136>

(25) **Ekonomik ve sosyal** kayıplar da insan kaybı kadar ağırdır.
"Economic and social loss is as destructive as human loss."

Multi-words with postpositional phrases and degree expression

Among the colligational patterns the ones that are constituted with postpositions give rise to MWs with adverbial function mostly. For instance, one of the most frequent string DT,bare_NN,nom_AV,bare in (26) contains MWs with PP *sonra* 'after', *önce* 'before' and it also involves MWs with degree adverbs such as *daha* 'more' as in below. Similar to this pattern, (27) consists of a variety of postpositions, such as *kadar* 'until', *için* 'for', *gibi* 'like' yet a formulaic expression *her ne kadar* 'although' does exist with a frequency of 648. For (26) and (27) the lexical frames ***bir * önce***, ***bir * sonra***, ***bir * için***, ***bir * gibi***, and ***bir * kadar*** can easily be generated. With subtle variation in structure such as, case assignment of nouns by postposition or the description of the noun in the PPs, we observe the production of MWs by colligational patterns as listed in (28), (29), (30) and (31) with decreasing rank order.

(26) rank.4 - DT,bare_NN,nom_AV,bare – <freq.4525>
bir süre sonra 'after a while'<768>
bir kez daha 'once more'<759>
bir an önce 'as soon as possible'<492>
bir kere daha 'once again'<137>

(27) rank.13 - DT,bare_NN,nom_PP,bare – <freq.1666>
her ne kadar 'although' <648>
bir süre için 'for a while'<133>
bir an bile 'not even a moment'<37>
bir çocuk gibi 'like a child'<47>

(28) rank.25 - NU,_NN,nom_AV,bare – <freq.996>
iki gün sonra 'after two days'<97>

(29) rank.49 - PN,bare_NN,dat_PP,bare – <freq.570>
o güne kadar 'till that day'<132>

(30) rank.68 - PN,bare_NN,nom_PP,bare – <freq.446>
bu iş için 'for this job'<77>

(31) rank.78 - PN,bare_NN,abl_AV,bare – <freq.372>
o günden sonra 'afte that day' <115>

Multi-words with additive -dA and postpositions

(32) rank.21 - DT,bare_NN,abl_AV,bare – <freq.1336>
bir yandan da 'besides'<547>
her şeyden önce 'first and foremost' <374>
diğer yandan da 'on the other hand' <131>
bir taraftan da 'in the mean time' <98>
diğer taraftan da 'on the other hand' <56>

The most frequent top 5 entry display that multi-words falling under the typology of additive *-dA* and postpositional phrase constitute formulaic expressions serving as conjunction or discourse connector in a text. Note that the interchangeability in the formation of MWs between the seemingly synonymous nouns *yan* 'side' and *taraf* 'side, way'. Yet corpus data shows that lexical phrases with *yan* (e.g. *bir yandan da* 'besides'<547>) are used more frequently than that of *taraf* (e.g. *bir taraftan da* 'in the mean time' <98>).

5.2.2.2. DISCONTINUOUS SEQUENCES

Multi-words in this category appear as NP fragments, part of conjunctive structures mostly occurring with additive –dA 'also' and as fragments of postpositional. The outstanding property of all the MWUs in this group is the absence of relevant components in either as the first segment or first and third segments of the sequence. Overall, discontinuous multi-word sequences either bridge two structural units (e.g. *için önemli bir* 'for an important'): they start at a clause or phrase boundary but the last words of the unit are the starting unit of a second grammatical structure or they link two phases (e.g. *çok büyük bir* 'a very big').

Multi-words with definite NPs or NP fragments

(33) rank.1 - AV,bare_AJ,bare_DT,bare – <freq.5292>

çok önemli bir ‘a very important’<496>
çok büyük bir ‘a very big’<312>
daha büyük bir ‘a more bigger’<164>

Missing units are head of NPs in almost all the occurrences. What is striking here is the following lexical frames are observed predominantly: ***çok * bir*** and ***daha * bir*** (40 units with *çok* ‘very’, 32 units involve *daha* ‘more’ out of 92). Almost all the adjectives in the variable slots are descriptors and a small number of them are classifiers.

(34) rank.7 -NN,nom_NN,nom_CJ,bare – <freq.2390>
ne var ki ‘however’ <745>
ne yazık ki ‘unfortunately’ <563>
temel hak ve ‘fundamental rights and’<144>
yer alan ve ‘to take place and’<68>
kamu kurum ve ‘state institutions and’<60>
anne baba ve ‘mother father and’ <28>

This sequence frequently produces nouns fragments which are part of 4-grams actually. For instance, the fragment *temel hak ve* is completed with *özgürlük* ‘freedom’ as a fixed expression. Or noun fragments with missing component which is filled by an element from a list reading structure as in *anne baba ve çocuk*. Out of 72 multi-words with this colligational sting 38 of them end with *ve* ‘and’. Along with incomplete MW production the same string can also produce complete MWs functioning as fixed expressions such as *ne var ki* ‘however’.

The following fragments act as a bridge in the construction of a sentence in which previous and following items of these fragments complete their meanings. As is noticed they are completed by subject NPs (e.g. in 36, *toplumla aile* (arasında bir ilişki); (in 40, *ekonomiye etkisi* (olan bir başka)) and VPs ((in 36, arasında bir ilişki) *kuruyorum*).

(35) rank.127 - NN,p3s+loc_DT,bare_NN,nom – <freq.265>
arasında bir ilişki ‘a relationship between’ <265>

(36) Toplumla aile arasında bir ilişki kuruyorum.
"I construct a relationship between society and the family."

(37) rank.169 - PN,bare_AV,bare_PN,bare– <freq.199>
biz de bu 'we also this' <50>
o da bu 's/he also this' <45>
biri de bu 'one of them also this' <21>
sen de bu 'you also this'<15>
bu da bu 'this also this' <15>
siz de bu 'you also this'<11>

(38) İşte biz de bu yarışmaya konuk olduk.
"Look, now we are also guests in this TV competition".

As is exemplified in (37) **** dA bu*** is appearing 6 times out of 10 entries so it can be treated as a discontinuous frame of PN,bare_AV,bare_PN,bare colligation.

(39) rank.181 - VB,pcan_DT,bare_DT,bare – <freq.187>
olan bir başka 'another … being ….'<45>
gereken bir diğer 'another … required to …'<22>

(40) Ekonomiye etkisi **olan bir başka** yanı vardır Gaziantep pasajlarının.
"Shopping malls of Gaziantep has another role, also influencing economy."

(41) rank.222 - PN,bare_NN,nom_PN,bare – <freq.160>
o zaman bu 'then/at that case this'<55>
o zaman o 'then/ at that case that' <30>

(42) Bırak **o zaman bu** mesleği
"Quit this job, then."

Out of 6 entries with this colligation ***o zaman ** appears with a variety of function words (e.g. bu, o etc.) 4 times so we consider it as a lexical frame of this discontinuous sequence.

Conjunctive patterns

Looking at the data below with fragments of multi-word sequences formed by conjunctive patterns, we detect that the most recurrent items are conjunctive *ve* 'and' patterns along with fragment from second conjuct (e.g. *ve daha sonra* 'and later'); disjunctive ya *da* 'or' patterns with demonstrative or determiner (e.g. *ya da daha* 'or more') and finally additive *-dA* patterns coming out as in adverbials.

(43) rank.6 CJ,bare_AV,bare_AV,bare <freq. 2680>
ve daha sonra 'and later' <245>
ya da daha 'or more' <174>
ya da çok 'or a lot' <119>

(44) rank.10 - CJ,bare_AJ,bare_DT,bare – <freq.1892>
ile ilgili bir 'with related to a'<123>
ve yeni bir 'and a new'<88>
ve belirli bir 'and a given'<84>

(45) Resim defterini açtı **ve yeni bir** sayfa çevirdi.
"She opened his sketch book and turned a new page."

(46) rank.14 - CJ,bare_AV,bare_DT,bare – <freq.1623>
ya da bir 'or a'<345>
ya da başka 'or different'<115>
ve böyle bir 'and such a' <62>

(47) Tarihçi, bir sorun **ya da bir** soruyla işe başlar.
"A historian starts with a problem or with a question."

(48) rank.15 -CJ,bare_AV,bare_AJ,bare – <freq.1575>
ve daha fazla 'and more' <77>
veya daha fazla 'or more' <57>
ya da olumsuz 'or negative' <57>

The majority of the discontinuous sequences with CJ,bare_AJ,bare_DT,bare colligation correspond to the frame ***ve * bir*** which appears 73 times of the 86 occurrences. Likewise with 50 out of the 81 occurrences ***ya da** **** is another lexical frame having the colligational pattern CJ,bare_AV,bare_AJ,bare.

(49) rank.124 - CJ,bare_PN,bare_NN,ins – <freq.268>
ve bu nedenle 'and because of this cause' <191>
ve bu amaçla 'and because of this purpose' <39>
ve bu suretle 'and because of this way' <20>
ve bu sebeple 'and because of this reason' <18>

The citations above is listing the all multi-word occurrences with relevant colligation and it is evident that ***ve bu **** constitute a frame with instrumental case marked NP fills the slot in fixed way. Semantically all the nouns in the slot refer to purpose, reason or cause of a reported events.

(50) rank.130 - NN,nom_AV,bare_AV,bare – <freq.262>
süre sonra da 'and after a while' <42>
yıl sonra da 'and after a year' <33>

(51) Bir **süre sonra da** Tercüme Bürosu üyeliğine getirildim.
"And after a while, I also became a member of the Translation Office."

Actually, the sequence *süre sonra da* is part of 4-grams which are *bir süre sonra da* 'and after a while' or *kısa süre sonra da* 'and after a short while'.

Postpositional phrases

(52) rank.20 - PP,bare_AJ,bare_DT,bare – <freq.1350>
için önemli bir 'for important a' <122>
kadar büyük bir 'as much big a' <75>
gibi önemli bir 'like important a'<38>

(53) Balkanlar ördek ve kazlar **için önemli bir** kışlama alanıdır.
"The Balkans is an important habitat for ducks and geese."

Half of the multi-words corresponding PP,bare_AJ,bare_DT,bare structure are formed by ***için * bir*** frame (24 out of 49).

(54) rank.30-PP,bare_AV,bare_AJ,bare – <freq.816>
için çok önemli 'for very important' <110>
için de geçerli 'valid for also' <84>
için en önemli 'for the most important'<63>

(55) rank.126-NN,nom_VB,pcdk+p3s_PP,bare – <freq.265>
zaman olduğu gibi 'as usual' <92>
ifade ettiği gibi 'as expressed' <46>

(56) Platon'un da **ifade ettiği gibi**, Felsefe Bilgisi'nin o çatı altında yer alması gerekiyor.
"As it is expressed by Plato, the knowledge of philosophy should be fall into that roof."

(57) Her **zaman olduğu gibi** sabahtan otele gittik.
"As we usually do, we arrived at the hotel in the morning."

Note that out of 92 citations 86 of them are starting with *her* 'every' and thus leading to a 4-gram as fixed expressions *her zaman olduğu gibi* 'as usual'.

The discontinuous multi-words above display the recurrent pattern that postpositions without complements combining following items: postposition+adjective+determiner/demonstative (e.g. *için önemli bir* 'for an important...'); postposition+quantifier+adjective (e.g. *için çok önemli* 'for very important) or participle +postposition (e.g. *ifade ettiği gibi* 'as it is said').

In considering the frames we identify, the structural property of them are determined by adopting the classification of Gray & Biber (2013, p.122). According this three-way classification, there are (i) Verb

based frames: frame contains one main verb or light verb (e.g. bir * olarak); (ii) Frames with other content words: frame contains one or more nouns, adjectives, adverbs but no verbs (e.g. ne kadar *, * bir süre, büyük bir *); (iii) Function word frames: frame consists of only function words such as prepositions, determiners, conjunctions, pronouns, etc. (e.g. bir * gibi, * dA bu). Table (10) summarizes the type of frames along with the corresponding continuous and discontinuous sequences determined in the peresent study.

Table 10. Types of frames in continuous and discontinuous sequences

Type of frame	Continuous sequence	Discontinuous sequence
Verb based	bir * var	—
	bir * olarak	
	daha * olarak	
	en * olarak	
Other content words	ne kadar *	çok * bir
	* bir süre	daha * bir
	* bir zaman	
	* bir şey	
	önemli bir *	
	başka bir *	
	diğer bir *	
	bir başka *	
	bir diğer *	
	tek bir *	
	bir tek *	
	* bir şekilde	
	* bir biçimde	
	başka bir *	
	bir başka *	
	büyük bir *	
	en önemli *	
Function words	bir * önce	* dA bu
	bir * sonra	ve * bir
	bir * için	ya da *
	bir * gibi	ve bu *
	bir * kadar	için * bir

6. CONCLUDING REMARKS AND FUTURE DIRECTIONS

In this study, we have provided a preliminary classification schema that can be applied prior to statistical ranking of the n-grams, collocations or MWU candidates extracted from a corpus. We have also demonstrated how to extract colligations from an annotated corpus and what kind of secondary data can be extracted from those colligational patterns. Moreover, we have argued that working on an annotated corpus, may significantly improve the precision of MWU extraction process in Turkish and contributed to the testing of hybrid, morphology involved approaches for MWU extraction in Turkish. The corpus-driven and frequency-based analysis that are followed in this paper lead us to examine a sample of colligational strings by utilizing the structural description of 3-grams to identify the prominent morphosyntactic tendencies and the lexical frames become apparent across the word-based colligations. Such analyses show tendencies for continuous and discontinuous MWU formations, but they are not enough to generate definitive rules. More in depth research should be done in the line that we have demonstrated in this paper. Followings are the suggestions for further studies.

- A MWU lexicon of Turkish should be formed by following an adequate and appropriate methodology.
- Systematic and quantitative research should be conducted to unveil the frames for discontinuous multi-word sequences with variable slots in Turkish.
- MWUs and their lexical frames should be studied and compared in both spoken and written registers of Turkish.
- MWU extraction studies should also include concerns on language teaching since Turkish language teaching without considering MWUs does not seem effective.
- Hybrid models including both statistical and structural/functional properties covering also intra-word components should be developed and tested.
- Studies on MWU extraction in Turkish may also help all other NLP studies such as disambiguation, word nets, machine translation, parallel corpora, NLP dictionary development, semantic tagging, text mining, speech recognition.

REFERENCES

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *Journal of English for Specific Purposes* 31, 81–92.

Aijmer, K. (1996). *Conversational routines in English: Convention and creativity*. London: Longman.

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.

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

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. In Z. Gavriilidou et al. (Eds.), *Proceedings of the Nooj 2010 International Conference and Workshop* (pp. 29-39). Komotini, Greece: Democritus University of Thrace.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word combinations. In A. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 101–122). Oxford: Oxford University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Fingan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

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

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *Journal of English for Specific Purposes* 26, 263–86.

Biber, D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics* 14 (3), 275–311.

Butler, C. (1998). Collocational frameworks in Spanish. *International Journal of Corpus Linguistics* 3, 1–32.

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

Chen, Y.H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology* 14, 30–49.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23, 397–423.

Cortes, V. (2013). *The purpose of this study is to*: Connecting lexical bundles and moves in research article introductions. *Journal of English for Specific Purposes* 12, 33-43.

Csomay, E. (2013). Lexical bundles in discourse structure: A corpus-based study of classroom discourse. *Applied Linguistics* 34, 369–88.

Demirhan, U.U. (2013). A description of the verb gel- with special reference to pattern grammar. (Unpublished M.A. Dissertation). Mersin University.

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

Eeg-Olofsson, M., & Altenberg, B. (1994). Discontinuous recurrent word combinations in the London–Lund Corpus. In U. Fries, G. Tottie, & P. Schneider (Eds.), *Creating and using English language corpora: Papers from the fourteenth international conference on English language research on computerized corpora* (pp. 63–77). Amsterdam: Rodopi.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition* 18, 91–126.

Firth, J.R. (1951). Modes of meaning *Essays and studies (The English Association)* (pp. 118-149).

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. Cowie (Ed.), Phraseology (pp. 145-160). Oxford: Oxford University Press.

Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics* 18, 109–135.

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.

Howarth, P. 1996. *Phraseology in English academic writing*. Tübingen: Max Niemeyer Verlag.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics* 19, 24-44.

Hunston, S. & Francis, G. (1999). *Pattern grammar*. Amsterdam: John Benjamins.

Hyland, K. (2008). As can be seen: Lexical bundels and disciplinary variation. *English for Specific Purposes* 27, 4-21.

Jespersen, O. (1924). *The philosophy of grammar*. London: George Allen & Unwin.

Jablonkai, R. (2010). English in the context of European integration: a corpus-driven analysis of lexical bundles in English EU documents. *Journal of English for Specific Purposes* 29, 253–267.

Kim, J. D., Ohta T., Tateishi Y., & Tsujii J. (2003). GENIA corpus- a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19, 180-182.

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

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward.* Hove: LTP.

Marco, M. J. L. (2000). Collocational frameworks in medical research papers: a genre-based study. *Journal of English for Specific Purposes* 19, 63–86.

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* (pp. 27-31). İzmir: Ege University.

Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach.* Oxford: Clarendon.

Nattinger, J. R., & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford: Oxford 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* (pp. 64-71). Spain: Barcelona.

Palmer, H. E. (1933). *Second interim report on English collocations*. Tokyo: Kaitakusha.

Pan, F., Reppen, R., & Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals. *Journal of English for Academic Purposes* 21, 60-71.

Partington, A. (1998). *Patterns and meanings*. Amsterdam: John Benjamins.

Partington, A., & Morley, J. (2004). From frequency to ideology: Investigating word and cluster/bundle frequency in political debate. In B. Lewandowska-Tomaszczyk (Ed.), *Practical application in language and computers-PALC 2003* (pp. 179-192). Franfurt a. Main: Peter Lang.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–225). London: Longman.

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*, (pp.131–133).

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: native like selection and native like fluency. In J. Richards, & R. Schmidt (Eds.), *Language and communication* (pp. 191–227). London: Longman.

Renouf, A., & Sinclair, J.M. (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 128–143). London: Longman.

Römer, U. (2010). Establishing the phraseological profile of a text type: the construction of meaning in academic book reviews. *English Text Construction* 3, 95–119.

Salem, A. (1987). *Pratique des segments répétés.* Paris: Institut National de la Langue Française.

Schmitt, N. (ed.). (2004). *Formulaic sequences: Acquisition, processing and use*. Amsterdam: John Benjamins.

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

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes* 12, 214–25.

Stubbs, M. (2007). An example of frequent English phraseology: Distributions, structures and functions. In R. Facchinetti (Ed.), *Corpus linguistics 25 years on* (pp. 89–105). Amsterdam/New York: Rodopi.

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

Vincent, B. (2013). Investigating academic phraseology through combinations of very frequent words: A methodological exploration. *Journal of English for Academic Purposes* 12, 44–56.

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

Weinert, R. (1995). The role of formulaic language in second language acquisition: A review. *Applied Linguistics* 16, 180–205.

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

Wray, A., & Perkins, M. (2000). The functions of formulaic language: An integrated model. *Language and Communication* 20, 1–28.

ABBREVIATIONS

3s	3rd person singular
AJ	adjective
AV	adverb
abl	case-ablative
avrek	adverbial
CJ	conjunction
DT	determiner
ins	case-instrumenal
loc	case-locative
NN	noun
NU	numeral
nom	case-nominative
p3s	possessive (-I)
pcan	adjectivel (-An)
pcdk	nominalizer (-DIk)
past	past tense
PN	pronoun
PP	postposition
VB	verb