Projecting Multiword Expression Resources on a Polish Treebank

Agata Savary[1] **and Jakub Waszczuk**[1,2]

[1]Université François Rabelais Tours, France
[2]Université d'Orléans, France
{agata.savary,jakub.waszczuk}@univ-tours.fr

Abstract

Multiword expressions (MWEs) are linguistic objects containing two or more words and showing idiosyncratic behavior at different levels. Treebanks with annotated MWEs enable studies of such properties, as well as training and evaluation of MWE-aware parsers. However, few treebanks contain full-fledged MWE annotations. We show how this gap can be bridged in Polish by projecting 3 MWE resources on a constituency treebank.

1 Introduction

Multiword expressions (MWEs) are linguistic objects containing two or more words and showing idiosyncratic behavior at different linguistic levels (Savary et al., 2015). For instance, at the morphological level they can have restricted paradigms, e.g., in Polish (PL) *zjadłbym konia z kopytami* (lit. *I would eat a horse with its hooves*) 'I am very hungry' can only occur in the conditional mood. At the syntactic level they can: (i) exhibit defective agreement, e.g., in French (FR) in *grands-mères* 'grandmothers' the adjective does not agree with the noun in gender unlike all regular adjectival modifiers, (ii) impose agreement constraints which do not apply to compositional structures, e.g., *to have one's heart in one's mouth* imposes agreement in person between both possessive pronouns and the subject, (iii) block some transformations typical for their structures, e.g., **the bucket was kicked by him*, (iv) prohibit or require modifiers, e.g., (FR) *germer dans le cerveau de quelqu'un* (lit. *to germinate in someone's brain*) imposes a pronominal or nominal modifier of *brain*, etc. At the semantic level, MWEs show a varying degree of non-compositionality, e.g., *to pull strings* is semantically opaque but can be understood compositionally if the components themselves are interpreted in an idiomatic way (*to pull* as 'to use', and *strings* as 'one's influence').

Treebanks in which MWE have been explicitly annotated are highly precious resources enabling us to study such more or less unpredictable properties. They also constitute basic prerequisites for training and evaluating parsers, which should best perform syntactic analysis jointly with MWE identification (Finkel and Manning, 2009; Green et al., 2013; Candito and Constant, 2014; Le Roux et al., 2014; Wehrli, 2014; Nasr et al., 2015; Constant and Nivre, 2016; Waszczuk et al., 2016).

However, few treebanks contain full-fledged MWE annotations, even for English (Rosén et al., 2015). Multiword named entities (MWNEs) constitute by far the most frequently annotated category (Erjavec et al., 2010; Savary et al., 2010). Continuous MWEs such as compound nouns, adverbs and prepositions and conjunctions are covered in some treebanks as in (Abeillé et al., 2003; Branco et al., 2010). Verbal MWEs (VMWEs) have been addressed for a fewer number of languages (Bejček et al., 2011; Eryigit et al., 2015; Seraji et al., 2014), and often restricted to some subtypes only, e.g., light-verb constructions (Vincze and Csirik, 2010).

Lexical MWE resources develop more rapidly than MWE-annotated treebanks (Losnegaard et al., 2016). They already exist for a large number of languages and are often distributed under open licenses. It is, thus, interesting to examine how far MWE lexicons can help in completing the existing treebanks with annotation layers dedicated to MWEs. Our case study deals with four Polish resources: (i) the named-entity annotation layer of a Polish reference corpus, (ii) an e-lexicon of nominal, adjectival and adverbial MWEs, (iii) a valence dictionary with a phraseological component, and (iv) a treebank with no initial MWE annotations.

Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pages 20–26,
Valencia, Spain, 4 April 2017. ©2017 Association for Computational Linguistics

We show how the 3 former resources can be automatically projected on the latter, by identifying syntactic nodes satisfying (totally or partly) the appropriate lexical and syntactic constraints.

2 Resources

The National Corpus of Polish (NCP) (Przepiórkowski et al., 2012) contains a manually double-annotated and adjudicated subcorpus of over 1 million words. Its **named entity layer** (**NCP-NE**), which builds on the morphosyntactic layer (relying in its turn on the segmentation layer), contains over 80,000 annotated NEs, 20% of which are MWNEs. Only the latter were used in the experiments described below. The annotation schema assumes notably the markup of nested, overlapping and discontinuous NEs, i.e., the annotation structures form trees (Savary et al., 2010).

SEJF (Czerepowicka and Savary, 2015) is a grammatical lexicon of Polish continuous MWEs containing over 4,700 compound nouns, adjectives and adverbs, where inflectional and word-order variation is described via fine-grained graph-based rules. It is provided in two forms – intensional (multiword lemmas and inflection rules) and extensional (list of morphologically annotated variants). The latter, generated automatically from the former, was used in our projecting experiments. Tab. 1 shows a sample extensional entry containing a MWE inflected form, its lemma and morphological tag: noun (`subst`) in singular (`sg`) genitive (`gen`) and feminine gender (`f`).

Inflected form	Lemma	Tag
drugiej połowy	*druga połowa*	`subst:sg:gen:f`

Table 1: An inflected form of *druga połowa* (lit. *second half*) 'one's husband or wife' in SEJF.

Walenty is a Polish large-scale valence dictionary of about 50000, 3700, 3000, and 1000 subcategorization frames (in its 2015 version) for Polish verbs, nouns, adjectives, and adverbs respectively. Its encoding formalism is rather expressive and theory-neutral, and includes an elaborate phraseological component (Przepiórkowski et al., 2014).[1] Thus, above 8,000 verbal frames contain lexicalized arguments of head verbs, i.e., they describe VMWEs. For instance the idiom highlighted in example (1) is described in Walenty as shown in Tab. 2. Each component separated by a '+' represents one required verbal argument with its lexical, morphological, syntactic, and (sometimes) semantic constraints. Here, the subject is compulsory and has a structural case (`subj{np(str)}`), which notably means that it normally occurs in the nominative, but turns to the genitive when realized as a numeral phrase (of a certain type). The subject being a required argument in a verbal frame does not contradict the fact that it can regularly be omitted in Polish, as in (1).[2]

(1) Nie umiem w tych sprawach **trzymać**
Not know.SG.PRI in these affairs hold.INF
języka **za** **zębami**.
tongue.SG.GEN behind teeth.
(lit. *I cannot hold my tongue behind my teeth in such cases*) 'I cannot hold my tongue in such cases'

The second required argument is a direct object realized as a nominal phrase in structural case, i.e., normally in the accusative but turning to the genitive when the sentence is negated, as in (1). The lexicalized object's head has the lemma *język* 'tongue', should be in singular (`sg`) and does not admit modifiers (`natr`). The second complement is a prepositional nominal phrase (`prepnp`) headed by the preposition *za* 'behind' governing the instrumental case (`inst`) and a lexicalized non-modifiable (`natr`) noun with the lemma *ząb* 'tooth' in plural (`pl`). Walenty's syntax is compact and meant to be easily handled by lexicographers but proved sufficiently formalized to be directly applicable to NLP tasks, such as automatic generation of grammar rules (Patejuk, 2015).

```
trzymać: subj{np(str)}+
  obj{lex(np(str),sg,'język',natr)}+
  {lex(prepnp(za,inst),pl,'ząb',natr)}
```

Table 2: Description of *trzymać język za zębami* 'hold one's tongue' in Walenty

Składnica is a Polish constituency treebank comprising about 9,000 sentences with manually disambiguated syntactic trees (Świdziński and Woliński, 2010). It was created by automatically generating all possible parses with a large-coverage DCG grammar, and then manually selecting the correct parse. It does not contain MWE

[1] Walenty and PDT-Vallex for Czech (Urešová et al., 2014) belong to the most elaborate and extensive endeavors towards the description of the valency of VMWEs (Przepiórkowski et al., 2016).

[2] This property is to be distinguished from impersonal verbs, which prohibit a subject, as in *dobrze mu z oczu patrzy* (lit. *looks him from eyes well*) 'he looks like a good person'.

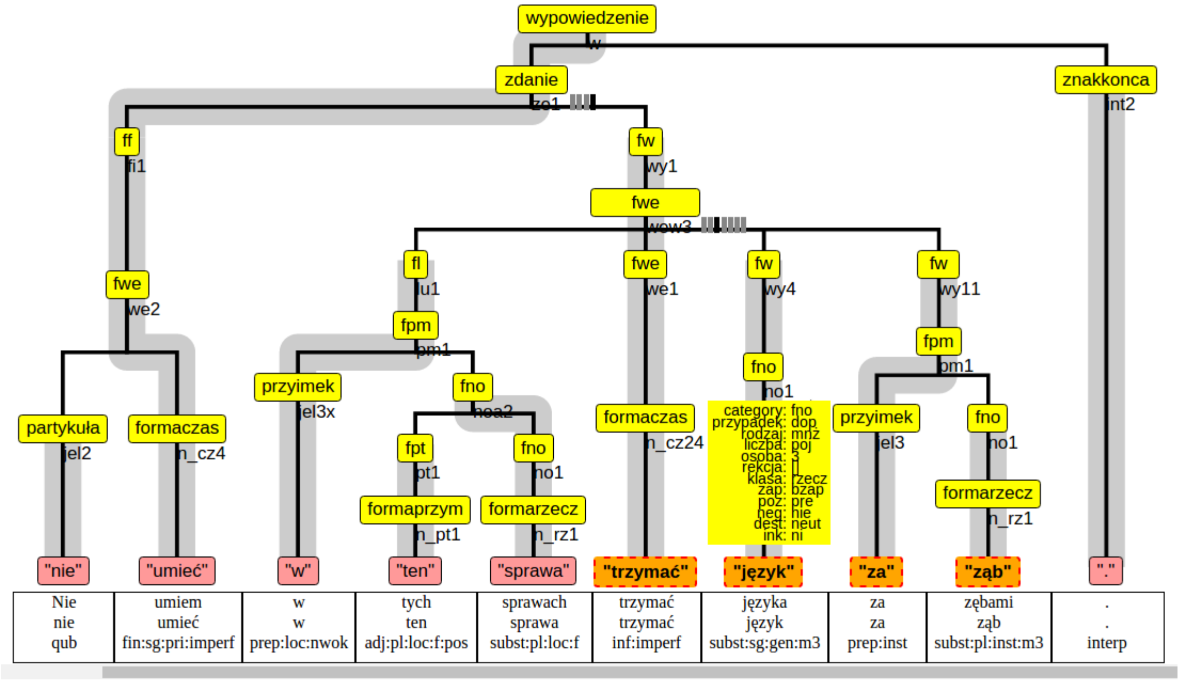

Figure 1: Syntax tree of example (1) in Składnica. The categories denote: `ff` 'finite phrase', `fl` 'adjunct', `fno` 'nominal phrase', `formaczas` 'verbal phrase', `formaprzym` 'adjectival phrase', `formarzecz` 'nominal phrase', `fpm` 'prepositional phrase', `fpt` 'adjectival phrase', `fw` 'required phrase', `fwe` 'verbal phrase', `partykuła` 'particle', `przyimek` 'preposition', `wypowiedzenie` 'utterance', `zdanie` 'sentence', `znakkońca` 'ending punctuation'. The feature structure of the `fno` node dominating the terminal *język* 'tongue' is highlighted. The feature codes include: `przypadek` 'case', `rodzaj` 'gender', `liczba` 'number', `osoba` 'person', `rekcja` 'case government', and `neg` 'negation'. The values denote: `dop` 'genitive', `mnz` 'human inanimate', `poj` 'singular', and `nie` 'negated'.

annotations. Its morphosyntactic tagset is mostly equivalent to the one used in Walenty, although it uses Polish terms: `mian`=*mianownik* 'nominative', `dk`=*dokonany* 'perfective aspect', etc.

Fig. 1 shows the correct syntax tree from Składnica for example (1). Each non-terminal node includes a feature structure (FS). Here, the FS of the node `fno` (nominal phrase), above the terminal *język* 'tongue', is highlighted. It includes the feature `neg=nie` meaning that this node occurs within the scope of a negated verb. This makes it easy to validate constraints from Walenty entries, such as the structural genitive of direct objects.

A notable feature of Składnica is that dependents of the verbs are explicitly marked as either arguments (`fw`) or adjuncts (`fl`), i.e., valency is accounted for. Note, however, that the valency of head verbs in VMWEs can differ from the one of the same verbs occurring as simple predicates.

3 Projection

Since Składnica contains no explicit MWE annotations, we produced them automatically by projecting NCP-NE, SEJF and Walenty on the syntax trees. The projection for NCP-NE was straightforward and did not require manual validation, since Składnica is a subcorpus of the NCP, whose NE annotation and adjudication were performed manually. The projection for SEJF and Walenty, followed by a manual validation, consisted in searching for syntactic nodes satisfying all lexical constraints and part of syntactic constraints of a MWE entry. The required lexical nodes were to be contiguous for SEJF but not for Walenty.

Here, we give more details on the Walenty-to-Składnica projection, which was the most challenging one. It required defining correspondences at different levels. Explicit morphological values and phrase types could be translated rather straightforwardly due to largely compatible tagsets (`np`→`fno` 'nominal phrase', `mian`→`nom` 'nominative', etc.). Context-dependent values like `str` (structural case) were encoded in conditional statements taking combination of features into account. For instance, the argument specification `obj(np(str))` translated into a feature structure containing one of the following: $[category = fno, przypadek = bier, neg = tak]$, $[category = fno, przypadek = dop, neg = nie]$ (nominal object, either in the accusative in an affirmative sentence or in the genitive in a negative one).

Once these correspondences were defined, identifying a Walenty entry in Składnica consisted in checking if the current sentence contained a subtree in which: (i) the lexically constrained arguments and adjuncts (and their own, recursively embedded, lexically constrained dependents) were present, (ii) selected syntactic constraints (those concerning `np` and `prepnp` phrases) were fulfilled. For instance in Fig. 1, a head verb, a direct object with a lexicalized head and a lexicalized prepositional complement were searched for, but an ellipsis of the subject was allowed.

Query language The MWE projection task is handled by: (i) a query language, providing an interface between the MWE resources and the treebank, (ii) procedures for compiling lexicon entries into the queries, and (iii) an interpreter which runs a query over treebank subtrees to check whether the corresponding MWE entry occurs in them.

Formally, we defined our core query language using the following abstract syntax:

$$
\begin{aligned}
&b\,\text{(Booleans)} ::= \texttt{true} \mid \texttt{false}\\
&n\,\text{(node queries)} ::= \; b \mid n_1 \wedge n_2 \mid n_1 \vee n_2\\
&\quad \mid \texttt{mark} \mid \texttt{satisfy}\,(\textit{node} \rightarrow b)\\
&t\,\text{(tree queries)} ::= \; b \mid t_1 \wedge t_2 \mid t_1 \vee t_2\\
&\quad \mid \texttt{root}\, n \mid \texttt{child}\;\, t \mid \ldots
\end{aligned}
$$

Thus, the properties of a given syntactic node or tree can be verified via an appropriate node query (NQ) or tree query (TQ), respectively. Both kinds of queries are recursive and TQs can additionally build on NQs. For instance, from the query interpretation point of view, the TQ `root` n is satisfied for a given tree iff its root satisfies the NQ n. Also, the TQ `child` t is satisfied iff at least one of its root's children trees satisfies the TQ t. Finally, particular feature values ($category$, $przypadek$, etc.) can be verified using the NQ `satisfy` ($node \rightarrow b$), which takes an arbitrary node-level predicate ($node \rightarrow b$) and tells whether it is satisfied over the current syntactic *node*.

The particularity of this query language is the `mark` construction, which marks a syntactic node as a part of a MWE. When a TQ t containing `mark` has been executed over a tree T, t's result contains all nodes matched with `mark`, provided that T satisfies all the constraints encoded in t.

`Mark` does not check any constraints by itself, but it can be easily combined with other NQs via query conjunction (i.e., $n \wedge \texttt{mark}$).

Note that, based on our core language, more complex queries can be expressed, for instance:

$$\texttt{member}\; n \overset{\text{def}}{=} \texttt{root}\;\; n \vee \texttt{child}\; \big(\texttt{member}\;\; n\big) \quad (2)$$

The query interpreter is defined over the core language only and handles MWE-related marking. For instance, given a query of type $t_1 \vee t_2$, while evaluating t_1, some subtree nodes may be `mark`ed as potential MWE components. But if t_1 finally evaluates to `false`, all these markings are wiped out. This behavior is guaranteed by the implementation of the core disjunction ($\vee$) operator.

Compiling MWE entries Let us focus on the Walenty-to-query compilation and on the entry from Tab. 2 in particular. Its querified version checks that (i) the base form of the lexical head, reached via the head-annotated edges (marked in grayed in Fig. 1), corresponds to the main verb of the entry (i.e., *trzymać*), and (ii) each of the lexically-constrained elements of the frame (i.e., noun phrase *język* and prepositional phrase *za zębami*) is realized by one of the `child`-ren trees of the queried tree. Part (i) of the query is implemented by the version of the `member` query (see Eq. 2) restricted to head-annotated edges. Implementation of (ii) depends on the particular frame element. Tree queries corresponding to (i) and (ii) are then combined using the $\wedge$ operator.

The `obj{lex(np(str),sg,'język',natr)}` frame element is also translated to a $\wedge$-combined set of tree queries, which individually check that all the given restrictions are satisfied: the lexical head is *język*, the number is singular, etc. The node query which verifies that *język* is the lexical head is combined with `mark`, so that it is designated as a part of the resulting MWE annotation, provided that all the other entry-related constraints are also satisfied. Modifiers, if specified, are recursively compiled into tree queries which are then applied over `child`-ren trees. Here, `natr` specifies that no modifiers are allowed, constraint compiled into a query which checks that the corresponding tree

Source	TP	FP	CRead	All	CRate
NKJP	1,304	n/a	n/a	1,304	n/a
SEJF	368	18	23	409	0.94
Walenty	365	78	18	452	0.95
Total	2,037	96	41	2,165	0.95

Table 3: Projection results including true positives (TP), false positives (FP), compositional readings (CRead), compositionality rate (CRate).

is non-branching (i.e., has no other children apart from its head, constraint satisfied in Fig. 1 by the subtree rooted with `fno` placed over the leaf *język*).[3] The other element of the frame, which describes the prepositional argument *za zębami*, is compiled into a query in a similar way.

4 Results

Table 3 shows the projection results. Among the 2165 automatically identified candidate MWEs, those 1,304 stemming from NCP-NE were supposed correct (since resulting from manual double-annotation and adjudication). The 861 remaining candidates were manually validated. They contained 733 true positives, 96 false positives, and 41 candidates with a compositional reading, as in examples (3)-(4). Thus, the precision of the SEJF/Walenty projection was equal to 0.85. The idiomaticity rate (El Maarouf and Oakes, 2015), i.e., the ratio of occurrences with idiomatic reading to all correctly recognized occurrences, is about 0.95. We expect that if NEs were taken into account, this ratio would be even higher, since NEs seem to exhibit compositional readings relatively rarely. Note also that false positives are much more frequent for entries stemming from Walenty than for those from SEJF, which shows the higher complexity of verbal MWEs as compared to other, continuous, MWEs.

(3) . . . w **drugiej połowie** XIX wieku
'. . . in the **second half** of the 19th century'
MWE: (lit. *second half*) 'one's husband or wife'

(4) Odetchnęła głęboko i **przymknęła oczy**.
'(She) breathed profoundly and **closed** her **eyes**.'
MWE: *przymknać oczy na coś* (lit. *to close one's eyes on sth*) 'to pretend not to see sth'

Notable errors in the projection procedure stem from allowing for the ellipsis of compulsory but non-lexicalized arguments. If all such arguments marked in Walenty were required in Składnica during the projection, correct MWEs occurrences with ellipted arguments would be missed, as in the case of the subject required in Tab. 2 but omitted in Fig. 1. Conversely, allowing for the ellipsis of such arguments results in some false positives, as in example (4), where the absence of the prepositional argument (headed by the preposition *na* 'on') excludes the idiomatic reading.

5 Summary and Perspectives

The automatic projection of MWEs resources on a treebank results in a manually validated resource containing over 2,000 VMWEs in about 9,000 constituency trees, and available under the GPL v3 license.[4] The results are represented in a simplified custom XML format, meant for an easy use, e.g., in automatic grammar extraction. This format refers to identifiers of sentences and tokens in the Składnica trees, which enables users to automatically project annotations on the original treebank.

We believe to have shown examples of fine-grained and high-quality MWE resources which might be promoted as standards for the international community. Adapting their formalisms to many languages should be possible with affordable efforts (already undertaken by us for French). In return, relatively reliable mapping procedures based on such resources may help bridge the gap towards large and comprehensive MWE-annotation in treebanks, which is currently a bottleneck in the MWE-oriented research.

Another interesting finding, worth confirming in other languages, is the high idiomaticity rate of MWEs. It is a hint that automated MWE identification based on purely syntactic methods and rich resources may achieve high accuracy, even in the absence of semantic non-compositionality models.

Future work includes repeating the experiments with the new version of Walenty released in 2016, as well as estimating the projection recall. We also wish to enhance the lexicon projection process, so as to account for more fine-grained constraints, and tune the degree of flexibility in constraint validation. Finally, an appropriate MWE annotation schema is needed in which each MWE occurrence would be linked to its corresponding entry in a MWE lexicon, and its required arguments, whether lexicalized or not, would be marked.

[3] The `non-branching` predicate is a part of the core language. We did not define it above for the sake of brevity.

[4] `http://zil.ipipan.waw.pl/Sk\%C5\%82adnicaMWE`

References

Anne Abeillé, Lionel Clément, and François Toussenel, 2003. *Building a treebank for French*, pages 165–187. Kluwer Academic Publishers.

Eduard Bejček, Pavel Straňák, and Daniel Zeman. 2011. Influence of Treebank Design on Representation of Multiword Expressions. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing - 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part I*, volume 6608 of *Lecture Notes in Computer Science*, pages 1–14. Springer.

António Branco, Francisco Costa, João Silva, Sara Silveira, Sérgio Castro, Mariana Avelãs, Clara Pinto, and João Graça. 2010. Developing a deep linguistic databank supporting a collection of treebanks: the cintil deepgrambank. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA).

Marie Candito and 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, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 743–753.

Matthieu Constant and 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)*, pages 161–171, Berlin, Germany, August. Association for Computational Linguistics.

Monika Czerepowicka and Agata Savary. 2015. SEJF - a Grammatical Lexicon of Polish Multi-Word Expressions. In *Proceedings of Language and Technology Conference (LTC'15), Poznań, Poland*. Wydawnictwo Poznańskie.

Ismail El Maarouf and Michael Oakes. 2015. Statistical Measures for Characterising MWEs. In *IC1207 COST PARSEME 5th general meeting*.

Tomaž Erjavec, Darja Fišer, Simon Krek, and Nina Ledinek. 2010. The JOS Linguistically Tagged Corpus of Slovene. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA).

Gulsen Eryigit, Kubra Adali, Dilara Torunoglu-Selamet, Umut Sulubacak, and Tugba Pamay. 2015. Annotation and Extraction of Multiword Expressions in Turkish Treebanks. In *Proceedings of NAACL-HLT 2015*, pages 70–76. Association for Computational Linguistics.

Jenny Rose Finkel and Christopher D. Manning. 2009. Joint Parsing and Named Entity Recognition. In *HLT-NAACL*, pages 326–334. The Association for Computational Linguistics.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing Models for Identifying Multiword Expressions. *Computational Linguistics*, 39(1).

Joseph Le Roux, Antoine Rozenknop, and Matthieu Constant. 2014. Syntactic Parsing and Compound Recognition via Dual Decomposition: Application to French. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1875–1885. Dublin City University and Association for Computational Linguistics.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. PARSEME Survey on MWE Resources. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Alexis Nasr, Carlos Ramisch, José Deulofeu, and 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)*, pages 1116–1126. Association for Computational Linguistics.

Agnieszka Patejuk. 2015. *Unlike coordination in Polish: an LFG account*. Ph.D. dissertation, Institute of Polish Language, Polish Academy of Sciences, Cracow.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*. Wydawnictwo Naukowe PWN, Warsaw.

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, and Marcin Woliński. 2014. Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, pages 83–91, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Adam Przepiórkowski, Jan Hajič, Elżbieta Hajnicz, and Zdeňka Urešová. 2016. Phraseology in two Slavic valency dictionaries: Limitations and perspectives. *International Journal of Lexicography*, 29. Forthcoming.

Victoria Rosén, Gyri Smørdal Losnegaard, Koenraad De Smedt, Eduard Bejček, Agata Savary, Adam Przepiórkowski, Petya Osenova, and Verginica Barbu Mitetelu. 2015. A survey of multiword expressions in treebanks. In *Proceedings of the 14th International Workshop on Treebanks & Linguistic Theories conference*, Warsaw, Poland, December.

Agata Savary, Jakub Waszczuk, and Adam Przepiórkowski. 2010. Towards the Annotation of Named Entities in the National Corpus of Polish. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA).

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Matthieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland, November.

Mojgan Seraji, Carina Jahani, Beáta Megyesi, and Joakim Nivre. 2014. A Persian Treebank with Stanford Typed Dependencies. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Zdeňka Urešová, Jan Štěpánek, Jan Hajič, Jarmila Panevova, and Marie Mikulová. 2014. PDT-vallex: Czech valency lexicon linked to treebanks. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Veronika Vincze and János Csirik. 2010. Hungarian Corpus of Light Verb Constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1110–1118. Coling 2010 Organizing Committee.

Jakub Waszczuk, Agata Savary, and Yannick Parmentier. 2016. Promoting multiword expressions in A* TAG parsing. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 429–439. ACL.

Eric Wehrli. 2014. The Relevance of Collocations for Parsing. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 26–32, Gothenburg, Sweden, April. Association for Computational Linguistics.

Marek Świdziński and Marcin Woliński. 2010. Towards a bank of constituent parse trees for Polish. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue: 13th International Conference, TSD 2010, Brno, Czech Republic*, volume 6231 of *Lecture Notes in Artificial Intelligence*, pages 197–204, Heidelberg. Springer-Verlag.