# MWEs in Treebanks: From Survey to Guidelines

Victoria Rosén, Koenraad de Smedt, Gyri Smørdal Smørdal Losnegaard, Eduard Bejček, Agata Savary, Sofia Osenova

# MWEs in Treebanks: From Survey to Guidelines

**Victoria Rosén[1], Koenraad De Smedt[1], Gyri Smørdal Losnegaard[1], Eduard Bejček[2], Agata Savary[3] and Petya Osenova[4]**
[1]University of Bergen, [2]Charles University in Prague, [3]François Rabelais University of Tours,
[4]Sofia University St. Kl. Ohridski and Bulgarian Academy of Sciences
[1]Bergen, [2]Prague, [3]Tours, [4]Sofia
victoria@uib.no, desmedt@uib.no, gyri.losnegaard@uib.no,
bejcek@ufal.mff.cuni.cz, agata.savary@univ-tours.fr, petya@bultreebank.org

**Abstract**

By means of an online survey, we have investigated ways in which various types of multiword expressions are annotated in existing treebanks. The results indicate that there is considerable variation in treatments across treebanks and thereby also, to some extent, across languages and across theoretical frameworks. The comparison is focused on the annotation of light verb constructions and verbal idioms. The survey shows that the light verb constructions either get special annotations as such, or are treated as ordinary verbs, while VP idioms are handled through different strategies. Based on insights from our investigation, we propose some general guidelines for annotating multiword expressions in treebanks. The recommendations address the following application-based needs: distinguishing MWEs from similar but compositional constructions; searching distinct types of MWEs in treebanks; awareness of literal and nonliteral meanings; and normalization of the MWE representation. The cross-lingually and cross-theoretically focused survey is intended as an aid to accessing treebanks and an aid for further work in treebank annotation.

**Keywords:** treebanks, MWEs, multiword expressions

## 1. Introduction

PARSEME (PARSing and Multiword Expressions) is an interdisciplinary scientific network on the role of multiword expressions (MWEs) in parsing.[1] Working Group 4 (WG4) in PARSEME is concerned with the enhancement of MWE-aware methodologies of treebank construction. A goal for the working group is to propose annotation guidelines for representing MWEs in treebanks. As a first step toward creating such guidelines, WG4 has conducted a survey of existing MWE annotations in treebanks (Rosén et al., 2015). The survey is open-ended and information on additional treebanks is being added. Preliminary results of the survey show that there is considerable variation in the types of MWEs that are annotated in various treebanks, and also that the annotations for the same type of MWE can vary a lot depending on the language and the treebank type. Proposing detailed annotation guidelines is therefore a daunting task. As a first step towards recommendations, we propose some general principles for MWE annotations in treebanks.

## 2. The MWE Annotation Survey

The survey was conducted by asking WG4 members with knowledge about particular treebanks to fill in some information in a wiki.[2] The special purpose wiki was created in a Wikimedia-like framework, with a simple markup language and easy hyperlinking. It contains a table with a row for each treebank and columns for various types of MWEs, as shown in Figure 1. Each blue cell in the table is clickable and leads to an embedded information page.

[1] http://parseme.eu

[2] http://clarino.uib.no/iness/page?page-id=MWEs_in_Parseme

### The MWE Types

The table headers show the types of MWEs to be described. The typology chosen is syntactically based, and the types are among the most common ones described in the literature (Baldwin and Su Nam Kim, 2010; Sag et al., 2002). We distinguish between nominal MWEs, verbal MWEs, prepositional MWEs, adjectival MWEs, MWEs of other categories, and proverbs. Nominal MWEs are further divided into the subtypes multiword named entities, NN compounds, and other nominal MWEs, while verbal MWEs are subdivided into phrasal verbs, light verb constructions, VP idioms, and other verbal MWEs. The table headers are clickable; each one leads to an embedded page with further information about the types of MWE that belong in that column. For instance, clicking on *Phrasal verbs* opens a page with the added information that there are three types of phrasal verbs: particle verbs such as *show up*, verbs with selected prepositions such as *think of*, and verbs with both particles and selected prepositions such as *come up with*.

### The Treebanks

The first column in the table lists the treebanks in the survey. They are grouped in the table according to annotation type. Clicking on the name of a treebank brings up a *treebank description page* with basic information such as: name, author(s), linguistic formalism, license, links to documentation, history (how the treebank was constructed), whether it is static or dynamic, etc.

The first group in the table is the dependency treebanks, represented by the following: The Estonian Dependency Treebank (Muischnek et al., 2014), the Latvian Treebank (Pretkalnina and Rituma, 2012), the META-NORD Sofie Swedish Treebank (Losnegaard et al., 2013), the Prague Dependency Treebank for Czech (Bejček et al., 2013), the ssj500k Dependency Treebank for Slovene (Erjavec et al.,

| Treebank | Language | Annotation type | Nominal MWEs | | | Verbal MWEs | | |
|---|---|---|---|---|---|---|---|---|
| | | | Multiword named entities | NN compounds | Other nominal MWEs | Phrasal verbs | Light verb constructions | VP idioms |
| The Estonian Dependency Treebank | Estonian | dep | NO | N/A | NO | YES | NO | NO |
| The Latvian Treebank | Latvian | dep | YES | YES | NO | N/A | NO | NO |
| META-NORD Sofie Swedish Treebank | Swedish | dep | YES | N/A | NO | NO | NO | NO |
| The Prague Dependency Treebank | Czech | dep | YES | YES | YES | N/A | YES | YES |
| The ssj500k Dependency Treebank | Slovene | dep | YES | NO | NO | NO | NO | NO |
| The Szeged Dependency Treebank | Hungarian | dep | YES | NO | NO | YES | YES | NO |
| Universal Dependencies Treebanks | many languages | dep | YES | YES | YES | YES | YES | NO |
| The PENN Treebank | English | const | YES | YES | NO | YES | NO | NO |

Figure 1: Screenshot of the upper left corner of the survey table

2010), and the Szeged Dependency Treebank for Hungarian (Vincze et al., 2010). After the individual dependency treebanks, there is also a row for the Universal Dependency Treebanks.[3]

The second group in the table is the constituency treebanks: The National Corpus of Polish (Głowińska and Przepiórkowski, 2010; Savary et al., 2010), the PENN Treebank for English,[4] the SQUOIA Spanish Treebank,[5] the TIGER Treebank for German (Brants et al., 2004), and the UZH Alpine German Treebank.[6]

Finally, there are six treebanks that cannot simply be classified as either dependency or constituency treebanks. BulTreeBank for Bulgarian (Simov et al., 2005) offers both constituency and dependency analyses, as does the French Treebank (Abeillé et al., 2003). The analyses in the Lassy Small Treebank for Dutch (van Noord, 2009) are a cross between dependency and constituency trees. The CINTIL Treebanks for Portuguese (Branco et al., 2010) and DeepBank for English (Flickinger et al., 2012) are both based on Head Driven Phrase Structure Grammar (Pollard and Sag, 1994), whereas NorGramBank for Norwegian (Dyvik et al., 2016) is based on Lexical Functional Grammar (LFG) (Dalrymple, 2001).

[3] http://universaldependencies.org/

[4] http://www.cis.upenn.edu/~treebank/

[5] http://www.cl.uzh.ch/research/maschinelleuebersetzung/hybridmt_en.html

[6] http://www.cl.uzh.ch/research/parallelcorpora/paralleltreebanks/smultron_en.html

## The MWE Descriptions

The cells in the table can be filled out with different values. The value *N/A* (for 'not applicable') means that the MWE type does not occur in the language. For example, many languages do not have phrasal verbs (such as Latvian, Bulgarian, French and Portuguese). The value *NO* means that the MWE type occurs in the language but that the treebank lacks annotation for it. A clickable *YES* means that the MWE type is annotated in the treebank. In some cases a language has a MWE type that is not annotated as such, but the wiki authors wanted to show how the MWE type is analyzed compositionally; these are marked by a clickable *COMP*.

For each MWE type that is annotated in a treebank, there is a *MWE description page* with a detailed description of the MWE. Each MWE description page contains information about (1) the type of MWE and the treebank name, (2) an example sentence containing the MWE, with interlinear glosses and an idiomatic translation, (3) a graphic (screenshot or similar) with a visualization of the analysis, (4) a prose explanation of the analysis, and (5) a search expression for the MWE and, if necessary for complicated search expressions, a prose description of what the expression does. An example of a MWE description page is given in Figure 2.

Since the survey is still being expanded, some work is yet to be done. When a new treebank is to be added to the table, all cells are by default filled in with *TBC* (for 'to be completed'); this label is changed to *YES* once the MWE description page for that MWE type has been filled out.

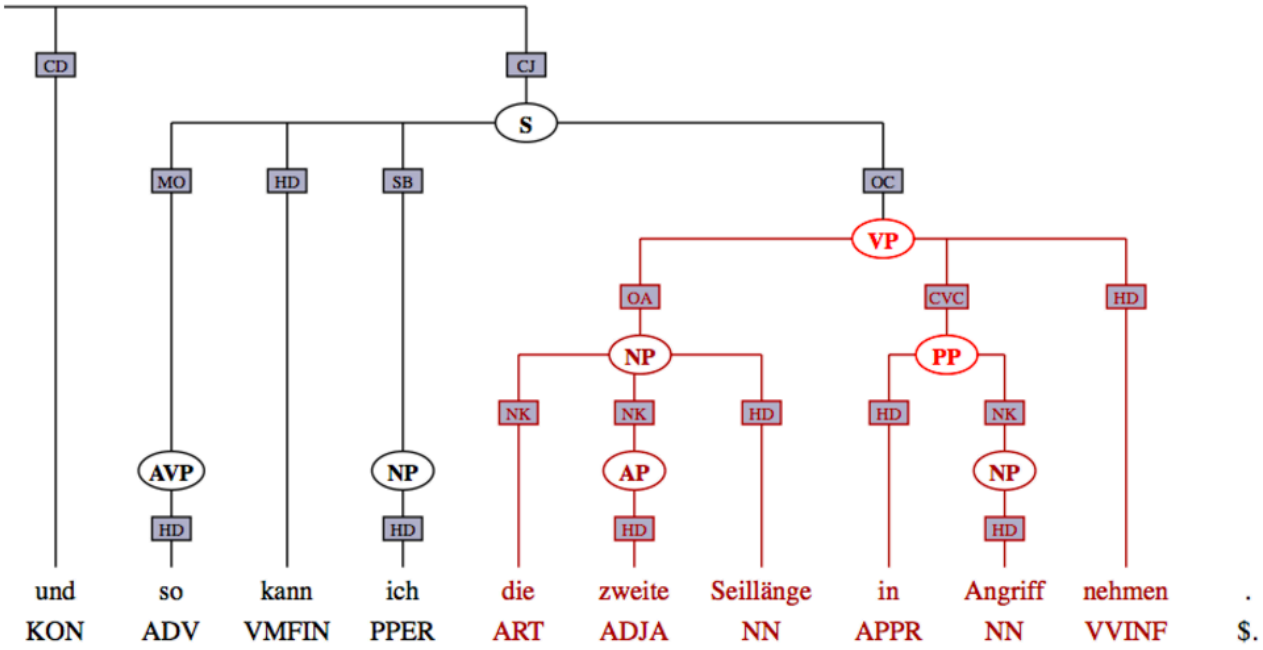
Light Verb Constructions (DE: Funktionsverbgefüge) in the UZH Alpine German Treebank

**Example**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ... | und | so | kann | ich | die | zweite | Seillänge | **in** | **Angriff** | **nehmen**. |
| ... | *and* | *so* | *can* | *I* | *the* | *second* | *rope_length* | *in* | *attack* | *take.* |

... and in this way I can tackle the second rope length.

**Analysis**




**About the analysis**

Following the German Negra/TIGER annotation guidelines we annotated light verb constructions by marking the NP or PP that goes with the light verb with the edge label CVC (collocational verb construction). Both the CVC-marked node and the verb must be children of the same VP constituent.

**Searching for light verb constructions**

Light Verb Constructions can be found by searching for the edge label CVC.

Figure 2: The MWE description page showing the annotation of light verb constructions in the UZH German Treebank

## 3. Survey Results

In an earlier study we examined the most commonly annotated MWE types in the treebanks in our survey: multiword named entities, prepositional MWEs and phrasal verbs (Rosén et al., 2015). We found that the annotations for prepositional MWEs and phrasal verbs shared many common properties, also across frameworks, whereas the annotations for multiword named entities were more diverse.

In the present study we have compared the analyses for light verb constructions and VP idioms, two important types of syntactically flexible constructions (Sag et al., 2002, p. 6–7). Based on the results, we have made some generalizations and recommendations for good practice in the annotation of MWEs in treebanks.

Light verb constructions involve a semantically bleached verb usually combined with an indefinite noun phrase, for example *make a wish*, *take a shower*, *have a nap*, or other kinds of phrases such as the prepositional phrase *in Angriff* in Figure 2. Typical light verbs in English are *do*, *give*, *have*, *make*, and *take*. The light verb contributes little to the meaning of the construction, which can often be paraphrased with a verbal form of the noun, as in *shower* rather than *take a shower*, *nap* rather than *take a nap*, etc. (Baldwin and Su Nam Kim, 2010, p. 277). It is, however, unclear which verbs in languages other than English should be considered light verbs. It is also not obvious how to delimit the class of light verbs in any single language.

VP idioms can be quite similar to light verb constructions

in that they can be composed of a verb plus a noun complement, but they are not restricted to a small set of verbs. They also differ in that their semantics is harder to predict from the combination of verb and noun. They may have only a noun complement, as in *shoot the breeze*, but additional constituents are also possible, as in *let the cat out of the bag*. Intransitive VP idioms, such as *go out on a limb*, also occur.

There are seven treebank rows in the table which indicate *YES* for light verb constructions and/or VP idioms. In the following we will examine the annotation of these constructions in six of these treebanks. (Since the analysis of light verb constructions is identical in the TIGER Treebank and the UZH Alpine Treebank, we show only the latter.)

### The UZH Alpine German Treebank

In the UZH Alpine German Treebank both light verb constructions and VP idioms are annotated. An example of a light verb construction for the phrase in (1) is shown in Figure 2. The annotation involves marking the NP or PP that goes with the light verb with the edge label `CVC` (for *collocational verb construction*). Both the `CVC` node and the verb must be children of the same VP node.

(1) *in Angriff nehmen*
in attack take
'tackle'

(2) *vergeht im Fluge*
passes in the flight
'flies by quickly'

VP idioms such as the phrase in (2) are annotated in the same way. Again the NP or PP that goes with the verb is annotated with `CVC`, and both this node and the verb must be children of the same VP constituent. However, in the VP idiom construction, there is no light verb. This means that VP idioms can only be distinguished from light verb constructions by excluding semantically light verbs such as *nehmen* 'take', *setzen* 'set', and *stellen* 'put' from the search results. In order for such a search to be successful, an exhaustive list of light verbs would be necessary, as well as a query language that allows the use of negation.

### The Prague Dependency Treebank

The Prague Dependency Treebank annotates both light verb constructions and VP idioms. The `CPHR` relation (for *Compound PHRase*) is used for light verb constructions. Figure 3 illustrates the analysis of the light verb construction in example (3). The light verb is marked by a green node in the dependency graph, while the predicative noun is marked by an orange node.

(3) *Výše sazby byla různá podle*
Amount of rate was different according to
*doby, kdy byla smlouva s klientem*
period when was the treaty with client
*uzavřena*.
concluded
'The amount of the rate differed with the period, when the treaty was concluded with a client.'

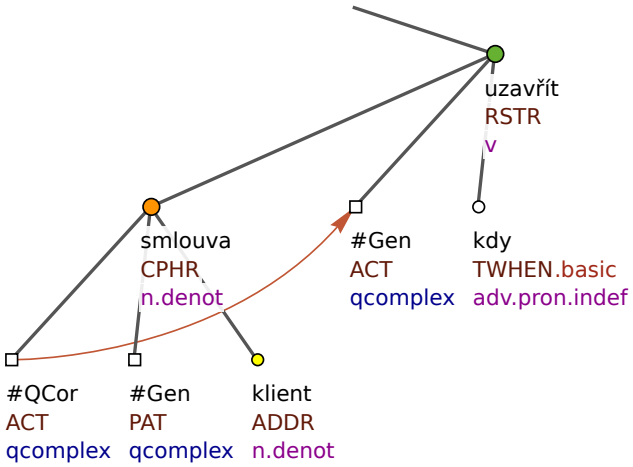



Figure 3: Example of the analysis of a light verb construction in the Prague Dependency Treebank

VP idioms are annotated in a similar way, but with the `DPHR` relation (for *Dependency part of a PHRase*). If the phrase is longer than a verb and one content word, the `DPHR` node represents all of them and its lemma is formed by those words joined together with underscores (e.g. `klacky_pod_nohy` in *házet klacky pod nohy*, literally 'to throw sticks under feet', meaning 'to crimp'). Both constructions can be easily found by searching for a verb with a dependent `CPHR` or `DPHR`.

### The Persian UD v1.2 Treebank

The UD annotation scheme uses the dependency relation `compound:lvc` for annotating light verbs in certain languages including Persian (Farsi), which frequently uses this construction. Figure 4 for example (4) shows that the light verb می‌کنند 'do.3PL' and the nominal part of the construction تعدیل 'dampening' are related in this way. While this annotation makes searching for light verb constructions straightforward, not all UD treebanks for other languages that have light verbs use the same relation.

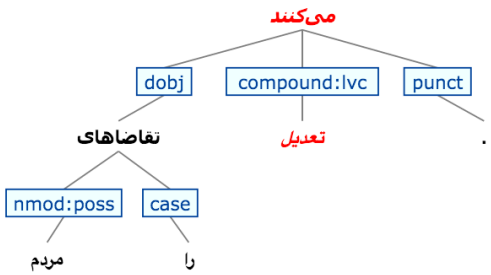



Figure 4: Example of the analysis of a light verb construction in the Persian UD v1.2 Treebank

(4) تقاضاهای مردم را تعدیل می‌کنند
demands people CASE dampening do.3PL
'They dampen people's demands.'

### The Szeged Dependency Treebank

The Szeged Dependency Treebank for Hungarian uses the `OBJ-LVC` relation for light verb constructions. Figure 5 illustrates the analysis of the light verb construction in example (5). The `OBJ-LVC` relation goes from the light verb *hoznunk* 'bring-INF-1PL' to the noun *döntést* 'decision-ACC'.

(5) *Holnap nagyon fontos döntést*
tomorrow very important decision-ACC
*kell hoznunk*.
must bring-INF-1PL
'Tomorrow we will have to make a very important decision.'

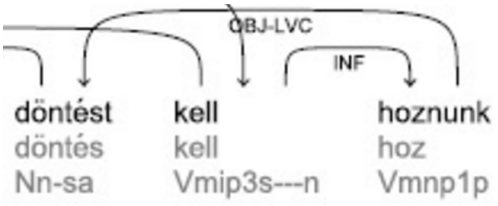



Figure 5: Example of the analysis of a light verb construction in the Szeged Dependency Treebank

### The French Treebank

The French Treebank annotates VP idioms, but not light verb constructions. The example of a VP idiom provided in the survey is the sentence in (6). The French expression *avoir lieu*, literally 'have place', means 'take place', 'occur'. Its analysis is shown in Figure 6. The head of the sentence is *eu*, the past participle of *avoir*, and the dependency `dep_cpd` links *lieu* to the first word in the MWE.

(6) *La réforme n' a pas encore eu lieu*.
the reform . has not yet had place
'The reform has not taken place yet.'

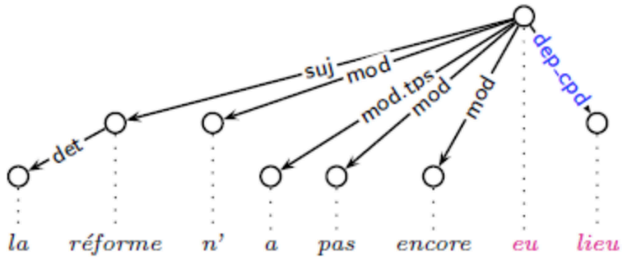



Figure 6: Example of the analysis of a VP idiom in the French Treebank

### NorGramBank

NorGramBank annotates light verb constructions compositionally. The sentence in (7) shows a construction with the light verb *ta* 'take', which is simply analyzed as taking the NP *en rask avgjørelse* 'a quick decision' as a direct object.

(7) *Eva tok en rask avgjørelse*.
Eva took a quick decision
'Eva made a quick decision.'

(8) *Et nytt mareritt fant sted hos Kitty*.
a new nightmare found place by Kitty
'A new nightmare occurred at Kitty's.'

VP idioms are annotated as MWEs in NorGramBank. In the sentence in (8) the VP idiom *finne sted*, literally 'find place', means 'take place', 'occur'. The analysis is shown in Figures 7 and 8. In the c-structure the verb and its complement are analyzed in the same way as for a non-idiomatic construction. The f-structure shows that the predicate of the sentence is `finne#sted`. The object argument `sted` is incorporated into the predicate, which is represented by the PRED feature in the f-structure. This argument is outside the angled brackets in the predicate argument list, meaning that it is not a semantic argument of the predicate. VP idioms may be searched for by searching for the character #, since this character is only used in the PRED values of VP idioms.

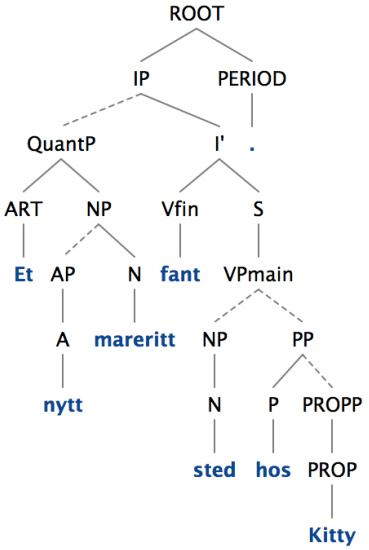



Figure 7: Example of the c-structure analysis of a VP idiom in NorGramBank

### Comparison of the Annotations

The treebanks that have special annotations for light verb constructions all analyze them in similar ways. The Prague Dependency Treebank uses the `CPHR` relation, the Szeged Dependency Treebank uses the `OBJ-LVC` relation, and the Persian UD v1.2 Treebank uses the `compound:lvc` relation. These are all simply different names for a dependency relation going from the light verb to the noun. In the UZH Alpine German Treebank, the `CVC` relation marks the NP or PP that goes with the light verb.

There is more variation in the VP idiom annotations. There are two dependency treebanks that have VP idiom annotations. The Prague Dependency Treebank uses a special `DPHR` relation for VP idioms. The French Treebank uses the `dep_cpd` relation, but this relation is actually used for all

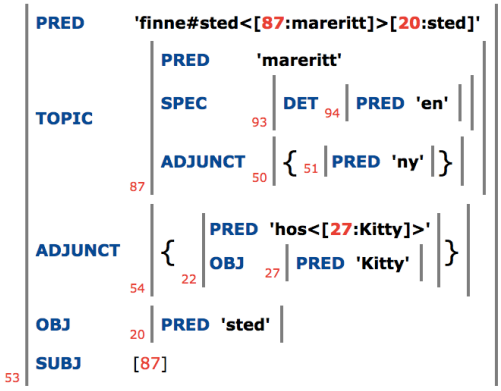



Figure 8: Example of the f-structure analysis of a VP idiom in NorGramBank

MWEs in this treebank, making it difficult to distinguish between VP idioms and other MWEs. A similar situation obtains with the UZH Alpine German Treebank, since the analysis of VP idioms is the same as the light verb analysis. NorGramBank provides an analysis for VP idioms where the complement of the verb is integrated into the verbal predicate, thus reducing the syntactic valency of the verb.
It is interesting to note that the light verb example for Czech, to conclude a treaty, does not correspond to the typical examples of light verbs in the literature for English. How the class of light verbs should be delimited crosslinguistically is unclear. VP idioms have meanings that cannot be derived compositionally, but it is not obvious that the same is true of light verb constructions. An example of a light verb construction from (Baldwin and Su Nam Kim, 2010) is *make amends*. Although the verb is one of the prototypical light verbs, it's not clear that it is used as a light verb in this construction. There is no paraphrase with the verb *amend*, and it seems to be as idiomatic as the above-mentioned French *avoir lieu* and Norwegian *finne sted*, both meaning 'take place'.

## 4. Towards Annotation Guidelines

We have shown that there are a various ways of treating MWEs in treebanks for different languages and treebank types. Given this situation, and the lack of full agreement on even what constitutes a MWE, how might it be possible to make guidelines for the annotation of MWEs? Although specific guidelines will need to be tuned to the treebank annotation type, we would like to formulate some general principles that might hold for all treebanks and languages. For linguistic research, as well as for the development of some language technology applications, it is important to be able to perform targeted searches for MWEs in treebanks. We argue that the following desiderata are beneficial to effective treebank search:

A. MWEs should be annotated as such, so that treebank queries can directly target them.

B. The annotation of noncompositional MWEs should distinguish them from homonymous strings with a compositional analysis.

C. Individual MWEs should be searchable even if they are discontinuous or variable in form.

D. It should be possible to search for various *types* of MWEs based on their characteristics.

**Principle A** is a general principle that aims at improving the ease with which MWEs can be identified in treebanks, without the need to be detected by heuristics. The recursive case of this principle is that MWEs which occur as part of other MWEs should also be annotated as such, so that embeddings of MWEs (e.g. in the complex name *Johann Wolfgang Goethe-Universität Frankfurt am Main*) can be discovered.
**Principle B** is a corollary: ease of identification implies that MWEs should be distinguished from homonymous constructions which are compositional.
For example, *under the knife* is an English idiom meaning "undergoing surgery". This idiom, illustrated in example (9), should be annotated in a way which distinguishes it from the compositional meaning in (10).

(9) The patient is *under the knife*.

(10) The napkin is *under the knife*.

The principle of marking the distinction should not prevent a treebank from having different levels, among which one may provide the same 'regular' syntactic analysis for examples (9) and (10).
**Principle C** will allow identification of non-fixed MWEs irrespective of their surface forms and word orders. For instance, the morphological and word order variants of the particle verb *shut down* in examples (11) and (12) should be searchable with a single query.

(11) The company is *shutting down* the power plant.

(12) The company has *shut* the power plant *down*.

In order to fulfill principle C, some normalization is recommended, i.e. each MWE occurrence in a corpus should be associated with its canonical form so as to conflate different morphosyntactic variants of the same MWE. In the simplest case a canonical form is a MWE lemma, e.g. *man servant* for *men servants*. Linking to a lexicon or knowledge base of MWEs, e.g. DuELME (Grégoire, 2010; Odijk, 2013) or dictionary storage for pre-annotation (Bejček and Straňák, 2010) should be considered. To the extent that a treebank is a parsed corpus, this should normally be achieved by having appropriate MWE entries in the lexicon used in parsing, as is the case in NorGramBank. Automatic lemmatization of MWEs is non-trivial in the general case, since components of a MWE lemma may not be lemmas themselves, as in *to spill the beans* but not *to spill the bean*. In highly inflected languages, automatic lemmatization of some MWE categories, such as person names, may be challenging (Piskorski et al., 2007); therefore assigning manually validated lemmas to named entities in a treebank may be an option (Savary et al., 2010).

**Principle D** implies that, to the extent possible and depending on the MWE ontology, all MWEs belonging to certain types will be retrievable as a set, for instance, all fixed expressions, all particle verb constructions or all VP idioms. The different types should not necessarily be annotated at the same level of linguistic analysis. Some may be annotated at word level, such as fixed expression (so-called words with spaces), some at one or more levels of syntactic structure (such as c-structure and f-structure, or analytical and tectogrammatical structure).

## 5. Conclusion and Future Work

The survey of MWE annotations in treebanks reported on in this paper is a useful tool for comparing how these expressions are treated in different languages and in treebanks of different types. The survey is open-ended and can accommodate entries for additional languages and treebanks during the PARSEME Action, which will continue until the spring of 2017. Based on our findings, we have suggested some general principles which may be helpful for a range of studies and applications which need to take into account the special status of MWEs. Nevertheless, these principles should be adopted or adapted in the light of the particular purpose of the treebank. On the basis of the general principles, WG4 will continue to work on developing more specific annotation guidelines that are attuned to the annotations in different types of treebanks.

## 6. Acknowledgments

This work has been supported in part by the PARSEME European COST Action (IC1207), the COST CZ grant of the MEYS of the Czech Republic (LD14117), and a grant to the INESS project by the Research Council of Norway. We thank Behrang QasemiZadeh for information on Persian.

## References

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for French. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, volume 20 of *Text, speech and language technology*. Kluwer Academic Publishers, Dordrecht.

Baldwin, T. and Su Nam Kim. (2010). Multiword expressions. In Nitin Indurkhya et al., editors, *Handbook of Natural Language Processing*, chapter 12. CRC Press, Boca Raton, FL, USA, 2nd edition.

Bejček, E. and Straňák, P. (2010). Annotation of multiword expressions in the prague dependency treebank. *Language Resources and Evaluation*, 44(1):7–21.

Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Šárka Zikánová. (2013). Prague Dependency Treebank 3.0. Data. Data, `http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3`.

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*.

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Dalrymple, M. (2001). *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics*. Academic Press, San Diego, CA.

Dyvik, H., Meurer, P., Rosén, V., De Smedt, K., Haugereid, P., Losnegaard, G. S., Lyse, G. I., and Thunes, M. (2016). NorGramBank: A 'Deep' Treebank for Norwegian. In *Proceedings of LREC 2016*, Portorož, Slovenia, May. ELRA.

Erjavec, T., Fiser, D., Krek, S., and Ledinek, N. (2010). The JOS linguistically tagged corpus of Slovene. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, page 1806–1809, Valletta, Malta, May. European Language Resources Association (ELRA).

Flickinger, D., Zhang, Y., and Kordoni, V. (2012). Deepbank: A dynamically annotated treebank of the Wall Street Journal. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories*, pages 85–96.

Grégoire, N. (2010). DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44(1):23–39.

Głowińska, K. and Przepiórkowski, A. (2010). The Design of Syntactic Annotation Levels in the National Corpus of Polish. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Losnegaard, G. S., Lyse, G. I., Gjesdal, A. M., De Smedt, K., Meurer, P., and Rosén, V. (2013). Linking Northern European infrastructures for improving the accessibility and documentation of complex resources. In Koenraad De Smedt, et al., editors, *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013, May 22–24, 2013, Oslo, Norway. NEALT Proceedings Series 20*, number 89 in Linköping Electronic Conference Proceedings, pages 44–59. Linköping University Electronic Press.

Muischnek, K., Müürisep, K., Puolakainen, T., Aedmaa, E., Kirt, R., and Särg, D. (2014). Estonian dependency treebank and its annotation scheme. In *Proceedings of 13th Workshop on Treebanks and Linguistic Theories (TLT13)*, pages 285–291.

Odijk, J. (2013). DUELME: Dutch electronic lexicon of multiword expressions. In Gil Francopoulo, editor, *LMF - Lexical Markup Framework*, page 133–144. ISTE/Wiley, London/Hoboken.

Piskorski, J., Sydow, M., and Kupść, A. (2007). Lemmatization of Polish Person Names. In *ACL 2007. Proceedings of the Workshop on Balto-Slavonic NLP 2007*, pages 27–34. Association for Computational Linguistics.

Pollard, C. and Sag, I. A. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

Pretkalnina, L. and Rituma, L. (2012). Syntactic issues identified developing the Latvian treebank. In *Baltic HLT*, pages 185–192.

Rosén, V., Losnegaard, G. S., De Smedt, K., Bejček, E., Savary, A., Przepiórkowski, A., Osenova, P., and Barbu Mititelu, V. (2015). A survey of multiword expressions in treebanks. In Markus Dickinson, et al., editors, *Proceedings of the Fourteenth Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 179–193, Warsaw, Poland. Institute of Computer Science, Polish Academy of Sciences.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Lecture Notes in Computer Science. Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, volume 2276, pages 189–206. Springer.

Savary, A., Waszczuk, J., and Przepiórkowski, A. (2010). Towards the Annotation of Named Entities in the Polish National Corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association, 17-23 May.

Simov, K., Osenova, P., Simov, A., and Kouylekov, M. (2005). Design and implementation of the Bulgarian HPSG-based treebank. *Journal of Research on Language and Computation*, Special Issue:495–522.

van Noord, G. (2009). Huge parsed corpora in LASSY. In Frank Van Eynde, et al., editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT7)*, pages 115–126. LOT.

Vincze, V., Szauter, D., Almási, A., Móra, G., Alexin, Z., and Csirik, J. (2010). Hungarian dependency treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 1855–1862. ELRA.