A Survey of Multiword Expressions in Treebanks

Victoria Rosén,[1] Gyri Smørdal Losnegaard,[1]
Koenraad De Smedt,[1] Eduard Bejček,[2] Agata Savary,[3]
Adam Przepiórkowski,[4,5] Petya Osenova[6,7] and Verginica Barbu Mititelu[8]

[1]University of Bergen, [2]Charles University in Prague,
[3]François Rabelais University of Tours,
[4]Institute of Computer Science, Polish Academy of Sciences,
[5]University of Warsaw, [6]Sofia University St. Kl. Ohridski,
[7]Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences,
[8]Romanian Academy Research Institute
for Artificial Intelligence "Mihai Drăgănescu"

E-mail: {victoria|gyri.losnegaard|desmedt}@uib.no,
bejcek@ufal.mff.cuni.cz, agata.savary@univ-tours.fr,
adamp@ipipan.waw.pl, petya@bultreebank.org, vergi@racai.ro

Abstract

We present the methodology and results of a survey on the annotation of multiword expressions in treebanks. The survey was conducted using a wiki-like website filled out by people knowledgeable about various treebanks. The survey results were studied with a comparative focus on prepositional MWEs, verb-particle constructions and multiword named entities.

1 Introduction

There is currently little agreement on how multiword expressions (MWEs) should be annotated in treebanks, and there is, in fact, not even agreement on what constitutes a MWE in NLP. This makes it difficult to study and exploit MWEs in language resources, including treebanks.

PARSEME[1] is a COST Action dedicated to the study of MWEs. PARSEME's working group 4 is concerned with the annotation of MWEs in treebanks. One of the intended outcomes of this working group is to make recommendations for common principles and guidelines for annotating MWEs in treebanks. As a step towards making such recommendations, we have made a survey of the ways in which

[1] http://www.parseme.eu/

different types of MWEs are currently annotated in a variety of treebanks. This survey was performed by asking people knowledgeable about particular treebanks to describe the annotation of different types of MWEs by filling out an online form. It has not been the goal of the present study to check to what extent the principles and guidelines for each treebank have been followed.

The paper is structured as follows: In section 2 the methodology of gathering and summarizing data is presented. Section 3 presents a summary of preliminary findings for three MWE types. Section 4 concludes the paper.

2 Methodology

A structured survey form was set up by establishing a wiki with editable pages written in a Wikimedia-like framework and featuring a simple markup language and easy hyperlinking. The main page of the wiki contains a table which we will call the 'survey table' and which is shown in Figure 1. The main page also presents detailed instructions for entering information.

There is a row in the survey table for each treebank for which information has been collected. The row name (in the first column of the table) is the name of the treebank. The second column contains the language, and the third the annotation type of the treebank. The remaining columns are for MWE types. All cells with blue in the survey table are clickable and lead to embedded information pages.[2] The next sections present the elements in the table in more detail.

2.1 The treebanks

The survey is open-ended and will continue to be updated with information about different treebanks until the end of the PARSEME action in the spring of 2017. Currently, information has been gathered about 17 treebanks for 15 languages. The two main types are dependency and constituency treebanks.

The dependency treebanks are (the language is shown in parentheses when it is not included in the name of the treebank):

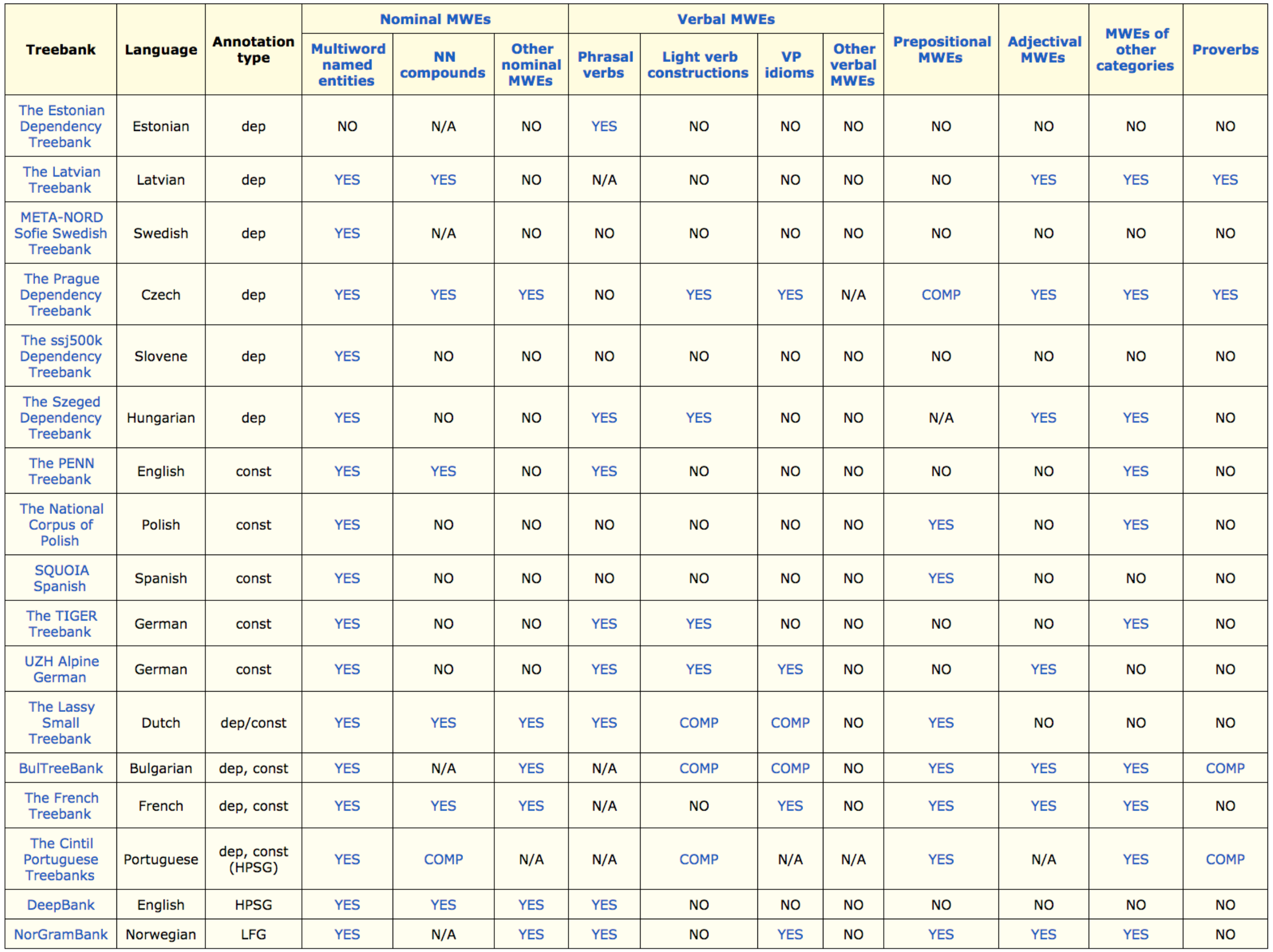
- The Estonian Dependency Treebank [11]
- The Latvian Treebank [13]
- The META-NORD Sofie Swedish Treebank [10]
- The Prague Dependency Treebank (Czech) [3]
- The ssj500k Dependency Treebank (Slovene) [7]
- The Szeged Dependency Treebank (Hungarian) [18]

The constituency treebanks include:

- The National Corpus of Polish [9, 15]
- The PENN Treebank (English)[3]

[2] For the online version, see `http://clarino.uib.no/iness/page?page-id=MWEs_in_Parseme`

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



Treebank	Language	Annotation type	Nominal MWEs			Verbal MWEs				Prepositional MWEs	Adjectival MWEs	MWEs of other categories	Proverbs
			Multiword named entities	NN compounds	Other nominal MWEs	Phrasal verbs	Light verb constructions	VP idioms	Other verbal MWEs				
The Estonian Dependency Treebank	Estonian	dep	NO	N/A	NO	YES	NO	NO	NO	NO	NO	NO	NO
The Latvian Treebank	Latvian	dep	YES	YES	NO	N/A	NO	NO	NO	NO	YES	YES	YES
META-NORD Sofie Swedish Treebank	Swedish	dep	YES	N/A	NO	NO	NO	NO	NO	NO	NO	NO	NO
The Prague Dependency Treebank	Czech	dep	YES	YES	YES	NO	YES	YES	N/A	COMP	YES	YES	YES
The ssj500k Dependency Treebank	Slovene	dep	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
The Szeged Dependency Treebank	Hungarian	dep	YES	NO	NO	YES	YES	NO	NO	N/A	YES	YES	NO
The PENN Treebank	English	const	YES	YES	NO	YES	NO	NO	NO	NO	NO	YES	NO
The National Corpus of Polish	Polish	const	YES	NO	NO	NO	NO	NO	NO	YES	NO	YES	NO
SQUOIA Spanish	Spanish	const	YES	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
The TIGER Treebank	German	const	YES	NO	NO	YES	YES	NO	NO	NO	NO	YES	NO
UZH Alpine German	German	const	YES	NO	NO	YES	YES	YES	NO	NO	YES	NO	NO
The Lassy Small Treebank	Dutch	dep/const	YES	YES	YES	YES	COMP	COMP	NO	YES	NO	NO	NO
BulTreeBank	Bulgarian	dep, const	YES	N/A	YES	N/A	COMP	COMP	NO	YES	YES	YES	COMP
The French Treebank	French	dep, const	YES	YES	YES	N/A	NO	YES	NO	YES	YES	YES	NO
The Cintil Portuguese Treebanks	Portuguese	dep, const (HPSG)	YES	COMP	N/A	N/A	COMP	N/A	N/A	YES	N/A	YES	COMP
DeepBank	English	HPSG	YES	YES	YES	YES	NO	NO	NO	NO	NO	NO	NO
NorGramBank	Norwegian	LFG	YES	N/A	YES	YES	NO	YES	NO	YES	YES	YES	NO

Figure 1: The survey table

- The SQUOIA Spanish Treebank[4]
- The TIGER Treebank (German) [5]
- The UZH Alpine German Treebank[5]

There are six treebanks which cannot be classified simply as either dependency or constituency treebanks. These are:

- BulTreeBank (Bulgarian) [16]
- The French Treebank [1]
- The Lassy Small Treebank (Dutch) [17]
- The CINTIL Treebanks (Portuguese) [4]
- DeepBank (English) [8]
- NorGramBank (Norwegian)[6]

BulTreeBank and the French Treebank offer both constituency and dependency analyses. The Lassy Small Treebank has analyses that are a cross between constituency and dependency graphs. The CINTIL Treebanks and DeepBank are both based on Head Driven Phrase Structure Grammar (HPSG) [12], while NorGramBank is based on Lexical Functional Grammar (LFG) [6].

Clicking on the treebank name (in the first column of the table) brings up a 'treebank description page'. Here information is given such as name, author, formalism, license, links to documentation, history (how the treebank was constructed), whether it is static or dynamic, etc.

2.2 The MWE types

The table headers show the types of MWEs described:

- Nominal MWEs
  - Multiword named entities
  - NN compounds
  - Other nominal MWEs
- Verbal MWEs
  - Phrasal verbs
  - Light verb constructions
  - VP idioms
  - Other verbal MWEs
- Prepositional MWEs
- Adjectival MWEs
- MWEs of other categories
- Proverbs

This typology was based on a discussion of more or less accepted types described in the literature [2, 14], taking into account the trade-off between offering major types as a guidance and allowing other types and subtypes that are found in

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

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

[6] http://clarino.uib.no/iness/

some treebanks. Clicking on a column header for a MWE type opens up a ‘MWE type description page’.

2.3 MWE information cells and MWE description pages

Each cell in a MWE type column has one of the following values:

- N/A (for ‘not applicable’): the MWE type does not occur in the language
- NO: the MWE type occurs in the language but the treebank lacks annotation for it
- YES: the MWE type is annotated in the treebank
- COMP: the MWE type is not annotated as such, but is analyzed compositionally

Clicking on the value YES or COMP brings up a ‘MWE example page’ with a detailed description of one or more examples of the MWE type in a particular treebank. Each MWE example page contains the following information (for each example):

- The type of MWE and the treebank name
- An example sentence containing the MWE, with interlinear glosses and an idiomatic translation
- A graphic (screenshot or similar) with a visualization of the analysis
- A prose explanation of the analysis
- A search expression for the MWE and a prose description of what the expression does

By way of illustration, the MWE example page for prepositional MWEs in NorGramBank is given in Figure 2.

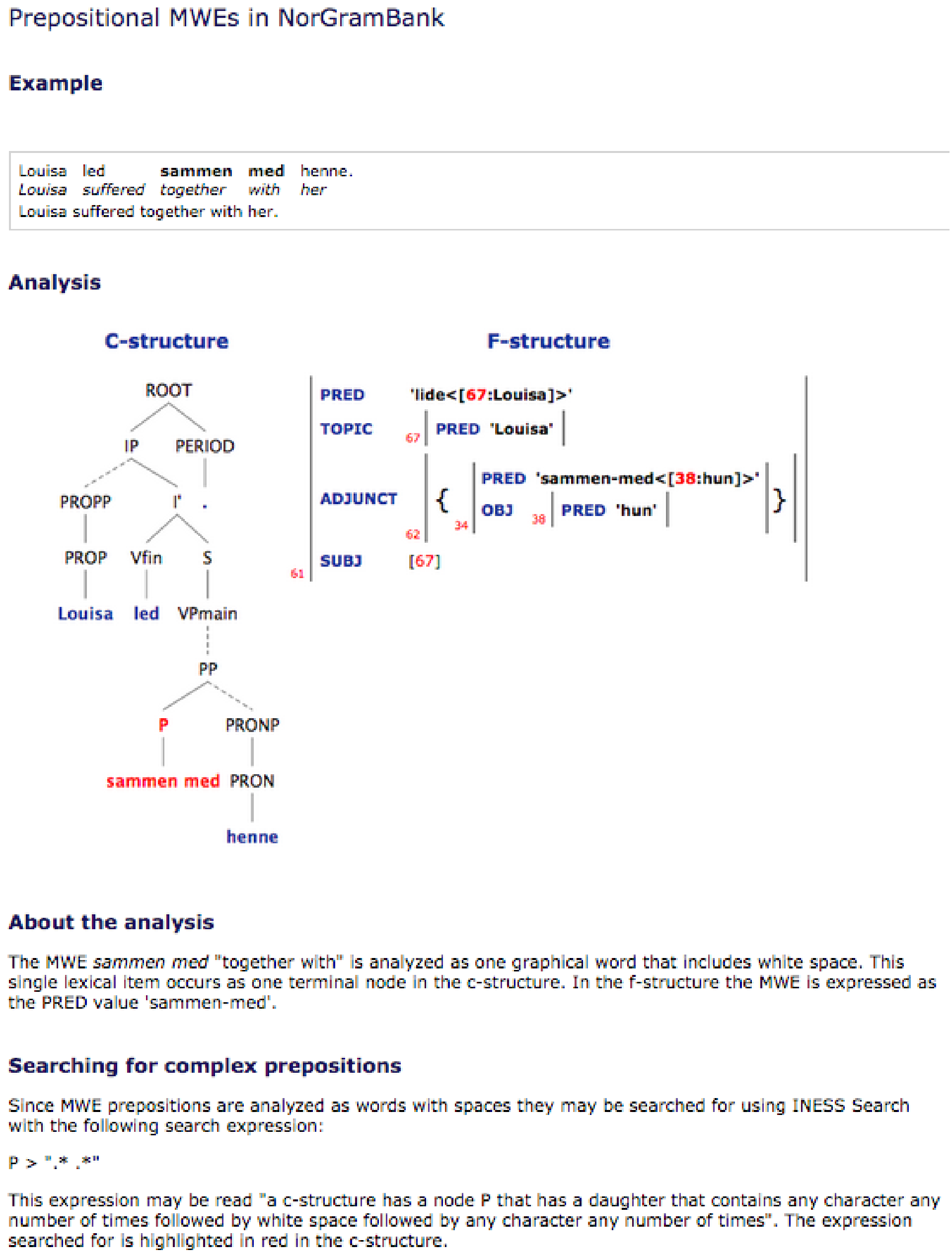
3 Results and discussion

The survey allows comparison of many different types of MWEs along several dimensions. Within the confines of the present paper, we will focus on comparisons for three of the most commonly annotated types of MWEs. Table 1 shows the number of MWEs of various types that are annotated in the survey.

3.1 Prepositional MWEs

Prepositional MWEs are often fixed expressions in Sag et al.’s terminology. Since fixed expressions are lexicalized and do not undergo morphosyntactic variation or internal modification, they can be handled with a words-with-spaces approach [14, p. 192].

Prepositional MWEs are annotated in somewhat different ways in the treebanks in our survey, as illustrated in Figure 3. BulTreeBank and NorGramBank treat them literally as words with spaces, in other words as single graphical words that include white space. The Bulgarian MWE Благодарение на “thanks to” is a terminal node in the tree dominated by *Prep*, while the Norwegian *sammen med* “together



Prepositional MWEs in NorGramBank

Example

Louisa led **sammen** **med** henne.
Louisa suffered together with her
Louisa suffered together with her.

Analysis

About the analysis

The MWE *sammen med* "together with" is analyzed as one graphical word that includes white space. This single lexical item occurs as one terminal node in the c-structure. In the f-structure the MWE is expressed as the PRED value 'sammen-med'.

Searching for complex prepositions

Since MWE prepositions are analyzed as words with spaces they may be searched for using INESS Search with the following search expression:

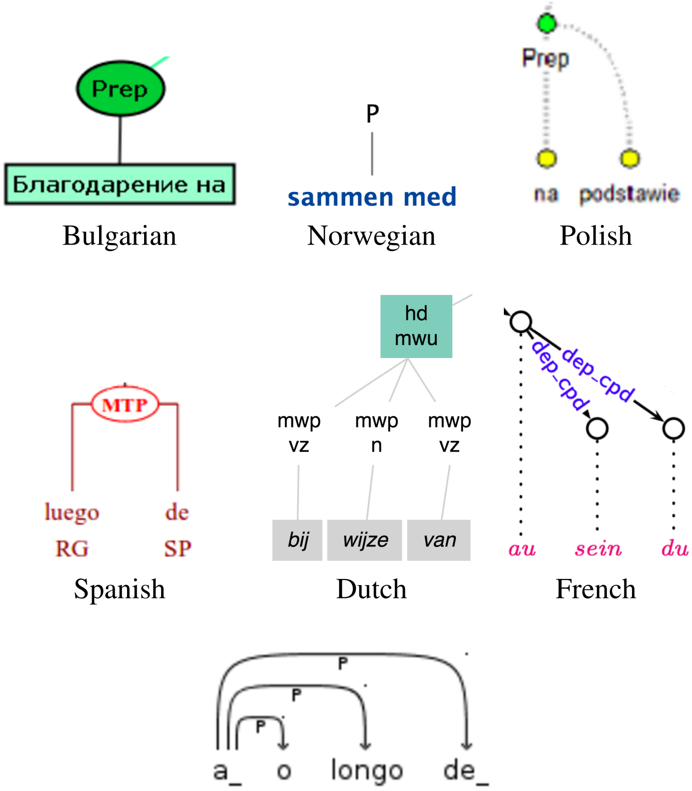
```
P > ".* .*"
```

This expression may be read "a c-structure has a node P that has a daughter that contains any character any number of times followed by white space followed by any character any number of times". The expression searched for is highlighted in red in the c-structure.

Figure 2: MWE example page for prepositional MWEs in NorGramBank

Nominal MWEs	Multiword named entities	16
	NN compounds	6 (+1)
	Others	6
Verbal MWEs	Phrasal verbs	8
	Light verb constructions	4 (+3)
	VP idioms	4 (+2)
	Others	0
Prepositional MWEs		7 (+1)
Adjectival MWEs		7
MWEs of other categories		10
Proverbs		2 (+2)

Table 1: Number of treebanks (out of all 17 treebanks in the survey) with annotations for the different MWE types, with the number of compositional analyses given in parentheses



Bulgarian

Norwegian

Polish

Spanish

Dutch

French

Portuguese

Figure 3: Overview of the annotations of prepositional MWEs in seven treebanks

with" is a terminal dominated by *P*. The National Corpus of Polish has a multi-layer annotation, not all of which is shown in the example. Parts of speech are assigned to individual components of a MWE preposition in the morphosyntactic annotation layer (*na* is a preposition and *podstawie* is a noun), and these components are joined into one unit (of type *Prep*) in the syntactic word layer. The SQUOIA Spanish treebank provides a phrasal analysis of the MWE *luego de* "after of", using a special node label *MTP*, and including the PoS labels for the constituents. The LASSY Small Treebank provides a similar analysis of the Dutch MWE *bij wijze van* "by way of", with *mwu* for the mother node and *mwp* for the daughter nodes in addition to the PoS labels for the constituents. The French Treebank provides a left-headed dependency analysis of the MWE *au sein du* "within" (literally "in-the breast of-the"), with *au* as the head and *sein* and *du* as dependents. The Cintil Portuguese Treebanks provide both constituency and dependency analyses; here we show the dependency analysis, which is similar to the one in the French treebank. The MWE *ao longo de* "along" is a left-headed dependency with *ao* as the head. As in French, there are contractions between prepositions and articles, so that the preposition *a* and the article *o* contract to the form *ao*.

Only two of the treebanks treat prepositional MWEs as words with spaces. The other treebanks that annotate these MWEs have separate nodes for their component words, and some of them include part of speech information for these component words. All of these treebanks treat them as prepositions on a syntactic level.

3.2 Verb-particle constructions

Sag et al. consider verb-particle constructions to be an important type of syntactically flexible expressions. These constructions cannot be treated as words with spaces since other words may intervene between the verb and the particle. They cannot simply be treated as compositional either, among other things because the particles often "assume semantics idiosyncratic to verb-particle constructions" [14, p. 194].

In the survey table there is one column for phrasal verbs. Clicking on the column header brings up a page with descriptions of the types of MWE annotations that should be entered in this column:

- Particle verbs such as *show up*
- Verbs with selected prepositions such as *think of*
- Verbs with both particles and selected prepositions such as *come up with*

Some of the languages in the survey do not have phrasal verbs of these three types; Bulgarian, Czech, French, Latvian and Portuguese have *N/A* for "not applicable" in the phrasal verbs column. Swedish, Slovene, Polish and Spanish have *NO* in this column, meaning that the language has the construction but that the treebank lacks annotation for it. Particle verbs are annotated in eight of the treebanks in various ways which reflect their MWE status. Figure 4 includes screenshots of the relevant parts of the analyses for these eight treebanks.

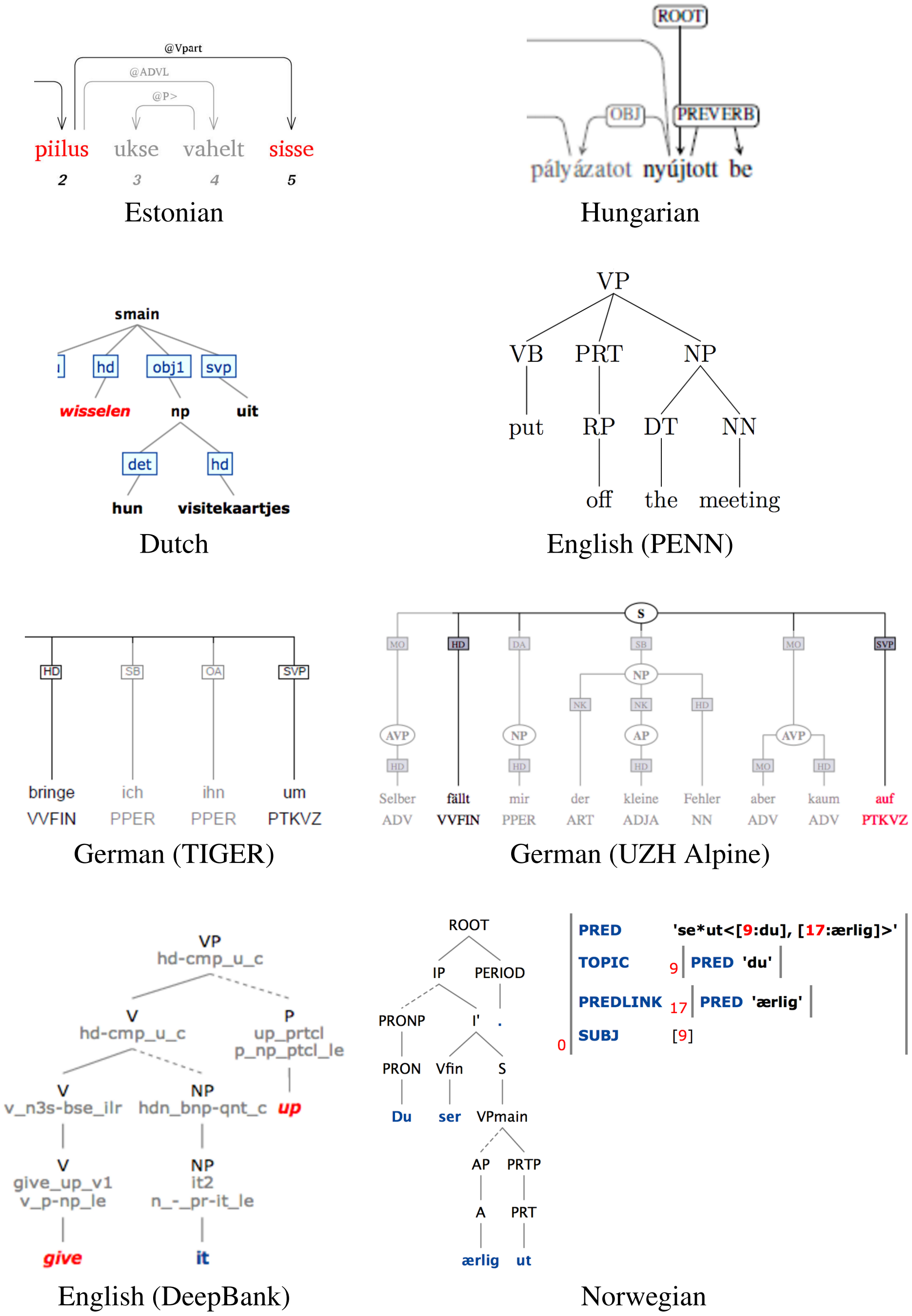


Figure 4: Overview of the annotations of particle verbs in eight treebanks

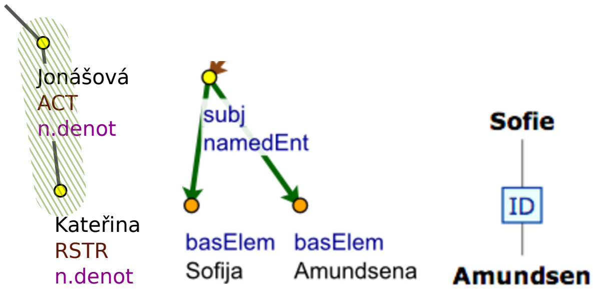
The annotations for particle verbs are quite similar across the treebanks that have them. In the Estonian treebank there is a VPART dependency from the verb to the particle. The Hungarian treebank has a PREVERB dependency from the verb to the particle; sometimes this particle is a prefix on the verb, and sometimes, as here, it is a separate graphical word. Dutch marks the verb particle as SVP for "separable verb prefix", since, as in Hungarian, it can sometimes form one word with the verb and sometimes, as in the example, occur as a separate word. In the three constituency treebanks with verb-particle constructions, the particle is annotated as a separate constituent in the S or VP that dominates it. The PENN treebank uses the PoS tag RP dominated by PRT; both of the German treebanks use the PoS tag PTKVZ for Partikel Verbzusatz dominated by an SVP node. DeepBank and NorGramBank do not only annotate the particle as a separate constituent, but also incorporate it into the verb in different ways. The DeepBank preterminal of the verb indicates that the verb *give* in this case has a lexical entry which specifies the complement *up*. In NorGramBank, the particle PRT is dominated by a particle phrase PRTP in the c(onstituent)-structure, but it does not contribute any predicate (PRED) of its own to the f(unctional)-structure. The particle is, however, integrated into the PRED for the verb, which is *se*ut*, meaning "look". These latter two annotations make more explicit that the predicate cannot simply be analyzed compositionally.

The annotations for particle verbs turn out to be surprisingly similar across treebanks. The challenge in annotating these constructions is not in how they should be annotated, but in finding the verb-particle constructions themselves.

3.3 Multiword named entities

Of sixteen treebanks for which information is provided for multiword named entities, twelve have examples of person names. In spite of the fact that person names themselves are very similar across the languages in the survey, we do see differences in their annotation. As an illustration, three examples from dependency treebanks are given in Figure 5. In Czech and Swedish there is a dependency between the first and last names, but in Czech the last name is the head, whereas in Swedish the first name is the head. In Latvian, there is a special node called 'namedEnt' which has both the first and the last names as dependents.

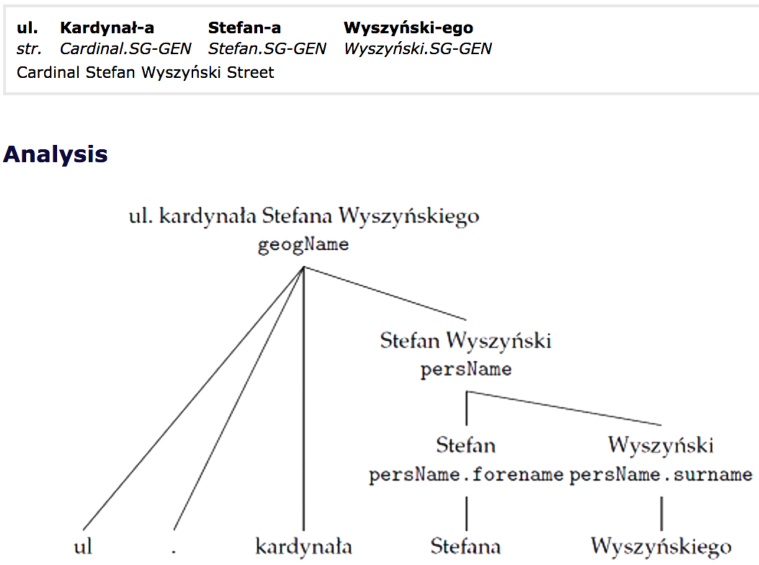
In addition to person names, there are several other types of multiword named entities which are exemplified: geographical names, names of institutions and organizations, temporal expressions such as dates and times, etc. Nine types of multiword named entities are distinguished in the Prague Dependency Treebank: person, institution, location, object, address, biblio, time, foreign and number. The National Corpus of Polish has six main types (persName, orgName, geogName, placeName, date and time), and there are eight subtypes. For most treebanks in the survey, however, only one or two examples are given, without it being clear if other types are also annotated.



Czech Latvian Swedish

Figure 5: Examples of MWE person names in three dependency treebanks

An example of a geographical named entity from the National Corpus of Polish is given in Figure 6. This is a complex example where the annotation of a person name is embedded inside the annotation of a geographical name. We note, however, that *Kardynała* 'Cardinal' is annotated as part of the geographical name, whereas it is actually a title that belongs hierarchically to a different level in the analysis. How such titles should be treated is an important question in itself. In the Dutch treebank, the title *drs.* is considered part of the named entity, as shown in Figure 7.



ul. **Kardynał-a** **Stefan-a** **Wyszyński-ego**
str. *Cardinal.SG-GEN* *Stefan.SG-GEN* *Wyszyński.SG-GEN*
Cardinal Stefan Wyszyński Street

Analysis

Figure 6: Example of a MWE named entity annotation in the Polish National Corpus

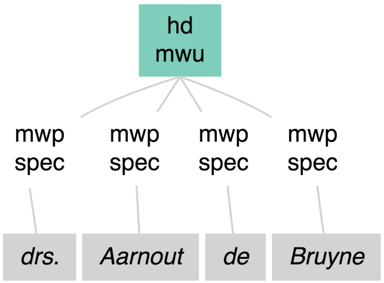


Figure 7: Example of a MWE named entity annotation in Lassy

A complex example from the ssj500k Dependency Treebank for Slovene is the analysis of the organization name *Odbor Združenih narodov za odpravo diskriminacije žensk* “The United Nations Committee on the Elimination of Discrimination against Women”. In this treebank, multiword named entities are annotated as chunks of connected tokens on the morphosyntactic layer. The whole entity is also labeled as a proper/organization name (*stvarno*). The dependencies are shown in Figure 8.

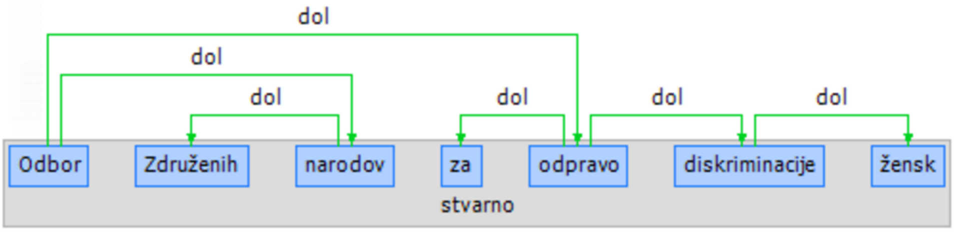


Figure 8: Example of a MWE named entity annotation in the ssj500K Dependency Treebank for Slovene

In conclusion, the annotation of multiword names ranges from very simple structures, similar to fixed expressions, to more complex structures, sometimes with other names embedded inside them. Treebanks may also vary considerably as to the types of named entities that they distinguish. This initial study shows that the survey should request more information about the range of possible annotations for multiword named entities in each treebank.

4 Conclusion and future work

We have reported on the first results from a focused survey on MWEs in various treebanks. We have developed a simple MWE typology, taking seminal works as a starting point. The survey includes treebanks with different annotation types.

While some MWEs are language specific (e.g. verb-particle constructions that are typical for Germanic languages), others occur in all the languages for which we have information (e.g. named entities).

The results indicate that for some MWE types (e.g. multiword named entities) there is more variation in annotation approaches than for other types (e.g. prepositional MWEs and verb-particle constructions).

Our study has also shown that better treebank documentation is important. It is often difficult to interpret the examples if there is no clear link to the tagset, the annotation guidelines, and similar information.

The survey is open-ended and can accommodate entries for additional languages and treebanks. The results of the survey are a step towards making recommendations for common principles and guidelines for annotating MWEs in treebanks.

Acknowledgments

This work has been supported in part by the PARSEME European Cost Action (IC1207) and the COST CZ grant of the MEYS of the Czech Republic (LD14117).

References

[1] Anne Abeillé, Lionel Clément, and François Toussenel. 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, 2003.

[2] Timothy Baldwin and Su Nam Kim. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, chapter 12. CRC Press, Boca Raton, FL, USA, 2nd edition, 2010.

[3] Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. Prague Dependency Treebank 3.0, 2013. Data, `http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3`.

[4] António Branco, Francisco Costa, João Silva, Sara Silveira, Sérgio Castro, Mariana Avelãs, Clara Pinto, and João Graça. Developing a deep linguistic databank supporting a collection of treebanks: the CINTIL DeepGramBank. In *LREC*, 2010.

[5] Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620, 2004.

[6] Mary Dalrymple. *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics*. Academic Press, San Diego, CA, 2001.

[7] Tomaz Erjavec, Darja Fiser, Simon Krek, and Nina Ledinek. The JOS linguistically tagged corpus of Slovene. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, page 1806–1809, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[8] Dan Flickinger, Yi Zhang, and Valia Kordoni. 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, 2012.

[9] Katarzyna Głowińska and Adam Przepiórkowski. The Design of Syntactic Annotation Levels in the National Corpus of Polish. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[10] Gyri Smørdal Losnegaard, Gunn Inger Lyse, Anje Müller Gjesdal, Koenraad De Smedt, Paul Meurer, and Victoria Rosén. Linking Northern European infrastructures for improving the accessibility and documentation of complex resources. In Koenraad De Smedt, Lars Borin, Krister Lindén, Bente Maegaard, Eiríkur Rögnvaldsson, and Kadri Vider, 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, 2013.

[11] Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen, Eleri Aedmaa, Riin Kirt, and Dage Särg. Estonian dependency treebank and its annotation scheme. In *Proceedings of 13th Workshop on Treebanks and Linguistic Theories (TLT13)*, pages 285–291, 2014.

[12] Carl Pollard and Ivan A Sag. *Head-driven phrase structure grammar*. University of Chicago Press, 1994.

[13] Lauma Pretkalnina and Laura Rituma. Syntactic issues identified developing the Latvian treebank. In *Baltic HLT*, pages 185–192, 2012.

[14] Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 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, 2002.

[15] Agata Savary, Jakub Waszczuk, and Adam Przepiórkowski. Towards the Annotation of Named Entities in the Polish National Corpus. In *Proceedings of LREC 10, Valletta, Malta*. European Language Resources Association, 17-23 May 2010.

[16] Kiril Simov, Petya Osenova, Alexander Simov, and Milen Kouylekov. Design and implementation of the Bulgarian HPSG-based treebank. *Journal of Research on Language and Computation*, Special Issue:495–522, 2005.

[17] Gertjan van Noord. Huge parsed corpora in LASSY. In Frank Van Eynde, Anette Frank, Koenraad De Smedt, and Gertjan van Noord, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT7)*, pages 115–126. LOT, 2009.

[18] Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. Hungarian dependency treebank. In *LREC*, 2010.