Annotation and Extraction of Multiword Expressions in Turkish Treebanks

Gülşen Eryiğit, Kübra Adalı, Dilara Torunoğlu-Selamet, Umut Sulubacak, Tuğba Pamay
Department of Computer Engineering,
Istanbul Technical University,
Istanbul, 34469, Turkey
{gulsen.cebiroglu|kubraadali|torunoglud|sulubacak|pamay}
@itu.edu.tr

Abstract

Multiword expressions (MWEs) present particular and distinctive semantic properties, hence their automatic extraction receives special attention from the natural language processing (NLP) and corpus linguistics community, and is still an active research area. Unfortunately, the creation of necessary resources for this task is quite rigorous and many languages suffer from the lack of these; as in the case for Turkish.

This study presents our MWE annotations on recently introduced Turkish Treebanks, which focuses on annotating various types of linguistic units and expressions, including named entities, numerical expressions, idiomatic phrases, verb phrases with auxiliaries and duplications. The paper aims to provide a benchmark and pave the way towards further MWE extraction research for Turkish. To this end, the paper also introduces our experimental results with seven baseline approaches, a dependency parser and a previously introduced rule-based extractor on these annotated corpora. Our highest performances achieved over these resources are about 60% F-scores.

1 Introduction

Automatic extraction of multiword expressions (MWEs) is an important and challenging task in natural language processing (NLP). They are introduced to be a key problem for the development of large-scale NLP technology (Sag et al., 2002). Multiword expressions are lexical items that can be decomposed into single words where these single words represent most of the time a totally different meaning compared to word sets within which they occur. Thus, MWEs pose significant problem for NLP and machine translation (MT) applications. The effect and the importance of MWE extraction techniques are being investigated by the NLP and CL communities. A recent ICT-Cost Action (IC1207-PARSEME "PARSing and Multi-word Expressions") focuses only on MWEs in a multidisciplinary level from different perspectives.

In the literature some studies are focused on deriving automatic MWE extraction techniques without using annotated data. Attia (2006) investigates the automatic acquisition of Arabic MWEs and proposes three complementary approaches to extract related MWEs automatically. Piao et al. (2006) propose similar approaches automatically identifying Chinese MWEs and achieve precision ranging from 61.16% to 93.96% for different types. Schone and Jurafsky (2001) seek a knowledge-free method for inducing MWEs from text corpora and provide two major evaluations of nine existing collocation-finders. Metin and Karaoğlan (2010) tries to explore Turkish collocations by using standard statistical methods (e.g Chi-square hypothesis test and mutual information). Tsvetkov and Wintner (2012) extract MWEs by using monolingual and parallel corpora (Hebrew-English), and then use the outcome to train a machine translation system. As mentioned in most of the aforementioned studies, although it might be feasible to automatically identify MWEs using these approaches, yet they need to be improved further. The need for and the importance of manually annotated large-scale data for MWE extraction purpose is not negligible. There exist many recent works on creating language resources for MWEs e.g. MWE databases, corpora and treebanks. The French corpora (Laporte et al., 2008a; Laporte et al., 2008b)

Proceedings of NAACL-HLT 2015, pages 70–76,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

and the Prague Dependency Treebank (Bejček and Straňák, 2010) may be given as examples of these studies among many others.

Dependency parsers are capable of providing quite acceptable performances for MWE extraction. Nivre and Nilsson (2004), Eryiğit et al. (2011), Vincze et al. (2013) and Candito and Constant (2014) investigate the impact of dependency parsers on Swedish, Turkish and Hungarian MWE extraction. Vincze et al. (2013) show that their results outperformed those achieved by state-of-the-art techniques for Hungarian LVC detection. Eryiğit et al. (2011) show that in the training stage, the unification of MWEs of a certain type, namely compound verb and noun formations, has a negative effect on parsing accuracy by increasing the lexical sparsity. In spite of their syntactic relations, MWEs still need special treatments in terms of semantic relations.

Inspired by these recent studies, to shed light and provide a direction for future studies on adequate MWE extraction techniques for Turkish, in this paper we present our annotation for MWEs on recently introduced Turkish Treebanks. We focus on annotating various types of linguistic units and expressions, including named entities, numerical expressions, idiomatic phrases, verb phrases with auxiliaries and duplications. The paper experiments with different lexical approaches together with automatic named entity recognition (NER). The results are compared with those of an available collocation extraction tool (Oflazer et al., 2004) and a dependency parser (Eryiğit et al., 2008). Although, the newly introduced methods improved the previous results by almost 20 percentage points (yielding ∼60% F-score), we treat these results as the state-of-the-art baselines for Turkish.

The paper is structured as follows: Section 2 introduces the used language resources, Section 3 discusses MWEs in Turkish, Section 4 presents models for MWE extraction, Section 5 gives the experimental results and discussions, Section 6 presents the conclusion.

2 Language Resources

We use four different treebanks in our experiments, three of which have been annotated within this study. The first treebank, METU-Sabancı Treebank, (MST) (Oflazer et al., 2003) is from Eryiğit et al. (2011) where the authors state that most of the MWEs in the original treebank are not annotated. They use a semi-automatic way for annotating these MWEs. To this end, they first extracted a MWE list consisting the 30150 MWEs available in the Turkish Dictionary (TDK, 2011) and then automatically listed the entire treebank sentences where the lemmas of the co-occurring words could match the lemmas of the MWE constituents in the list. They then manually marked the sentences where the co-occurring words may be actually accepted as a MWE (but somehow missed during the construction of the original treebank). This semi-automatic annotation approach is incapable of detecting non-adjacent MWE constituents. IMST, IVS and IWT are recently introduced Turkish treebanks annotated with a new dependency scheme (Sulubacak and Eryiğit, 2014).

IMST contains exactly the same sentences thus the same MWEs as MST. But differing from the previous work, the annotation of MWEs are done fully manually without using a semi-automatic selection as explained above. The MWEs are annotated by the use of a specific dependency label (MWE) regardless of their category. In this study, we present our MWE annotations on these three treebanks: IVS with 300 sentences, IMST with 5,635 sentences collected from formally-written data and IWT with 5,009 sentences collected from Web 2.0.

Table 1 presents the resulting MWE statistics on each of these datasets. Since a MWE may consist of two or more words, the table provides both the exact number of MWEs (in the second line) and the total number of MWE relations between MWE constituents (in the first line). As may be noticed from this table, IMST contains almost 50% more MWE annotation than MST of Eryiğit et al. (2011) due to the full manual annotation. Finally the last line of the table gives the number of MWEs with different lengths.

3 MWEs in Turkish

Due to its morphological typology, MWE annotation and extraction methodologies developed for most prominent languages are not suitable for Turkish. Whereas the most well-researched European lan-

	MST			IMST			IVS			IWT		
# of MWE relations	2432			3544			295			2780		
exact # of MWEs	2038			3069			269			2597		
exact # of MWEs	L=2	L=3	L>3	L=2	L=3	L>3	L=2	L=3	L>3	L=2	L=3	L>3
with Word Lengths	1792	159	87	2757	205	107	247	18	4	2444	127	26

Table 1: MWEs in Turkish Treebanks

guages are typically fusional or analytic, Turkish is an agglutinative language, meaning that it is possible to derive and inflect words indefinitely through cascading suffixes. In fact, the derivation is so common that most sentences contain several derived words incorporating one or more suffixes, even in the colloquial language. The constituents of MWEs also commonly undergo inflection (Oflazer et al., 2004; Savary, 2008), giving way to numerous forms of the same expression each appropriate for a different syntactic function. Furthermore, many idiomatic MWEs may also be interpreted literally—that is, there are permissible expressions used in their literal meaning that are morphosyntactically identical to a MWE. Another point is that the constituents of a MWE may occur at nonadjacent positions in the sentence. Figure 1 gives an example for the MWE “ekmeğini yemek” (*to gain one’s livelihood from (someone)*). In the given sentence, the words composing the MWE are both inflected (the first word “ekmek” (*bread*) with 1st person possessive agreement suffix in accusative form and the second word “yemek” (*to eat*) in past tense with 2nd singular person agreement) and written separately from each other.

For these reasons, ordered surface word form matches do not suffice in properly assessing the semantic quality of expressions. Therefore, the disambiguation of MWEs is a more complicated problem than could be resolved by use of look-up tables.

In the rest of this section, we describe the extent of MWEs we specified in our framework. We specify six major categories for MWEs, considering common idiosyncratic formations in Turkish in addition to well-recognized global conventions. We consider any word falling under these categories to be a MWE, as we later build our extraction models around them. The categories are given below:

Named Entities: Proper names and titles of unique persons such as “Genel Sekreter Ban Ki-moon” (*Secretary-General Ban Ki-moon*), organizations such as “Avrupa İnsan Hakları Mahkemesi” (*European Court of Human Rights*) and locations such as “Papua Yeni Gine” (*Papua New Guinea*) occur very frequently in both edited and unedited texts. Commonly recognized as named entities, these expressions often span multiple words, thereby forming a category of MWEs.

Numerical Expressions: We mark any group of contiguous tokens denoting a numerical expression as MWEs, including spelled out numbers, quantities such as currency values and percentages, and temporal expressions such as date and time phrases. Such expressions are often considered to be a subgroup of named entities, but since they are among the most frequently encountered MWEs, we handle them under a separate category to emphasize their importance.

Idiomatic Phrases: Many common idiomatic phrases in Turkish are also occasionally used in their literal meanings, such as “yola düşmek” (*hit the road*, *or lit. fall on the road*). Since both meanings of the phrase would appear morphosyntactically similar, such cases lead to ambiguities in meaning that must be resolved using contextual information. For this reason, we consider idiomatic phrases to be a most challenging category of MWEs.

Light verb constructions: Turkish has a way of forming verb phrases using auxiliary verbs such as “olmak” (*to be*), “etmek” (*to do*), “yapmak” (*to make*) and “kılmak” (*to render*). Among the examples, especially the first two are extremely productive and often used in very common expressions like “teşekkür etmek” (*to thank*, *or lit. to do thank*). Although the figurative meanings of such phrases are usually predictable, they still comprise idiomatic phrases. We handle these outside the previous category due to their prevalence, much like numerical

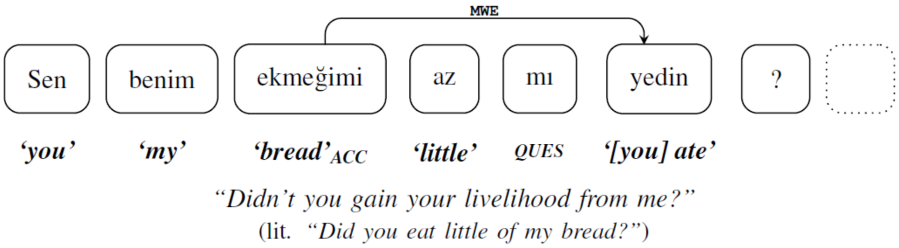
MWE
Sen
benim
ekmeğimi
az
mı
yedin
?
'you'
'my'
'bread'ACC
'little'
QUES
'[you] ate'
"Didn't you gain your livelihood from me?"
(lit. "Did you eat little of my bread?")

Figure 1: A sample Turkish MWE

expressions.

Compound Function Words: We include any compound particles, multi-word interjections and other function word compounds under MWEs. This category excludes function words modified by intensifiers such as "de" and "ise", which also regularly modify content words, as in "ya da" (*or*). Ultimately, there are few permissible function word compounds in Turkish, but they are often commonly used phrases, and warrant a category of MWEs.

Duplications: It is common to use word duplication as a grammatical mechanism in both formal and informal Turkish. Duplicating an adjective allows the word to be used as an adverb much like affixation, such as in "yavaş yavaş" (*slowly*, *or lit. slow slow*). Onomatopoeic or gibberish (and usually rhyming) pairs of words such as "allak bullak" (*topsy-turvy*) are also used fairly often to the same effect. Furthermore, there is the 'm'-duplication, which is a common mechanism in colloquial Turkish, where a word is repeated and an 'm' is prefixed to the duplicate (replacing the initial consonant) in order to add the '*and so on*' meaning, like in "form morm" (*forms and so*). We evaluate all such duplications as MWEs.

4 Models for MWE Extraction

For our MWE extraction experiments, we test with a Turkish dependency parser from Eryiğit et al. (2008), an existing collocation extraction tool (Oflazer et al., 2004) (which we call MorphoColl from this point on), and seven lexical models. The lexical models are based on the previous work by Eryiğit et al. (2011), three of which are identical to the models described in the study and the rest integrate different lexical approaches and a NER module into these models. The rest of this section gives the details about our extraction models and their methodologies.

4.1 Dependency Parser

This model comprises a generic dependency parser which includes **MWE** as one of the dependency relations. We extract MWEs by traversing these relations represented in the output dependency graphs.

4.2 MorphoColl

This model attempts to automatically extract collocations making use of lexical information and morphosyntactic rules. It is composed of three sequential layers, where each layer has its own set of rules and produces the input to the next layer as its output.

4.3 Lexical Models

We first filtered MWEs from a Turkish dictionary (TDK, 2011) into a list and used this list as a look-up table. We used the list in three elementary models with different validation criteria, as introduced previously in Eryiğit et al. (2011).

Model #0: The first MWE extraction model selects the sequences of words whose surface forms match those of the constituents of a MWE in the referenced list. Thus, this model extracts lexicalized collocations which are considered fixed MWEs (Oflazer et al., 2004). An example for this case is given below:

- "**Arka arkaya** iki operasyon geçirdi."
  lit. *(*Back to back) (two) (operations) (he/she had).
  (*He/she had two operations* ***consecutively***.)

Model #1: The second model selects the sequences of words whose surface forms except the last word (which may go under inflection) are the same as the constituents of a MWE in the referenced list. For the last constituent, the stem of the word is required to match. This model extracts collocations belonging to the semi-lexicalized category as stated in (Oflazer et al., 2004). Below is an example for this case:

- "Geleceğini **haber vermedi**."
  lit. (that he/she was coming) (he/she didn't give) (news).
  (*He/she didn't* ***inform***)

Model #2: The third model checks only the stems of the words and select the sequences of words matching the stems of a MWE in the referenced list. Non-lexicalized collocations (Oflazer et al., 2004) each of whose constituents can undergo inflection are extracted by this model. The following example demonstrates this case:

- "Asla **umudunu kesmeyeceksin**."
  lit. (Never) (your hope) (you will cut)
  (*You will never* ***despair***)

As a summary, Model 0 doesn't allow any inflections or derivations in the MWE candidate whereas Model 1 allows for only the last word, and Model 2 allows for all of its words. Since the used dictionary does not include proper names, the models introduced above are incapable of detecting named entities. Thus, our following two models which we name "**Model #1 + NER**" and "**Model #2 + NER**" use a Turkish named entity recognizer (Şeker and Eryiğit, 2012) on top of the mentioned models. Since the NER module may also return single word entities, only the extracted entities with multiple words are accepted as MWEs in these models. Below are some examples of the MWEs which are extracted by the NER in both models:

- "**Milli Savunma Bakanlığı'nın** toplantısı bugün yapılacak."
  lit. (National) (Defense) (of the Ministry) (the meeting) (today) (is to be held)
  (*The* ***Ministry of National Defense*** *meeting is to be held today.*)

- "**Bayındır Sokak'taki** evimden çıktım."
  lit. (Bayındır) (located in Street) (from my house) (I left)
  (*I left my house located in* ***Bayındır Street***.)

The used NER tool which is trained on a data set following the MUC guidelines (Chinchor and Robinson, 1997) for named entity annotation does not extract the titles of the proper names as part of the entity such as in "Başkan Barack Obama" (*President Barack Obama*) where the word 'president' is not extracted as part of the MWE. On the other hand, in our annotations on Turkish Treebanks, these words are also annotated as part of the MWEs. The **Model #1 + Enlarged_NER** implicates the previous and/or the next word of the proper name to the extracted MWE if their first characters are in uppercase letter with the aim to detect the missing title words. The following example shows a MWE consisting of titles and proper names as would be extracted by this model:

- "**Kaymakam Arif Beyi** davet ettik."
  lit. (Mister) (Arif) (Governor) (invite) (we have made)
  (*We have invited* ***Mister Governor Arif***.)

It is impractical to expect from a dictionary list to contain duplications (especially for m-duplications) because there is a theoretically infinite number of duplications (Section 3). Our last model **Model #1 + Enlarged_NER + Dup** contains an additional module which detects these repetitions on top of the previous model. Below is an example showing a MWE formed by word repetition handled by this model:

- "Onu **yavaş yavaş** sakinleştirdi."
  lit. (him/her) (slow slow) (he/she calmed down).
  (*He/she* ***slowly*** *calmed him/her down*)

5 Experimental Results and Discussions

Table 2 gives the precision, recall and F-scores (based on the number of MWEs) for the evaluation of the presented models on the introduced datasets. As stated previously, IMST, which contains higher number of annotated MWEs (Section2) yields lower recall scores compared to MST for all of the models. This is because of the newly annotated MWEs with non-adjacent constituents (Section3). On the other hand, all of the models give higher precision scores on IMST where the missing MWE annotations of MST are eliminated due to careful manually annotations on IMST.

Although, Model #1 is a very straightforward lexical matching approach, it outperforms Morpho-Coll and the dependency parser on newly annotated

	MST			IMST			IVS			IWT		
	P	R	F	P	R	F	P	R	F	P	R	F
Dependency Parser	38.77	44.7	41.52	42.04	32.16	36.44	37.41	19.33	25.49	43.05	39.74	41.33
MorphoColl	80.77	22.67	35.40	77.5	15.71	26.12	82.93	12.64	21.94	86.24	15.44	26.19
Model #0	20.89	9.47	13.32	39.92	12.38	18.1	52.06	14.13	22.22	43.72	13.94	21.14
Model #1	32.65	42.89	37.07	46.73	40.27	43.26	63.13	42.01	50.45	51.5	39.7	44.84
Model #2	27.62	45.93	34.49	39.23	43.99	41.48	45.59	44.24	44.91	43.95	44.17	44.06
Model #1+NER	42.85	56.28	48.65	57.69	49.72	53.41	70.95	47.21	56.7	59.49	45.86	51.79
Model #2+NER	35.67	59.32	44.56	47.66	53.44	50.38	50.96	49.44	50.19	50.08	50.33	50.2
Model #1+Enlarged_NER	51.92	68.2	58.96	66.59	57.38	61.64	71.51	47.58	57.14	67.38	51.95	58.67
Model #1+Enlarged_NER+Dup	52.74	70.46	60.32	67.25	59.14	62.93	72.43	47.58	57.44	68.13	53.75	60.1

Table 2: Baseline System Results

datasets. The reason is because, the literal interpretation of MWEs with adjacent constituents is less probable compared to idiomatic usage. Such as the MWE "ayvayı yemek" which is close in meaning to *to be in hot water* (*slang to be in trouble*) may also be used literally in the case of *eating a quince* which is a much less probable usage.

The impact of adding a NER layer improves the results almost 10 percentage points. Our Enlarged_NER adds almost 10 percentage points on top of this, and the impact (∼2 percentage points) of duplication detection is also promising although not as high as the previous two. Our best performed model **Model #1 + Enlarged_NER + Dup** achieves 60.32%, 62.93%, 57.44% and 60.1% F-scores in MST, IMST, IVS and IWT respectively.

The extractors that we presented in this paper are limited to an individual dependency parser, a rule-based model and dictionary-based models with rule-based additions. Since these models do not go beyond considering the lexical forms and syntactic structures of constituents, they have an equally limited performance in determining MWEs, which are essentially semantic entities. As such, our models should only be considered baseline models. We expect the models to be a benchmark for future work on more sophisticated MWE extraction systems for Turkish and facilitate comparison with studies on other languages analogous to Turkish in their morphosyntactic structure, such as other agglutinative languages like Finnish and Hungarian, as well as various morphologically rich languages like French and Arabic.

Our premise is that, in order to properly pick out MWEs from within texts, a model needs to integrate morpho-lexical, syntactic and semantic modules all in one, in order to respectively extract critical constituents, appoint the grammatical relations between them, and determine the nature of the extracted phrases. One of our future plans is to design and implement such a model following this study, making use of machine learning and incorporating sequential modules, each working out a separate aspect of the candidate expressions. Additionally, we aim to expand our survey and test our new model on other languages besides Turkish for a more thorough performance evaluation.

6 Conclusion

In this study, we described the various challenges in annotating and extracting MWEs in Turkish, due to the typology and certain idiosyncratic features of the language. We outlined the framework we established on what constitutes a MWE, along with the exceptional cases that have been considered. Afterwards, we discussed our elementary approach to extracting MWEs in Turkish, then presented the basic extraction models we developed and tested on four Turkish treebanks. Our best model which uses a lexical look-up approach allowing the inflection of the final MWE constituent, an enhanced named entity recognition module and a duplication extraction module obtains about 60% F-measure in these treebanks.

Acknowledgment

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

Mohammed A. Attia. 2006. Accommodating multiword expressions in an arabic lfg grammar. In *Proceedings of the 5th International Conference on Advances in Natural Language Processing*, FinTAL'06, pages 87–98, Berlin, Heidelberg. Springer-Verlag.

Eduard Bejček and Pavel Straňák. 2010. Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation*, 44(1-2):7–21.

Marie Candito and Matthieu Constant. 2014. Strategies for contiguous multiword expression analysis and dependency parsing. In *ACL 14-The 52nd Annual Meeting of the Association for Computational Linguistics*. ACL.

Nancy Chinchor and Patricia Robinson. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, page 29.

Gülşen Eryiğit, Tugay İlbay, and Ozan Arkan Can. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages (IWPT)*, pages 45–55, Dublin, Ireland, October.

Gülşen Eryiğit, Joakim Nivre, and Kemal Oflazer. 2008. Dependency parsing of Turkish. *Computational Linguistics*, 34(3):357–389.

Eric Laporte, Takuya Nakamura, and Stavroula Voyatzi. 2008a. A French corpus annotated for multiword expressions with adverbial function. In *Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop*, pages 48–51.

Eric Laporte, Takuya Nakamura, and Stavroula Voyatzi. 2008b. A French corpus annotated for multiword nouns. In *Language Resources and Evaluation Conference. Workshop Towards a Shared Task on Multiword Expressions*, pages 27–30.

Senem Kumova Metin and Bahar Karaoğlan. 2010. Collocation extraction in Turkish texts using statistical methods. In *Advances in Natural Language Processing*, pages 238–249. Springer.

Joakim Nivre and Jens Nilsson. 2004. Multiword units in syntactic parsing. *Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*.

Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. 2003. Building a Turkish treebank. In *Treebanks*, pages 261–277. Springer.

Kemal Oflazer, Özlem Çetinoğlu, and Bilge Say. 2004. Integrating morphology with multi-word expression processing in Turkish. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 64–71. Association for Computational Linguistics.

S Piao, Guangfan Sun, Paul Rayson, and Qi Yuan. 2006. Automatic extraction of Chinese multiword expressions with a statistical tool. In *Workshop on Multi-word-expressions in a Multilingual Context (EACL 2006)*.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.

Agata Savary. 2008. Computational inflection of multi-word units. *A contrastive study of lexical approaches*, 1(2).

Patrick Schone and Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 100–108.

Gökhan Akın Şeker and Gülşen Eryiğit. 2012. Initial explorations on using CRFs for Turkish named entity recognition. In *Proceedings of COLING 2012*, Mumbai, India, 8-15 December.

Umut Sulubacak and Gülşen Eryiğit. 2014. A redefined Turkish dependency grammar and its implementations: A new Turkish web treebank & the revised Turkish treebank. under review.

TDK. 2011. *Turkish Language Association Turkish dictionary*. http://www.tdk.gov.tr.

Yulia Tsvetkov and Shuly Wintner. 2012. Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(04):549–573.

Veronika Vincze, János Zsibrita, and TI Nagy. 2013. Dependency parsing for identifying Hungarian light verb constructions. In *Proc. of the Sixth International Joint Conference on Natural Language Processing*, pages 207–215.