# Multiword expressions at length and in depth

## Extended papers from the MWE 2017 workshop

Edited by

Stella Markantonatou

Carlos Ramisch

Agata Savary

Veronika Vincze



Phraseology and Multiword Expressions 99

language science press

Phraseology and Multiword Expressions



In this series:

1. Manfred Sailer & Stella Markantonatou (eds.). Multiword expressions: Insights from a multilingual perspective.

ISSN: 2625-3127

# Multiword expressions at length and in depth

## Extended papers from the MWE 2017 workshop

Edited by

Stella Markantonatou

Carlos Ramisch

Agata Savary

Veronika Vincze

language science press

Markantonatou, Stella, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.). 2018. *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop* (Phraseology and Multiword Expressions 99). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/00

ISBN: 978-0-000000-00-0 (Digital)
no print ISBNs!
ISSN: 2625-3127
no DOI
ID not assigned!

Cover and concept of design: Ulrike Harbort
Fonts: Linux Libertine, Libertinus Math, Arimo, DejaVu Sans Mono
Typesetting software: XeLaTeX

Language Science Press
Unter den Linden 6
10099 Berlin, Germany
langsci-press.org

Storage and cataloguing done by FU Berlin

# Contents

# Chapter 1

# Identifying verbal multiword expressions with POS tagging and parsing techniques

Katalin Ilona Simkó
University of Szeged

Viktória Kovács
University of Szeged

Veronika Vincze
University of Szeged
MTA-SZTE Research Group on Artificial Intelligence

The chapter describes an extended version (USzeged+) of our previous system (USzeged) submitted to PARSEME's Shared Task on automatic identification of verbal multiword expressions. USzeged+ exploits POS tagging and dependency parsing to identify single- and multi-token verbal MWEs in text. USzeged competed on nine of the eighteen languages, where USzeged+ aims to identify the VMWEs in all eighteen languages of the shared task and contains fixes for deficiencies of the previously submitted system. Our chapter describes how our system works and gives a detailed error analysis.

## 1 Introduction

Multiword expressions (MWEs) are frequent elements of all natural languages. They are made up of more than one lexeme, but their meaning is not predictable from the meaning of their components. There are different types of MWEs such

Katalin Ilona Simkó, Viktória Kovács & Veronika Vincze. 2018. Identifying verbal multiword expressions with POS tagging and parsing techniques. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 1–14. Berlin: Language Science Press. DOI:??

as stereotyped similes (*as white as snow*), collocations (*strong tea*), or idioms (*to kick the bucket*). This chapter deals with verbal MWEs (VMWEs) where the head element of the MWE is a verb, for example verb-particle constructions (*look after*), or light-verb constructions (*take a shower*).

This chapter describes our system for verbal MWE recognition. It was built for the PARSEME Shared Task 1.0 (Savary et al. 2017), USzeged and its extension, USzeged+. Both systems use POS tagging and dependency parsing and are capable of identifying single- and multi-token verbal MWEs. They are language-independent: USzeged was submitted for nine of the eighteen languages of the Shared Task, while for this extended version, USzeged+, we present results for all eighteen languages.

In this chapter, we first describe the original USzeged system and give our results submitted to the Shared Task with detailed error analysis. This part of the chapter builds heavily on our workshop paper (Simkó et al. 2017). Then, we describe the details of the updated, USzeged+ version and give the results we achieved with this new system. Last, we give a comparison of the results achieved using our approach in the original, USzeged system and the new USzeged+ one in an experiment using the available Hungarian data.

## 2 USzeged - The original system

The USzeged system was built for the shared task on automatic identification of verbal multiword expressions organized as part of the 2017 MWE workshop (Savary et al. 2017).[1] The shared task's aim is to identify verbal MWEs in multiple languages. In total, eighteen languages are covered that were annotated using guidelines taking universal and language-specific phenomena into account.

The guidelines identify five different types of verbal MWEs: idioms (ID), light-verb constructions (LVC), verb-particle constructions (VPC), inherently reflexive verbs (IReflV) and "other" (OTH). Their identification in natural language processing is difficult because they are often discontinuous and non-compositional, the categories are heterogeneous and the structures show high syntactic variability.

The precise definitions of MWE, VMWE and the VMWE types can be found in Taslimipoor et al. (2018 [this volume]), as well as details on the different languages' databases used.

Our team created the Hungarian shared task database and VMWE annotation. Our system is mostly based on our experiences with the Hungarian data in this

[1] http://multiword.sourceforge.net/sharedtask2017

annotation phase. Our goal was to create a simple system capable of handling MWE identification in multiple languages.

## 2.1 System description

The USzeged system exploits the syntactic relations within MWEs, i.e. it directly connects MWEs and parsing, an approach described in many sources (Constant & Nivre 2016; Nasr et al. 2015; Candito & Constant 2014; Green et al. 2011; 2013; Wehrli et al. 2010; Waszczuk et al. 2016) and one of the basic ideas behind the work done by the PARSEME group.[2] The core of our system is directly based on the work described in Vincze et al. (2013): using dependency parsing to identify MWEs. That system uses complex dependency relations specific to the given syntactic relation and MWE type. We note that a high number of the languages of the shared task are morphologically rich and have free word order, which entails that syntactically flexible MWEs might not be adjacent. Hence, a syntax-based approach seems a better fit for the task than sequence labeling or similar strategies.

The USzeged system uses only the MWE type as a merged dependency label, i.e. no clue is encoded to the syntactic relation between two parts of the MWE. Moreover, it also treats single-token MWEs. As multiple languages had single-token MWEs as well as multi-token ones that are dealt with in dependency parsing, we expanded the approach using POS tagging. Frequent single-token MWEs are, for example, German and Hungarian VPCs: when the particle directly precedes the verb, German and Hungarian spelling rules require that they are spelled as one word, however, it still remains a construction made up of two lexemes with non-compositional meaning (e.g. (HU) **kinyír** (ki+nyír) 'out+cut' ⇒ 'kill' or (DE) **aufmachen** (auf+machen) 'up+do' ⇒ 'open').

MWEs have specific morphological, syntactic and semantic properties. Our approach treats multi-token MWEs on the level of syntax – similarly to the `mwe` dependency relation in the Universal Dependency grammar (Nivre 2015) – and single-token MWEs on the level of morphology.

The USzeged system works in four steps, and the main MWE identification happens during POS tagging and dependency parsing of the text. Our system relies on the POS tagging and dependency annotations provided by the organizers of the shared task in the companion CoNLL files and the verbal MWE annotation of the texts and is completely language-independent given those inputs.

[2] http://typo.uni-konstanz.de/parseme/

In the first step, we prepared the training file from the above mentioned inputs. We merged the training MWE annotation into its morphological and dependency annotation for single- and multi-token MWEs, respectively. The POS tag of single-token MWEs got replaced with their MWE type, while for the multi-token MWEs the dependency graphs' label changed: the label of the dependent node in the tree was replaced with a label denoting the MWE type.

Figure 1, Figure 2 and Figure 3 show the single-token MWE's change in POS tag and multi-token MWE dependency relabeling for VPCs and LVCs in a Hungarian example.

| | original label | relabeled | (HU) |
|---|---|---|---|
| **bekezdés** | NOUN | VPC | (HU) |
| in+starting, 'paragraph' | | | |
| **határozathozatal** | NOUN | LVC | (HU) |
| decision+bringing, 'decision-making' | | | |

Figure 1: Adding the VPC and LVC single-token MWE POS tags to (HU) **bekezdés** (be+kezdés) 'in+starting' ⇒ 'paragraph' and (HU) **határozathozatal** (határozat hozatal) 'decision+bringing' ⇒ 'decision-making'.

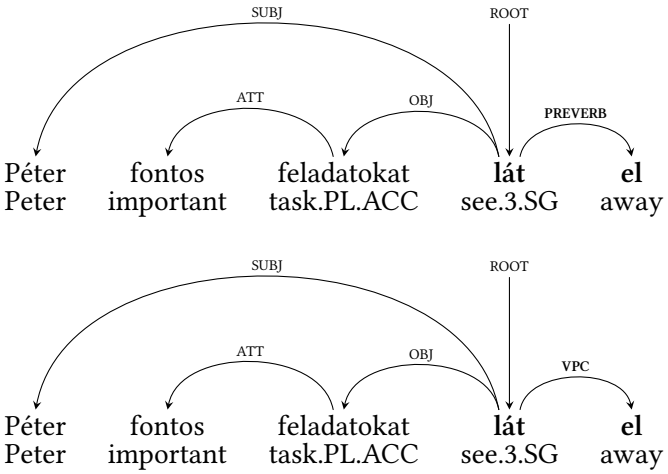


Figure 2: Adding the VPC multi-token MWEs label to the dependency graph in (HU) *Péter fontos feladatokat **lát el**.* 'Peter important tasks sees away' ⇒ 'Peter takes care of important tasks'.

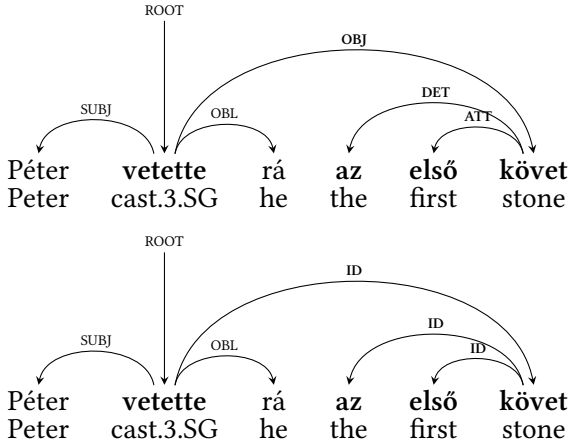


Figure 3: Adding the ID multi-token MWE label to the dependency graph in (HU) Péter **vetette** rá **az első követ**. 'Peter cast the first stone on him'.

For multi-token MWEs our approach is based on our hypothesis that the dependent MWE elements will be directly connected to the other MWE element(s). We do not change the structure of the dependency relations in the tree, but change the dependency label of the dependent MWE element to the MWE type, therefore making the MWE element retraceable from the dependency annotation of the sentence. For example *lát* and *el* in Figure 2 make up a VPC (**ellát** 'take care'), so the dependency relation label of the dependent element, *el* changes from the general syntactic label **PREVERB** to the MWE label **VPC**, with this **VPC** label now connecting the two elements of the MWE.

For MWEs of more than two tokens, the conversion replaces the dependency labels of all MWE elements that depend on the head. In Figure 3, the head of the idiom (**az első követ veti** 'casts the first stone') is the verb, *vetette* (cast.Sg3.Past). All other elements' dependency labels are changed to **ID**.

The second step is training the parser: we used the Bohnet parser (Bohnet 2010) for both POS tagging and dependency parsing. For the single-token MWEs, we trained the Bohnet parser's POS tagger module on the MWE-merged corpora and its dependency parser for the multi-token MWEs. The parser would treat the MWE POS tags and dependency labels as any other POS tag and dependency label.

We did the same for each language and created POS tagging and dependency parsing models capable of identifying MWEs for them. For some languages in the

shared task, we had to omit sentences from the training data that were overly long (spanning over 500 tokens in some cases) and therefore caused errors in training due to lack of memory. This affected one French, one Polish, two Italian, five Romanian and nine Turkish sentences.

Third, we ran the POS tagging and dependency parsing models of each language on their respective test corpora. The output contains the MWE POS tags and dependency labels used in that language as well as the standard POS and syntactic ones.

The fourth and last step is to extract the MWE tags and labels from the output of the POS tagger and the dependency parser. The MWE POS tagged words are annotated as single-token MWEs of the type of their POS tag. From the MWE dependency labels, we annotate the words connected by MWE labels of the same type as making up a multi-token MWE of that type (see Figure 4).

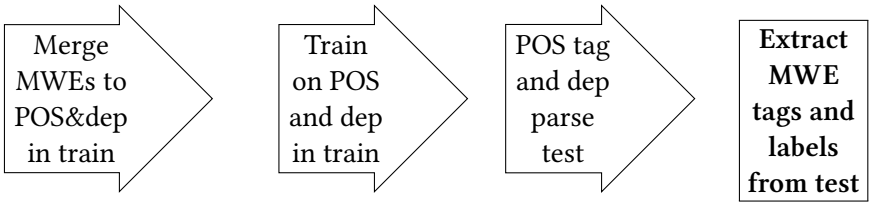


Figure 4: Steps of the USzeged system.

There are arguments for and against our approach. The system cannot handle multi-token MWEs where the elements are not connected in the tree and replacing the POS tags and dependency labels can have a negative effect on the accuracy of POS tagging and parsing. However, as our end goal is not the POS tagging or dependency parse of the data, we believe that this side effect is negligible since higher-level applications (e.g. machine translation) can profit from more accurate MWE identification. On the other hand, the approach has low technical requirements and it is very easily adaptable to other languages.

## 2.2 Results

We submitted the USzeged system for all languages in the shared task with provided dependency analysis and POS tagging. We attempted to use just the POS tagging component of our system on the languages that only had POS tagging available to give partial results (i.e. identifying only single-token MWEs), but we

found that these languages incidentally had no or very few single-token MWEs (Farsi 0, Maltese 4, Romanian 44, Slovene 3, Turkish 22), therefore we had no access to adequate training data and did not submit results for these languages.

Our results on the nine languages are reported in Simkó et al. (2017). Our system was submitted for German, Greek, Spanish, French, Hungarian, Italian, Polish, Portuguese, and Swedish. For the evaluation, we employed the metrics used for the evaluation of the shared task (Savary et al. 2017).

The F-scores show great differences between languages, but so did they for the other systems submitted. Compared to the other, mostly closed-track systems, the USzeged system ranked close to or at the top on German, Hungarian, and Swedish. For the other languages (except for Polish and Portuguese, where ours is the worst performing system), we ranked in the mid-range.

## 2.3 Error analysis

After receiving the gold annotation for the test corpora, we investigated the strengths and weaknesses of our system.

Our error analysis showed that the USzeged system performs by far best on single-token MWEs, which in this dataset are mostly made up of the verb-particle construction category, correctly identifying around 60% of VPCs, but only about 40% of other types on average. It is probably due to the fact that single-token MWEs are identified by POS tagging techniques, which are known to obtain more accurate results in most languages than dependency parsing.

German, Hungarian, and Swedish were also the languages with the highest proportions of the VPC type of verbal MWEs in the shared task, which also correlates with why our system performed best on them. Romance languages contain almost no VPCs and the remaining ones have much less also. In this way, the frequency of VPCs strongly influences our results on the given language.

For French and Italian, our system also performed worse on IReflVs. In general, we had some trouble identifying longer IDs and LVCs and MWEs including prepositions. A further source of error was when there was no syntactic edge in between members of a specific MWE, for instance, in German, the copula **sein** 'be' was often indirectly connected to the other words of the MWE (e.g. ***im Rennen sein*** 'in race be' ⇒ 'to compete'), hence our method was not able to recognize it as part of the MWE. As our system does not restructure the syntactic trees, if the elements of a multi-token VMWE are not connected (i.e. they do not form a graph) in their dependency annotation, we cannot identify the full MWE, however, we can still identify tokens of it correctly if at least two tokens within the MWE are attached.

## 3 The extended system - USzeged+

The primary aim of our extension was to be able to use our system for the languages in the shared task without any available POS and dependency data. We achieved this by parsing the annotated set in a preprocessing step. For the languages with gold POS and dependency data already available, we did not use this extra step (see Figure 5).

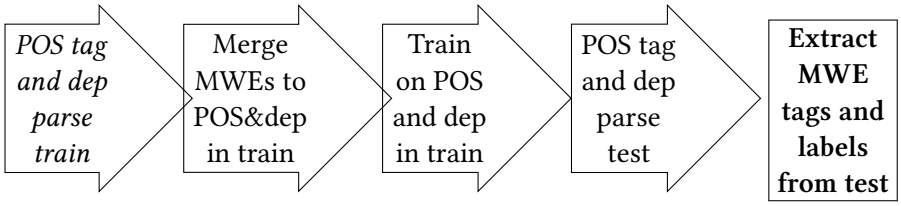


Figure 5: Steps of the USzeged+ system.

We used data for the remaining languages from the Universal Dependencies Project release 2.0 (Nivre et al. 2016) to train the Bohnet parser for POS-tagging and dependency parsing and parsed the VMWE annotated shared task's training sets. We should note that for some languages, the VMWE corpus and the Universal Dependencies corpus are overlapping. This influences our dependency parse to some degree as the training data might partially include the test data, but as our end goal here is not the full dependency parse of the texts (moreover, we already use gold dependency annotations for the languages which have it directly available), we feel that this factor is negligible. Henceforward, we exploited the very same processes as before: we merged the parsed data with the VMWE annotations and once again, trained the Bohnet parser on the VMWE merged data. We then parsed the test sets for the shared task and extracted the MWE POS-tagged and MWE dependency labeled words and phrases.

### 3.1 Results

Table 1 shows the USzeged+ results for all shared task languages. The languages covered by USzeged can be found in the upper part of the table, and the ones covered by USzeged+ are in the lower part. The "upper" languages of this table show differences to the results presented in (Simkó et al. 2017). This is due to two main factors: the Bohnet parser was updated between our USzeged and USzeged+

versions of the system and we also corrected some bugs in our conversion tool. The basic working principles of our system are the same as described above.

Table 1: USzeged+ results: Languages covered by the previous system also are on top.

| | P-MWE | R-MWE | F1-MWE | P-token | R-token | F1-token |
|---|---|---|---|---|---|---|
| DE | 31.16 | 40.20 | 35.11 | 40.65 | 43.05 | 41.82 |
| EL | 37.01 | 30.20 | 33.26 | 49.14 | 32.65 | 39.23 |
| ES | 25.67 | 52.00 | 34.37 | 32.13 | 55.20 | 40.62 |
| FR | 31.23 | 31.60 | 31.41 | 43.57 | 39.44 | 41.40 |
| HU | 62.02 | 71.34 | 66.36 | 58.45 | 69.08 | 63.32 |
| IT | 9.21 | 6.80 | 7.83 | 33.29 | 18.70 | 23.94 |
| PL | 35.96 | 59.40 | 44.80 | 41.33 | 63.68 | 50.13 |
| PT | 33.29 | 52.80 | 40.84 | 40.76 | 58.96 | 48.20 |
| SV | 14.96 | 22.88 | 18.09 | 20.55 | 28.21 | 23.77 |
| BG | 53.26 | 43.13 | 47.66 | 77.79 | 49.25 | 60.32 |
| CS | 44.95 | 57.93 | 50.62 | 57.60 | 64.46 | 60.84 |
| FA | 69.58 | 46.20 | 55.53 | 85.78 | 53.14 | 65.63 |
| HE | 41.18 | 8.40 | 13.95 | 55.22 | 8.63 | 14.93 |
| LT | 33.33 | 7.00 | 11.57 | 40.48 | 6.97 | 11.89 |
| MT | 0.00 | 0.00 | 0.00 | 4.65 | 0.31 | 0.59 |
| RO | 46.29 | 67.40 | 54.89 | 53.01 | 71.68 | 60.95 |
| SL | 59.49 | 18.80 | 28.57 | 66.46 | 18.77 | 29.27 |
| TR | 39.34 | 37.92 | 38.62 | 42.07 | 39.49 | 40.74 |

Using gold or parsed POS and dependency data as the starting phase does not have a significant impact on the results (as we will show in another experiment in §4), with the exception of Maltese. As Maltese currently has no available Universal Dependencies treebank, we used cross-language training to train our parser. As a Semitic language, Maltese is basically related to Arabic but spelt with Latin characters and about half of its vocabulary originates from Italian. Thus, we selected the available Italian Universal Dependencies treebank to train the parser and parse the VMWE train data. This had a very bad effect on our results: no full MWE could be correctly identified in the VMWE test set. Hence, for Maltese, a more suitable solution is still to be found for our approach. For all other languages – where the parser for the VMWE train data was trained on the same

language – the final results are much more comparable to those of the languages with gold trees.

Besides Maltese, one of the languages where our system performed poorly is Italian. We investigated the Italian training corpus and found that its annotation has different underlying principles than most of the other corpora. Namely, it allows sentences to have multiple roots (which is prohibited in other dependency theories), hence it confuses the parser's training to a high degree and therefore very few valuable results (i.e. MWE annotations) can be converted from the parsed sentences. Finally, Swedish results are probably due to the small size of the training corpora.

Table 2 and Table 3 show our results in F-score for the different MWE types; crossed out cells indicate that the type was not present for the given language.

Overall, the USzeged+ system performs best on inherently reflexive verbs (IReflV). IReflVs contain irreflexive pronouns, which show little variability, thus they can be relatively easily recognized by the system. However, the system performs worst on idioms and the "other" category due to their bigger variability and the longer MWEs in these types. Light-verb constructions and verb-particle constructions show varied results depending on the variability of the category in the given language. VPCs could be easily recognized in Hungarian (an F-score over 80), while LVC identification was most successful in Romanian (an F-score of 70).

There are also differences in the annotations of the languages: for instance, Farsi contains only VMWEs for the "other" category, which makes it hard to make any comparisons with the other languages on the effective identification of VMWE categories.

The results also show that while our system has very similar average results on the other language group of the shared task (interestingly even the "other" category, which is probably due to Farsi), results are much lower on Romance languages on average. This is most probably due to the issues on the Italian dependency data (see above), which resulted in poor performance for almost all of the VMWE categories in Italian.

We did some error analysis based on the languages we can speak. This revealed that as for LVCs, our system usually marks as false positives those verb-object pairs where the verb is an otherwise frequent light verb in the given language (e.g. (PT) **ter** 'to have'). Also, participle forms of LVCs were often missed in (FR) **études menées** 'studies conducted'. As for VPCs, many compositional instances of verbs with particles were falsely marked in German and in Hungarian, like (DE) **anheben** 'hang up'. The same is true for IReflVs: compositional ones like (PT) **encantar-se** 'enchant' were sometimes falsely identified as VMWEs. A fur-

Table 2: USzeged+ MWE-level F-score results for the different MWE types.

| | all | VPC | LVC | ID | IReflV | OTH |
|---|---|---|---|---|---|---|
| BG | 47.66 | - | 20.44 | 18.79 | 60.70 | - |
| CS | 50.62 | - | 26.96 | 2.51 | 61.71 | 0.00 |
| DE | 35.11 | 54.55 | 10.39 | 16.16 | 17.50 | - |
| EL | 33.26 | 56.00 | 36.68 | 9.52 | - | 8.00 |
| ES | 34.37 | - | 32.40 | 9.70 | 43.04 | 0.00 |
| FA | 55.53 | - | - | - | - | 55.53 |
| FR | 31.41 | - | 27.67 | 16.48 | 52.94 | 0.00 |
| HE | 13.95 | 9.35 | 24.20 | 0.00 | - | 13.76 |
| HU | 66.36 | 80.59 | 35.47 | - | - | - |
| IT | 7.83 | 22.22 | 5.13 | 5.05 | 0.00 | 0.00 |
| LT | 11.57 | - | 23.53 | 0.00 | - | - |
| MT | 0.00 | - | 0.00 | 0.00 | - | 0.00 |
| PL | 44.80 | - | 25.38 | 0.00 | 65.63 | - |
| PT | 40.84 | - | 45.89 | 11.57 | 40.99 | - |
| RO | 54.89 | - | 70.59 | 20.51 | 54.92 | - |
| SL | 28.57 | 0.00 | 14.29 | 1.90 | 45.71 | 0.00 |
| SV | 18.09 | 22.11 | 8.70 | 0.00 | 2.90 | 0.00 |
| TR | 38.62 | - | 39.77 | 32.71 | - | 34.34 |
| Average | 34.08 | 34.97 | 26.32 | 9.06 | 40.54 | 10.15 |

ther source of errors could also be some inconsistencies in the data: in a few cases, annotators missed to mark some clear examples of VMWEs in the test data, which resulted again in false positives. Finally, the German corpus contained some English sentences, e.g. *[...] if Proporz were to be taken out of the Austrian economy, actual unemployment would be ... higher?* In this sentence, *be* and *higher* are marked as an instance of VPC. The word *be* is a typical particle in German, while last words of the sentences are often verbs in German due to word order reasons. Probably this is the reason why the system gave this analysis.

As our system uses different methods to assign single- and multi-token MWE labels, we also investigated our results for these separately. We found that most languages only contain no or very few single-token MWEs, with the exception of German and Hungarian. Approximately 12% of VMWEs are single-token in the

Table 3: USzeged+ system's token-level F-score results for the different MWE types.

| | all | VPC | LVC | ID | IReflV | OTH |
|---|---|---|---|---|---|---|
| BG | 60.32 | - | 32.63 | 24.30 | 74.05 | - |
| CS | 60.84 | - | 31.78 | 17.87 | 72.44 | 0.00 |
| DE | 41.82 | 56.14 | 12.58 | 32.41 | 24.85 | - |
| EL | 39.23 | 56.00 | 40.65 | 19.97 | - | 4.04 |
| ES | 40.62 | - | 35.86 | 22.47 | 44.27 | 0.00 |
| FA | 65.63 | - | - | - | - | 65.63 |
| FR | 41.40 | - | 30.17 | 36.11 | 52.94 | 0.00 |
| HE | 14.93 | 13.82 | 23.21 | 10.62 | - | 10.83 |
| HU | 63.32 | 80.59 | 40.82 | - | - | - |
| IT | 23.94 | 27.78 | 12.60 | 23.78 | 0.00 | 0.00 |
| LT | 11.89 | - | 23.30 | 3.28 | - | - |
| MT | 0.59 | - | 0.71 | 0.33 | - | 0.00 |
| PL | 50.13 | - | 28.18 | 13.19 | 69.44 | - |
| PT | 48.20 | - | 50.73 | 27.80 | 42.07 | - |
| RO | 60.95 | - | 73.03 | 48.88 | 55.75 | - |
| SL | 29.27 | 1.70 | 15.87 | 9.36 | 46.29 | 0.00 |
| SV | 23.77 | 26.80 | 8.79 | 2.26 | 2.90 | 0.00 |
| TR | 40.74 | - | 41.32 | 34.43 | - | 36.36 |
| Average | 39.87 | 37.55 | 29.54 | 20.44 | 44.09 | 10.62 |

German data, while they make up of 40% of VMWEs in the Hungarian data. Table 4 shows our system's accuracy on single- and multi-token MWEs for German and Hungarian.

Table 4: Accuracy on single- and multi-token MWEs for German and Hungarian.

| | single-token | multi-token | overall |
|---|---|---|---|
| DE | 64 | 36 | 46 |
| HU | 83 | 43.3 | 68.9 |

These results confirm that our system achieves better results on single-token MWEs than on multi-token ones.

## 4 Gold or parsed?

In this last section, we describe a small experiment comparing our new addition to the system: the parsed POS and dependency data. We compare our results on Hungarian using the gold POS and dependency data with an experimental setup mirroring that of the languages without this gold data. We used the Hungarian Universal Dependencies treebank to train the Bohnet parser for POS tagging and dependency parsing and exploited these trained models to parse the VMWE train sentences.

Table 5 shows the results of this experiment; the results for HU-GOLD are the same as the ones for Hungarian in the above tables. The results show that gold and parsed methods in our system can provide very comparable results. Interestingly, in both MWE-level and token-level results, the parsed method provides much higher precision but lower recall than the gold method.

For MWE types, LVCs are causing the main difference in the two systems. Many VPCs in the Hungarian data are single-token, so our system deals with them on the level of POS tagging, which is not affected by gold or parsed dependency trees.

Overall, both options achieved approximately the same results in the automatic VMWE recognition task.

Table 5: Gold and parsed results for Hungarian.

| | HU-GOLD | HU-PARSED |
|---|---|---|
| P-MWE | 62.02 | 74.94 |
| R-MWE | 71.34 | 62.93 |
| F-MWE | 66.36 | 68.41 |
| P-token | 58.45 | 74.74 |
| R-token | 69.08 | 53.85 |
| F-token | 63.32 | 62.60 |
| F-VPC-MWE | 80.59 | 78.59 |
| F-LVC-MWE | 35.47 | 25.41 |
| F-VPC-token | 79.05 | 77.66 |
| F-LVC-token | 40.82 | 25.71 |

## 5 Conclusions

In our chapter, we presented our system for verbal MWE recognition. The system uses POS tagging and dependency parsing as a means of finding verbal MWEs in multiple languages.

Apart from parsing-based solutions (Al Saied et al. (2017), Nerima et al. (2017) and our system), the shared task hosted a number of other approaches, like neural networks (Klyueva et al. (2017)) or sequence labeling based models (Boroş et al. (2017), Maldonado et al. (2017)). In the final results, parsing-based systems achieved the best results for almost all languages, showing that this approach works very well for language independent MWE identification.

Our chapter further shows that it is possible to build a highly language independent MWE detection methodology that makes use of a limited amount of language-specific data and achieve reasonable results.

## Acknowledgements

We would like to thank the anonymous reviewers and the volume editors for their invaluable comments on our work.

Veronika Vincze was supported by the UNKP-17-4 New National Excellence Program of the Ministry of Human Capacities, Hungary.

## Abbreviations

| | | | |
|---|---|---|---|
| ID | idiom | OTH | other |
| IREFLV | inherently reflexive verb | POS | part of speech |
| LVC | light-verb construction | VPC | verb-particle construction |
| MWE | multiword expression | VMWE | verbal multiword expression |

## References

Al Saied, Hazem, Matthieu Constant & Marie Candito. 2017. The ATILF-LLF system for parseme shared task: A transition-based verbal multiword expression tagger. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 127–132. Association for Computational Linguistics. http://aclanthology.coli.uni-saarland.de/pdf/W/W17/W17-1717.pdf. DOI:10.18653/v1/W17-1717

Bohnet, Bernd. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics* (COLING 2010), 89–97. http://www.aclweb.org/anthology/C10-1011.

Boroş, Tiberiu, Sonia Pipa, Verginica Barbu Mititelu & Dan Tufiş. 2017. A data-driven approach to verbal multiword expression detection. PARSEME shared task system description paper. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 121–126. Association for Computational Linguistics. http://www.aclweb.org/anthology/W17-1716. DOI:10.18653/v1/W17-1716

Candito, Marie & Matthieu Constant. 2014. Strategies for contiguous multiword expression analysis and dependency parsing. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, 743–753. Association for Computational Linguistics. http://www.aclweb.org/anthology/P14-1070.

Constant, Matthieu & Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 161–171. Association for Computational Linguistics. http://www.aclweb.org/anthology/P16-1016.

Green, Spence, Marie-Catherine de Marneffe, John Bauer & Christopher D. Manning. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with French. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 725–735. Association for Computational Linguistics. http://www.aclweb.org/anthology/D11-1067.

Green, Spence, Marie-Catherine de Marneffe & Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics* 39(1). 195–227. DOI:10.1162/COLI_a_00139

Klyueva, Natalia, Antoine Doucet & Milan Straka. 2017. Neural networks for multi-word expression detection. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 60–65. Association for Computational Linguistics. http://aclanthology.coli.uni-saarland.de/pdf/W/W17/W17-1707.pdf. April 4, 2017. DOI:10.18653/v1/W17-1707

Maldonado, Alfredo, Lifeng Han, Erwan Moreau, Ashjan Alsulaimani, Koel Dutta Chowdhury, Carl Vogel & Qun Liu. 2017. Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 114–120. Association for Computational Linguistics. http://www.aclweb.org/anthology/W17-1715. DOI:10.18653/v1/W17-1715

Nasr, Alexis, Carlos Ramisch, José Deulofeu & André Valli. 2015. Joint dependency parsing and multiword expression tokenization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1116–1126. Association for Computational Linguistics. http://www.aclweb.org/anthology/P15-1108.

Nerima, Luka, Vasiliki Foufi & Eric Wehrli. 2017. Parsing and MWE detection: Fips at the PARSEME shared task. In *Proceedings of The 13th Workshop on Multiword Expressions* (MWE '17), 54–59. Association for Computational Linguistics. http://aclanthology.coli.uni-saarland.de/pdf/W/W17/W17-1706.pdf. DOI:10.18653/v1/W17-1706

Nivre, Joakim. 2015. Towards a universal grammar for natural language processing. In Alexander Gelbukh (ed.), *Computational Linguistics and intelligent text processing*, 3–16. Cham: Springer. http://stp.lingfil.uu.se/~nivre/docs/nivre15cicling.pdf.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), 1659–1666. European Language Resources Association (ELRA). 23-28 May, 2016.

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 31–47. Association for Computational Linguistics. http://aclanthology.coli.uni-saarland.de/pdf/W/W17/W17-1704.pdf. DOI:10.18653/v1/W17-1704

Simkó, Katalin Ilona, Viktória Kovács & Veronika Vincze. 2017. USzeged: Identifying verbal multiword expressions with POS tagging and parsing techniques. In *Proceedings of The 13th Workshop on Multiword Expressions* (MWE '17), 48–53. Association for Computational Linguistics.

Taslimipoor, Shiva, Omid Rohanian, Ruslan Mitkov & Afsaneh Fazly. 2018. A fresh look at modelling and evaluation of multiword expression tokens. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.),

*Multiword expressions at length and in depth. Extended papers from the MWE 2017 workshop*. Berlin: Language Science Press.

Vincze, Veronika, János Zsibrita & István Nagy T. 2013. Dependency parsing for identifying Hungarian light verb constructions. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, 207–215. Nagoya, Japan: Asian Federation of Natural Language Processing. http://www.aclweb.org/anthology/I13-1024.

Waszczuk, Jakub, Agata Savary & Yannick Parmentier. 2016. Promoting multiword expressions in a* TAG parsing. In Nicoletta Calzolari, Yuji Matsumoto & Rashmi Prasad (eds.), *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers* (COLING-16), 429–439. Association for Computational Linguistics. http://aclweb.org/anthology/C/C16/C16-1042.pdf. December 11-16, 2016.

Wehrli, Eric, Violeta Seretan & Luka Nerima. 2010. Sentence analysis and collocation identification. In *Proceedings of the Workshop on Multiword Expressions: From Theory to Applications* (MWE '10), 27–35. Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W10/W10-??04.

# Did you like this book?

This book was brought to you for free

# Multiword expressions at length and in depth

The annual workshop on multiword expressions takes place since 2001 in conjunction with major computational linguistics conferences and attracts the attention of an ever-growing community working on a variety of languages, linguistic phenomena and related computational processing issues. MWE 2017 took place in Valencia, Spain, and represented a vibrant panorama of the current research landscape on the computational treatment of multiword expressions, featuring many high-quality submissions. Furthermore, MWE 2017 included the first shared task on multilingual identification of verbal multiword expressions. The shared task, with extended communal work, has developed important multilingual resources and mobilised several research groups in computational linguistics worldwide.

This book contains extended versions of selected papers from the workshop. Authors worked hard to include detailed explanations, broader and deeper analyses, and new exciting results, which were thoroughly reviewed by an internationally renowned committee. We hope that this distinctly joint effort will provide a meaningful and useful snapshot of the multilingual state of the art in multiword expressions modelling and processing, and will be a point of reference for future work.