Comprehensive Annotation of Multiword Expressions in Turkish

Kübra Adalı
Dep. of Computer Engineering
Istanbul Technical University
Maslak, Istanbul 34369
Email: kubraadali@itu.edu.tr

Tutkum Dinç
Dep. of Linguistics
Istanbul University
Beyazıt, Istanbul
Email: tdinc@iu.edu.tr

Memduh Gökırmak
Dep. of Computer Engineering
Istanbul Technical University
Maslak, Istanbul 34369
Email: gokirmak@itu.edu.tr

Gülşen Eryiğit
Dep. of Computer Engineering
Istanbul Technical University
Maslak, Istanbul 34369
Email: gulsenc@itu.edu.tr

Abstract—**Multiword expressions (MWEs) are pervasive in Turkish, as in many other languages. There are many challenges related to MWEs in Natural Language Processing. The scarcity of annotated language resources is one of the most prominent for lesser-studied languages and as always development of these resources requires a noteworthy effort. This paper is the first study which specifically focuses on the development of Turkish MWE resources for the purpose of 1) the categorization of different MWE types in Turkish 2) use in MWE identification, and 3) use in research focusing on interleaving between MWE identification and parsing. For these purposes, we annotated two Turkish treebanks (IMST and IWT) with 11 MWE categories and 8 subcategories for the MWE category Named Entity.**

1. Introduction

As the name implies, multiword expressions are composed of multiple words that together produce an idiosyncratic meaning or have a distinctive syntactic role. They pose several challenges for natural language processing tasks as well as in language acquisition for non-native speakers. As a result, they have been an important issue covered in many studies since the inception of the field of NLP. The reader may consult many comprehensive studies for a complete discussion of MWEs ([1], [2], [3]). Their extraction and processing within NLP applications is still a very active research topic as may be seen by many recent workshops ([4], [5]) and research initiatives (e.g. EU PARSEME Cost Action [6]).

Annotated data sets and lexicons are very valuable resources for MWE processing tasks. A comprehensive annotation of MWEs is a troublesome and exhaustive process. Many languages including Turkish suffer from lack of MWE annotated language resources. Manually annotated treebanks are syntactically annotated corpora and are valuable resources for parsing research. The annotation of MWEs on treebanks would undoubtedly help investigations on the integration of MWE identification and parsing studies. As a result, there are many efforts to annotate MWEs on treebanks. Unfortunately, there is as of yet no common standard on how to annotate them. The aim of WG4 of PARSEME is to establish such standards for treebanks. [7] makes a survey of MWE annotated treebanks. Some of these are the Prague Dependency Treebank [8], French Dependency Treebank [9], Penn Treebank (a constituency treebank for English) [10].

Although there have been some previous attempts [11], [12] to build MWE annotated treebanks for Turkish, this study is the first comprehensive annotation of MWEs on Turkish treebanks, being a fully manual annotation with detailed fine categories. This study is also a first attempt to define suitable categories for the MWE annotation of Turkish, and we believe this will also aid the creation of multi-lingual MWE annotation guidelines. Two existing Turkish dependency treebanks (IMST [13] a treebank of well-edited texts and IWT [14] a Web treebank) are annotated with 11 main MWE categories (nominal compounds, duplications, verbal compounds, light verb constructions, compounds constructed with determiners, conjunctions, formulaic expressions, idiomatic expressions, proverbs and named entities) and 8 named entity sub-categories (Person, Location, Organization Names, Date and Time Expressions, Percentage, Monetary Expressions, Miscellaneous Numerical Expressions).

The remainder of the paper is structured as follows: Section 2 gives information about previous MWE studies in Turkish and introduces our proposed MWE categories, Section 3 presents the annotation process and the statistics, and Section 4 is the conclusion.

2. MWEs in Turkish

There are a couple of studies which focus on MWE discovery [15], MWE annotation [11], [12] and MWE identification [12], [16] in Turkish. [15] employs two simple statistical methods, a Chi-square hypothesis test and mutual information in order to discover Turkish collocations. [11] reveals that the performance of parsing is affected differently by the concatenation of different MWE types' components. The most recent study on MWEs is [12] in which a coarse, undifferentiated annotation of MWEs took place and different lexical models for MWE identification including automatic named entity recognition were tested, demonstrating that their extraction model improves the accuracy of MWE extraction by a dependency parser [17] and the extraction tool of [16].

Similar to other languages, MWEs poses interesting challenges for Turkish. Especially, the variability of MWE instances are very high due to the agglutinative and morphologically very rich nature of this language. The constituents of a MWE may be inflected, resulting in a high number of different surface forms [16], [18]. To give an example the MWE "aklına gelmek" (*to come to mind*) may appear in different forms by taking personal agreement, tense, aspect and modality suffixes. In the sentence "Aklıma gelmedi" (*It didn't come to my mind.*), both of the components underwent inflection and are different from their lemma forms: the first word "aklına" (*to the mind*) with 1st person possessive agreement suffix in dative form and the second word "gelmek" (*to come*) in past tense with 3rd person singular agreement. Non-compositionality and discontinuity are common challenges of MWEs which also appear in Turkish.

In this section, we introduce the categories that we defined for MWE types in Turkish which we believe will provide the opportunity to address the problems of different types separately. The sub-categorization of MWEs will also pave the way for further investigations on hierarchical approaches for MWE identification and its integration into parsing. With this aim, we define 11 categories of MWEs which we detail in the remaining of this section.

2.1. Nominal Compound MWEs

As described in [19], noun compounds are word like units made up of two nominals. Our definition of nominal compound MWEs differs from this general definition in that they comprise only a subset of noun compounds used commonly enough to express a wide concept or class. These consist of bare compounds (the components do not take extra suffixes to mark the relation between them) and -(s)I compounds (the first component has no suffixes while the second one is marked with the third person possessive suffix -(s)I) [19] . To give some examples, "kadın çorabı" (*hosiery*), "hakem heyeti" (*arbitration court*) ,"kredi kartı" (*credit card*), "diş macunu" (*toothpaste*). As may be observed from the examples, the overall sense of this type of MWE may be discerned from its components.

2.2. Duplication MWEs

Duplications are linguistic units that are formed mainly by duplicating a nominal or modifier. The production of the second word can be done in several ways: the reproduction of the exact words, synonymous words, antonymous words, onomatopoeic words, gibberish words. The examples that refers to each are "çabuk çabuk" (*very quickly*, *or lit. quick quick*), "mal mülk" (*property*, *or lit. property property*), "aşağı yukarı" (*almost, nearly*, *or lit. down up*), "adı sanı" (*public profile*, *or lit. name and fame*), "paldır küldür" (*pell-mell*, *or lit. pell mell*). Duplications with an interrogative particle in between are also considered to be duplication MWEs (e.g., "güzel mi güzel" (*so beautiful*)). Duplications can strengthen the meaning of the main word, turn an adjective into an adverb, or add an idiomatic meaning. We decided not to include the 'm'-duplication (where a word is repeated with the first letter replaced with 'm' in the second occurrence) as a type of duplication MWE.

2.3. Verbal Compound MWEs

In this type, the components form the MWE without undergoing a significant semantic change. They are formed with a noun and a verb[1]. This type of MWEs may be inflected more frequently than other types due to the verbal nature of their constructions. Examples of this pattern can be like: "karar vermek" (*to decide*), "söz vermek" (*to promise*).

2.4. Light Verb Construction MWEs

Light Verb Construction MWEs are formed by six auxiliary verbs which are "olmak" (*to be*), "etmek" (*to do*), "yapmak" (*to make*) , "kılmak" (*to render*), "eylemek" (*to make*) and "buyurmak" (*to order*). Together with a preceding nominal, these auxiliary verbs behave as a finite verb. The verb phrase is a construction which has its own meaning, which can be idiomatic or relatively similar to that of its components. These MWEs can be easily detected using morphosyntactic information such as the existence of an auxiliary verb at the end of a verb phrase. Some examples are: "aşık olmak" (*fall in love*), "sinir etmek" (*to aggrevate*), "veda etmek" (*to bid farewell*), "yemek yapmak" (*to cook*), "geçersiz kılmak" (*to revoke*), "emir buyurmak" (*to give order*). However, not every construction with the aforementioned auxiliary verbs falls under this category. For example, MWEs like "aforoz etmek" (*to excommunicate*) and "ah etmek" (*to sigh*) are considered idiomatic expressions and will be handled under that category.

2.5. Compound MWEs Constructed with Determiners

This category consists of compounds having at least one determiner component. The compounds "her şey" (*everything*), "şu an" (*now*), "bir daha" (*again/never*) may be given as examples for this category. Differing from the previous compound MWE categories, MWEs of this category type may be used in different roles (nominal, adjectival or adverbial) in a sentence.

2.6. Conjunction MWEs

Conjunction MWEs are a sort of transition phrase and are used to concatenate two sentences. Some examples of this category may be given as the followings: "bu arada" (*by the way*), "bu yüzden" (*therefore*), "o halde" (*then*),

[1] Excluding light verb constructions which are also a special type of verbal MWEs collected under a separate category.

"bu sebeple" (*for this reason*) etc. While exhibiting some semantic flexibility, the components of MWEs in this category largely retain their original meaning. This category excludes constructions formed by the addition of an enclitic intensifier such as "de", "ise" , "ki" (e.g., "öyle ki'(*so that*), '"ya da"(*or*)).

2.7. Formulaic Expression MWEs

MWEs in this category satisfy the following semantic and syntactic conditions. As the semantic condition, the MWE should carry the meaning of well wishing or gratitude. For the syntactic condition, the MWE is an independent clause, mostly with an elided verb implied to be in a subjunctive mood. Some examples are : "Ellerine sağlık (olsun)" (*May God bless your hands*), "Görüşmek üzere" (*See you soon*), "Hoşça kal" (*Good Bye*). MWEs in this category may rarely resemble light verb constructions that also carry a sense of gratitude, such as "teşekkür etmek" (*to thank*), "rica etmek" (*to request*)

2.8. Idiomatic Expression MWEs

Idiomatic expressions are MWEs with non-compositional meanings; i.e., the meaning of the MWE differs from the literal meaning of its components. For example: "etekleri zil çalmak" (*to be very happy*, *or lit. ring the bells on the skirt*), "gemi azıya almak" (*to get out of control or lit. to scratch the bit with grinders*) etc. This type of MWEs are quite challenging for MWE identification due the ambiguity between idiomatic and literal use. To give some examples: "ayvayı yemek" (*to be in a worrisome and bad situation*, *or lit. to eat the quince*) and "ayağa kalkmak" (*to protest*, *or lit. to stand up*). In these cases, there is no morphosyntactical difference between the two utilizations of the word group as an idiom or as an ordinary phrase carrying literal meaning, hence it could be difficult to detect the MWE using the contextual information.

2.9. Simile Expressions MWEs

Similes are expressions comparing two things, in an often striking manner, using a connecting word (e.g., the word "gibi" (*like*) in Turkish). We include under this category not every comparison but only those in frequent use. The syntactic construction has two main parts: the figurative part and post-positional particle which refers to only one word "gibi" (*like/alike*). Here are some examples: "Agop'un kazı gibi" (*voraciously*), "damdan düşer gibi" (*out of the blue*), "Avcunun içi gibi" (*well known*), "kedinin ciğere baktığı gibi" (*anxiously*) etc.

2.10. Proverb MWEs

Proverbs are idiomatic and frozen sentences [20] with no words changing or undergoing inflections. Consequently, the category can be considered the easiest one to identify as an MWE. They often describe some observation or experience with didactic intent.

Some examples are given below:

- "Damlaya damlaya göl olur."
  lit. (By dribbling) (a lake) (composes) .
  (*Many a little makes a mickle.*)
- "Güneş balçıkla sıvanmaz."
  lit. (The sun) (with mud) (can not be covered) .
  (*The truth can not be hidden.*)
- "(Yalancının mumu) (yatsıya) (kadar) (yanar)."
  lit. (The candle of the lier) (until isha) (burn) .
  (*The truth can not be hidden.*)

2.11. Named Entities

In our annotation we consider a Named Entity to be a set of tokens denoting some unique entity in the real world. Their syntactic patterns and semantic properties are fixed, and they are not necessarily multi word expressions. Since most of the time they consist of two or more words, they are also treated as an MWE category. Named entities include 8 subcategories, namely; ENAMEX types (Person, Location, Organization Names), TIMEX types (Date and Time) and NUMEX types (Percentage, Monetary Expressions, Miscellaneous Numerical Expressions). We follow the MUC-6 [21] guidelines for our named entity definitions.

Person

This tag denotes persons, referred to by name, and excludes any titles or alternate references other than the name of the person in question. The examples: "Başbakan *Turgut Özal*" (Prime Minister *Turgut Özal*), "Maliye Bakanı *Ali Babacan*" (Finance Minister *Ali Babacan*).

Location

Denotes the proper name of a location. For example: "*Amerika Birleşik Devletleri*'nden mektup geldi" (A letter came from the *United States of America*).

Organization

This subcategory is used for the name or the group of names of an organization such as "*Birleşmiş Milletler* kararı uyguladı." (*United Nations* enforced the judgment.).

Date

Expresses an absolute date. As an example: "Doğum tarihi *25 Temmuz 1987* 'di." (Her birth date is *25th of July in 1987*).

Time

In this category, the named entity states an absolute time. The examples are : “Saat *6:30*’da film başlıyor. ” (The film starts at *6:30*.), “Sınavı bugün *10:30*’daymış.” (Her exam is today at *10:30*)

Percentage

This category is used to represent percentage information e.g. “Devrelerin *yüzde yirmisi* arızalı.” (*Twenty percent* of the circuits are defective.), “Adayların *yüzde sekseni* sınavdan kaldı.” (*Eigthy percent* of the candidates have failed the examination).

Money

For this category, the word group denotes an expression of money or monetary value. The example is : “O kitaba *altmış lira* verdim.” (I paid *sixty liras* for that book.)

Miscellaneous Number

We have diverged from the MUC-6 guidelines in this tag, and marked cardinal numbers with their own named entity tag. To give an example: “*Altı yüz bin* araba satılacak” (*Six hundred thousand* cars will be sold.).

3. Annotation

The annotation process was carried out in two stages on both treebanks, with two annotators carrying out both on each treebank. The stages are as follows:

- The Annotation of NE categories
- The Annotation of MWE categories except Named Entities

3.1. NE Annotation

In the NE annotation process we have annotated the entities according to the categories we described in the previous section. Figure 2 shows an example dependency tree consisting an organization named entity. In our annotation we have largely followed the MUC-6 [21] guidelines for the annotations, with the addition of a single extra category for miscellaneous numerical expressions. The MUC-6 guidelines establish a standard for marking plain text sentences with XML tags, however, we have annotated sentences in CoNLL format in which morphological information and dependency relations are marked. We have added two extra columns to the data, one marking the type of the named entity, and another marking possible following items in a collocative named entity. This way of annotating the named entities is particularly well suited to Turkish as named entities tend to be adjacent, and their dependencies relations are overwhelmingly organized left to right from dependent to head. Inflectional suffixes are excluded from the named entity in the plain text format marked with XML tags, but the entire token in the CoNLL file is marked as a part of the named entity. As the lemma is given in each CoNLL token, this does not result in a loss of data. Figure 1 shows an example CoNLL annotation for the sentence “Ben Arçelik’e sordum 31 Aralık’a kadarmış.” (*I asked Arçelik, it’s until December 31st.*) which examplifies such a case on the word “Aralık” (December) inflected with a dative case marker.

TABLE 1. THE NUMBERS OF TYPES OF NEs AND THE KAPPA COEFFICIENTS

NE Type	IMST	IWT	IWT	Kappa Co.	Total
		An.-1	An.-2		
Person	1071	385	426	0.88	1497
Organization	418	401	503	0.64	921
Location	491	260	274	0.79	765
Money	54	45	48	0.98	102
Percentage	44	8	7	0.99	51
Misc. Number	427	-	317	-	744
Date	106	59	76	0.87	182
Time	20	10	17	0.95	37
Total	2631	1168	1668	-	4299

Table 1 shows the numbers of NE categories in two treebanks. We annotated IMST [13] with detailed NE types for the first time, however IWT [14] was annotated for MWEs previously in a recent study [22]. This made possible to calculate the Cohen’s Kappa coefficient[2] [23] in order to evaluate the inter-annotator agreement between the current and the previous annotation [22]. From the scores it is seen that there is sufficient agreement between our annotator and the previous annotator.

3.2. MWE Annotation

During the original dependency annotations of both treebanks, the annotators were asked to annotate the interrelations of a multiword expression with a single catch-all dependency type (named as MWE as well) [12]. But the annotation was limited to only this dependency relation without any extra information on types of the MWEs. In this work, we refine previous annotations by inspecting all the treebank sentences and reannotating all the MWEs with finer categories.

We have done the annotation with the participation of the linguistics student who oversaw the categorization of MWEs. The annotation of MWEs was performed in a number of iterations of an annotation and check cycle. We automatically checked the annotation for dependency-related errors, and manually examined cases marked in previous annotations ([12], [13], [24]) were not marked and vice versa. This iteration was carried out until all problematic cases were handled. The first MWE annotation of the treebanks is complete. We plan to have the MWEs annotated again by

[2] During the calculation of the Kappa coefficient we saw that in an unmodified Kappa value the agreement rate was too high to be meaningful. We used a weight value of 0.01 for the number of tokens both annotators did not annotate, resulting in a much more meaningful statistic.

```
Ben Arçelik’e   sordum         31 Aralık’a     kadarmış.
I   Arçelik.DAT ask.PAST.1-SG  31 December.DAT until.EVID.3-SG.
I asked Arçelik, it’s until December 31st.
```

ID	Surface Form	Dependency Head	Dependency Relation	NE Type	Next Word
1	Ben	3	SUBJECT		
2	Arçelik’e	3	MODIFIER	ORGANIZATION.ENAMEX	
3	sordum	8	COORDINATION		
4	31	5	MWE	DATE.TIMEX	5
5	Aralık’a	8	MODIFIER	DATE.TIMEX	
6	_	7	DERIV		
7	_	8	DERIV		
8	kadarmış	0	PREDICATE		
9	.	8	PUNCTUATION		

Figure 1. The annotation format of an example sentence

```
Maliye  Bakanlığı       konuyla     ilgili  açıklama  yaptı.
Finance Ministry.3-POSS subject.INS related statement make.PAST.3-SG
The Ministry of Finance made a statement on the issue.
```

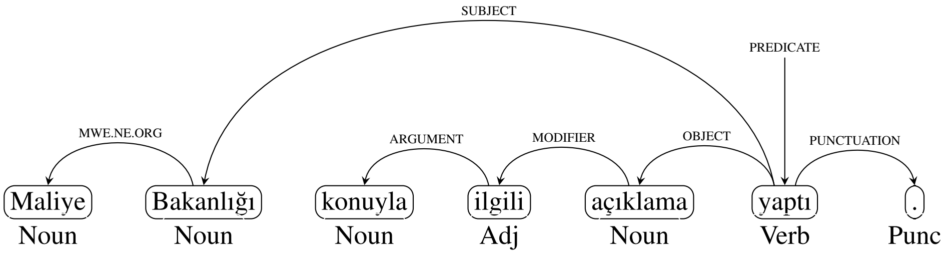

Figure 2. An example dependency tree showing an organization named entity

TABLE 2. THE DISTRIBUTION OF THE NUMBERS OF CATEGORIES OF MWES IN TWO TURKISH TREEBANKS

MWE Type	IMST	IWT	Total
Named Entities	910	439	1349
Compound	525	545	1070
Conjunction	32	41	73
Duplication	209	130	339
Formulaic Expression	22	221	243
Idiomatic Expression	773	598	1371
Lightverb Construction	537	648	1185
Nominal Compound	136	156	292
Proverb	3	4	7
Simile Expression	12	7	19
Total	3159	2789	5948

another linguistics student, and calculate the agreement as in the named entity annotation.

Table 2 gives the results of distribution of MWE categories in two treebanks. As seen on the Table 2, one of the biggest categories of MWE is named entities, which means the performance of a Named Entity Recognition system used in MWE extraction will substantially affect the performance of the system. The other large category is idiomatic expressions, which makes MWE extraction a challenging issue, as we are obliged to deal with the particular challenges of idiomatic expressions to build a high performance system.

4. Conclusion

In this paper, we proposed a basis for Turkish MWE and NE categorization to be used as a working guide in annotation. The categorization framework, which was prepared by taking into account the idiosyncratic features of Turkish, consists of 11 categories of MWEs. We performed annotations on two Turkish treebanks using the proposed framework. We annotated the categories of MWEs as the first annotation task and the annotation of NEs and their subcategories as the second on the Turkish treebanks. For the annotation task, we enlisted the aid of linguistics researchers that have expertise on the morphosyntactic and semantic features of Turkish.

The categorization framework that we defined in this study and the annotated treebanks will hopefully be used in future studies in the annotation and identification of MWEs in Turkish.

Acknowledgments

We would like to acknowledge that this work is part of a research project entitled "Parsing Web 2.0 Sentences" subsidized by the TUBITAK (Turkish Scientific and Technological Research Council) 1001 program (grant number 112E276) and part of the ICT COST Action IC1207 PARSEME (PARSing and Multi-word Expressions).

References

[1] I. A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger, "Multiword expressions: A pain in the neck for NLP," in *Computational Linguistics and Intelligent Text Processing*. Springer, 2002, pp. 1–15.

[2] I. Arnon and N. Snider, "More than words: Frequency effects for multi-word phrases," *Journal of Memory and Language*, vol. 62, no. 1, pp. 67–82, 2010.

[3] C. Ramisch, *Multiword Expressions Acquisition*, ser. Theory and Applications of Natural Language Processing. Springer, 2015.

[4] *Proceedings of the 11th Workshop on Multiword Expressions*. Denver, Colorado: Association for Computational Linguistics, June 2015. [Online]. Available: http://www.aclweb.org/anthology/W15-09

[5] V. Kordoni, M. Egg, s. t. o. Agata Savary, s. t. o. Eric Wehrli, and S. Evert, Eds., *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*. Gothenburg, Sweden: Association for Computational Linguistics, April 2014. [Online]. Available: http://www.aclweb.org/anthology/W14-08

[6] A. Savary, M. Sailer, Y. Parmentier, M. Rosner, V. Rosén, A. Przepiórkowski, C. Krstev, V. Vincze, B. Wójtowicz, G. S. Losnegaard *et al.*, "Parseme–parsing and multiword expressions within a european multilingual network," in *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, 2015.

[7] V. Rosén, G. S. Losnegaard, K. De Smedt, E. Bejcek, A. Savary, A. Przepiórkowski, P. Osenova, and V. B. Mititelu, "A survey of multiword expressions in treebanks," in *International Workshop on Treebanks and Linguistic Theories (TLT14)*, 2015, p. 179.

[8] E. Bejček and P. Straňák, "Annotation of multiword expressions in the Prague dependency treebank," *Language Resources and Evaluation*, vol. 44, no. 1-2, pp. 7–21, 2010.

[9] A. Abeillé, L. Clément, and F. Toussenel, "Building a treebank for french," in *Treebanks*. Springer, 2003, pp. 165–187.

[10] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[11] G. Eryiğit, T. İlbay, and O. A. Can, "Multiword expressions in statistical dependency parsing," in *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages (IWPT)*, Dublin, Ireland, October 2011, pp. 45–55. [Online]. Available: http://www.aclweb.org/W11-3806

[12] G. Eryiğit, K. ADALI, D. Torunoğlu-Selamet, U. Sulubacak, and T. Pamay, *Proceedings of the 11th Workshop on Multiword Expressions*. Association for Computational Linguistics, 2015, ch. Annotation and Extraction of Multiword Expressions in Turkish Treebanks, pp. 70–76. [Online]. Available: http://aclweb.org/anthology/W15-0912

[13] U. Sulubacak and G. Eryiğit, "Imst: A revisited turkish dependency treebank," in *TurCLing 2016, The First International Conference on Turkic Computational Linguistics at CICLING 2016*, Konya, Turkey, April 2016.

[14] T. Pamay, U. Sulubacak, D. Torunoglu-Selamet, and G. Eryigit, "The annotation process of the itu web treebank," in *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, 2015, p. 95.

[15] S. K. Metin and B. Karaoğlan, "Collocation extraction in Turkish texts using statistical methods," in *Advances in Natural Language Processing*. Springer, 2010, pp. 238–249.

[16] K. Oflazer, O. Çetinoğlu, and B. Say, "Integrating morphology with multi-word expression processing in Turkish," in *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*. Association for Computational Linguistics, 2004, pp. 64–71.

[17] G. Eryiğit, J. Nivre, and K. Oflazer, "Dependency parsing of Turkish," *Computational Linguistics*, vol. 34, no. 3, pp. 357–389, 2008.

[18] A. Savary, "Computational inflection of multi-word units," *A contrastive study of lexical approaches*, vol. 1, no. 2, 2008.

[19] C. K. Asli Göksel, *Turkish: A Comprehensive Grammar (Comprehensive Grammars)*, bilingual ed., ser. Comprehensive Grammars. Routledge, 2005.

[20] J. Baptista, A. Correia, and G. Fernandes, "Frozen sentences of portuguese: Formal descriptions for nlp," in *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, ser. MWE '04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004, pp. 72–79. [Online]. Available: http://dl.acm.org/citation.cfm?id=1613186.1613196

[21] R. Grishman, "The nyu system for muc-6 or where's the syntax?" in *Proceedings of the 6th conference on Message understanding*. Association for Computational Linguistics, 1995, pp. 167–175.

[22] G. A. Şeker and G. Eryiğit, "Initial explorations on using CRFs for Turkish named entity recognition," in *Proceedings of COLING 2012*, Mumbai, India, 8-15 December 2012.

[23] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.

[24] U. Sulubacak and G. Eryiğit, "A redefined Turkish dependency grammar and its implementations: A new Turkish web treebank & the revised Turkish treebank," 2014, under review.