PARSEME Shared Task and PDT: Comparison and Data Conversion

Dubrovnik poster proposal, Working Group 4, work in progress

Eduard Bejček, Pavel Straňák and Zdeňka Urešová
Charles University in Prague, MFF, ÚFAL
{bejcek,stranak,uresova}@ufal.mff.cuni.cz

Abstract

We show that verbal MWEs similar to those specified in PARSEME Shared Task have already been annotated in Prague Dependency Treebank and all their individual categories can be extracted and used within the Shared Task.

1 Motivation

PARSEME Shared Task (PST) on automatic detection of verbal multiword expressions (VMWEs) aims at wide range of languages from different language families. While the annotation of training and testing data for the task is necessary for many languages, reusing existing annotated data is preferred whenever it is possible.

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) (Bejček et al., 2013) 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 seven groups of VMWEs: light verb constructions (`LVC`), idioms (`ID`), verb particle combinations (`VPC`), inherently pronominal verbs (`IPronV`), inherently prepositional verbs (`IPrepV`), possibly other language specific category 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 PDT and the information is spread across several levels of annotation. Thus we first have to relate the PDT annotation to the Shared Task 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 Shared Task format can take place.

On the top of automatic checks of the converted data, we expect a certain amount of manual work as the annotation guidelines for the Shared Task (Savary et al., 2015) differ in details from the annotation guidelines for the VMWEs annotation in the PDT.

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.

The treatment of VMWEs in the PDT is related to valency, as the valency formalism allows for morphological, syntactic and semantic description of VMWEs in the treebank. VMWEs are recorded in the related valency lexicon, PDT-Vallex, as specific verb senses. PDT-Vallex (Urešová, 2011) has been available already with the original PDT 2.0 treebank (Hajič et al., 2006). For the general MWEs annotation in the PDT see (Straňák, 2010), 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 (Urešová et al., 2013).

In this section, we describe how the proposed seven types of VMWEs recognized in the PST are encoded in the PDT annotation and how their conversion into the common format is done.

3.1 Light Verb Constructions

In the PDT annotation Light Verb Constructions (`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 both in the valency lexicon PDT-Vallex and in the PDT itself labeled by the CPHR functor (Compound PHRase). For example: *to carry on a conversation*$_{CPHR}$, *to undertake preparations*$_{CPHR}$.

The same functor (CPHR) is used also for a specific type of phrases with the verb "to be" (*Je třeba odejít.* = *Is necessary*$_{CPHR}$ *to-leave.*). These phrases can be easily excluded using the information about the verbal lemma. Result: `LVC`s can be converted from PDT without manual annotation work.

3.2 Verbal Idioms

Idioms (`ID`s) are understood similarly in 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*). Verbal `ID`s always consist of two nodes: the governing verb part and the dependent node (with the DPHR functor = Dependent part of PHRaseme) that can represent all other lexical components of `ID`, should there be more than one. Similarly to the `LVC` case, `ID`s can be easily extracted, in this case based on the DPHR functor.

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 a different single lexical unit, not a MWE).

3.4 Inherently Pronominal Verbs

Inherently Pronominal Verbs (`IPronV`, or a "reflexive verb") contains 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) and both its parts are assigned to just one node in the deep syntactic layer of the PDT. The lexical unit contains the appropriate particle as part of the lemma. Using this annotation, all `IPronV`s can be extracted from the PDT texts and converted to the PST dataset.

There are some borderline cases where PDT annotation differs from the PST guidelines. These cases have been manually checked and corrected when necessary.

3.5 Inherently Prepositional Verbs

Inherently Prepositional Verbs (`IPrepV`) are a category which does not have a straightforward counterpart in the PDT. However, we can use the surface information about a required complement form, which is recorded at each complement slot in the PDT-Vallex lexicon. We use a hypothesis that if there exists only a single allowed surface form for a complement, and such form requires a prepositional phrase, then it fulfills the PST definition of a `IPrepV`.

3.6 Others

For this category (`OTH`), the special annotation dedicated to MWEs (Straňák, 2010) is useful. The annotation is linked to the SemLex lexicon, where we can find all verbal MWEs. All of them that do not fall into one of the previous categories are annotated as `OTH`.

3.7 Language Specific Category

No language specific categories are defined for Czech.

References

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. 2013. Prague dependency treebank 3.0. Data.

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

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

Pavel Straňák. 2010. *Annotation of Multiword Expressions in The Prague Dependency Treebank*. Ph.D. thesis, Charles University in Prague.

Zdeňka Urešová, Eva Fučíková, Jan Hajič, and Jana Šindlerová. 2013. 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. Association for Computational Linguistics, Association for Computational Linguistics.

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