Mapping a MWE lexicon on a treebank [WG1,WG4]

Jakub Waszczuk, Agata Savary
Université François Rabelais Tours, France
first.last@univ-tours.fr

Treebanks annotated with multiword expressions (MWEs) are important linguistic resources in NLP. They allow one to study the syntactic properties of MWEs, which are usually partly regular and partly idiosyncratic. They also constitute basic prerequisites for training and evaluating parsers, which should best perform syntactic analysis jointly with MWE identification (Finkel and Manning 2009, Wehrli *et al.* 2010, Green *et al.* 2011, 2013, Candito and Constant 2014, Wehrli 2014, Nasr *et al.* 2015).

However, few treebanks contain a full-fledged range of MWE annotations, even for English (Rosén *et al.* 2015). Multiword named entities constitute by far the most frequently annotated category, e.g., in (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 fever 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).

Lexical resources of MWEs develop more rapidly than MWE-annotated treebanks. As shown by a recent PARSEME survey (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 in this respect deals with two Polish resources: a valence dictionary containing a phraseological component, and a treebank with no initial MWE annotations. We show how the former can be automatically mapped on the latter, by identifying syntactic nodes satisfying (totally or partly) the appropriate lexical and syntactic constraints. We focus on VMWEs, since they belong to the most interesting and challenging MWE types due to the complex constraints that they impose on their arguments, and to the fact that their lexicalized components often occur in text in a discontinuous manner.

Walenty is a Polish large-scale valence dictionary of about 50,000, 3,700 3,000, and 1,000 subcategorization frames for Polish verbs, nouns, adjectives, and adverbs respectively. Its encoding formalism is rather expressive and theory-neutral[1], and includes an elaborate phraseological component (Przepiórkowski *et al.* 2014). 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 Fig. 1. 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 nominative, but turns to genitive when the head verb is nominalized. The subject being a required argument in a verbal frame does not contradict the fact that it can regularly be omitted in Polish sentences[2], as in example 1.

(1) Nie umiem w tych sprawach **trzymać** **języka za zębami**.
Not know.SG.PRI in these affairs hold.INF 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 accusative but turning to genitive when the sentence is negated as in example 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`).

[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* 'looks him from eyes well'⇒ 'he looks like a good person'.

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

Figure 1: Description of *trzymać język za zębami* ('hold one's tongue') in Walenty

Walenty's syntax is very 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).

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 Chomskian large-coverage grammar, and then manually selecting the correct parse. It does not contain MWE 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. 3 shows the correct syntax tree from Składnica for example (1). Each non-terminal node includes a feature structure (FS). For instance, the FS of the node `fno` (nominal phrase) above the terminal *język* 'tongue' shown in Fig. 2, includes the feature `neg=nie` meaning that this node occurs within the scope of a negated verb. This enables an easy validation of some 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 obviously differ from the one of the same verbs occurring as simple predicates.

Mapping Walenty entries on Składnica trees required defining correspondences at different levels. Explicit morphological values and phrase types could be translated rather straightforwardly due to largely compatible tagsets (e.g., `np`→`fno` 'nominal phrase', `mian`→`nom` 'nominative'). Context-dependent values like `str` (structural case) or `agr` (agreeing case) had to be 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 phrase object, either in accusative in an affirmative sentence or in genitive in a negative one).

Once these correspondences in morphosyntactic descriptions were defined, the procedure of identifying a Walenty MWE entry in Składnica consisted in checking if the current sentence contained a subtree in which the corresponding constraints were fulfilled. For instance in Fig. 3, 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. For the first experiments, we implemented a relaxed version of the mapping where only the lexically constrained arguments and adjuncts (and their own, recursively embedded, lexically constrained dependents) were taken into account and only selected syntactic constraints were verified in the mapping process[3].

Results As a result of the mapping, 499 occurrences of candidate verbal MWEs were automatically identified in the treebank and manually validated: 390 of them were true positives[4], 27 were compositional occurrences (cf. Appendix B), and 82 were false positives (resulting mainly from relieving too many constraints in the mapping procedure). The idiomaticity rate (El Maarouf and Oakes 2015), i.e. the ratio of occurrences with idiomatic reading to all correctly recognized occurrences in this sample, is equal to 0.93. This data set has already been used for an automatic extraction of a Lexicalized Tree Adjoining Grammar of Polish. Each phrase containing a MWE yielded notably an elementary tree with multiple co-anchors.

Futures work includes enhancing the Walenty mapping procedures so as account for more fine-grained constraints, and tuning the degree of flexibility in constraint validation so as to obtain optimal precision and recall. We also wish to produce more complete MWE annotations of Składnica including named entities and compounds, whose density in corpora is usually much higher than of verbal MWEs. Existing resources such as the named entity layer of the National Corpus of Polish (Savary *et al.* 2010) or SEJF, a Polish extensional lexicon of nominal, adjectival and adverbial MWEs (Czerepowicka and Savary 2015), could be used to this aim. Finally, we will work towards defining an appropriate MWE annotation schema in which each MWE occurrence is linked to its corresponding entry in a MWE lexicon, and its

[3] Namely, syntactic constraints for the `np` and `prepnp` phrases were verified, while for the other types of phrases only lexical constraints were checked.

[4] This rather low density of VMWEs confirms previous observations from the pilot corpus annotation within the PARSEME shared task on automatic identification of VMWEs.

required arguments, whether lexicalized or not, are clearly marked.

References

Abeillé, A., Clément, L., and Toussenel, F. (2003). *Building a treebank for French*, pp. 165–187. Kluwer Academic Publishers.

Bejček, E., Straňák, P., and Zeman, D. (2011). Influence of Treebank Design on Representation of Multiword Expressions. In A. F. Gelbukh, ed., *Computational Linguistics and Intelligent Text Processing - 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part I*, pp. 1–14. Springer.

Branco, A., Costa, F., Silva, J., Silveira, S., Castro, S., Avelãs, M., Pinto, C., and Graça, J. (2010). Developing a deep linguistic databank supporting a collection of treebanks: the CINTIL DeepGramBank. In *LREC*.

Candito, M. and Constant, M. (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*, pp. 743–753.

Czerepowicka, M. and Savary, A. (2015). SEJF - a Grammatical L @misc11858/00-097C-0000-0023-4338-F, title = PDT-Vallex: Czech Valency lexicon linked to treebanks, author = Urešová, Zdeňka and Štěpánek, Jan and Hajič, Jan and Panevova, Jarmila and Mikulová, Marie, url = http://hdl.handle.net/11858/00-097C-0000-0023-4338-F, note = LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, year 2014 exicon of Polish Multi-Word Expression. In *Proceedings of Language and Technology Conference (LTC'15), Poznań, Poland*. Wydawnictwo Poznańskie.

El Maarouf, I. and Oakes, M. (2015). Statistical Measures for Characterising MWEs. In *IC1207 COST PARSEME 5th general meeting*.

Erjavec, T., Fiser, D., Krek, S., and Ledinek, N. (2010). The JOS linguistically tagged corpus of Slovene. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, p. 1806–1809, Valletta, Malta. European Language Resources Association (ELRA).

Eryigit, G., Adali, K., Torunoglu-Selamet, D., Sulubacak, U., and Pamay, T. (2015). Annotation and Extraction of Multiword Expressions in Turkish Treebanks. In *Proceedings of NAACL-HLT 2015*, pp. 70–76. Association for Computationa @misc11858/00-097C-0000-0023-4338-F, title = PDT-Vallex: Czech Valency lexicon linked to treebanks, author = Urešová, Zdeňka and Štěpánek, Jan and Hajič, Jan and Panevova, Jarmila and Mikulová, Marie, url = http://hdl.handle.net/11858/00-097C-0000-0023-4338-F, note = LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, year 2014 l Linguistics.

Finkel, J. R. and Manning, C. D. (2009). Joint Parsing and Named Entity Recognition. In *HLT-NAACL*, pp. 326–334. The Association for Computational Linguistics.

Green, S., de Marneffe, M.-C., Bauer, J., and Manning, C. D. (2011). Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French. In *EMNLP*, pp. 725–735. ACL.

Green, S., de Marneffe, M.-C., and Manning, C. D. (2013). Parsing Models for Identifying Multiword Expressions. *Computational Linguistics*, **39**(1), 195–227.

Losnegaard, G. S., Sangati, F., Escartín, C. P., Savary, A., Bargmann, S., and Monti, J. (2016). Parseme survey on mwe resources. In N. C. C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Nasr, A., Ramisch, C., Deulofeu, J., and André, V. (2015). Joint Dependency Parsing and Multiword Expression Tokenisation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'15)*.

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

Przepiórkowski, A., Hajnicz, E., Patejuk, A., and Woliński, M. (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)*, pp. 83–91, Dublin, Ire-

land. Association for Computational Linguistics and Dublin City University.

Przepiórkowski, A., Hajič, J., Hajnicz, E., and Urešová, Z. (2016). Phraseology in two Slavic valency dictionaries: Limitations and perspectives. *International Journal of Lexicography*, **29**. Forthcoming.

Rosén, V., Losnegaard, G. S., De Smedt, K., Bejček, E., Savary, A., Przepiórkowski, A., Osenova, P., and Barbu Mitetelu, V. (2015). A survey of multiword expressions in treebanks. In *Proceedings of the 14th International Workshop on Treebanks & Linguistic Theories conference*, Warsaw, Poland.

Savary, A., Waszczuk, J., and Przepiórkowski, A. (2010). Towards the Annotation of Named Entities in the Polish National Corpus. In *Proceedings of LREC 10, Valletta, Malta*. European Language Resources Association.

Seraji, M., Jahani, C., Megyesi, B., and Nivre, J. (2014). A persian treebank with stanford typed dependencies. In N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Urešová, Z., Štěpánek, J., Hajič, J., Panevova, J., and Mikulová, M. (2014). PDT-vallex: Czech valency lexicon linked to treebanks. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Wehrli, E. (2014). The relevance of collocations for parsing. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pp. 26–32, Gothenburg, Sweden. Association for Computational Linguistics.

Wehrli, E., Seretan, V., and Nerima, L. (2010). Sentence analysis and collocation identification. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pp. 27–35, Beijing, China. Association for Computational Linguistics.

Świdziński, M. and Woliński, M. (2010). Towards a bank of constituent parse trees for Polish. In P. Sojka, A. Horák, I. Kopeček, and K. Pala, eds., *Text, Speech and Dialogue: 13th International Conference, TSD 2010, Brno, Czech Republic*, pp. 197–204, Heidelberg. Springer-Verlag.

Appendix A

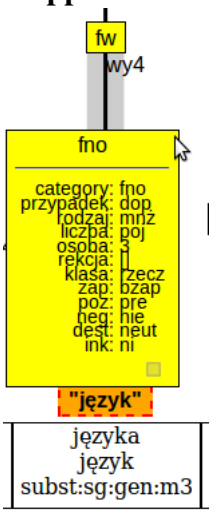

Figure 2: Feature structure of the node `fno` (*fraza nominalna* 'nominal phrase') dominating the terminal *język* 'tongue' in the syntax tree from Fig. 3. The feature codes include: `przypadek` 'case', `rodzaj` 'gender', `liczba` 'number', `osoba` 'person', `rekcja` 'case government', and `neg` 'negation'. The values denote: `fno` 'nominal phrase', `dop` 'genitive', `mnz` 'human inanimate', `poj` 'singular', and `nie` 'negated'.

Appendix B

Sample compositional reading occurrences in Składnica of verbal MWEs from Walenty

(2) Dobrze **mieć** takie jedno **zdanie** w swoim dorobku pisarskim.
'It is good to have one such sentence in one's writing outcome.'
(MWE: *to have a sentence*⇒'to have an opinion'

(3) Mała lampka rozpraszała mrok, **rzucając** nikłe **światło** na błękitną tapetę.
'The small lamp was dispelling the darkness, shedding weak light on the blue wallpaper.'
(MWE: *to shed light*)

(4) Nie **podał** Klossowi **ręki**, wskazał mu tylko krzesło.
'He did not give Kloss his hand, he just pointed at a chair.'
(MWE: *give someone a hand*⇒'help')

(5) Zrobiłem krok do przodu i **pociągnąłem** Dorę **za sobą**.
'I took a step forward and pulled Dora behind me.'
(MWE: *to pull someone behind oneself*⇒'to inspire someone so as to make them follow you')

Appendix C

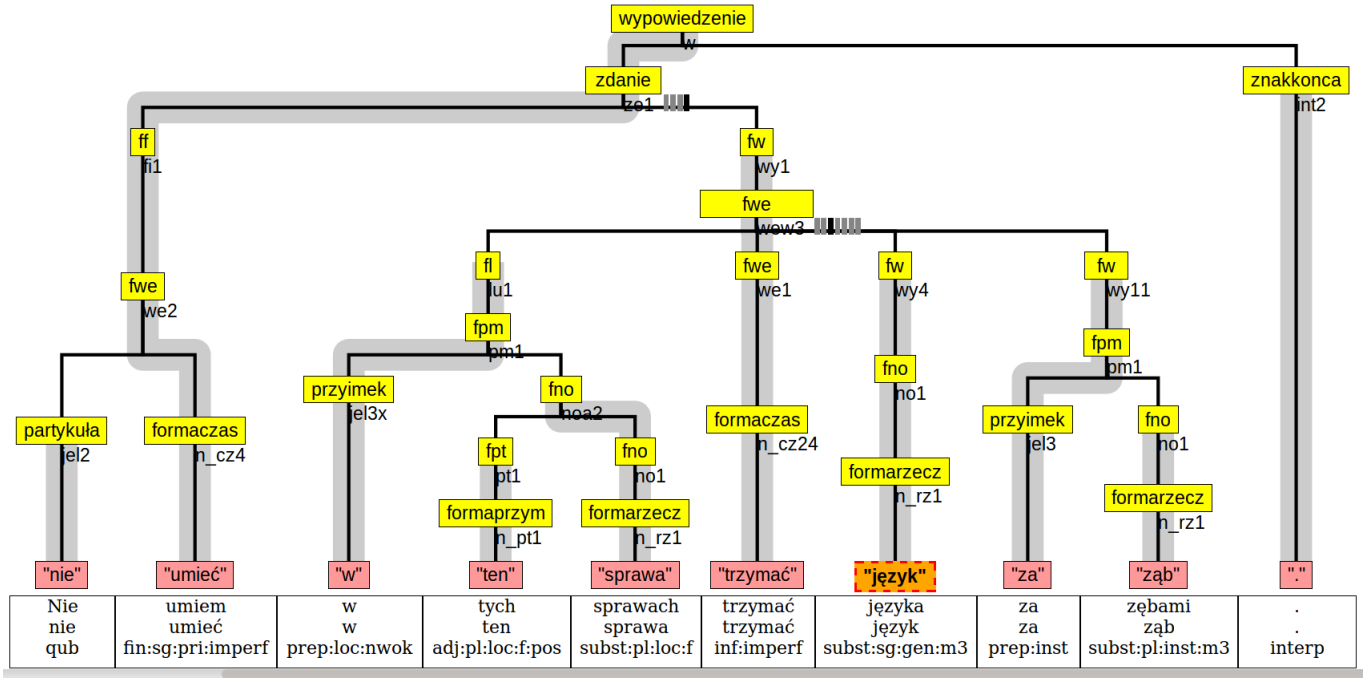

Figure 3: Syntax tree of example (1) in Składnica. The categories denote: `ff` 'finite phrase', `fl` 'adjunct', `fno` 'nominal phrase', `formaczas` 'verbal phrase', `formaprzym` 'adjectival phrase', `formarzecz` 'nominal phrase', `fpm` 'prepositional phrase', `fpt` 'adjectival phrase', `fw` 'required phrase', `fwe` 'verbal phrase', `partykuła` 'particle', `przyimek` 'preposition', `wypowiedzenie` 'utterance', `zdanie` 'sentence', `znakkońca` 'ending punctuation'. The categories `formaczas`, `formaprzym`, `formarzecz` seem redundant with `fwe`, `fpt` and `fno`, but they are distinguished since the do not appear in the original grammar used for pre-parsing the treebank.