Mersin Üniversitesi Dil ve Edebiyat Dergisi, MEUDED, 2015; 12 (1), 1-42.

MULTI-WORD EXPRESSIONS IN GENRE SPECIFICATION

Mustafa Aksan[1], Yeşim Aksan[2]

Mersin University

Abstract: Corpus analyses of lexical structures have uncovered different functions that they come to serve in textual organisation. Frequently occurring patterns of lexical items, the multi-word units, display different distributional properties across different genres and contribute to particular discourse structure. This study first presents a typology of tri-grams in Turkish extracted from the Turkish National Corpus. Second, it discusses the distributional properties of most frequently occurring trigrams in fiction and non-fiction texts. The formal and functional typologies presented here help specify genre properties of Turkish texts.

Keywords: *Multi-word expressions, tri-grams, corpus linguistics, Turkish National Corpus*.

TÜR ÖZELLİKLERİNİN BELİRLENMESİNDE ÇOKSÖZCÜKLÜ BİRİMLER

Öz: Derlem dilbilimi incelemeleri, birden çok sözcükten oluşan sözcük

[1] Mersin Üniversitesi, Fen Edebiyat Fakültesi, İngiliz Dili ve Edebiyatı Bölümü, mustaksan@gmail.com

[2] Mersin Üniversitesi, Fen Edebiyat Fakültesi, İngiliz Dili ve Edebiyatı Bölümü yesim.aksan@gmail.com

Makale gönderim tarihi: 2 Şubat 2015 ; Kabul tarihi: 20 Mart 2015

birimlerinin metin kuruluşunda önemli işlevleri olduklarını göstermiştir. Rastlantı ötesi sıklıkla bir arada kullanılan çoksözcüklü birimlerin metin türlerinin özelliklerinin ve farklılıklarının belirlenmesinde etkili oldukları görülmüştür. Bu çalışma ilk kez Türkçede çoksözcüklü birimlerinin yapısal türlerinin ve işlevsel görünümlerinin saptanmasını amaçlamaktadır. Özellikle üçlü birimlerin kurgusal ve bilgilendirici metinlerdeki sayısal dağılımlarının tür özelliklerinin belirlenmesindeki işlevlerini saptayan bu çalışma, Türkçe Ulusal Derlemi'nde derlem araçları ile çıkardığı veriyi değerlendirmektedir.

Anahtar Sözcükler: *Çoksözcüklü birimler, üçlü birimler, derlem dilbilimi, Türkçe Ulusal Derlemi.*

1. INTRODUCTION

The studies on patterns of use observed in corpus data have changed our understanding of lexical items and their role in textual organization. As the number of corpora increased over the years and tools of textual analysis became more and more sophisticated, research on lexical structures introduced new perspectives into their form and function. In simplest terms, it is argued that meaning of a lexical item is rarely determined by the item itself but rather by the patterns it establishes with other meanings in the text. Thus, a lexical item is redefined as "a higher rank of lexical structure above the word" (Sinclair, 2004, p. 24).

The recurrent units or patterns of lexical items that are above the word have been subject of analysis in terms of their internal construction, their role in organizing texts, and their function in information processing. A quick look at the existing terms like, multi-word expressions/units/constructions, lexical bundles, clusters, chunks, and n-grams among others, indicates that there are different conceptualizations of these units. Research on such units also varies from their identifications and extractions from corpus to their role and function in human cognition as pre-made easy-to-process information packages to minimize the cognitive effort or their function in structuring the discourse.

In this study, we will present an analysis of tri-grams in Turkish and their role in identifying the genre properties of texts. The data of the

study is obtained from two specially designed corpora representing texts from informative and imaginative domains of contemporary Turkish. The paper will proceed as follows: in the first section, the method of analysis, extraction process and the properties of the datasets are introduced. The structural and functional typology of the tri-grams will be presented in section two. In the third section, we will present statistical distribution of tri-grams across the two corpora. Here, the tri-grams will be listed in terms of their basic frequencies in respective genres. Finally, we will present variations among the selected tri-grams in terms of their role in genre identification over the citations in fiction and non-fiction texts.

2. WORDS AND RECURRENT PATTERNS

Biber, Conrad and Cortes (2004, p. 371) provide a very simple definition of a multi-word expression[3]: "the most frequent sequence of words in a register". Most often, these units are not conventional linguistic forms that are syntactically and semantically coherent. They are part of well-defined grammatical phrase or a clause, in other words, a fragment in which some constituent of a full structure is missing. In texts, they frequently occur between full phrases or clauses acting as bridging forms and incorporating a constituent either or both ends. Typically a bi-gram is part of larger tri-gram and a tri-gram is often a part of a four-gram expression. They are systematically structured in use as they recur in patterns in texts and as Sinclair (1991, p. 108) puts it:

> By far the majority of text is made of the occurrence of common words in common patterns, or in slight variants of those common patterns. Most everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up the text.

It is evident that such units have significant textual or discourse functions as they structure the text through recurrent patterns. Since these established units are used repeatedly and consistently in text or

[3] We will use the term n-gram throughout the paper since we will discuss tri-grams only.

discourse, they also signal conventionalized expressions that occur in genre specific manifestations. Thus, it is possible to determine genre specific properties of a text or a discourse by distinguishing multi-word expressions that typically occur in that particular text.

2.1. METHODOLOGY

2.1.1. CORPORA USED

Tri-grams emerging in two different sub-corpora drawn from datasets of Turkish National Corpus (http://www.tnc.org.tr)[4] are analyzed both automatically and manually. Two equal size sub-corpora, Corpus of Contemporary Turkish Fiction (CCTF) and Corpus of Contemporary Turkish Non-fiction (CCTNF), covering a period of 20 years (1990-2009) were constructed for the purposes of this study. CCTF is a 1,009.548 word corpus and it consists of samples from the novels and short stories of contemporary Turkish authors. CCTNF is a 999,102 word corpus containing non-academic text samples complied from a variety of domains. They include social sciences (e.g., sociology, linguistics), art (e.g., architecture, design), commerce-finance (e.g., business, industry), belief-thought (e.g., religion, philosophy), world affairs (e.g., history, politics), applied science (e.g., engineering, computing), and leisure (e.g., travel, gardening).[5] Both corpora include samples taken from 200 different texts published in books. Including a wide range of texts through equally sized samples (approximately 5,000 words) ensures representativeness and balance of two corpora. Thus, CCTF and CCTNF represent their respective genres.

[4] Turkish National Corpus (TNC) is designed to be a balanced, large scale (50 million words) and general-purpose corpus for contemporary Turkish. It has benefited from previous practices and efforts for the construction of corpora. It generally follows the framework of British National Corpus (BNC) yet necessary adjustments in corpus design of TNC is made whenever needed. In selecting written texts for TNC, representativeness of the corpus is achieved through balance and sampling of varieties of contemporary Turkish. Accordingly, written texts included in TNC are selected using three criteria (text domain, time and medium of text) set in during the construction of BNC (Aksan, Aksan, Koltuksuz et al., 2012).

[5] We follow the criteria proposed by Lee (2001) in determining the genre categories, such as fiction and non-fiction (non-academic prose) and domain classifications.

2.1.2. EXTRACTIONS OF MWES

Ngram Statistical Package (NSP) software tool (Banerjee & Pedersen, 2003) is used to generate rank order frequency lists of tri-grams. Text-NSP code is modified to operate on files in the local encoding of the operating system. In our case, the default encoding on Windows is Turkish. The corpus input is all lower-cased and sentence-splitted. A text file to declare which punctuation marks need to be excluded from the tokenization process is prepared. In short, in identifying tri-grams, orthographic word units form the basis of extraction and their arbitrary combination of separate words, for instance *siyah-beyaz* 'black-white', are regarded as a single word. To extract the tri-grams from the corpora and to apply the associative measures, a set of commands is executed as an MS-DOS command prompt.

To include tri-grams into the rank-order frequency lists, a frequency cut-off is determined as 14 times per million words. It means that the lists contain tri-grams used at least 14 times per million.[6] As stated in various studies on MWEs, to avoid idiosyncratic uses by individual writers, a word sequence must recur frequently across numerous texts must be counted as MWEs (Biber et al., 1999; 2004; Biber and Barbieri, 2007; O'Keeffe, McCarthy, Carter, 2007; Hyland, 2008, Liu, 2012 among many others). For the analysis in this study, tri-grams occurred at least 5 different texts. A total of 110 tri-grams identified on the basis of their frequency cut-off point were analyzed. Tri-grams being part of four or five-grams were ignored. As it will be discussed in the following parts of this study, most of the tri-grams are semantically transparent and formally regular. They are commonly used as essential elements in constructing coherent discourse.

To identify and analyze discourse functions of 3-word strings, concordance lines are extracted and sorted via AntConc 3.2.5 for quantitative and qualitative analyses (Anthony, 2010). These concordance lines show the extended discourse context of searched tri-grams and the pragmatic functions they perform. To find out

[6] Actual cut-off points are arbitrary. Previous research uses cut-off ranges between 10 and 40 instances per million words (see, for example, Biber et al. 2004; Hyland, 2008; O'Keeffe, McCarthy and Carter, 2007; Simpson-Vlach and Ellis, 2010). As Greaves and Warren (2010, p. 213) underscore that "This decision is partly driven by the size of the corpus being examined, especially when researchers want to analyse larger n-grams."

genre-specific use of tri-grams, statistical analysis was also done via Minitab 16. The proportion test was conducted between the observed frequencies of tri-grams extracted from two corpora to determine the statistical significance of tri-gram occurrences. To the best of our knowledge, this is the first study that uses such a statistical measure to analyze MWEs, and results of which are given in section 3.

3. TRI-GRAMS IN TURKISH

3.1. A STRUCTURAL TYPOLOGY OF TURKISH TRI-GRAMS

A review of previous work on multi-words in Turkish reveals that there are only a few studies available to researchers.[7] The typology proposed here will be the first that defines types of structures realized in corpora. Since we will concentrate on differences of tri-grams in realizations across genres, we will not go into formal aspects of these lexical structures. At this point, we will simply note that majority of these units, despite their different basic formal properties, function as adverbials in text or discourse in both imaginative and informative domains. In most cases, it is not easy to distinguish conjunctions from discourse connectives. Furthermore, given their various adverbial functions and different contexts of use, it is even harder to decide whether some of these are coordinators or linking adverbs (Biber et al., 1999) as they can combine with coordinators to form multi-word patterns.

Our initial observations suggest that MWEs in Turkish are not very much different than the MWEs identified in corpus studies in English.

[7] Oflazer, Çetinoğlu, and Say (2004) classify multi-word expressions in Turkish including productive reduplicative forms and light verb constructions. The primary aim of this study is to develop an algorithm that will automatically identify and extract such units. Although it has a significant coverage of data and includes potential types, the study neither presents formal analysis nor their textual functions. 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 (whole word vs. morpheme dichotomy). While some of the trigrams that we discuss in this paper also include inflectional affixes on one or more elements in their construction, we confine ourselves to "words" excluding a more fined-grained analysis to include the functional categories. One practical reason for this exclusion is that in the identification of these units, the software commonly picks items that appear between spaces in the written form.

What comes out from the frequency-driven analysis are mostly noun phrases (NPs) or noun phrase (NP) fragments, a similar situation with English. The following types that we have identified are almost exclusively NPs, yet we have identified more categories to underlie their 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.

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 Phrasess

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 şeyden önce 'first of all'

o günden sonra 'ever after'

her şeye rağmen 'despite all'

o güne kadar 'until that day'

her zamanki gibi 'as usual'

başta olmak 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/extential 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'

The data shows that tri-grams with a verbal element are quite rare in Turkish (excluding light verb constructions). 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.

All forms of tri-grams are composed of 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 of what in Turkish as it appears to be the case in English as well.

3.2. FUNCTIONS OF MWES

Functions of tri-grams and their sub-categories are determined on the basis of the taxonomies proposed by Biber, Conrad & Cortes (2004); Cortes (2004); Carter & McCarthy (2006); Hyland (2008). Biber, Conrad & Cortes (2004) identify three main categories as referential expressions, discourse organizers and stance expressions for the functions of MWEs in English. Referential expressions make direct reference to physical and abstract entities to identify the entity or to single out some particular aspects of the entity as important; discourse organizers show relationships between prior and coming discourse; and stance expressions convey the writer's attitudes and evaluations. In addition to them, tri-grams in our study are employed to convey features of conversational interaction such as reporting, questioning. This particular category is named as conversational features.

Tri-grams obtained in fiction and non-fiction texts are classified according to their functions in context. Table (1) manifests examples of the tri-grams and their functional categories each of which contains some relevant sub-categories.

Table 1. MWEs classified according to their functions in context

Category	Sub-category	Example
Referential expressions	Time reference	bir süre sonra 'after a while'

	Place reference	dünyanın her yerinde ‘in all over the world’
	Vague expression	gibi bir şey ‘something like X’
	Quantification	çok daha fazla ‘even more’
	Description	bir hali vardı ‘seemed /as if s/he was’
Text organizers	Transitional signals	ne yazık ki ‘unfortunately’
	Resultative signals	bunun sonucu olarak ‘as a result’
	Framing signals	başta olmak üzere ‘ being in the first place’
Stance expressions	Epistemic stance	ve belki de ‘and perhaps/maybe’
	Other stance expressions	sadece ve sadece ‘only and only’
Conversational features	Interactional markers	öyle değil mi ‘don’t you agree / isn’t it so’
	Reporting	dedim kendi kendime ‘said to myself’

4. COMPARING GENRES

The observed frequencies of tri-grams in fiction and non-fiction texts are found to be statistically significant via proportion test. On the basis of this analysis, we developed a scale of ratio which indicates genre-specific use of MWEs. The ratio scale consists of 5 groups ordered on the basis of the interval value extending from the least similar use of tri-grams to the most similar use of them across the genres. According to this ratio scale, the low ratio (i.e., 0-40%) between the use of tri-grams in fiction and non-fiction texts signals a difference. In other words, it indicates a genre-sensitive use of MWEs. Group 1 comprises the expressions whose interval value is between 0-18% and Group 2 consists of tri-grams having 20-39% value. These two groups represent respectively the least similar and less similar

tri-grams identified in two genres, as given below in Figure (1) and (2).

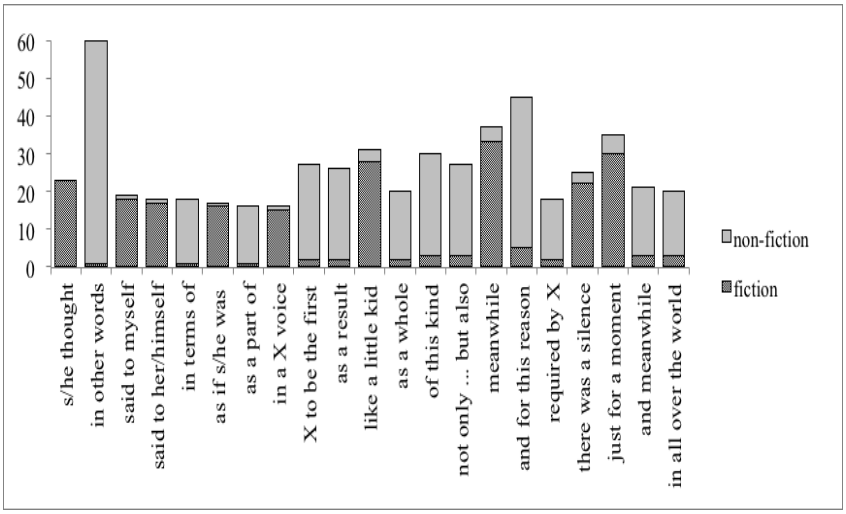

Figure 1. The least similar tri-grams in fiction and non-fiction texts (Group 1)

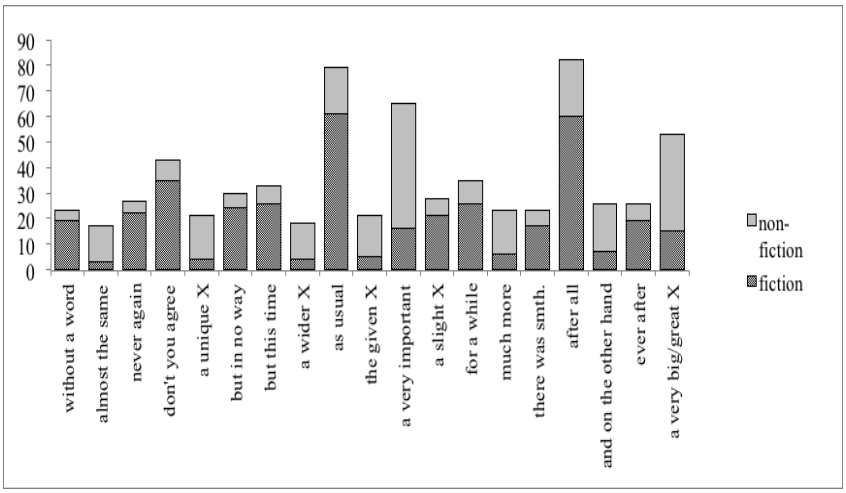

Figure 2. Less similar tri-grams in fiction and non-fiction texts (Group 2)

Among the automatically identified expressions, the occurrence of tri-grams across the genres is determined almost 50%. This figure displays that the use of a MWE in one genre is almost the half of its use in the other genre (see the Figure 3).

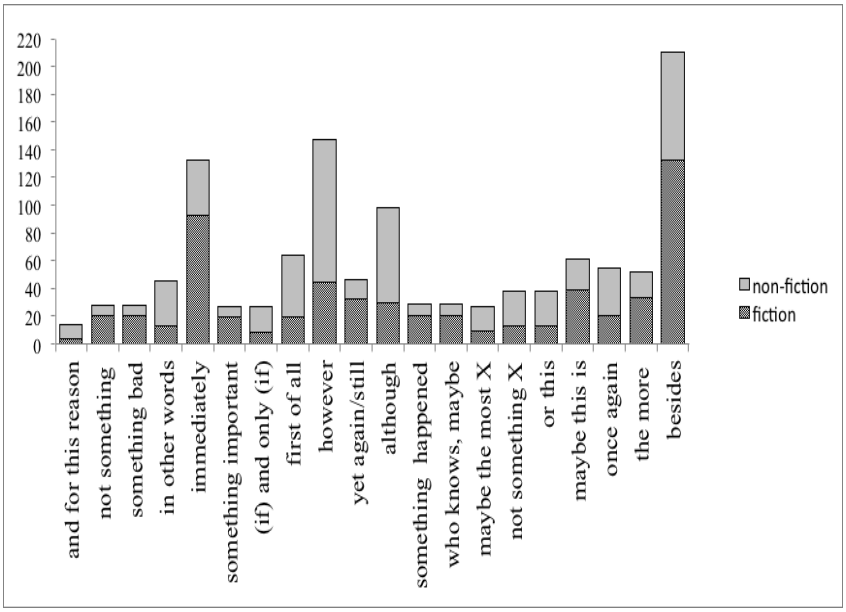

Figure 3. Tri-grams representing the half of the figure across the genre (Group 3)

High ratio between the employment of tri-grams (i.e., 60-100%) across the fiction and non-fiction texts signals similarity in the use of 3-word strings over the genres. In ratio scale, Group 4 contains expressions with a proportion value between 60-73% and Group 5 is composed of tri-grams with 78-96% interval value. In this respect, Group 4 and 5 demonstrate the most similar uses of MWEs in our corpus data (see Figure (4) and (5) respectively).

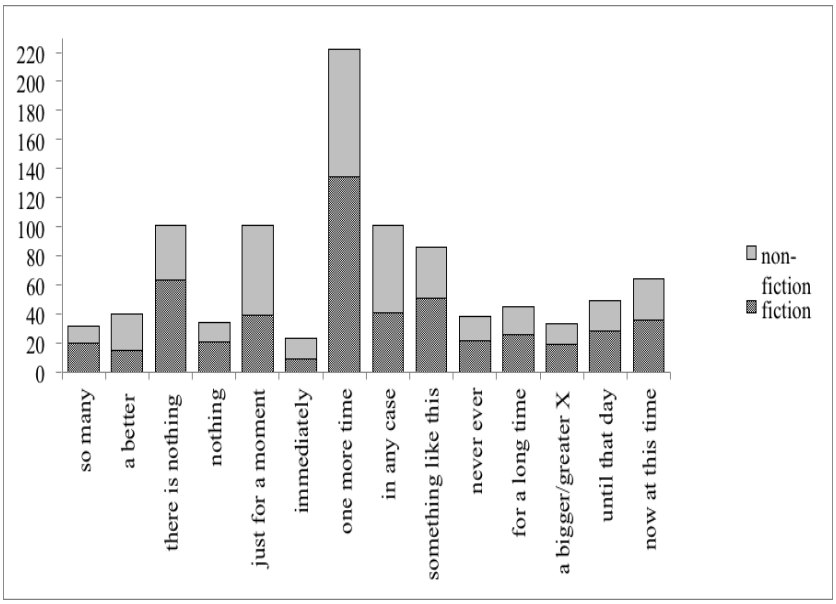

Figure 4. More similar tri-grams in fiction and non-fiction texts (Group 4)

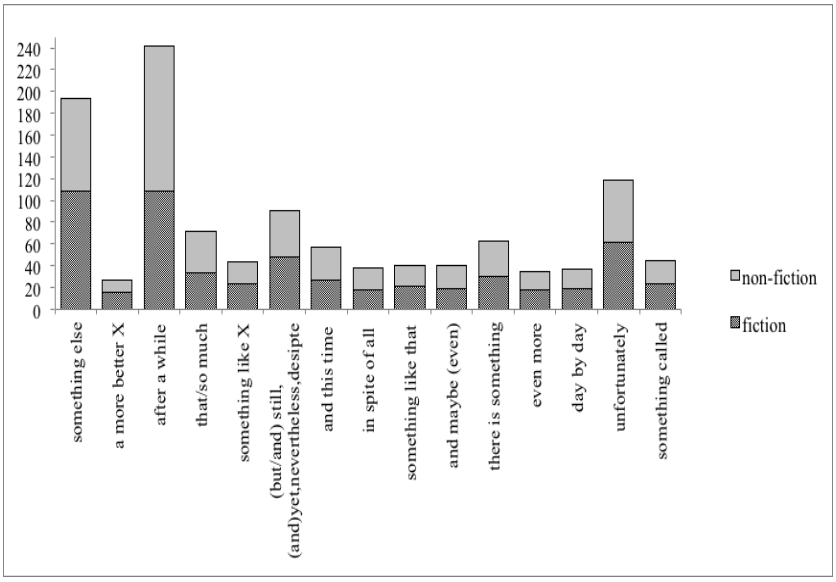

Figure 5. The most similar tri-grams in fiction and non-fiction texts (Group 5)

4.1. GENRE-SPECIFIC MWES AND THEIR FUNCTIONS

As maintained above, Group (1) and (2) contain tri-grams that designate genre-specific uses. Given this fact, we focus on some of these tri-grams and highlight their discourse functions and their roles in characterizing features of fiction and non-fiction texts. Table (2) below demonstrates the tri-grams chosen from the Group (1) and (2) that reflect the characteristic properties of fictional world.

Table 2. Tri-grams specific to fiction

Trigram	Freq. in fiction	Freq. in non-fiction	Ratio	Gloss	English equivalent
diye geçirdi içinden	23	0	0.00	as passed from inside	s/he thought
dedim kendi kendime	18	1	0.06	said one oneself	said to myself
dedi kendi kendine	17	1	0.06	said he himself	said to her/himself
bir hali vardı	16	1	0.06	a state existed	as if s/he was
bir ses tonuyla	15	1	0.07	a sound tone	in a... voice
bir çocuk gibi	28	3	0.11	a child like	like a little kid
tam o sırada	33	4	0.12	exactly that time	meanwhile / at that moment
bir sessizlik oldu	22	3	0.14	a silence occurred	there was a silence
bir an için	30	5	0.17	a moment for	just for a moment
hiçbir şey söylemeden	19	4	0.21	nothing without saying	without a word
bir daha hiç	22	5	0.23	a again never	never again
öyle değil mi	35	8	0.23	as not Q	don't you agree? /Isn’t it so?
ama bir türlü	24	6	0.25	but one way	but in no way
ama bu kez	26	7	0.27	but this time	but this time

her zamanki gibi	61	18	0.30	all times like	as usual/as always
belli belirsiz bir	21	7	0.33	clear unclear one	a slight X
bir süre daha	26	9	0.35	a period more	for a while
bir şey vardı	17	6	0.35	a thing being	there was something
ne de olsa	60	22	0.37	what more being	after all
o günden sonra	19	7	0.37	that day after	ever after/after that day

The most striking difference between two genres is observed on the reliance of conversational features in fictional prose. Direct/indirect speech and thought representation of characters are represented though tri-grams *diye geçirdi içinden* 's/he thought', *dedi kendi kendine* 'said to her/himself'. Again, the tri-gram *öyle değil mi* 'don't you agree / isn't it so', a syntactic hedge in its essential function, serves as a marker of "relational language" (O'Keeffe, McCarthy, Carter, 2007:162). It serves a variety of discourse functions and establish interpersonal meaning through confirming or not confirming the proposition stated by the addressee, self-confirmation through the elaboration of the topic via list of activities, or elaborating the topic with examples, or acting as summative or text-closing signal (Aksan & Aksan, 2013).

Through the use of these tri-grams a concrete, actual language use in fictional world is created. Tri-grams functioning as markers of conversational features in fiction provide this genre with some sort of "fingerprint of everyday conversation" (O'Keeffe, McCarthy, Carter, 2007:68).

In terms of referential expressions, tri-grams signaling time reference unique to fiction texts are found to be relatively less formal realization of temporal references, such as *tam o sırada* 'meanwhile / at that moment', *bir an için* 'for a moment', *ve bu arada* 'and meanwhile', *bir daha hiç* 'never again'. From the evaluative perspective of the speaker (author or character in the narrative), chain of events in the narrative time and evaluative perspective of the speaker are conveyed through these tri-grams, as seen in the excerpt below.

bir an için 'for a moment'

(1) İçeri girince bir an için öyle durdu. CCTF-İskender Savaşır-*Tutku 2000*

"After arriving in, he hesitated for a moment."

In the category of referential expressions, descriptions in fiction texts are mostly evaluative expressions on the part of the speaker as they bring into the description vivid and lively use of language.

bir çocuk gibi 'like a little kid'

(2) Hüsrev Bey yeniden sustu, utangaç bir çocuk gibi önüne baktı. CCTF-Ahmet Altan-Yalnızlığın Özel Tarihi

"Hüsrev Bey silenced again and looked down, like an embarrassed little kid."

Table (3) demonstrates the tri-grams classified under the Group 1 and 2 in non-fiction texts.

Table 3. Tri-grams specific to non-fiction

Trigram	Freq. in non-fiction	Freq. in fiction	Ratio	Gloss	English equivalent
bir başka deyişle	59	1	0.02	a other saying	in other words
ile ilgili olarak	17	1	0.06	with related being	about/related to
bir parçası olarak	15	1	0.07	a part being	as a part of
başta olmak üzere	25	2	0.08	heading being	being in the first
bunun sonucu olarak	24	2	0.08	this result being	as a result
bir bütün olarak	18	2	0.11	a whole being	as a whole
bu tür bir	27	3	0.11	this kind one	of this kind

bu nedenle de	40	5	0.13	this reason PRT	because of this
için gerekli olan	16	2	0.13	for required being	required by X
ve bu arada	18	3	0.17	and this interval	by the way/ and meanwhile
dünyanın her yerinde	17	3	0.18	world every place	in all over the world
hemen hemen aynı	14	3	0.21	almost same	almost the same
kendine özgü bir	17	4	0.24	self specific one	a unique …
daha geniş bir	14	4	0.28	more wide one	broader than
söz konusu olan	16	5	0.31	word about being	the given …/ is about
çok önemli bir	49	16	0.33	very important one	a very important
çok daha fazla	17	6	0.35	very more additional	even more
diğer yandan da	19	7	0.37	other side PRT	and on the other hand
çok büyük bir	38	15	0.39	very big one	a very big/great

What is peculiar in tri-grams observed in non-fiction texts is that they are frequently used as text organizers. In Group 1, it is possible to find representative tri-grams for all types of transitional signals in the non-fiction texts, while there is no such tri-gram use in the fiction. In Group 1 and 2, almost all the instances of transitional signals are detected in non-fiction. Tri-grams being transitional signals serve as to establish expansive (*bir başka deyişle* 'in other words') additive (*bir yandan da* 'in addition to'), resultative (*bunun sonucu olarak* 'as a result') and causative (*bu nedenle de* 'as a consequence') links between the prior and forthcoming discourse.

Functioning as text organizers in non-fiction, framing signals, which are not found in fiction texts, are used to qualify a certain element of discourse (e.g., *Enigma operation, the financial capital Shangai* in the

excepts (3) and (4) below) in terms of its procedures or its significance and thus, they connect preceding and forthcoming discourse under a given condition.

ile ilgili olarak 'about, related to'

(3) Ancak Almanlar Enigma operasyonu ile ilgili olarak titiz prosedürler oluşturmalarına rağmen,… CCTNF-Süleyman Sevinç-Engima

"Although Germans formulated strict procedures about Operation Enigma …"

başta olmak üzere 'being in the first place'

(4) Ekonomik başkent Şangay başta olmak üzere Çinlilerin de son aylarda dolar satışını hızlandırdığı belirtildi. CCTNF-Cengiz Özakıncı-Euro Dolar Savaşı

"The financial capital Shangai being in the first place, it is reported that China accelerated dollar sales in recent months."

About the referential expressions in non-fiction, the most obvious case is the use of tri-grams to describe state of affairs in the discourse. Descriptions are more concrete in the sense that they either refer to the described objects as a whole (*bir bütün olarak* ' as a whole') or as a part of something (*bir parçası olarak* 'as a part of') tangible. While non-fiction descriptive tri-grams target concrete properties in the size or volume of the object (*daha geniş bir* 'borader than'), those in the fiction texts do not specify or target any concrete property.

bir bütün olarak 'as a whole'

(5) Bu gözden geçirme ödevi *bir bütün olarak* ele almanızı ve bölümler arası ilişkilerin uygun kurulup kurulmadığını görmenizi sağlayacaktır. CCTNF-Türker Baş-Bilimsel bir Çalışma Ödevi Nasıl Hazırlanır

"This review will let you consider the assignment *as a whole* and figure out if proper links between the chapters exist."

daha geniş bir 'broader than'

(6) Aile çevresinden *daha geniş bir* sosyal çevreye geçtiğimizde yine

beden dilimizin temel anlaşma aracımız olmaya devam ettiğini görürüz. CCTNF-Zuhal Baltaş-Acar Baltaş-Bedenin Dili

"When you move to a social environment *broader than* your family, again we observe that our body language becomes the basic mean for communication."

The citations above suggest that genre-sensitivity of tri-grams displaying the least similar use in fiction and non-fiction texts can be explained conclusively. Tri-grams that are members of the sub-categories of referential expressions, time markers, descriptive expressions and conversational features (reporting and interactional marker) reflect the properties of fictional technique, namely "degree of specification, fictional point of view and fictional sequencing" as stated by Leech and Short (1981, p. 185). Descriptive focus (vividness and objectivity in description) is said to be "limited to literature… since literature through its medium of language can range over the whole data of human experience" (Leech and Short 1981, p. 183). With all these properties, fiction as a genre is conveying the every day language through more metaphorical and creative descriptions, informal choices of referential expressions and text organizers. Non-fiction texts, on the other hand, aim to state, explain or explore facts on a variety of entities. Thus, descriptive tri-grams, tri-grams serving the function of all types of text organizers and a small amount of tri-grams functioning as stance markers represent the straightforward and concrete style of non-fiction genre. In more general terms, while most of the tri-grams specific to fiction depict "speaker-listener world", tri-grams sensitive to non-fiction reflect "content-proposition world" (O'Keeffee, McCarthy, Carter, 2007, p. 71). We may also draw parallels between fiction texts and "spontaneity and interactiveness of speaking", and non-fiction texts and "deliberateness and detachment of writing" (Chafe, 1986, p. 272) being two modes of interaction.[8]

4.2. MWES IN BOTH GENRES: FICTION AND NON-FICTION

The proportion analysis described above, reveals that Group (4) and (5) include tri-grams that are almost equally distributed across both

[8] Similar issue about the use of MWEs is also expressed in Biber and Barbieri (2007, p. 282) as "(…) the use of lexical bundles is influenced by both mode and communicative purpose."

corpora (Table 4). In this section, we will analyze functions of tri-grams from these groups in both corpora to highlight possible differences of use over genres. For this purpose, we have selected six tri-grams representing three main functional taxonomy, identified by Biber, Conrad & Cortes (2004).

(i) Referential expressions (time); *uzun bir süre* 'for a long time', *o güne kadar* 'until that day', *o kadar çok* 'that/so much', *bir o kadar* 'at least that much'
(ii) Discourse organizers: *ama gene/yine de*, 'but still, yet, nevertheless'
(iii) Stance expressions: *ve belki de*, 'and perhaps/maybe'

The aim here is to conduct a qualitative analysis complemented with quantitative analysis of the corpus data. We have already seen the role of quantitative analysis and its use in determining genre properties in the use of tri-grams across genres. We will now pursue a qualitative analysis to account for differences across genres that are not clear from quantitative data.

Table 4. Tri-grams similar in fiction and non-fiction

Trigram	Freq. in fiction	Freq. in non-fiction	Ratio	Gloss	English equivalent
uzun bir sure	26	19	0.73	long one time	for a long time
daha büyük bir	19	14	0.74	more big one	a bigger/greater X
o güne kadar	28	21	0.75	that day until	until/to that day
işte o zaman	36	28	0.78	that time	now at this time
başka bir şey	108	86	0.80	other one thing	something else
çok daha iyi	15	12	0.80	much more better	a more better X
bir süre sonra	108	134	0.81	one time after	after a while
o kadar çok	33	38	0.87	that much	that/so much
gibi bir şey	23	20	0.87	like one thing	something like X

ama yine de	48	43	0.90	but again PRT	but still, yet, (but/and) still, (and) yet, nevertheless, despite this
bu kez de	27	30	0.90	this time PRT	but this time
her şeye rağmen	18	20	0.90	all things despite	in spite of all
öyle bir şey	21	19	0.90	such one thing	something like that
ve belki de	19	21	0.90	and maybe PRT	and perhaps/maybe
bir şey var	30	32	0.94	one thing exit	there is something
bir o kadar	18	17	0.94	one that much	at least that much
her geçen gün	19	18	0.95	each passing day	day by day
ne yazık ki	61	58	0.95	what pity REL	unfortunately
diye bir şey	23	22	0.96	called one thing	something called

4.2.1. INTERNAL STRUCTURE OF SAMPLE TRI-GRAMS

4.2.1.1. *O GÜNE KADAR* 'UNTIL THAT DAY', *O KADAR ÇOK* 'THAT /SO MUCH', *BIR O KADAR* 'AT LEAST THAT MUCH'

One of the most frequent postpositions in Turkish is *kadar* 'until' and it occurs in three of the sample tri-grams in this study. It is the only postposition that can occur with complements marked with diffferent cases, nominative, genitive and the dative. Each of these different case markings signal for different functions of *kadar* as can be followed from English equivalents in the following discussion. Furthermore, *kadar* also combines with certain pronouns to form "lexical items" that fulfill various discourse functions.

A bare postposition in Turkish, *kadar* heads postpositional phrases that are used as adverbial or adjectival modifiers. Almost all postpositional phrases can be used in adverbial function in Turkish and the type of modification can be determined from the case of the complement nominal. In the case of the tri-gram *o güne kadar* 'until that day' the

noun complement of the prepositional phrase is marked with the dative. The dative marked forms express temporal or spatial meanings and in both uses the phrases indicate a terminal point. In all our citations in both corpora, the tri-gram *o güne kadar* functions as a time reference and signals a point of termination in the course of events in the discourse. More specifically, there is a turning point in the events, a strong contrast between what happened before and what happens now or will happen in the future. The remaining two other tri-grams represents combination of *kadar* with demonstrative *o* 'that' forming a quantification expression *o kadar* 'that/so much'. In *o kadar çok* 'so much/that much' a degree expression is added and in bir o kadar,' at least that much' the indefinite article *bir* 'a/one' is added to form yet another indefinite quantifing or more specifically an equative expression.

4.2.1.2. *UZUN BIR SÜRE* 'FOR A LONG TIME'

The tri-gram *uzun bir süre* 'for a long time' is headed by a deverbal noun which combines with indefinite article *bir* 'a/one' and the adjective *uzun* 'long' to constitute a full phrase with a compositional semantics. The citations in both corpora indicate that the phrase has an adverbial function of time reference with relatively stable meaning, measuring the duration of events in the discourse. This measure is expressed rather vaguely as can be deduced from the overall meaning of the phrase.

4.2.1.3. *AMA GENE/YINE DE*, 'BUT STILL, YET, NEVERTHELESS', *VE BELKI DE*, 'AND PERHAPS/MAYBE'

The final two tri-grams are both discourse connectives (Kerslake, 1992; Ruhi, 1992; 1998; Zeyrek, Turan and Demirşahin, 2008; Zeyrek, forthcoming). All three constitutive units in each tri-gram are individual lexical items and have their own meaning and use in the language. The clitic *de* is mainly an emphatic particle, appears in both tri-grams and it is required for the same purposes. In *ama yine de* 'yet again/despite', the adversative conjunct *ama* 'but' can be said to be the head of the structure providing the fundamental meaning and it combines with *yine* 'again' which further combines with clitic *de*. The addition of *de* prevents potential ambiguity in the tri-gram since without it the overall meaning of the structure would be 'but again'. The stance marker *ve belki de* 'and maybe even /and perhaps' can also

be analyzed as a special case of very common additive conjunctive *ve de* ‘and (even)’ in which an epistemic modal adverbial is inserted. The additive connective (Göksel and Kerslake, 2005; Kerslake, 1992) *ve de* ‘and (even)’ is an emphatic form of the simple additive *ve* that functions as a coordinator. The emphasis is on the importance of the comment that will follow the first clausal conjunct.

4.2.2. SIMILAR TRI-GRAMS IN FICTION AND NON-FICTION

The tri-grams do not display major differences in their basic functions across the corpora. The cited formal variations, however, signal major differences in their use. Differences are observed mostly in the combinations of tri-grams with particular adverbials in the discourse. For example, the quantitative comparative *ne kadar X o kadar Y* ‘the more X the more Y’ occurs only once in fiction, however, it has at least five more citations in non-fiction. On other hand, in the citations of the tri-gram *o güne kadar* ‘until that day’ it combines more frequently with adverbs in fiction texts that function as discourse connective than those in non-fiction. While there appear to be no further systematic differences across the corpora with respect to functions of these two tri-grams, the observed differences significantly point to specific genre properties.

4.2.2.1. REFERENTIAL EXPRESSION: TIME REFERENCE

There are two major types of *uzun bir süre* ‘for a long time’ in the non-fiction texts. The first type includes those that refer to individual’s particular experience that is mostly negatively evaluated. In such cases, even though the duration of the event is short and not extending over longer stretches of time, the unfavorable nature of the experience causes the individual to feel the pace of time otherwise. In other words, the unpleasant experience is expected to end in the shortest time possible yet it does not.

(7) “Kim o?” Yanıt kesin ve netti: “Polis”. “Bir dakika,” dedi Müjdat içeriden. Uzun bir süre bekledik. Neden sonra kapı açıldı. CCTNF-Tarık Akan-*Anne Kafamda Bit Var*

“Who is it?” The reply was clear and sure: “Police”. Just a minute said Müjdat from inside. We waited *for a long time*. Finally the door opened.”

In this particular citation, an actor who became one of the torture victims of fascist military coup in 1980, is taken under custody and is brought to one of his friends' house by the police. The friend who is also an actor responds to the knock on the door by asking for a second while the experiencer is waiting outside the door, handcuffed and not knowing what is going to happen next. The duration between asking for a second and opening of the door by the fellow actor cannot possibly be longer otherwise the police would not hesitate to break the door and turn the individual inside into another torture victim. Further evidence for the evaluation of duration longer than expected by the individual comes in the following sentence. The opening of the door is expressed with adverbial modifier *neden sonra* 'finally' that functions as marker of time span judged as unnecessarily longer than it should be.

In its other use in non-fiction texts, the vague temporal expression *uzun bir süre* 'for a long time' refers to longer periods of time that are not necessarily experienced by the user. While some of these uses can be definite or almost definite periods of time, in other words, the user has the option of stating the period in terms of more concrete terms or measures out the period, does not prefer to do so. Instead, the period mentioned is left vague because the user is not sure about either the beginning or the end of the process he is referring to.

(8) Bu gibi değişimlerle şimdi var olan karma ekonomi sisteminin oldukça uzun bir süre yaşayacağı, ama bu arada çok önemli bir evrim geçireceği kanısındayım. CCTNF- Murat Belge-*Türkiye Dünyanın Neresinde*?

"With such changes, I believe that this heterogenious economy will *for a long time* but will undergo very significant evolution in the meantime."

In a small number of citations in non-fiction texts, the trigram is further premodified by *çok* 'very', and in some others followed by *sonra* 'after' rendering these citations as examples of four-grams including the trigram in question which is a very frequent. Only in one citation, we find a definite measure expression, as in *bin yıl gibi uzun bir süre* 'a longer period of time like a thousand years', referring to linguistic change that took place in Turkish under heavy influence of Arabic and Persian:

(9) Bin yıl gibi uzun bir süre İslam uygarlığı içinde olunması dolayısıyla Türkçeye Arapça ve Farçadan binlerce kelime… CCTNF-İsa Kayaalp-*İletişim ve Dil*

"Thousands of words flooded into Turkish as a result of being within the Islamic civilization *for a long time* like thousand years…"

As in the above cases, even the presence of definite measure expression, the tri-gram still represents a vague time reference.

The citations of this tri-gram in the fiction texts do not introduce any different use than those listed above (i.e., experienced vs. non-experienced events). Only in two cases we find comparative *daha* 'more' combining with the tri-gram with no significant change in time reference. In both genres, the tri-gram occurs in clause initial position while in only one citation in fiction, it occurs in clause final position which may be attributed to stylistic variation among genres.

The second trigram in the category of vague time reference is *o güne kadar* 'until/to that day'. It refers to a termination point in course of events in discourse. In the majority of the cases, it combines with the clause that is negated where the negated clause introduces the events that are used to or habitually happen under normal conditions. The termination in the course of events is further modified by the use *hiç* 'never' or its derivatives (e.g., *hiçbir* 'none') in a number of cases in both genres. In the following sample, though the habit of tipping for appreciated services continues, for the first time tips handed exceed the previous amounts. Thus, termination applies to 'previous' amounts given until that day, where the day has marked importance as the day of wedding:

(10) Kız tarafı berberin çıraklarına o güne kadar hiç vermediği bahşişleri verir, aristokrat ve zengin görünmeye çalışır. CCTNF-Harun Yahya-*Adamlık Dini*

"The bride's family tips amounts that they have never did before to the aide of the hair-dresser *until that day* and try to appear as if they are aristocrats and rich."

In the citation below, the emphasis is on the initiation rather termination. Here, the addresser marks the turning point from what has previously taken to be ordinary to something that will happen for the first time in his entire life. It is not possible that the individual suddenly develops a habit but the use of trigram attached at the beginning of the negative clause announces the end of previous routine:

(11) O güne kadar ağzıma sigara koymamıştım. Ama arabada bizim zabitlerden bir tane istedim. CCTNF-Soner Yalçın-*Teşiklatın İki Silahşörü*

"I have had never smoked *until that day*. Yet I asked for one from the officers in our car."

Attachment of *o güne kadar* 'until that day' to the negative clause below extends the period of time as opposed to previous citations. Here, again the emphasis is on the ever first occurrence of an event, in this case a discovery:

(12) … Leif Ericson adlı bir Viking kaşif, o güne kadar ayak basılmamış bazı Karayip adalarını keşfetti; … CCTNF-Ali Poyrazoğlu-*Ödünç Yaşamlar*

"A Viking explorer named Leif Ericson has discovered the Carabian Islands that no one has stepped *until that day*."

As noted above, *o güne kadar* 'until/to that day' combines with discourse connectives that signal various pragmatic functions, including *meğer* 'in fact', *sanki* 'as if', *hatta* 'even', and *yoksa* 'or', among others. In none of the citations in the non-fiction, this tri-gram combines with such discourse connectives or any other adverb. *Kimse* 'nobody' also contributes to the overall meaning of 'never happened before', a polarity item required in the negative context:

(13) Bir akşam muhtar, fakir köylünün kapısını çalmış. Meğer o güne kadar adamcağızın kapısını kimse çalmamış. CCTF-Sara Gürbüz Özeren-*Sihirli Leylekler*

"One night the village governer knocks the door of the poor villager. In fact *until that day* no one has knocked the man's door."

The ever first happening of an event in the discourse is also overtly expressed in the following:

(14) O güne kadar çoğu disiplinsiz çetelerle çatışmış ve kolayca ilerlemiş olan Yunan ordusu, yeniden kurulmakta olan Türk ordusu ile ilk kez karşı karşıya gelir. CCTF- Turgut Özakman-*Şu Çılgın Türkler*

"The Greek army that advanced easily fighting only against unruly bands *until that day* faces recently assembled Turkish army for the first time."

There is no negative marking here and this represents a rare occurrence of presupposition that *ordu* 'army' vs. *disiplinsiz çete* 'unruly horde' provides the necessary background for the proper opposition.

Another difference in the citations of *o güne kadar* 'until/to that day' in both corpora can be followed from their respective positions in the clause. While the tri-gram does not occur clause-finally in non-fiction, it repeatedly occurs in this position in the fiction. The ramifications of this positional difference might have pragmatic consequences, however, we will not pursue this in this study.

4.2.2.2. REFERENTIAL EXPRESSION: QUANTIFICATION

The tri-gram *o kadar çok* 'that/so much' comes in at least in three different formal representations. In structural terms, we find that the tri-gram combines with loan subordinator *ki*. Except from corpora representing "correlative comparison"[9] (Kornfilt, 1997), where a correlative frame is added before a conditional phrase, all citations in informative domain include *ki*. In most citations, the structure is composed of two parts where the tri-gram is separated from the

[9] The quantitative adverbial *kadar* is also frequent in exclamations comparable to English exclamations with *what* or *so*.

following *ki* by either a verbal or a nominal element that is quantified. This is an expected situation, since in Turkish *kadar*-phrases act as adverbial or adjectival depending on the context of modification: in (15) below, the *kadar*-phrase quantifies over the degree of wanting and in (16) it quantifies over the nominal, the excessive amount of "crook" people.

(15) Biliyor musun Yeliz ben ailenin tek erkek çocuğuyum ve babam o kadar çok istemiş ki erkek evladı olmasını altı kızdan sonra ben doğmuşum. CCTNF-Yeliz Güllü-*Özürlü Olduğum için Kimseden Özür Dilemiyorum*.

"You know, I'm the only boy in the family, and my father wanted a boy *so much that* after six daughters I was born."

(16) Pislikler o kadar çok ki saymakla bitmiyor. CCTNF-Erdal Şekeroğlu-*Tırtıl Yazıları*

"The crooks are *so much that* I cannot stop counting."

Since *ki* appears as a separate item in most citations, MWE extraction process has captured *o kadar çok* 'that/so much' as a tri-gram excluding *ki* which would otherwise derive a four-gram expression. Lewis (1969) indicates that in clauses expressing consequence, different than English where that is deleted, in Turkish *ki* is retained. In *ki* attached citations with no other element intervening, the post-modified element is a nominal and hence the tri-gram quantifies over the amount of entities represented by this nominal. Since combination of the postposition *kadar* with demonstrative *o* 'that' commonly expresses negative evaluation (Göksel & Kerslake, 2005), *o kadar çok* 'that/so much' tri-gram also conveys a sense of negativity, though not in all cases.[10] The separated citations are overwhelmingly cases of adverbial modification since the intervening element is almost always a verbal constituent.

The correlative frame citations represent another case of two part constructions. In the correlative comparison in Turkish, a regular

[10] Kerslake (1992) observes that "sentence final *ki*" has repudative function in discourse. In this function, *ki* attaches to a negative statement or a question. The repudative function is available only when *ki* combines with negative statements.

tensed verb follows the conditional, and a correlative frame is added. In the following citation, the conditional precedes the tensed verb and it is preceded by vague quantifier expression *ne kadar çok* 'the more …' that functions as the standard of comparison for correlation. [11]

(17) Bir erkek ne kadar çok kadınla yatmışsa o kadar çok kadın tarafından tepe tepe kullanılmıştır. CCTNF-Sinan Akyüz-*Etekli İktidar Erkek Hakları Kitabı*

"The more a man slept with many a woman the more he has been used to the end."

The same subtypes of vague quantifier is found in both corpora except the fact that there is one correlative comparison citing in fiction as opposed to five citations in non-fiction.

Certain forms of use, on the other hand, are confined to fiction. There are citations in which *o kadar çok* 'that/so much' is used in contexts that are understood as a comparative quantification with no conditional or preceding standard of comparison. In the following citations, there is no overt marking of the "condition" on the consequence yet *o zaman* 'only then' and *öyle olunca* 'when it happens in this manner' contribute to sense of condition expressed in the citation:

(18) Hiç aklına getirmeyeceksin korkuyu, o zaman o kadar çok korkmuyorsun. CCTF-Ferhan Şensoy-*Rum Memet*

"Never bring that subject to your mind, only then you do not fear *that much*."

(19) Öyle olunca kazalar da o kadar çok olmaz, çok canlar yanmaz… CCTF-Savaş Ay-*Ay Hikayeleri*

"When so happens, accidents do not happen *so much*, people do not hurt… "

[11] Note that in this form there is coordination rather than subordination hence *ki* is not a part of the tri-gram.

We may conclude that the observed variations in form and their distribution in corpora, the tri-gram in question also denote differences in genres.

As in the case of previous vague quantifier, the next *kadar* phrase also comes in different forms. *Bir o kadar* 'at least that much' appears in its own form as well as combination with clitic *de* either preceding (de bir o kadar) the tri-gram or following (*bir o kadar da*). Furthermore, there are citations in which there is an intervening element between the tri-gram and the following clitic *de*.

The pattern formed in this trigram mainly expresses an equative meaning with emphasis or intensity of measured entity or event. [12] As in the previous *kadar*-phrase, *bir o kadar* 'at least that much' also functions as adverbial and adjectival. Thus, there are citations where the tri-gram quantifies vaguely over the nominals that are evaluated as in equal size or amount, and there are citations in which a modifying adjective is measured as equal in terms of the standard of comparison.

There are cases where the amount in question is overtly stated as the basis for further comparison. When we find overt expression of quantity in one clause, this amount is not stated in the following:

(20) … sayıları o tarihte üç yüz bini bulmuş Macar mülteciler, bir o kadar Romen ve Yugoslav kaçak… CCTF-Adalet Ağoğlu-*Romantik*

"… the number of Hungarian refugees reaching three hundred thousand at that time, *at least that much* Rumenian and Yugoslavian fugitives…"

Qualities as measured properties are also represented in the following equative expression with *bir o kadar* 'that/so much as'. The additive

[12] Another use of *o kadar* phrase is noted in Lewis (1969, p. 269) "a single conditional verb followed by de is concessive. He cites the following sentence: *Ne kadar yukarıdan inerse o kadar derine girer* 'the greater the height it falls from, the deeper it goes in (of a pile-driver lit. 'from whatever amount high it descends, to that amount deep it enter'). It calls for a detailed analysis, however we simply note that in all citations representing this tri-gram, it is possible to insert conditional without any significant change in the overall meaning of the discourse segment.

da here is understood as an additional positive quality of a person whose human qualities are already established and appreciated.

(21) Şefkatli, titiz ve bir o kadar da hassas bir yapının insanıydı. CCTF-Ahmet Günbay Yıldız-*Benim Çiçeklerim Ateşte Açar*

"Affectionate, tedious, and equally *that much* sensitive natured person."

The same tri-gram can also be coordinated as in the following citation. In this particular case, the addresser is still evaluates the measure of events as equal, however, equal applies to opposing values of easy and hard simultaneously. Simple juxtaposition combines both tri-grams:

(22) Bir o kadar zor, bir o kadar kolay. CCTF-Cenk Babaeren-*Bir Erkeğin Güncesi*

"That much hard, that much easy."

The opposing qualities are also measured in the following, this time coordinated with adversative *but*:

(23) Çok çirkin ama bir o kadar da tatlı dilli. CCTF-Dilek Dündar-*El Alem Ne Der*

"A very ugly but also *that much*/equally a sweet talker."

It is evident that the comparison of quantities or qualities is measured vaguely in both corpora. There are citations in which the amount to be compared is stated explicitly, yet the measure is vague and in most cases subjectively evaluated.

4.2.2.3. DISCOURSE ORGANIZER

The final two trigrams are a discourse organizer and a stance marker that display less contrast in terms of their uses in the corpora. Both incorporate the clitic *de* in their form as an intensifier, and while *ama gene/yine de* '(but/and) still, (and) yet, nevertheless, despite this' *de* functions as adversative, in the other *ve belki de* 'and maybe (even)', as an additive.

English equivalents of *ama gene/yine de* '(but/and) still, (and) yet, nevertheless, despite this' may vary due to the context of use in discourse as suggested in the above. *Ama* is the most common adversative connective and in the trigram it combines with *gene/yine de* which is also an adversative connective.[13] Neither *ama* nor *gene /yine de* can replace the trigram *ama gene/yine de* in use, and if they do there will be significant difference in emphasis. The 'combined' form is more emphatic than other adversatives or their other possible combinations.

In discourse, the connective appears with adverbial function and expresses a concessive relationship among the discourse segments it connects. In simple terms, the adverbial signals the relation that holds between the segments of the discourse is 'counter-expectation'. In other words, what is presented in the first proposition that triggers an expectation in the addressee is cancelled in the following segment that occurs with the concessive marker (to use a more specific term, "concessive cancellative discourse marker" (Bell, 2010)). Thus, whatever the expectation with respect to events introduced in the antecedent segment is very clearly and emphatically cancelled out the concessive segment.

In terms of its position, *ama gene/yine de* '(but/and) still, (and) yet, nevertheless, despite this' always occurs clause initially. In the following citation from non-fiction, it either occurs after a comma separating the cancellation marker from the introducing clause within the same sentence or at the beginning of a following sentence that expresses a counter-expectation:

(25) Sınırlı düzeyde bir ilişkim olmuştu, ama yine de yakınen tanımıştım.

"I had a confined relationship *yet/but still* I came to know him/her closely." CCTNF-Faik Başbuğ-*Devleti Nasıl Soyuyorlar*

[13] Kerslake (1992, p. 87) defines the adversative as 'contrary to expectation'. The adversative marks a turning point in discourse, reversal or partial reversal, "What I am about to say will cancel out the adversative effect of what I have just said'.

In (25) above, the addresser refers to a previous relationship and defines it as relatively limited. The following concessive clause provides information that the addresser came to know the partner during the relationship very closely or intimately. Hence, while the state described in the introduction depicts a superficial quality of the relationship, the concessive pictures a contradictory state where the addresser asserts that he acquainted with the other closely.

In (26) below, the contradiction expressed in the following discourse segment cancels the expectations evoked, no matter how emphatically, in the preceding segments (fair amount of heat, not like a summer day):

(26) Oda iyice ısındı, yaz günü gibi değil. Ama yine de sıcak. CCTNF-İbrahim Yıldırım-*Müşteki Aşklar*

"The room heated fair amount, it is not as in a day in summer. *And yet* it is hot."

The sense of disappointment is implicated in (27). Cancellation of expectations has serious consequence as implied in cancellative clause ('no measure taken') despite the fact that warnings are reported repeatedly:

(27) Defalarca bunu üst düzey komutanlarına bildirdi. Ama yine de bir önlem alınmadı. CCTNF-Murat Yıldırım, Cemalletin Emeç-*Yeşil*

"He has reported this to his superirors over and over again. *And yet* no measure has been taken."

Here, counter-expectation is further signaled by a negative in the following clause.

The unfavorable condition stated in the first discourse segment is mitigated, contrary to expectation, as there is a cause for tolerable acceptance:

(28) Giriş çıkışı çamur içindeydi. Ama gene de ilk evimiz olduğu için bize saray gibi görünmüştü. CCTNF-Zülfü Livaneli-*Sevdalım Hayat*

"The entrance was covered with mud. *But still* it appeared as a place to us as it was the first house we owned."

In (28), the addresser's subjective evaluation is expressed. The first segment indicates that the house is in mass, covered with mud all over, and yet the residents likened it to a palace. The adverse relation between these discourse segments is mitigated relatively by the insertion of a reason for counter-evaluation that states that the residents became house owners for the first time.

The citations in fiction do not display significant differences. However, there are certain uses that are not found in the non-fiction, pointing to peculiarities of the genres in question. For example, in (29) below both clauses incorporate the same participle (*bulabildiğim* 'that was able to find') and there is no item that would implicate reasons for contradictory or cancelling relation between the two events in the presence of the marker for the purpose:

(29) … dünyaya çekinmeden bakan kadının yüzünde güçlükle bulabildiğim, ama gene de bulabildiğim teyzem … CCTF-Bilge Karasu-*Lağımlaranası ya da Beyoğlu*

"… with difficulty I was able to find (my aunt) at the face of woman who could look out at the world with no reservation, *and still* I was able to find, my aunt."

This citation is different than other cancellatives in the sense that what is underlined is not counter-expectation but rather, an effort on the part of the addresser against all odds, an accomplishment. It is valued dearly by the addressor, no matter how hard it was and against all odds, it finally happened and it is worth of all the efforts.

The cancellative meaning is further highlighted in the following sample (30) via the use of another adversative with concessive meaning, *gerçi* 'actually, admittedly' in the previous clause:

(30) Otomobilde kaç kişi olduğunu bilmiyorum gerçi, ama gene de önce şöförü vurmak istiyorum. CCTF-Hasan Ali Toptaş-*Sonsuzluğa Nokta*

"Actually, I don't know how many are there in the car, *and yet* I want start to shoot the driver first."

The contradiction to expectation may not be expressed strongly (the use of 'perhaps') in (31). The very same individual with the very same qualities, no matter how unexpected a quality it may have, it is still what it is:

(31) … ayıların dilinden anladığını iddia eden gerçek bir ayıydı belki ama gene de ince bir ayı. CCTF-Vivet Kanetti-*Kurabiye Saatinde*

"Perhaps a real bear that claims to understand the language of the bear, *yet still* a delicate bear."

In some uses as in (32) and (33) however, there is no expectation to be cancelled. The addresser is well aware of the concessions nevertheless poses the question:

(32) Biliyorum sormam bir şeyi değiştirmeyecek ama yine de merak ediyorum. CCTF-Serdar Özkan-*Kayıp Gül*

"I know that my questioning will change nothing *nevertheless* I wonder."

(33) Aman aman, ne hırs o! Farkındayım bunun ama yine de siz bilirsiniz. CCTF-İzzet Harun Akçay-*Mavi Şehir*

"Well well, what an ambition there is! I am aware of this *yet* it is your decision."

4.2.2.4. STANCE MARKER

The final tri-gram connective is formed by addition of modal adverb *belki* "perhaps" and emphatic clitic *de* to the most common additive connective *ve* 'and'.[14] The modal adverb *belki* marks non-factual probability in Turkish and it "… expresses the strength of speaker's

[14] Ruhi (1992) discusses in detail various discourse functions of this additive. Göksel and Kerslake (2004) note that 've de' is an emphatic form of *ve* and is used to highlight the significance of the comment that follows. Kerslake (1992, p. 86) "… the combination *ve* de (the *de* being continuative) is frequent in the colloquial in the internal additive sense 'and here's another point."

confidence in the soundness of the assumption" (Göksel and Kerslake, 2005, p. 298). There is no grammatical marker of non-factual modality in Turkish and in the absence of modal adverbs, there is a decrease in modal strength expressed in the sentence.

In citations in both corpora, the use of the connective is always optional. In other words, the connective introduces an optional clause or phrase that provides a potential alternative to what is being introduced into the discourse in immediately preceding clause or phrase. The addresser is aware of the probability of an alternative event or state to what he has just asserted, and feels it necessary to provide the alternative to the addressee in the context. In the following citation, the phrase connected to the previous one is option and it expresses the addresser's uncertainty.

(34) Çocukluğumda ve belki de bütün hayatım boyunca, bu evin bana verdiği huzur, emniyet ve istikrarı başka hiç bir yer… CCTNF-Gündüz Vassaf-*Annem Belkıs*

"In my childhood *and perhaps* all through my life, I could not find the peace, security and stability in nowhere else other than this home."

In some cases, the options presented are in contradiction that the connective *or* may is used:

(35) Şimdi bu sanatçı, ya bireyi olduğu toplumsal sınıf ile uyum içindedir, ya da ona ters düşmekte ve belki de karşı çıkmaktadır.

"Now this artist is either is in tune with the social class he belongs or disputes the class (values) *and even* opposes (them)."

In this particular citation, there are opposite alternatives introduced. Addition of the third alternative forces a contradiction rather than a probability.

The uncertainty of the situation is further explicated by the addresser in (36) by the use of *who knows* and *where* appear in the previous discourse segment. Further contribution to non-factuality are provided

by *–mIş* (evidentiality marker) and *–DIr* (non-factuality marker) in respective conjoined clauses.

(36) Şimdi bu ülkenin kim bilir neresinde yerleşmiş *ve belki de* çok mutlu bir şekilde yaşıyorlardır… CCTF-Hayri Ersoy-*Sürdüler Sürgün Oldum*

"Now who knows where in this country they have settled *and may be even* living quite happily."

The addresser implies an option in the following citation, a deduction in fact, from the given prior knowledge of the circumstances. In this case, the additive combined with the modal adverbial come to express an afterthought. The addresser is aware of the events leading to the death of the beloved artist on stage, and concludes that this may well be a suicide:

(37) … heyecanına yenildi ve hayata sahnede veda etti. Ve belki de o intihar etti. CCTF-Pınar Çekirge-*Nokta*

"…overwhelmed by his excitement and said his farewell to life on stage. *And may be* committed suicide."

5. CONCLUSION

In this study, we have presented a detailed formal and functional analysis of tri-grams in Turkish for the first time over datasets extracted form specially constructed corpora. In accordance with corpus analysis, we have presented quantitative and qualitative aspects of these units in imaginative and informative domains, particularly their role in genre specification.

We observed that compared to English, the structural types of multi-word expressions in Turkish are far less in number. This is definitely related to structure of word in Turkish. The agglutinative morphology allows stacking of morphemes in a word; grammatical categories "written" as separate words in different languages, are functional morphemes attached to stems in Turkish. The typologies presented here should be considered as initial proposals and are subject to change as future studies and new data highlights other properties multi-word expressions.

It is evident from the data of tri-grams that their distributional differences across two corpora signal their genre properties. The quantificational differences themselves are sufficient enough to conclude that different genres make use of different repertoire of recurrent patterns, conventionalized and specific to a particular genre. The distributional data indicated that some patterns are definitely genre-specific as they do not occur in other genres. On the other hand, a group of tri-grams can be found in both though with varying degrees, implicating they are part of the users' stock of common expressions. In the final group of tri-grams, there are equally distributed numbers of tri-grams in both genres, still providing evidence for genre specification though not quantificationally but also by qualitative analysis. The citations in the corpora, in most cases manifest slight structural alternations, signal differences in use.

This study concludes that in genre identification and in determining genre specific properties of texts, corpus tools are very helpful. The tri-gram analysis presented here exemplifies how such tools can identify conventional and established recurrent patterns in different genres as well their pragmatic functions.

Abbreviations

ADJ: Adjective
ADV: Adverb
CCTF: Corpus of Contemporary Fiction
CCTNF: Corpus of Contemporary Non-Fiction
CONJ: Conjunctive
Dem: Demonstrative
InArt: Indefinite Article
N: Noun
NP: Noun Phrase
P: Postposition
PRT: Particle
PP: Postposition Phrase
REL: Relativizer

REFERENCES

Aksan, Y., Aksan, M. & Koltuksuz, A. et al. (2012, May). Construction of the Turkish National Corpus (TNC). *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, (pp. 3223-3227), Istanbul, Turkey.

Aksan, M. & Aksan, Y. (2013, September). Multi-word units and pragmatic functions in genre specification. Paper presented at 13th IPrA Conference 08-13 September 2013. New Delhi, India.

Anthony, L. (2010). AntConc (Version 3.2.2.5w) [Computer Software]. Tokyo, Japan:Waseda University. http://www.antlab.sci.waseda.ac.jp/

Banerjee, S. & Pederson, T. (2003). The design, implementation and use of the (N)gram (S)tatistic (P)ackage. In … (Eds.), *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, (pp. 370-381).

Bell, D.M. (2010). Nevertheless, still and yet: Concessive cancellative discourse markers, *Journal of Pragmatics 42*, 1912-1927.

Biber, D. Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). The Longman grammar of spoken and written English. Harlow: Pearson Education.

Biber, D., Conrad, S. & Cortes, V. (2004) If you look at ... : Lexical bundles in university teaching and textbooks, *Applied Linguistics 25*(3), 371–405.

Biber, D. & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes 26*, 263-286.

Chafe, W. (1986). Evidentiality in English conversation and academic writing. In Chafe, W. & J. Nichols (Eds), Evidentiality: The linguistic coding of epistemology, (pp. 261–72). Norwood, N.J.: Ablex.

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

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes 23*(4), 397–423.

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

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

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4–21.

Minitab 16 Statistical Software 2010. [Computer Software]. State College, PA: Minitab, Inc. www.minitab.com.

Kerslake, C. (1992). The role of connectives in discourse construction in Turkish. *Modern Studies in Turkish: Proceedings of the 6th International Conference in Turkish Linguistics*, 12-14 August 1992, (pp. 77-103). Eskişehir: Anadolu Üniversitesi.

Kornfilt, J. (1997) *Turkish*. London: Routledge.

Lee, D. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology, 5*(3), 37-72.

Leech, G. & Short, M. (1981). *Style in fiction*. London: Longman.

Lewis, G. (1969). *Turkish*. Oxford: Oxford University Press.

Liu, D. (2012). The most frequently used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes 31*, 25-35.

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 processing in Turkish. *Second ACL Workshop on Multiword Expressions: Integrating Processing*, July 2004, (pp. 64-71).

Ruhi, Ş. (1992) Ve'yi nerede inceleyelim: Sözdizimde mi? Anlambilimde mi, yoksa ...? [Where to study 'and': In syntax? Semantics, or…?] In C. Aksoy, G. Doğan & A. Kocaman (Eds.), *20. Yıl Yazıları,*104-132. Ankara: Karaca Dil Kursu,.

Ruhi, Ş. (1998) Restrictions on the interchangeability of discourse connectives: A study on *ama* and *fakat*. In L. Johanson et al. (Eds.), *The Mainz Meeting: Proceedings of the Seventh International Conference on Turkish Linguistics*, *Turcologica 32*, 135-153. Wiesbaden: Harrassowitz.

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

Sinclair, J. (2004). *Trust the text*. Oxford: Oxford University Press.

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

Zeyrek, D., Turan, Ü.D. & Demirşahin, I. (2008). Structural and presuppositional connectives in Turkish. In Benz, A., Kühnlein, P. (Eds.), *Constraints in discourse*, 131-137 Amsterdam: John Benjamins.

Zeyrek, D. (forthcoming). *On the distribution of the contrastive-concessive discourse connectives ama 'but/yet' and fakat 'but' in written Turkish*. Amsterdam: John Benjamins.

APPENDIX

The 20 top-ranked tri-grams in fiction and non-fiction texts

Rank	**Fiction Text**	English equivalent	**Freq.**	**Non-fiction Text**	English equivalent	**Freq.**
1	bir kez daha	once more	134	bir süre sonra	after a while	134
2	bir yandan da	besides	132	ne var ki	however	103
3	başka bir şey	something else	108	bir kez daha	once more	88
4	bir süre sonra	after a while	108	başka bir şey	something else	86

5	bir an önce	immediately	93	bir yandan da	besides	78
6	bir şey yok	there is nothing	63	ya da bir	or a	73
7	her zamanki gibi	as usual	61	her ne kadar	although	68
8	ne yazık ki	unfortunately	61	kısa bir süre	for a short time/period	62
9	ne de olsa	after all	60	ne olursa olsun	in any case	60
10	böyle bir şey	something like this	51	bir başka deyişle	in other words	59
11	ama yine de	‘(but/and) still, (and) yet, nevertheless, despite this’	48	ne yazık ki	unfortunately	58
12	ne var ki	however	44	çok önemli bir	a very important	49
13	ya da bir	or a	42	bu nedenle de	for this reason	47
14	ne olursa olsun	in any case	41	her şeyden önce	before everything	45
15	kısa bir süre	for a short time/period	39	ama yine de	‘(but/and) still, (and) yet, nevertheless, despite this’	43
16	belki de bu	maybe/perhaps this	39	bir an önce	immediately	39
17	en ufak bir	even the smallest/slightest	38	bir şey yok	there is nothing	38
18	işte o zaman	at that moment	36	o kadar çok	that/so much	38
19	öyle değil mi	don’t you agree / isn’t it so	35	çok büyük bir	a very big/great	38
20	ne kadar çok	the more	33	daha sonra da	and then after	37