Enhancing the PARSEME Turkish Corpus of Verbal Multiword Expressions

Yağmur Öztürk[1], Najet Hadj Mohamed[2], Adam Lion-Bouton[2], Agata Savary[1]

Paris-Saclay University - LISN[1], University of Tours - LIFAT[2]

{yagmur.ozturk, agata.savary}@universite-paris-saclay.fr

{najat.hadjmohamed, adam.lion-bouton}@etu.univ-tours.fr

Abstract

The PARSEME (Parsing and Multiword Expressions) project proposes multilingual corpora annotated for multiword expressions (MWEs). In this case study, we focus on the Turkish corpus of PARSEME. Turkish is an agglutinative language and shows high inflection and derivation in word forms. This can cause some issues in terms of automatic morphosyntactic annotation. We provide an overview of the problems observed in the morphosyntactic annotation of the Turkish PARSEME corpus. These issues are mostly observed on the lemmas, which is important for the approximation of a type of an MWE. We propose modifications of the original corpus with some enhancements on the lemmas and parts of speech. The enhancements are then evaluated with an identification system from the PARSEME Shared Task 1.2 to detect MWEs, namely Seen2Seen. Results show increase in the F-measure for MWE identification, emphasizing the necessity of robust morphosyntactic annotation for MWE processing, especially for languages that show high surface variability.

Keywords: multiword expressions, morphosyntax, agglutinative languages, lemmatization, natural language processing

1. Introduction

Natural language processing tasks come with the challenge of working across languages. Meeting this challenge, both Universal Dependencies (UD)[1] and PARSEME[2] are multilingual projects with the aim of unifying linguistic descriptions across languages. Languages that are typologically distant from some of the high-resourced Germanic and Romance languages can be a challenge to adapt into this unified typology made for linguistic annotation. Since PARSEME annotations are done on previously annotated UD treebanks, problems occurring in the treebanks are persistent in the PARSEME corpora.

In the scope of this case study, we have examined the Turkish corpus (Erden et al., 2018) of PARSEME due to its rich morphology realized with inflectional and derivational suffixes. This corpus was automatically annotated for morphosyntax with UDPipe (Straka, 2018) and manually annotated (Berk et al., 2018) for verbal multiword expressions (VMWEs) with PARSEME's annotation guidelines[3]. In the PARSEME Shared Task 1.1 (Ramisch et al., 2018), a MWE is defined as a group of lexicalized words displaying lexical, morphological, syntactic and/or semantic idiosyncrasy. Traditionally, lexicalisation refers to the process in which a word acquires the status of an autonomous lexical unit. Expanding the scope of this definition, in MWEs, PARSEME considers lexicalisation applying not only to the whole unit of an MWE, but also its individual components. The reason for this is the need of precising the span of an MWE. Thus, we only annotate the lexically fixed components of an MWE, and these components are referred to as lexicalized within a given MWE (Markantonatou et al., 2018).

MWEs are represented as multisets of lemmas of their components, e.g. the English MWE *let bygones be bygones*

is represented as {be, bygone, bygone, let} and the Turkish ***geri adım attılar*** (lit. 'they took a step back') 'they retreated' as {adım, at, geri}. Many VMWEs exhibit a certain degree of morphosyntactic flexibility, which is displayed by their various forms. For instance, ***geri adım attılar*** 'they took a step back', and ***geri adım atabilirlerdi*** 'they could have taken a step back' are occurrences of the VMWE, represented as {geri, adım, at}{back, step, throw}. A type is the set of all occurrences of the same MWE, and is formally represented as a multiset of lemmas of its lexicalized components. In practice, approximating types as multisets of lemmas can be helpful for MWE identification by neutralizing morphosyntactic variability, e.g. by conflating different forms and occurrences of the same MWE. For MWE types to be correctly identified, correct lemmatization of occurrences is important. In languages with higher rate of inflection and derivation, this issue can be more visible.

For this case study, examination of the Turkish corpus was made using a corpus visualization tool provided by PARSEME. This tool enables the user to see all occurrences of annotated VMWE types and their categories defined by the PARSEME annotation guide. The most notable issue observed in the Turkish corpus is the frequent incorrect/incomplete lemmatization of highly inflected verbs. Issues are further discussed in section 4. After the manual enhancement of most of these lemmas, one of the best performing systems of the PARSEME shared task 1.2, called Seen2Seen, was trained and tested on the enhanced data to show the impact of enhanced lemmatization.

2. Related Works

One of the first works on annotating Turkish MWEs was done by (Adalı et al., 2016) to define a comprehensive annotation guide. The authors mention specific constructions such as duplication and named entities. This guide is referenced in the first edition of PARSEME. Annotation from this edition was adapted to the updated version of the unified guidelines on the same corpus in the following editions of PARSEME.

Pertaining to the morphosyntactic annotation of treebanks

---

in Turkish, it is important to mention manual annotation work in progress, which aims to increase the number of derivational representations. With this manual annotation, (Türk et al., 2019) aim to increase the accuracy of annotation not only for Turkish but also for other agglutinative languages. For future works, this corpus can be annotated for MWEs, which will help build a better corpus for PARSEME. In the PARSEME project, most of the language data is annotated for morphosyntax using automatic tools trained on treebanks, such as UDPipe[4]. The Turkish corpus is also automatically annotated for morphosyntax with UDPipe, and manually annotated for VMWEs. Turkish data have existed in PARSEME since edition 1.0 and went through changes in terms of annotation guidelines. Edition 1.1 uses the ITU NLP Tool[5] for morphosyntactic annotation. In the latest edition, re-parsing was executed relying on a model in UDPipe version 2.4 based on IMST[6] (Sulubacak et al., 2016). This treebank was first manually annotated in non-UD style and then converted to UD. Manual annotations of VMWEs from edition 1.1 were updated to match the new annotation guidelines of PARSEME edition 1.2. The raw[7] Turkish corpus of PARSEME consists of newspaper articles, which is the same genre of text used in the IMST.

In addition, UD currently has 9 Turkish treebanks. These treebanks have followed slightly different annotation processes from one another, therefore teams focusing on Turkish are working on the unification (Türk et al., 2019) of annotation guidelines for UD. Moreover, shortcomings of UD in expressing the derivational nature of languages as a more general problem have been studied by (Bedir et al., 2021).

In UD 2.0, the lemmatizer works with a guesser that produces (lemma rule, UPOS) pairs, where the lemma rule generates a lemma from a word by stripping some prefixes and suffixes and prepending and appending new prefixes and suffixes. The lemmatization rules look at the last four characters of a word, but also at the word prefix, and the disambiguation is performed by an averaged perceptron tagger (Straka and Straková, 2017). However, we cannot exactly know where this system fails to perform optimally in Turkish lemmatization.

Lemmatization is especially important for languages like Turkish, which have rich inflectional morphology, with possibly many inflectional suffixes agglutinated to a single verb or noun. It is also possible to come across verb constructions of inflected forms that were not observed in the training and development corpora. This property was touched upon by (Oflazer et al., 2004), where Turkish word forms that consist of morphemes concatenated to a root morpheme or to other morphemes were compared to beads on a string. Works have been made to contribute to the rep-

resentation of Turkish and other agglutinative languages in UD-based treebanks, which in turn helps to develop more accurately annotated datasets for such languages.

3. Verbal Multiword Expressions in Turkish

VMWEs are the main type of MWEs in question for the PARSEME project. VMWEs are seen as a bigger challenge than non-verbal MWEs since they exhibit higher surface variability (Pasquer et al., 2020). Taking on this challenge, the PARSEME framework defines a VMWE as an expressions: (i) with at least two lexicalized components, including a head word and at least one other syntactically related word, (ii) whose head (in a canonical form) is a verb, and (iii) which functions as a verbal phrase (Ramisch et al., 2018). In the final annotation guidelines, PARSEME also defines VMWE categories for better identification of their occurrences. Three of the main five categories exist in Turkish, more frequently Light Verb Constructions (LVC) and Verbal Idioms (VID), more rarely Multi Verb Constructions (MVC)[8] as in examples (1)–(3), respectively[9] Differences between these categories lie in the role of the component words. Fundamentally, LVCs occur with light verbs and predicative nouns, VIDs have at least two lexicalized components including a head verb and at least one of its dependents, and MVCs are constructed with more than one verb.

(1) **şüphe et-ti** (category: *LVC*)
şüphe et-PAST
doubt do-PAST
'(someone) doubted'

(2) **kulak as-ma-dı** (cat.: *VID*)
kulak as-NEG-PAST
ear   hang-NEG-PAST
'(someone) did not pay attention'

(3) **gid-ip   gel-ir-ken** (cat.: *MVC*)
git-CONV gel-HAB-CONV
go-CONV come-HAB-CONV
'going in between'

MWE occurrences do not have a balanced distribution among all types and most of the types rarely occur in a given corpus, since it only represents a small part of the natural language.

In a language such as Turkish, we can see a lot of inflection and derivation which causes some issues for this task. In the VID example (2), both of the components can get suffixes in various occurrences of this type, such as
kulak assalardı 'ear hang-CND-PLUR-PAST'. In the case of incorrect lemmatization, this MWE can be found more than once with different lemmas, therefore erroneously increasing the number of VMWE types (as approximated by

PARSEME) in the Turkish corpus. This wrong representation can also endanger future studies that use this corpus. Specific issues are explained in more detail in the following section.

4. Issues in the Morphosyntactic Annotation of the Turkish Corpus

This section provides an overview of the issues observed in the PARSEME Turkish corpus regarding morphosyntactic annotation realized by UDPipe. Three main issues were observed and are discussed below with examples.

Sound change produced in the stem by suffixation Some stems ending with voiceless consonants, go through the process of sound change produced in the stem by suffixation. In this process, the ending ('p', 't', 'k', 'ç') changes to its voiced counterpart ('b', 'd', 'g', 'c') before adjoining a suffix that commences with a vowel (Goksel, 2005).
For instance the lemma *et* 'to do', when receiving the suffix *-ecek* signalling future tense, yields *edecek* 'will do'. This lemma has a very high surface variability, both inside and outside of VMWEs. Since it is more of a challenge for automatic systems to identify lemmas that go through this type of change, we frequently see incorrectly lemmatized word forms in the corpus caused by this issue.
Consider examples (4)–(6) showing various morphological variants of the same LVC. In (4), the suffix *-ti* does not start with a vowel, so the voicing does not take place and the lemmatization is correct (cf. the UD lemma in the second line of the example). In (5), conversely, the voicing of the lemma does take place and the lemmatizer fails to restore the correct lemma. This challenge may be even harder when several suffixes are adjoined, as in (6).

(4) **istifa   et-ti**
    istifa   et-PAST (UD lemma: *et*)
    resignation do-PAST
    '(someone) resigned'

(5) **istifa   ed-ecek**
    istifa   et-FUT (UD lemma: **ed*)
    resignation do-FUT
    '(someone) will resign'

(6) **istifa   ed-ebil-ir-di**
    istifa   et-POT-HAB-PAST (UD lem.: **edebil*)
    resignation do-POT-HAB-PAST
    '(someone) could have resigned'

Suffixation MWE occurrences vary in inflectional forms and in Turkish we observe high surface variability. In general, suffixation can be observed in all components of a Turkish MWE. An example of suffixation in one component can be the occurrences of *dava aç* 'to sue' in the examples (8) and (9).

(7) **dava   aç-tı**
    dava   aç-PAST (UD lemma: *aç*)
    lawsuit open-PAST
    '(someone) commenced lawsuit'

(8) **dava   aç-ıl-abil-ir**
    dava   aç-PASS-POT-HAB (UD lemma: **açılab*)
    lawsuit open-PASS-POT-HAB
    'lawsuit could be commenced'

(9) **dava   aç-ıl-acak**
    dava   aç-PASS-FUT (UD lemma: **açıla*)
    lawsuit open-PASS-FUT
    'lawsuit will be commenced'

In (8)–(9), we observe wrongly stripped series of suffixes in the verb component of the VMWE. We can also note that there can be more than one example of insufficient suffix stripping in a given verb, as illustrated above.
We occasionally came across the opposite issue, namely with too many, rather than too few, suffixes hypothesized by the lemmatizer. For instance, the first component of the VMWE *rehin alındı* 'hostage take-PASS-PAST', was lemmatized as **reh*, instead of the correct *rehin*.The reason for this can be the resemblance between the ending of this word with the possessive suffix *-in* '-yours' in Turkish.

Nominalization Some commonly used derived nouns are components of LVCs and they play the roles of predicative nouns (i.e. describe actions or states). In the IMST, these nominal derivations are mostly assigned the VERB POS and lemmatized into infinitives. Their nominal nature is retrievable from the morphological feature VERB-FORM=VNOUN, as in example (10).

(10) **açıkla-ma   yap-tı**
    açıkla-VNOUN yap-PAST (UD. lem.: *açıkla*)
    to.state-VNOUN make-PAST
    '(someone) made (a) statement'

Here, the first component *açıklama* is the result of a derivation realized with suffix *-ma*, which turns the verb *açıkla* 'to state' into a noun *açıklama* 'statement'. This analysis, notwithstanding its defensibility, is incompatible with the PARSEME definition of an LVC as a verb-noun combination, since (10) is represented as a combination of verbs instead.

5. Enhancement Process

The issues described in the preceding section were observed via the PARSEME annotation consistency checker[10]. All problems of voicing and suffixation, illustrated in section 4., spotted in this way, were manually corrected both within VMWE components and in other occurrences of the same verbs. Cases like (10) were more difficult to decide on since they lie on the fuzzy border between inflection and derivation. Ideally, on the one hand, we would expect a more elaborate morphosyntactic representation of nominalizations in UD, and a more flexible definition of LVCs in PARSEME on the other. In the meantime, we changed the lemmas of only the clearly lexicalized nominalizations, functioning as standalone nouns independently of the verbs they stem from, like *açıklama* 'statement'. The enhancements were made manually on the training, development and test corpora, which are in

[10]https://grew.fr/download/PARSEME/tr.html

Data	Global MWE-based			Global Token-based		
	P	R	F1	P	R	F1
ST TRAIN/DEV/TEST	61.69	65.33	**63.46**	63.1	65.69	**64.37**
Enhanced TRAIN/DEV, ST TEST	61.33	63.94	62.61	62.86	64.52	63.68
Enhanced TRAIN/DEV/TEST	61.43	70.98	**65.86**	62.90	71.60	**66.97**

Table 1: The overall results of Seen2Seen evaluation for the shared task and the enhanced data.

Data	LVC			VID		
	P	R	F1	P	R	F1
ST TRAIN/DEV/TEST	59.87	65.57	**62.59**	61.77	63.41	**62.58**
Enhanced TRAIN/DEV, ST TEST	58.54	65.93	62.02	62.33	60.26	61.28
Enhanced TRAIN/DEV/TEST	59.36	74.91	**66.23**	61.72	65.40	**63.50**

Table 2: The results of global-MWE based Seen2Seen evaluation per MWE category.

the CUPT[11] format. In total, 3116 tokens were affected by the enhancements. As a result we obtained a more accurate count of VMWE types. Namely, previously there were 2826 types of VMWEs, with 2.74 occurrences per type on average, whereas the enhancement reduced the number of types to 2310 and increased the occurrences per type to 3.34. This is because, with incorrect lemmas, occurrences of the same VMWE, as in examples (5)–(6), could be wrongly split into different clusters of types.

6. System Results

To see the impact of the corrections, we used one of the MWE identification systems from the edition 1.2 of the PARSEME shared task. This system, namely Seen2Seen[12], ranked first in the global F-measure in the closed track (where no external resources were allowed) and second across both the closed and the open track.[13] Seen2Seen reads all MWEs annotated as such in the training corpus, and extracts all candidate occurrences of the same multisets of lemmas in the test corpus. These candidates then go through a set of morphosyntactic filters. In total, 8 filters are defined, and the training phase allows us to decide which filter to activate for which language, based on the performances on the development corpus.

Seen2Seen was used to annotate the original and the enhanced data. Tables 1 and 2 show the results of this experiment. The first line of each table corresponds to the system trained and evaluated on the original shared task (ST) data. In the second line, the system is trained on the enhanced TRAIN and DEV files but tested on the original TEST. In the last line, the system is both trained and tested on the enhanced files. Note that, while results of the three scenarios of testing are shown next to each other in tables 1 and 2, scores shown in line 3 cannot be directly compared to those shown in lines 1 and 2 since they are computed on different version of the TEST.[14] Table 1 shows the macro-average results for Turkish on the general metrics,

namely the MWE-based (correctly identifying a VMWE as a whole) and the token-based (correctly identifying the individual components of a VMWE), recall, precision and F-measure. In Table 2, the results of the MWE-based metrics can be compared per category: LVC and VID.[15]

The results show a difference of 2.5 and 2.6 in the global MWE-based and token-based F-measure, respectively. This is mainly due to two factors. Firstly, with enhanced lemmas, the number of VMWE occurrences considered seen[16] in TEST grows from 812 to 911, while Seen2Seen has a stable performance on the seen VMWEs[17] Secondly, while precision slightly drops between line 1 and 3, the recall significantly increases. This is probably because the variants of seen VMWEs were previously omitted by the system if their lemmas spuriously diverged from seen VMWEs. Now, with more accurate lemmas, the system does see them as valid VMWE candidates. When the lines 1 and 2 are compared, we see a minor decrease in the F-measure by 0.85 and 0.69, which was expected since it is not optimal for a system to be tested on a data set which was annotated according to different principles than those in the training data.

Per-category results show that the Recall for LVCs increased by 9.34 which was expected since our enhancements were very frequent in light verbs. This also resulted in an increase of 3.64 in the F-measure. Conversely, we observe, only an increase of 1.99 in the Recall and 1.8 in the F-measure of VIDs, which might point out that components of VIDs might not vary in surface forms as much as LVCs due to their idiomatic nature.

7. Conclusion

We examined a corpus of VMWEs in Turkish, annotated for the PARSEME project. We detected some shortcomings in terms of morphosyntactic annotation. We focused on enhancing the lemmas in the corpus for better MWE processing. One of the best performing systems from the

---

[11]http://multiword.sourceforge.net/cupt-format
[12]https://gitlab.com/cpasquer/st_2020
[13]http://multiword.sourceforge.net/sharedtaskresults2020
[14]For the same reason statistical significance tests would not be truly meaningful either.

[15]MVCs are ignored in this table since only one such MWE occurs in both corpora, and it was not affected by our corrections.
[16]An expression is defined as seen if the multiset of lemmas of its lexicalized components was annotated in the training and development sets (Ramisch et al., 2020).
[17]F=0.7329 and F=0.7303 for the ST and the enhanced data, respectively.

PARSEME shared task was trained and tested on the enhanced data to compare the impact of our corrections. The results showed an increase of F-measure for MWE identification when the system was trained and tested on the new corpus when compared to the ST results. We also observed an increase of F-measure in the LVCs, which emphasized the amount of enhancement made in the LVC components.

Our results and the new data establish a new benchmark for the Turkish MWE identification. They also show the necessity for a high-quality morphosyntactic annotation for better MWE processing, especially in morphologically rich corpora. Our observations can also pave the way to some future studies with the examination of other agglutinative languages for MWE processing to see if enhancements of the same nature can be made.

8. Bibliographical References

Adalı, K., Dinç, T., Gökırmak, M., and Eryiğit, G. (2016). Comprehensive annotation of multiword expressions in turkish. In *TurCLing 2016 The First International Conference on Turkic Computational Linguistics at CICLING*.

Bedir, T., Şahin, K., Gungor, O., Uskudarli, S., Özgür, A., Güngör, T., and Ozturk Basaran, B. (2021). Overcoming the challenges in morphological annotation of Turkish in universal dependencies framework. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 112–122, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Berk, G., Erden, B., and Güngör, T. (2018). Turkish verbal multiword expressions corpus. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.

Goksel, A. (2005). *Turkish: A comprehensive grammar*. Routledge Comprehensive Grammars. Routledge, London, England, May.

Stella Markantonatou, et al., editors. (2018). *Multiword expressions at length and in depth*. Number 2 in Phraseology and Multiword Expressions. Language Science Press, Berlin.

Oflazer, K., Çetinoğlu, Ö., and Say, B. (2004). Integrating morphology with multi-word expression processing in Turkish. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 64–71, Barcelona, Spain, July. Association for Computational Linguistics.

Pasquer, C., Savary, A., Ramisch, C., and Antoine, J.-Y. (2020). Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., and Walsh, A. (2018). Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A., and Xu, H. (2020). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online, December. Association for Computational Linguistics.

Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, October. Association for Computational Linguistics.

Türk, U., Atmaca, F., Özateş, Ş. B., Köksal, A., Ozturk Basaran, B., Gungor, T., and Özgür, A. (2019). Turkish treebanking: Unifying and constructing efforts. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 166–177, Florence, Italy, August. Association for Computational Linguistics.

9. Language Resource References

Erden, B., Berk, G., and Gungor, T. (2018). Turkish verbal multiword expressions corpus. In *26th IEEE Signal Processing and Communications Applications Conference, SIU 2018*, pages 1–4, İzmir, Turkey, May.

Sulubacak, U., Gokirmak, M., Tyers, F., Coltekin, C., Nivre, J., and Eryigit, G. (2016). Universal dependencies for turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 3444–3454, Osaka, Japan, December. The COLING 2016 Organizing Committee.