Methodological Considerations for Multi-word Unit Extraction in Turkish

Ümit Mersinli
Mersin University
Mersin, Turkey
umitmersinli@gmail.com

Yeşim Aksan
Mersin University
Mersin, Turkey
yesimaksan@gmail.com

***Abstract*— Multi-word Unit (MWU) extraction in Turkish has its own challenges due to the agglutinative nature of the language and the lack of reliable tools and reference datasets. The aim of this study is to share the hands-on experience on MWU extraction in the ongoing projects using Turkish National Corpus (TNC) as the data source. Since Turkish still does not have a reference MWU set, the primary purpose of these projects is to form a reference MWU dictionary of Turkish which will serve as a resource to evaluate the performance of any extraction tool or technique. In this paper we will discuss methodological considerations for clarifying appropriate processes for Turkish MWU extraction. Techniques or suggestions compiled in this paper form an overall proposal for further Turkish-specific computational or statistical work. The linguistic perspective underlying the choices of a valid methodology is described in the first part of the study. In the second part, important methodological considerations are discussed through real examples from the TNC. In the conclusion, suggestions for an interdisciplinary approach and a hybrid methodology are summarized.**

Keywords—MWU extraction; multi-word; Turkish phraseology; Turkish National Corpus

I. INTRODUCTION

As Mel'čuk [1] states, "people speak not in words but in phrases" or in Firth's [2] words, as a well-known statement among linguists, "you shall know a word by the company it keeps". The importance of MWUs in any language-related area leads to a huge amount of work done especially for English.

For Turkish, on the other hand, the lack of a preliminary, well-documented, reference MWU lexicon to evaluate the performance of any linguistic, statistical or computational extraction methodology seems to be the basic challenge to overcome. Works of Oflazer et al. [3], Eryiğit et al. [4], Kumova & Karaoğlan [5], Aksan & Aksan [6], Durrant & Mathews-Aydınlı [7], Aksan, Mersinli & Altunay [8] and Mersinli & Demirhan [9] covers some aspects of Turkish phraseology but unfortunately, Turkish NLP literature is far from providing a comprehensive, reference MWU lexicon. In this respect, the purpose of this paper is to share the hands-on experience on MWU extraction projects using the Turkish National Corpus (TNC) [10] as the data source, rather than to provide finalized software, resources or methodology. The following sections will summarize the crucial points of the study in progress. In each section, sample data is provided for illustrative purposes only. They should not be regarded as finalized data sets of the ongoing study.

II. METHODOLOGICAL CONSIDERATIONS

According to Pecina [11], eliciting the best methodology for MWU extraction depends heavily on data, language, and the notion of MWU itself. However, these concerns are underestimated in current Turkish NLP literature. Thus, the methodological considerations discussed in this paper will emphasize the importance of some neglected aspects of MWU extraction in Turkish.

A. Choosing The Corpus

Most of the current studies on Turkish MWU extraction, focus on optimizing the statistical or computational processes or optimizing the sorting procedure of the outcome. The importance of the input, or corpus in our case, is often underestimated. In this part of the paper, we will deal with the necessary qualifications of a corpus to be used as input for MWU extraction in Turkish.

First, the difference between a linguistic corpus and a text archive needs to be clarified [12]. According to Sinclair [13], "a corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language" but not a random text collection of any available type. Second, a reference corpus should cover naturally occurring, contemporary language data and have a design to represent the language, if not a historical or specialized corpus. Third, a corpus should cover, if applicable, a variety of text-types and mediums of that language. In other words, the corpus should be a well-balanced and representative one to be used in MWU extraction.

In this respect, it is crucial to rely on a reference corpus like Turkish National Corpus in order to extract true rankings of the n-grams. The size of the TNC is 50,997,016 running words, representing a wide range of text categories spanning a period of 23 years (1990-2013). It consists of samples from textual data representing 9 different domains (98%) with 4978 documents and transcribed spoken data (2%) with 434 documents. Table (1) shows the distribution of texts in the written part of the TNC.

In addition, the annotation system of the TNC covers over 90 inflectional morphemes, all of which are compatible with modern Turkish linguistics studies. Analysis and tagging of

derivational morphemes are in progress and will provide insights for the relationship between word and multi-word forming processes of Turkish.

TABLE I. DISTRIBUTION OF TEXTS ACCORDING TO DOMAINS IN TNC-WRITTEN

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

Table (2) shows the MWU candidates derived from the written part of the TNC including 49,983,288 words. The top 5 multi-word candidates obtained from the written part of the TNC and from the newspaper articles section of it demonstrate how serious the differences between data extracted from a reference corpus and the data from a specialized corpus are.

TABLE II. THE 5 TOP-RANKED 3-GRAMS IN A REFERENCE CORPUS AND A SPECIAL CORPUS

Rank	**TNC_all**[a]	**Freq.**	**TNC_Newspapers**	**Freq.**
1	**bir süre sonra**	4419	recep tayyip erdoğan	555
2	**bir kez daha**	4000	**bir kez daha**	506
3	**ne var ki**	3360	başbakan recep tayyip	449
4	**başka bir şey**	3238	yönetim kurulu başkanı	442
5	**ne yazık ki**	3020	şöyle devam etti	367
6	**her ne kadar**	3012	**bir an önce**	367
7	**bir yandan da**	2993	genel başkan yardımcısı	323
8	**bir an önce**	2413	ahmet necdet sezer	316
9	**kısa bir süre**	2300	cumhurbaşkanı ahmet necdet	288
10	**ne olursa olsun**	2182	düzenlediği basın toplantısında	263

[a.] MWUs are in bold

As seen in Table (2), multi-word units are not only language specific but also text-type specific. Thus, relying on a text archive derived from the Web or a specialized corpus covering newspapers, for instance, is not a relevant approach to extract MWUs of Turkish, but it is a kind of approach used for extracting the MWUs of that specific text type. If the purpose of the extraction is to derive Named Entities, on the other hand, a Web-based, newspaper corpus may be the appropriate option in terms of choosing the corpus.

B. Optimizing the Input

As stated above, choosing and optimizing the input is an important part of our proposal. The basic shift from the conventional approaches is to make use of punctuation marks as a natural delimiter for MWU candidates. Thus, all punctuation marks and numericals in the corpus are replaced with line-breaks which serve as a splitter for n-grams. Since the primary concern of this study is not to extract proper nouns, all the corpus text is also lowercased to avoid duplicate n-grams. Table (3) is a sample raw text and its optimized version.

TABLE III. CORPUS OPTIMIZATION FOR MWU EXTRACTION

Raw text
Günlerden bir gün , okuldan evine dönen Hetzer,sırt çantasından çıkardığı yepyeni bir kitabı, babasına gösterir.
Optimized text
günlerden bir gün okuldan evine dönen hetzer sırt çantasından çıkardığı yepyeni bir kitabı babasına gösterir

After the optimization, the lower-cased, sentence-splitted, punctuation-delimited, ASCII-coded TNC texts are processed in Text-NSP [14], for obtaining all the sample lists presented in this paper. Moreover, for the sake of simplicity, no associative measures are used for extracting MWUs, and all the values represent the observed frequencies of the data. A detailed discussion on associative measures applied on Turkish MWU candidates can be found in Kumova-Metin & Karaoğlan [5] and Mersinli [15].

C. Looking Beyond Words

It is a well-known phenomenon that an inflected Turkish verb is actually a sentence in English, in most cases. The same is also true for other phrases like postpositions or connectives. We can easily observe that most of the connectives in English are actually suffix-word pairs in Turkish such as *-mAk için* "in order to", *-A göre* "according to" etc. The point here is that any multi-word in any language may appear as single words, multi-words, suffixes or suffix-word pairs in any other language and vice versa. Thus, especially dealing with an agglutinative language, suffix-word pairs need to be taken into serious consideration. Postpositional phrases, for instance, requires specific suffixations in the preceding word in Turkish.

Below are the most frequent suffix-word pairs of Turkish, extracted with the help of the annotation framework of the TNC. The suffixes are annotated according to their functions as nominalizers, case markers or person/number agreements in the table. The frequencies are extracted from bigrams including the first word ending with the given suffix and the second word as a whole.

TABLE IV. MOST FREQUENT SUFFIX-WORD PAIRS IN TURKISH

Suffix_type	Freq.	Example	English
nzmk__için	58535	etmek_için	in order to
dat__göre	37850	buna_göre	according to
abl__sonra	36515	olduktan_sonra	after
p3s__için	33514	olduğu_için	since
p3s__gibi	31306	olduğu_gibi	as it is
dat__kadar	28429	bugüne_kadar	until
nzmk__üzere	17728	olmak_üzere	almost
gen__için	15336	bunun_için	for this
acc__olarak	11895	sonucu_olarak	as a result of
pl__için	9990	onlar_için	for them

As Table (4) demonstrates, the term 'multi-word' in Turkish should also cover "suffix-word" pairs as a term which we may call a "multi-morpheme unit". Looking for in-word or intra-word units in Turkish may be the solution for most of the challenges encountered in MWU extraction processes.

Also the inflectional patterns in Turkish should be considered as multi-words or, in a more appropriate terminology, multi-morpheme units, since their distribution among different text-types provides evidence for their functional unity specific to certain text-types. Below are the 6-morphgrams and their distribution among 3 text-types in the TNC. The tagset includes the functions such as causative, passive, auxiliary verb, aorist, nominalizer, adverbial, negation, verb I, necessity, perfective, imperfect, person agreement, possessive, accusative, locative, copular etc. in their abbreviated forms. Almost all 6-morphgrams start with some voice suffixes and end with 3rd person singular suffix as seen in the table.

TABLE V. SAMPLE MORPHGRAMS AND THEIR DISTRIBUTION AMONG TEXT-TYPES IN TURKISH

6-morphgrams	Academic	Fiction	Newspapers
caus+pasv+va1+nzma+p3s+acc	27	0	1
caus+pasv+va1+neg+aor+3s	444	76	63
caus+pasv+aor+vi+avsa+3s	386	25	46
caus+pasv+imprf+vi+past+3s	277	164	47
caus+pasv+imprf+vi+perf+3s	4	16	4
caus+pasv+neg+necc+cop+3s	220	3	12
caus+pasv+neg+nzma+p3s+acc	24	4	13
caus+pasv+neg+perf+cop+3s	172	5	6
caus+pasv+nzma+p3s+cop+3s	838	11	29
caus+pasv+nzma+p3s+loc+kia	85	2	6

Table (5) clearly demonstrates that causative+passive inflection is specific to academic Turkish and can be regarded as a multi-morpheme unit in itself. Although very rare in usage, these verbal morphgrams can be extended to 9 morphemes in Turkish as in the inflected verb, *çıkartılabilinirdi* which starts with the verb *çık-* and includes the suffixes causative, causative, passive, auxiliary_verb, passive, aorist, verb_i, past_tense, 3rd_person_singular in the given order. The inflected verb can be translated as "it could be made possible to extract" which is a full sentence in English and thus, again, blurs our notion of 'word' in the term 'multi-word unit'.

D. Bidirectional Sorting

Another common practice in MWU extraction can be summarized as sorting n-grams using associative measures or a combination of them, providing a cut-off point and regarding the remaining top n-grams as MWUs. As discussed in Mersinli [15], the relevance of relying only on sorting the n-grams without any linguistic filtering is questionable. A hybrid approach combining quantitative sorting and qualitative filtering techniques, as in Seretan et al. [16], seems more productive for Turkish if the purpose is to prepare a reference MWU set and to describe multi-word formation processes in Turkish.

Below are the associative measures stated as linguistically relevant for the given n-grams in Turkish [15]. Since the 2-grams include most of the sub-MWUs in Turkish, although most of the measures are for these candidates, it seems reasonable to rely on observed frequencies of 3-grams for extracting MWUs in Turkish.

TABLE VI. RELEVANT ASSOCIATIVE MEASURES FOR TURKISH

n-grams	Measures
2-grams	T-score, Fisher's Exact Test (left-sided), Log-likelihood,True Mutual Information, Poisson-Stirling Measure
3-grams	Poisson-Stirling Measure
4-grams	Log-likelihood

With that concern in mind, in order to measure the fixedness of 3-grams, since they are more likely to include as MWUs in a Turkish dictionary, we have used the frequencies of inner components, such as the frequency of the first two words and the last two words of 3-grams. If the difference between those values are high, then it is regarded as an evidence declaring that the given 3-grams is not a MWU but includes 2-grams that are more fixed than the whole 3-grams.

To be more specific, Table (7) shows the ranking of the values gained by subtracting the frequency of the last two words from the frequencies of the first two, in a given 3-gram. The MWUs within the given 3-grams are in bold shows the fixedness of the ones in the center of the ranking.

TABLE VII. BIDIRECTIONALLY SORTED SAMPLE 3-GRAMS

ABC	Freq	Freq.AB	Freq.BC	Freq.(AB - BC)
korkacak **bir şey**	50	50	15360	-15310
konuda **bir şey**	51	51	15360	-15309
aklına **bir şey**	51	51	15360	-15309
yapabileceği **bir şey**	51	51	15360	-15309
bildiğim **bir şey**	54	54	15360	-15306
……				
ne yazık ki	3020	3020	3020	0
her zamanki gibi	992	992	992	0
en ufak bir	849	849	849	0
her ikisi de	804	804	804	0
ittihat ve terakki	649	649	649	0
……				
ya da bunun	51	13650	51	13599
ya da siyasi	50	13650	50	13600
ya da karşı	50	13650	50	13600
ya da üçüncü	50	13650	50	13600
ya da kültürel	50	13650	50	13600

As seen in Table (7), a bidirectional sorting reveals the MWUs in the center even without applying any statistical associative measure and provides evidence for the 2-gram MWUs within the given candidates. The results of setting double thresholds based on such a simple measure points out that the relevance of any sorting practice does not rely on the complexity of the formulae we use.

E. Lexico-grammatical Filtering

'Colligation' is another key term that is important in identifying the MWUs in a given set of candidates. As defined by Baker [17], a colligation is "a form of collocation which involves relationships at the grammatical rather than the lexical level". For rich morphology languages, then, grammatical relations between two or more words becomes important since they actually declare the constraints that prevent some frequent n-grams from becoming multi-words, or letting some less frequent ones become multi-word units.

Thus, in a hybrid approach, sorting and filtering are two basic processes, being the first statistical and the later rule-based. In order to provide the filtering rules for MWUs and non-MWUs linguistically, we have classified grammatical or colligational patterns of the MWU candidates into 3 categories, presented with examples from the TNC, below.

TABLE VIII. CLASSIFICATION OF COLLIGATIONAL PATTERNS OF N-GRAMS

Category 1 – Complete structures: MWU patterns		
Sample colligational pattern	*n-gram*	*English*
AJ+bare DT+bare NN+nom	kısa bir süre	*(in) a short time*
AJ+bare DT+bare NN+loc	etkin bir şekilde	*in an efficient manner*
Category 2 – Sub-patterns: Non-closed, potential sub-MWUs		
Sample colligational pattern	*n-gram*	*English*
AV,bare_AJ,bare_DT,bare	çok önemli bir	*a very important*
Category 3 –Incomplete structures: non-MWU patterns		
Sample colligational pattern	*n-gram*	*English*
PP,bare_AJ,bare_DT,bare	için önemli bir	*an important ... for*
PP,bare_AJ,bare_DT,bare	kadar geniş bir	*as a broad ... as*

The categories in Table (8) allow filtering the MWUs and non-MWUs as well as reserving partial ones that may be used to identify sub-MWUs. In brief, Category 3 candidates are filtered out, Category 1 is filtered in and Category 2 candidates are reserved for identifying 4-gram MWUs. Since the identification of sub-MWU strings is problematic not only for MWU extraction but also for all lexical frequencies in any language, it requires separate techniques, and it is out of the scope of current study.

Extracting colligations also provides a general ranking based on grammatical patterns of MWU candidates and makes the filtering process more linguistically relevant. Below is the top ten 3-gram colligations in the TNC. Table (9) demonstrates that 3-word units in Turkish mostly provides a closed projection including a specifier, a modifier and a head, making 3-grams worth extracting more than 2-grams including mostly light verb constructions or reduplications.

TABLE IX. CLASSIFICATION OF COLLIGATIONAL PATTERNS OF N-GRAMS

	Colligation	**Sample 3-grams**	**English**
1	AV,bare_AJ,bare_DT,bare	çok önemli bir	a very important
2	AJ,bare_DT,bare_NN,nom	kısa bir süre	a short time
3	NN,nom_CJ,bare_NN,nom	radyo ve televizyon	radio and television
4	DT,bare_NN,nom_AV,bare	bir süre sonra	after a while
5	AJ,bare_CJ,bare_AJ,bare	ekonomik ve sosyal	economic and social
6	CJ,bare_AV,bare_AV,bare	ama yine de	but still
7	NN,nom_NN,nom_CJ,bare	ne var ki	however, yet
8	AJ,bare_DT,bare_NN,loc	etkin bir şekilde	efficiently
9	AV,bare_DT,bare_NN,nom	böyle bir şey	such a thing
10	CJ,bare_AJ,bare_DT,bare	ile ilgili bir	a … related to

III. CONCLUSION

The methodological considerations discussed in this paper show that MWU extraction is rather a trial-and-error process for a given language. Thus, any attempt, be statistical, computational or linguistic is worth sharing in an inter-disciplinary manner to fill the gap in this area. A reference MWU set or a MWU dictionary, for that purpose, will serve as an input not only for linguistics but also for all related areas of study. Fig.1 summarizes a sample recursive process followed in the proposed strategies.

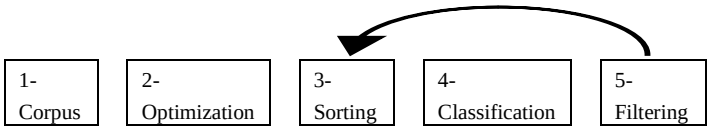

Fig. 1. Basics of the proposed strategy

Considering the fact that Turkish is an agglutinative language and has little to do with words but rather operates on suffixes, the term 'multi-morpheme unit' (MMU) seems more operational for further cross-linguistic studies. In addition, lexico-grammatical constraints in MMU forming are as important as the observed frequencies of any MMU candidate and thus colligational analysis and filtering of n-grams should be a part of any strategy that includes statistical ranking of MMU candidates.

This paper briefly summarized some methodological considerations for multi-morpheme unit (MMU) extraction in Turkish. The purpose of the study is to discuss some ignored aspects of MMU extraction in Turkish and provide an overall idea on the methodological considerations we faced with. Turkish lexicon includes more MMUs than already documented. Any technical or linguistic contribution will be of great importance and a hybrid, inter-disciplinary approach may be the answer to most of the questions in the field.

MMU extraction is some reverse engineering of the MMU forming processes in our minds. Only a process-based approach may provide data for linguistics of Turkish. A product-based approach, or extracting a reference MMU set, however, can serve as an initial step for identifying the grammatical constraints that governs the MMU forming processes in Turkish. Interdisciplinary studies conducted by engineers and linguists are of great importance in this sense, that, not only MMUs but also the rules underlying the process of forming them can only be described by such collaborative studies.

ACKNOWLEDGMENT

This work is supported by a grant from Scientific and Technological Research Council of Turkey (TÜBİTAK, Grant No: 115K135).

REFERENCES

[1] Mel'čuk, I. A.: Phrasemes in language and phraseology in linguistics. In: Everaert, M., van der Linden, E.J, Schenk, A. and Schreuder, R. (eds.) Idioms: Structural and Psychological Perspectives. Lawrence Erlbaum, Hillsdale, NJ (1995)

[2] Firth, J.R.: A Synopsis of Linguistic theory 1930-1955. In: Palmer, F. (ed). Selected Papers of J. R. Firth, Longman, Harlow (1968)

[3] Oflazer, K., Çetinoğlu, Ö. and Say, B.: Integrating morphology with multi-word expression processing in Turkish. In: Proceedings of the Workshop on Multiword Expressions: Integrating Processing (MWE '04). Association for Computational Linguistics, pp. 64-71 (2004)

[4] Eryiğit, G. et.al. Annotation and Extraction of Multiword Expressions in Turkish Treebanks. In Proceedings of the 11th Workshop on Multiword Expressions: MWE 2014. June 4, 2015 Denver, Colorado, USA. pp.70-76. (2015)

[5] Kumova-Metin, S., Karaoğlan B.: Collocation extraction in Turkish texts using statistical methods. In: 7th International Conference on Natural Language Processing (LNCS-ISI) IceTAL 2010, pp. 238-249 (2010)

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

[7] Durrant, P., Mathews-Aydınlı, J.: A function first approach to identifying formulaic language in academic writing. English for Specific Purposes, 30, 58-72 (2011)

[8] Aksan, Y., Mersinli, Ü. and Altunay, S.: Colligational analysis of Turkish multi-word units. Paper presented at CCS-2015, Corpus-Based Word Frequency: Methods and Applications. 19-20 February 2015. Mersin University, Turkey (2015)

[9] Mersinli, Ü. and Demirhan, U.: Çok sözcüklü kullanımlar ve ilköğretim Türkçe ders kitapları. In: Aksan, M. ve Aksan, Y. (eds.). Türkçe Öğretiminde Güncel Çalışmalar.: Mersin Üniversitesi, Mersin (2012)

[10] Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T., Mersinli, Ü., Demirhan, U. U., Yılmazer, H., Kurtoğlu, Ö., Atasoy, G., Öz, S., Yıldız, İ.: Construction of the Turkish National Corpus (TNC). In: Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), pp. 3223-3227 (2012)

[11] Pecina, P.: Lexical association measures and collocation extraction. Language Resources and Evaluation, 44, 137-158 (2010)

[12] Aksan, M. and Aksan, Y.: Linguistic corpora: A view from Turkish. In: Oflazer, K. and Saraçlar, M. (eds.) Studies in Turkish Language Processing. Springer Verlag, Berlin (forthcoming)

[13] Sinclair, J. McH. and Renouf, A.J.: A lexical syllabus for language learning. In: McCarthy, M.J. and Carter, R.A. (eds.) Vocabulary in Language Teaching. Longman, London. (1987)

[14] Banerjee, S. and Pedersen, T.: The design, implementation, and use of the ngram statistics package. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, pp. 370-381 (2003)

[15] Mersinli, Ü.: Associative measures and multi-word unit extraction in Turkish. Journal of Language and Literature 12 (1), 43-61 (2015)

[16] Seretan, V., Nerima, L., and Wehrli, E.: Multi-word collocation extraction by syntactic composition of collocation bigrams. Amsterdam Studies in the Theory and History of Linguistic Science. Series Iv, Current Issues in Linguistic Theory, 260, 91-100 (2004)

[17] Baker, P., Hardie, A., & McEnery, T. A glossary of corpus linguistics. Edinburgh: Edinburgh University Press. (2006)