Extracting Verbal Multiword Data from Rich Treebank Annotation

Eduard Bejček, Jan Hajič, Pavel Straňák and Zdeňka Urešová

Charles University in Prague, Faculty of Mathematics and Physics, ÚFAL
{bejcek,hajic,stranak,uresova}@ufal.mff.cuni.cz

Abstract

The PARSEME Shared Task on automatic identification of verbal multiword expressions aims at identifying such expressions in running texts. Typology of verbal multiword expressions, very detailed annotation guidelines and gold-standard data for as many languages as possible will be provided. Since the Prague Dependency Treebank includes Czech multiword expression annotation, it was natural to make an attempt to automatically convert the data into the Shared Task format. However, since the Czech treebank predates the Shared Task annotation guidelines, a prior examination was necessary to determine to which extent the conversion can be fully automatic and how much manual work remains.

In this paper, we show that information contained in the Prague Dependency Treebank is sufficient to extract all of the Shared Task categories of verbal multiword expressions relevant for Czech, even if these categories are originally annotated differently; nevertheless, some manual checking and annotation would still be necessary, e.g. for distinguishing borderline cases.

1 Motivation

The goal of the PARSEME [11] Shared Task (PST)[1] is to develop automatic detection of verbal multiword expressions (VMWEs) for a wide range of languages from different language families. It includes data preparation for the task participants, based on annotation guidelines that were tested on real data for almost twenty languages [16].[2] The training and testing data for the PST (3,500 instances per language) are being annotated; while manual annotation is necessary for many languages, reusing existing annotated data is preferred whenever possible.

This preference led us to explore the Prague Dependency Treebank (PDT, [1, 4]), which includes quite a rich annotation of MWEs.[3] However, the anno-

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

[2] Also at http://parsemefr.lif.univ-mrs.fr/guidelines-hypertext.

[3] Some VMWEs categories were annotated during the creation of the original PDT 2.0, others were annotated particularly for PDT 2.5; PDT 3.0 contains all of them.

tation of the PDT preceded the PARSEME typology of VMWEs and thus it is understandable that the information encoded there is not straightforwardly transformable into the PST categories and format. Nevertheless, we hoped that the PDT annotation did contain all the necessary information. If confirmed, it would prove that the original scheme of rich annotation was well conceived, and in particular, that the MWE annotation in PDT in fact followed the principles recommended in [10].

2 Introduction

We believe that for the Czech language, annotation of VMWEs already encoded in the data of the Prague Dependency Treebank 3.0 (PDT) [1] presents suitable material for the PST and satisfies the task needs in both (i) the amount of annotated data and (ii) the types of VMWEs that correspond to the types proposed in the PST.

The PARSEME Shared Task identifies six groups of VMWEs: light verb constructions (`LVC`), idioms (`ID`), verb particle combinations (`VPC`), inherently reflexive verbs (`IReflV`), language specific types and other verbal MWEs (`OTH`).

All the various types of VMWEs required by the PST are annotated in quite a number of diverse ways in the PDT and the information is spread across several layers of annotation. Thus we first had to relate the PDT annotation to the PST guidelines in order to confirm that the PDT data can be reused for the Shared Task and only then the extraction of all types of VMWEs (relevant for Czech) and their conversion into the PST format could take place.

At the same time (or even more importantly), we were testing the following four principles for good-quality MWE treebank design published in [10], which are based on a survey of as many as 23 different treebanks (dependency-based, constituency-based, HPSG, LFG, mixed):
Principle A: *to annotate MWEs as such*,
Principle B: *to mark MWEs in a distinctive and specific way*,
Principle C: *to annotate even discontinuous MWEs and MWEs of varying forms*,
Principle D: *to allow for searching MWEs by their type*.

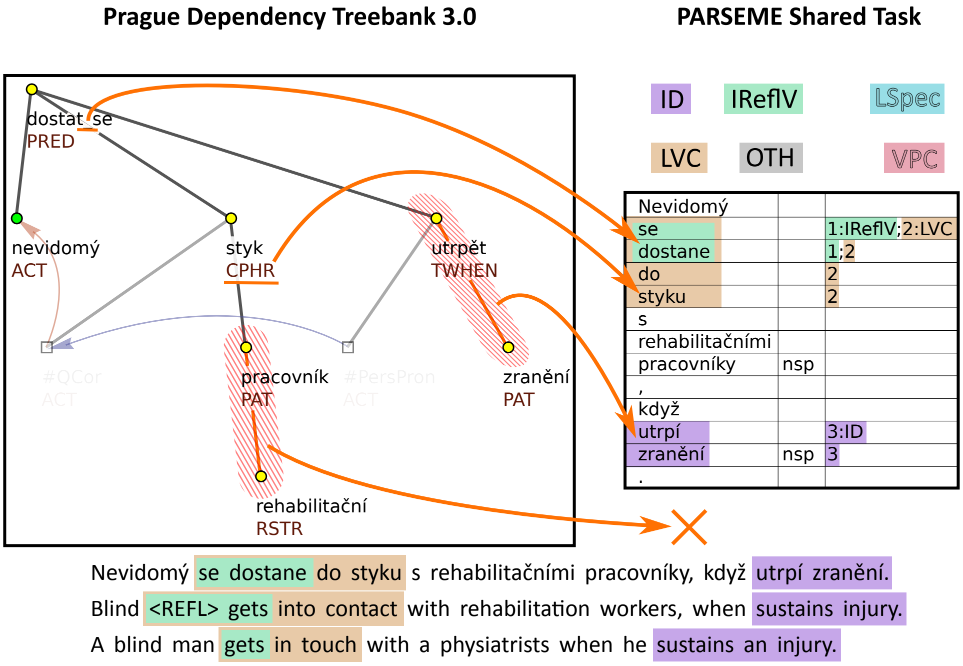
After thorough analysis of the PDT we have concluded that Principles A and B are clearly fulfilled in the PDT due to its explicit MWE annotation. Principle C is also followed thanks to the explicit links between the PDT's annotation layers. Principle D is, from the PST point of view, followed only partially, since the respective typologies do not match one-to-one.

Thorough inspection of the PDT annotation scheme resulted in an automatic conversion procedure with rules formulated for each of the PST types. Manual checks and some amount of manual annotation is still necessary, even if for only a fraction of the data.

3 Conversion of Czech data

As already explained, the creation of the Czech language data for the PST takes advantage of the existing rich annotation of the PDT, including explicit annotation of VMWEs.

The treatment of verbal idioms (part of the ID category) and LVCs in the PDT is related to valency, as the valency formalism allows for morphological, syntactic and semantic description of VMWEs in the treebank [2, 3, 13]. These VMWEs are recorded in the related valency lexicon, PDT-Vallex [14], as specific "senses" of the base lemma. For the annotation of verb-noun idiomatic combinations and some other types of MWEs in the PDT style treebanks and in the associated valency lexicons see [9, 15]. PDT-Vallex has been available already with the original PDT 2.0 treebank [4]. Afterwards, explicit general annotation of MWEs including verbal phrases which now correspond to the ID, LVC and OTH categories has been carried out (see [12]). The MWE annotation became part of later PDT releases, including the most recent, PDT 3.0 [1].[4] Reflexive verbs (IReflV) are treated as "words with spaces" on the deep syntactic annotation layer, with the particle being part of such words.



Nevidomý		
se		1:IReflV;2:LVC
dostane		1;2
do		2
styku		2
s		
rehabilitačními		
pracovníky	nsp	
,		
když		
utrpí		3:ID
zranění	nsp	3
.		

Nevidomý se dostane do styku s rehabilitačními pracovníky, když utrpí zranění.
Blind <REFL> gets into contact with rehabilitation workers, when sustains injury.
A blind man gets in touch with a physiatrists when he sustains an injury.

Figure 1: Extracting VMWE information: PDT annotation on the left, PST format with three VMWEs identified on the right (numbers distinguish VMWE occurences, "nsp" stands for "no space after", colours are ours). Only four types are relevant for Czech – neither VPC nor any language specific type is used.

[4] http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3

To sum up, different PST types of VMWEs are obtained from various information sources available at the different layers of annotation in the PDT. See Figure 1 for an illustration of three of them (the annotation view is simplified only to cover MWE-related phenomena); an annotation of the non-verbal MWE “rehabilitační pracovník” (*rehabilitation worker*) which is not being converted for PST is also shown.

In this section, we describe the PDT-style annotation of the proposed six types of VMWEs recognized in the PST as well as their conversion into the common PST format (Sections 3.1–3.6). Two special aspects are discussed, namely deverbative variants (Section 3.7) and cases of overlapping annotation (Section 3.8).

3.1 Light Verb Constructions

In the PDT annotation, `LVC`s consist of two lexical units: a semantically empty (or “light”) verb and a noun carrying the main lexical meaning of the entire phrase. The nominal part of the `LVC`s is labeled by the CPHR functor (Compound PHRase). For example: *to come*$_{\text{PRED}}$ *into force*$_{\text{CPHR}}$, *to undertake*$_{\text{PRED}}$ *preparations*$_{\text{CPHR}}$. `LVC`s are identified as depicted in Figure 2.

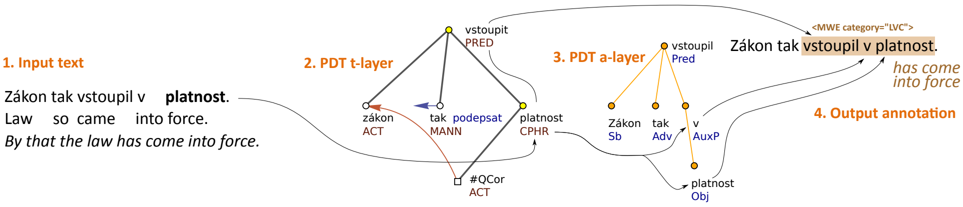


Figure 2: Identifying an `LVC` containing a preposition using two layers of PDT annotation. Deep syntactic layer provides a CPHR node and its governing verb (step 2). The preposition, in this case a part of the `LVC`, is represented by a node between the light verb node and the predicative noun node in the surface syntactic tree (step 3). The preposition node is (also) referenced from the CPHR node.

Three more things have to be taken into account:

1. Prepositions, if they are part of the `LVC`, must be retrieved from the surface syntactic layer, since they are not present on the deep layer. If there is any extra node between a node for a predicate and a node for a CPHR, it is part of the `LVC`.
2. If reflexive particles are part of the verb lemma (see `IReflV` in Section 3.4), they also have to become part of the `LVC`.
3. The CPHR functor is also used for a specific type of phrases with the verb “to be” (*it is necessary*$_{\text{CPHR}}$ *to leave*). These phrases, not assumed by the PST guidelines, are excluded by checking the lemma of the verb.

There are 2496 `LVC`s in the PDT extracted by the above rules. Minor details aside, `LVC`s as defined for the PST can be identified on the basis of the existing PDT annotation without any additional manual annotation.

3.2 Verbal Idioms

These VMWEs, denoted as ID in the PST guidelines, compose quite a large group containing not only traditional idioms. We have to process it in two steps.

Part of the VMWEs defined as IDs, namely those which are quite fixed idioms, are understood similarly in the PDT and in the guidelines for the Shared Task, e.g.: “házet klacky pod nohy” lit. *to-throw sticks under feet* (= *to put obstacles in one’s way*), “brát vítr z plachet” (= *to take the wind out of someone’s sails*). These verbal idioms (similarly to LVCs) always consist of two nodes in the PDT: the governing verb part and the dependent node (with the DPHR functor = Dependent part of PHRaseme). These idioms can be thus easily extracted by looking for the DPHR functor. The DPHR node represents all other lexical components of the idiom, should there be more than one (lemma of the deep syntactic layer is e.g. “klacky_pod_nohy” or “vítr_z_plachet”), since these are quite fixed expressions in terms of (the impossibility of) insertion or other modification. Even prepositions are part of it and their detection is even easier than with a CPHR. See an example in Figure 3.

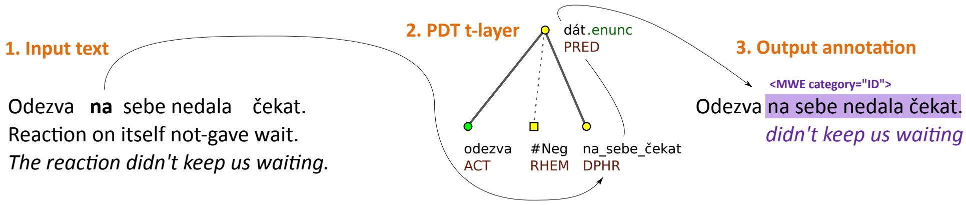


Figure 3: Identifying a four-word ID using a DPHR node and its governing node. (The VMWE here is negated: “na sebe **ne**dala čekat” instead of canonical “na sebe dala čekat”. It does not interfere with the extraction process, since negation is annotated separately; thus the two phrases themselves look the same.)

The other group of VMWEs categorized as ID in PST is not so fixed. VMWEs from this group do not fulfill the criteria for DPHR annotation in the PDT, but they still qualify to be an IDin the PST. They have been annotated together with all other MWEs in PDT 3.0 [12]. The problem is they are marked neither as idioms, nor even as verbal expressions. Moreover, they are recorded on the deep syntactic layer as a set of nodes (i.e. content words), neglecting auxiliary words.

Our approach finds a head in the syntactic tree of such a set. If it is a verb, the MWE is a verbal one (Figure 4). Then other auxiliary nodes (e.g., prepositions) referred to by the annotated content words are added. (The exception is a conjunction introducing the whole phrase: it does not belong to the VMWE.) The resulting VMWE gets the ID mark, unless it overlaps with CPHR or DPHR annotation (see Section 3.8).

We have identified 2107 IDs using either the PDT 3.0 MWE or DPHR annotation.

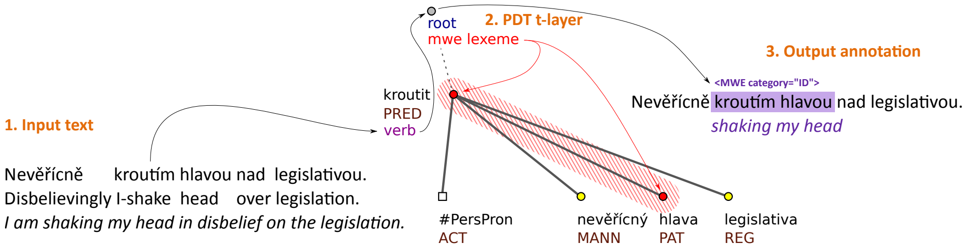

1. Input text
Nevěřícně kroutím hlavou nad legislativou.
Disbelievingly I-shake head over legislation.
I am shaking my head in disbelief on the legislation.
root
2. PDT t-layer
mwe lexeme
kroutit
PRED
verb
#PersPron
ACT
nevěřícný
MANN
hlava
PAT
legislativa
REG
3. Output annotation
<MWE category="ID">
Nevěřícně kroutím hlavou nad legislativou.
shaking my head

Figure 4: Identifying an `ID` using PDT 3.0 annotation. Such a MWEs is visualized here as a hatched area and is in fact recorded in the tree root with links to appropriate nodes (red arrows). The dependent node here is not marked as DPHR nor CPHR but as a regular PAT; it is however part of the PDT 3.0-annotated MWE.

3.3 Verb-particle Combinations

Verb-particle combinations (`VPC`) are not present in Czech. A phenomenon similar to `VPC`s is in Czech realized by verbal prefixes (the result being another *single* lexical unit, i.e., not a MWE).

3.4 Inherently Reflexive Verbs

Inherently Reflexive Verbs (`IReflV`) contain one of two possible clitics in Czech: “se” or “si”, e.g. “bát se” (= *to be afraid*), “hledět si” (= *to mind sth*). Such verb is considered a separate lexical unit (different from the verb appearing without the particle if such verb exists at all) and both its parts are represented by just one node at the deep syntactic layer of the PDT, and the node’s lemma matches the PDT-Vallex lexical unit, which includes the appropriate particle as part of the headword in the lexicon. This annotation was used for exactly the two types qualified as `IReflV` in the PST guidelines, namely, for the case when the non-reflexive counterpart verb does not exist or when its meaning is markedly changed. Using this annotation, all `IReflV`s can be extracted from the PDT texts and converted, see Figure 5.

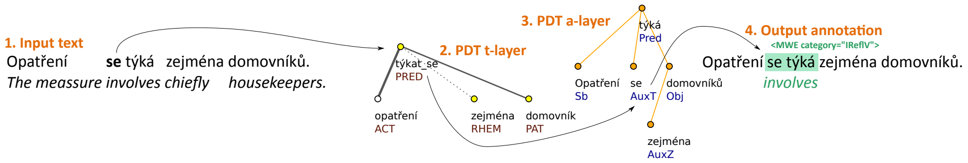


Figure 5: Identifying an `IReflV` using a lemma on deep syntactic layer (step 2) together with an analytical function AuxT on a surface syntactic layer (step 3).

`IReflV`s should be possible to extract also without the deep syntactic layer; an analytical function of an `IReflV` reflexive particle should be either AuxT or AuxO on a surface syntactic layer; other values (AuxR, Obj, or Adv) are reserved for reflexive particles used in other than IReflV contexts, e.g. in passive constructions. Suspicious cases (705 verb occurrences) in which the information from the two layers of annotation clashes have been detected by looking for discrepancy between the lemma and the corresponding analytical function and manually checked and corrected when necessary (330 cases). There are some borderline cases where the PDT annotation differs from the PST guidelines; however, these are mainly errors in annotation and not a true difference between the PST and PDT guidelines.

By this approach, 10,266 VMWEs of the IReflV type were extracted from the PDT. The conversion was automatic except for the 705 manually checked occurrences.

3.5 Others

This category (OTH) is specified in the PST guidelines as a VMWE that does not fit into any of the other categories, as described in the previous sections. Namely, it applies to "coordinations of verbs, e.g. *to drink and drive*, and compound verbs, e.g. *to short-circuit, to pretty-print, to voice act*". The second subtype usually results in a one-word expression in Czech, so we need to search only for coordinated verbs.

For this category, the PDT 3.0 MWEs annotation [12] is useful again. All MWEs containing two verbs connected by a coordinating conjunction are marked as an OTH, see Figure 6. This is a very marginal category; we have found only two OTHs in the data.

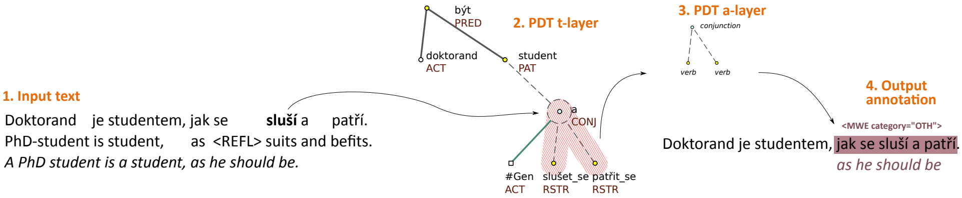

1. Input text
Doktorand je studentem, jak se **sluší** a patří.
PhD-student is student, as <REFL> suits and befits.
A PhD student is a student, as he should be.
být
PRED
2. PDT t-layer
doktorand
ACT
student
PAT
a
CONJ
#Gen
ACT
slušet_se
RSTR
patřit_se
RSTR
3. PDT a-layer
conjunction
verb
verb
4. Output annotation
<MWE category="OTH">
Doktorand je studentem, jak se sluší a patří.
as he should be

Figure 6: Identifying an OTH by the pattern "two coordinated verbs" (step 3). (Both coordinated verbs also qualify as IReflVs, which is not shown in the figure.)

3.6 Language Specific Category

No language specific categories are defined for Czech.

3.7 Deverbative variants

PARSEME Shared Task guidelines also recognize other, non-verbal variants of verbal MWEs, such as relative clauses (*heart which he broke*), gerunds (*heart break-*

ing), nominal groups (*heart-breaking*), or adjectival groups (*breaking her heart*). In Czech, nominalization is a common way of verbal MWE variation, see [7, 6, 8].

There is no nominal group annotated as CPHR in the PDT and thus no LVC variant. There are several nominal MWEs annotated as DPHR, but only seven of them are made from verbal MWEs. We have picked them manually. During the PDT 3.0 MWE annotation project [12], annotators were asked to mark deverbative variants with the verbal lexicon entry. This annotation, although it is not frequent, is also used.

The situation is quite different for IReflV where many non-verbal lemmas also contain reflexive particles "se" or "si". These cases qualify themselves as nominal or adverbial variants of inherently reflexive verbs.

To sum up, there are deverbative MWEs in the PST Czech data, however they are not frequent.

We are also preparing other deverbative MWEs using data by an idiom recognizer based on a database, upgraded for deverbatives by Milena Hnátková [5].

3.8 Overlaps

Since the data for PST are extracted from various pieces of annotation, it can easily happen they are duplicated or that they overlap. All these cases have to be solved properly, as described below.

3.8.1 Coordination

Part of a VMWE can be coordinated while the other part is used only once, as in "Ministerstvo poskytuje malým podnikatelům informační služby a poradenskou činnost." (*The ministry provides information services and counselling activities to small businesses.*), where two LVCs are present: *to provide services* and *to provide activities*. Such a case is correct and both VMWEs should be preserved and marked in the output data, with the verb "provide" being part of both.

3.8.2 Duplicates due to added nodes in the PDT

Since a large part of the MWE annotation in the PDT is encoded at the deep syntactic layer, sometimes a VMWE is found that has no direct realization in the surface form of the sentence, although it is present in its deep structure. For example, *The measure can be taken for six month at most and only for selected items.*, which in fact means *The measure can be taken for six month at most* and *the measure can be taken only for selected items.* In the PDT, two light verb constructions are annotated and both of them are linked to the same words. This would result in duplicate annotation of the words "measure" "be" and "taken" in the sentence. Such duplicates are detected and removed before the data are exported.

3.8.3 Overlapping different types of VMWEs

As described previously, we combine explicit idiomatic annotation (DPHR), explicit light verb annotation (CPHR) and the verbal MWE annotation from PDT 3.0. If they overlap, the type of the MWE (ID or LVC) is always determined by the explicit DPHR/CPHR annotation. If only the PDT 3.0 MWE annotation is present, it always gets ID type as the most probable case; however, this could be checked manually in future.

Whenever IReflV overlaps with any other, usually larger MWE, both are correct and should remain in the output. Other overlaps of different types of VMWEs are not possible due to the source data we work with.

It is yet to be determined what to do with cases where an ID from DPHR and from PDT 3.0 MWE annotation overlaps with different word range.

3.9 Results

After removing the overlaps, there are over 14,000 verbal multiword expressions exported in the PST format. Table 1 presents the numbers of individual types of VMWEs.

VMWE type	number of instances	without overlaps
ID	2,107	1,611
LVC	2,496	2,437
IReflV	10,266	9,982
OTH	2	2
Total	14,871	14,032

Table 1: Number of VMWEs extracted from the PDT and prepared for PARSEME Shared Task. The first number is a raw number of VMWEs found, the second one is after removal of duplicates and overlapped expressions that should not overlap.

4 Conclusions

It can be concluded that due to a well-founded, rich annotation scheme used in the Prague Dependency Treebank, which also conforms to most of the four PARSEME MWE annotation principles, we can almost fully automatically transform the original MWE annotation into the PARSEME Shared Task verbal MWE types. By that, we can extract 14,032 VMWEs.

In the near future, we still want to manually check some borderline cases mentioned above, e.g. whether an isolated verbal PDT 3.0 MWE should be always an ID, or how to solve overlapping annotation of the same type but of a different range. We will include deverbative MWEs from separate automatic lexicon-based annotation.

5 Acknowledgement

The work described here has been supported by the project PARSEME, LD14117, by the Ministry of Education, Youth and Sports of the Czech Republic, and carried out within the framework of the project COST IC1207 PARSEME. The project used data distributed by the LINDAT/CLARIN repository, supported by the Ministry of Education, Youth and Sports of the Czech Republic (projects LM2010013, LM2015071). We also thank our colleague Milena Hnátková who kindly extracted deverbative variants of VMWEs using her phraseme database and we are working on incorporating this data into our outputs.

References

[1] 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 available from LINDAT/CLARIN, http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3.

[2] Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. In Joakim Nivre and Erhard Hinrichs, editors, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68, Vaxjo, Sweden, 2003. Vaxjo University Press.

[3] Jan Hajič and Zdeňka Urešová. Linguistic annotation: from links to cross-layer lexicons. In Joakim Nivre and Erhard Hinrichs, editors, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 69–80, Vaxjo, Sweden, 2003. Vaxjo University Press.

[4] Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková Razímová, and Zdeňka Urešová. Prague Dependency Treebank 2.0, 2006. LDC2006T01. Philadelphia, PA, USA.

[5] Milena Hnátková. Značkování frazémů a idiomů v Českém národním korpusu s pomocí Slovníku české frazeologie a idiomatiky. *Slovo a slovesnost*, 2002.

[6] Veronika Kolářová. *Valence deverbativních substantiv v češtině (PhD thesis)*. PhD thesis, Univerzita Karlova v Praze, Matematicko-fyzikální fakulta, Praha, Czechia, 2005.

[7] Veronika Kolářová. Valency of Deverbal Nouns in Czech. *The Prague Bulletin of Mathematical Linguistics*, 86:5–20, 2006.

[8] Veronika Kolářová. *Special valency behavior of Czech deverbal nouns*, chapter 2, pages 19–60. Studies in Language Companion Series, 158. John Benjamins Publishing Company, Amsterdam, The Netherlands, 2014.

[9] Adam Przepiórkowski, Jan Hajič, Elżbieta Hajnicz, and Zdeňka Urešová. Phraseology in two slavic valency dictionaries: limitations and perspectives. *International Journal of Lexicography*, (1):1–38, 2016.

[10] Victoria Rosén, Koenraad De Smedt, Gyri Losnegaard, Eduard Bejček, Agata Savary, and Petya Osenova. MWEs in treebanks: From survey to guidelines. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2323–2330, Paris, France, 2016. European Language Resources Association.

[11] 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. 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 2015.

[12] Pavel Straňák. *Annotation of Multiword Expressions in The Prague Dependency Treebank*. PhD thesis, Charles University in Prague, 2010.

[13] Zdeňka Urešová. *Valence sloves v Pražském závislostním korpusu*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, 2011.

[14] Zdeňka Urešová. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, 2011.

[15] Zdeňka Urešová, Eva Fučíková, Jan Hajič, and Jana Šindlerová. An analysis of annotation of verb-noun idiomatic combinations in a parallel dependency corpus. In *The 9th Workshop on Multiword Expressions (MWE 2013)*, pages 58–63, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics, Association for Computational Linguistics.

[16] Veronika Vincze, Agata Savary, Marie Candito, Carlos Ramisch, and Fabienne Cap. Annotation guidelines for the PARSEME shared task on automatic detection of verbal multiword expressions, version 6.0, 2016. http://typo.uni-konstanz.de/parseme/images/shared-task/guidelines/PARSEME-ST-annotation-guidelines-v6.pdf.