Semi-automatic Detection of Multiword Expressions in the Slovak Dependency Treebank

Daniela Majchráková,[*] **Ondřej Dušek,**[‡] **Jan Hajič,**[‡] **Agáta Karčová**[*] and **Radovan Garabík**[*]
[*]Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava
[‡]Charles Univ. in Prague, Fac. Math. & Phys., Inst. of Formal and Applied Linguistics
danam@korpus.sk, {odusek,hajic}@ufal.mff.cuni.cz,
agatak@korpus.sk, garabik@kassiopeia.juls.savba.sk

Abstract

We describe a method for semi-automatic extraction of Slovak multiword expressions (MWEs) from a dependency treebank. The process uses an automatic conversion from dependency syntactic trees to deep syntax and automatic tagging of verbal argument nodes based on a valency dictionary. Both the valency dictionary and the treebank conversion were adapted from the corresponding Czech versions; the automatically translated valency dictionary has been manually proofread and corrected. There are two main achievements – a valency dictionary of Slovak MWEs with direct links to corresponding expressions in the Czech dictionary, PDT-Vallex, and a method of extraction of MWEs from the Slovak Dependency Treebank. The extraction reached very high precision but lower recall in a manual evaluation. This is a work in progress, the overall goal of which is twofold: to create a Slovak language valency dictionary paralleling the Czech one, with bilingual links; and to use the extracted verbal frames in a collocation dictionary of Slovak verbs.

1. Introduction

This work is primarily aimed at building a Slovak valency lexicon interlinked with a dependency treebank, and in this paper we focus on multiword expressions (MWEs). The prospective valency lexicon is inspired by the Czech PDT-Vallex, a lexicon based on the Prague Dependency Treebank (PDT). We exploit here the fact that Czech and Slovak are very closely related, mutually intelligible languages that show a direct 1:1 relation in a greater part of their grammatical and lexical inventory, including MWEs.

Following the definitions of MWEs for PDT annotation, here we understand by MWEs those lexical combinations "that contain some idiosyncratic element that differentiates them from normal expressions" (Bejček et al., 2012: 234). There are two types of MWEs we focused on: *light verb constructions* and *verbal phrasemes*. The valency frames of both groups are marked with special semantic labels (functors) in the deep-syntax/semantic annotation of the PDT (*tectogrammatical layer*): Compound Phraseme (CPHR) for light verb phrases and Dependent Phraseme (DPHR) for phrasemes.

In the first stage of our work, PDT-Vallex was automatically translated into Slovak and valency frames for Slovak verbs were automatically created based on their Czech counterparts. Subsequently, the translations of verbs and their valency frames were manually proofread to ensure correctness, especially those related to MWEs. The result of this process is a preliminary version of the Slovak Valency Lexicon (SVL).

The second stage involves linking the SVL to the Slovak Dependency Treebank (SDT) (Šimková and Garabík, 2006). We developed an automatic procedure to convert the SVL to a deep-syntactic representation parallel to the PDT. Here we used a list of MWE candidates extracted from the SVL to automatically identify the individual occurrences of MWEs. We evaluated the precision and recall of the automatic MWE detection by manual assessment on a small part of the SDT.

The paper is structured as follows: In Section 2., we introduce the SDT. We describe the creation of the SVL in Section 3., contrasting MWE usage in Czech and Slovak. Section 4. details our automatic procedure for the conversion of the SDT to a deep-syntactic representation. Section 5. presents the evaluation of the automatic MWE detection in the treebank and Section 6. concludes the paper.

2. Slovak Dependency Treebank

The Slovak Dependency Treebank (SDT) (Šimková and Garabík, 2006) is a manually annotated dependency treebank of contemporary written Slovak. The annotation follows the methodology of the Prague Dependency Treebank (PDT) (Hajič et al., 1999). However, the SDT contains only surface dependency (*analytical*) trees, it does not include the deep-syntax/semantic (*tectogrammatical*) layer (see Section 4.), where valency and MWEs are annotated in the PDT.

The SDT contains 1,159,462 tokens in 71,672 sentences, 50,313 sentences (846,967 tokens) out of which were annotated by two independent annotators. Most texts in the treebank include manual morphological annotation (lemmas and morphological tags) based on the Slovak National Corpus tagset (Garabík and Šimková, 2012).[1]

The selection of the texts aims at a somewhat balanced corpus – there are professional texts (scientific articles, theses), fiction, and journalistic texts.

3. Building the Slovak Valency Lexicon: PDT-Vallex Translation

The PDT-Vallex (Hajič et al., 2003; Urešová, 2011a; Urešová, 2011b) is a valency lexicon interlinked with the Prague Dependency Treebank. It consists of over 11 thousand valency frames for more than 7,000 verbs. The verbs, their senses, and their valency frames are collected from sentences in the PDT.

Although Czech and Slovak are close languages, the translation of PDT-Vallex was not straightforward. The automatic translation consists of simple lexical substitution of verbs and their complementations. We then manually checked all entries relevant to the MWE extraction (261 light verbs/CPHR nodes and 480 phrasemes/DPHR nodes). The manual proofreading of the automatic translation and contrastive analysis of equivalent Czech and Slovak MWEs proved that given the closeness of both languages, there was a huge overlap of MWEs in Czech and Slovak. However, we found several cases where identical semantic content was represented by very different lexical and/or syntactic means, mainly in phrasemes.

For the purpose of obtaining the list of Slovak MWEs for automatic annotation, we mention only briefly some similarities and differences between Czech and Slovak equivalent expressions we encountered in the translation of the valency dictionary.

3.1. Similarities of Czech and Slovak MWEs

The similarities of Czech and Slovak CPHR and DPHR structures can be summarized as follows:

- Most verbs and nouns from PDT-Vallex expressing the same semantic content are etymological cognates – e.g., *podat/podať*,[2] ("hand over"), *obracet/obracať* ("turn over"), *dojem/dojem* ("impression"), *zřetel/zreteľ* ("consideration").

- Slovak and Czech verbal aspects are identical in almost all cases[3] and reflexive verbs in Czech are also reflexive in Slovak – e.g., *dát se/dať sa* ("be possible"), *udělat si/urobiť si* ("make").

- The structure of light verbs and phrasemes is identical in both languages, with just a few exceptions.

3.2. Differences between Czech and Slovak MWE Equivalents

The differences between Czech and Slovak MWEs include grammatical and/or lexical distinctions, which are reflected in the component structure of some MWEs.

[1] There are some short texts in the treebank which were tagged automatically, but these were excluded for the purpose of this article.

[2] In these examples, the Czech word is displayed first, followed by the Slovak equivalent separated by a slash.

[3] Both Czech and Slovak verbs form aspectual pairs for incompletive/processual and completive aspect, e.g., *hádzať/hodiť* ("be throwing"/"throw").

Grammatical differences. According to the grammatical features, some MWE equivalents vary in the noun case; this is usually connected to the absence of or the preference for a different preposition: *přicházet v úvahu/prichádzať do úvahy* ("come into consideration"; accusative vs. genitive), *zažít na vlastní kůži/prežiť na vlastnej koži* ("experience on one's own"; accusative vs. locative).

As PDT-Vallex consists only of MWEs occuring in PDT, some of the phrases were not covered by verbs in both verbal aspects. In some cases, the aspect variant included in PDT-Vallex is less frequent or outright rare in the Slovak equivalent. In order to obtain better coverage, we decided to use both verb aspects in the Slovak translation: *zavádět řeč na jiné téma → zavádzať reč na inú tému*, *zaviesť reč na inú tému* ("steer to another topic")

Differences in lexical component. Some MWEs differ in lexical components in the use of synonymic equivalent, e.g., *vzít nohy na ramena/vziať nohy na plecia* ("run away"), *shodit pod stůl/zmietnuť zo stola* ("drop from the table"). Significant differences are present in idioms like *vyšly navrch/vyšli na povrch* ("come out"), in which the Czech adverb corresponds to Slovak noun in accusative form. There were also differences in verbal components. In some cases we preferred more frequent and neutral synonyms instead of the equivalents perceived as marked (e.g., archaic, poetic etc.) *učinit/urobiť*, *náležet/prislúchať*.

Differences in component structure of MWE equivalents. There were not many structural differences between Czech and Slovak MWEs. They can be illustrated by the following schematics (with Czech MWE structures on the left and Slovak on the right):

- Adding/removal of a grammatical component (preposition):

V + S	V+ Prep + S
zírat údivem	*civieť súdivom* ("gape in awe")

V+ Prep + S	V+ S
dát za vyučenou	*dať príučku* ("give a lesson")

- Adverbs change to a prepositional phrase (petrified in the second example, cannot be split into separate components):

V+ Adv	V+ Prep + S
vyjít navrch	*vychádzať na povrch* ("come out")

V+ Adv	V+ [Prep + S]
vycházet vstříc	*vychádzať vústrety* ("to be acommodating")

- Absence of a Slovak equivalent for the Czech particle *co*:

V+ Part + [Prep + S]	V+ [Prep + S]
mít co do činění	*mať do činenia* ("have something to do with")

- Partial disagreement arising from the nature of the Slovak particle *treba* ("is needed"). The difference is only apparent in the present tense where the particle *treba* does not require the auxiliary verb *byt*[4]; this is different from past and future tense:

V+ Adv	Adv
je třeba	*treba* ("is needed"; present tense)

[4] The present tense also occurs with the auxiliary verb *byť (je treba)*; this is, however, considered colloquial. We still included this variant in the dictionary to increase coverage.

- the phrase *bůh vám zaplať/pánboh zaplať* ("God bless you") has a different structure and lexical components in Slovak:

S+ Pron + V	S + V
bůh vám zaplať	*pánboh zaplať*

In some cases, the translation of a Czech MWE is not possible at all; either it contains lexical lacunae or the phrase as a whole is not used in Slovak. Examples of light verb constructions without an equivalent in Slovak are: *dát/dávat preferenci* ("give preference"), examples of phrasemes are: *vydat všanc* ("submit to risk"), *vzít roha* ("run away"), *být na štíru* ("have a problem with").

4. Automatic Tectogrammatical Annotation

To link SDT to a valency lexicon paralleling PDT-Vallex, we created a procedure for the conversion of the SDT from surface dependency trees to tectogrammatial trees, a deep-syntactic/semantic representation based on the Functional Generative Description (Sgall, 1967; Sgall et al., 1986). The tectogrammatical representation of a sentence is a dependency tree which only consists of nodes that carry lexical meaning; auxiliary words are no longer included. Each tectogrammatical node is marked with a lemma, a *functor* (semantic role label) and a set of *grammatemes*, which carry grammatical meanings, such as number, tense, or modality.

The surface dependency trees are automatically converted into tectogrammatical trees by a set of small, rule-based modules implemented within the Treex NLP framework (Popel and Žabokrtský, 2010). Since the conversion makes heavy use of morphology information and was primarily developed with the Czech positional morphological tagset (Hajič, 2004) used in PDT in mind, it also includes a morphological tagset conversion step.

4.1. Morphological Tagset Conversion

For morphological tagset conversion, we make use of the Interset framework (Zeman, 2008). This framework contains a common list of various morphological properties across languages and their values to support conversion among different tagsets. One can either use directly the morphological information stored in Interset, or convert the source morphological tag into a different framework.

We have created an Interset driver (converter) for the Slovak National Treebank morphological tagset. We use both the information stored directly in Interset and a conversion to the PDT tagset. This allows us to reuse both language-independent and Czech-specific modules in the conversion process.

4.2. From Analytical to Tectogrammatical

The Treex modules for the conversion from analytical (surface dependencies) to tectogrammatical representation (deep syntax/semantics) closely follow the modules used for a similar conversion in Czech and English within the CzEng parallel corpus (Bojar et al., 2012) and the TectoMT machine translation system (Žabokrtský et al., 2008). However, unlike in CzEng and TectoMT, we apply the conversion to manually annotated analytical trees.

The conversion consists (roughly) of the following steps:

1. Auxiliary and grammatical words, such as prepositions and auxiliary verbs, are identified in the analytical tree. A new tectogrammatical tree is built that does not contain the auxiliary words as separate nodes, but retains links to the multiple analytical nodes for a single tectogrammatical node, including all auxiliaries.

2. Coordination and apposition functors (such as CONJ, DISJ, ADVS for conjunctive, adversative, and disjunctive relation) are identified.

3. Links to auxiliaries are distributed through coordination structures, i.e., if a preposition applies to multiple coordinated nouns, tectogrammatical nodes for all nouns will have a link to its analytical node.

4. Finite clause heads, relative clause heads, and relative clause co-reference are marked.

5. Tectogrammatical lemmas are normalized. In the current implementation for Slovak, this applies to personal and possessive pronouns, which all obtain a technical lemma *#PersPron*, and to reflexive tantum verbs, where the reflexive particle *sa/si* becomes part of the lemma (e.g., *smiať_sa* for "laugh").

6. All nodes are assigned grammatemes. In the current version, all nodes obtain semantic part-of-speech (noun, adjective, verb, adverb), and semantic verbs further obtain diathesis information (active, passive, reflexive diathesis).[5]

7. Functors are assigned to all nodes. We use rules based on lexical meaning, auxiliary words linked from a given node, and part-of-speech of the lexical word to estimate its semantic function.

   This step also includes detection of multiword expressions – light verb constructions and phrasemes, which are given functors *CPHR* and *DPHR*, respectively. These are detected based on candidate lists gathered from the Slovak Valency Lexicon (SVL, see Section 3.).[6]

8. Special tectogrammatical nodes are generated for actors not expressed on the surface — pro-dropped pronominal subjects and generic actors in reflexive passive constructions, such as *Dom sa stavia* (lex. *A-house itself builds*).

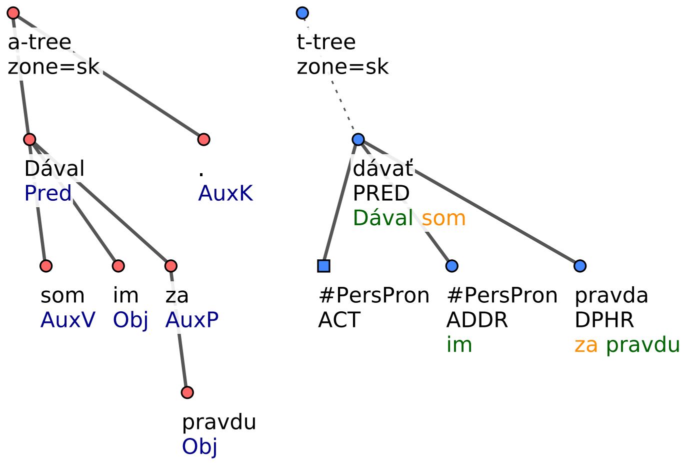

Figure 1: An original dependency tree from the Slovak Dependency Treebank (left, with dependency labels given in blue) and a tectogrammatical tree after conversion (right, with functors on the second line). The DPHR functor marks a dependent part of the phraseme *Dával som im za pravdu.* ("I agreed with them.").

Figure 1 shows a comparison of the original dependency tree with the result of the tectogrammatical conversion.

4.3. The Result of the Tectogrammatical Conversion

While the tectogrammatical layer conversion is almost equivalent to automatic tectogrammatical annotation used for English and Czech, it is missing some of the attributes present in the manual annotation of PDT:

- Generated nodes for other semantic participants than actors,

[5] Cf. Urešová and Pajas (2009) for more information on diathesis.

[6] The detection algorithm checks for the presence of all dependent parts of a MWE in the surface dependency subtree governed by its verb, then assigns MWE functors to corresponding tectogrammatical nodes. It abstracts from particular inflection forms by checking base word forms (lemmas) only. While such an abstraction may possibly result in lower precision, our experiments in Section 5. show that it is sufficient in practice.

- Full pronominal co-reference,
- Generated nodes for cases of ellipsis,
- Explicit valency frame assignment, i.e., sense disambiguation for verbs and some nouns,
- Focus-topic articulation and discourse structure.

However, even this level of annotation is suitable for linguistic inquiry and automated tasks such as machine translation, and can be used as a starting point for full manual tectogrammatic annotation.

5. Evaluation

In order to estimate the performance of the automatic MWE annotation, we randomly selected about one thousand sentences out of the tectogrammatical conversion of the Slovak Dependency Treebank,[7] where we annotated CPHR and DPHR nodes for light verbs and phrasemes manually. We then compared this sample to the result of the automatic conversion.

Table 1 shows estimates of precision and recall for three main types of text – newspaper texts, professional texts (i.e., scientific), and fiction. The ratio columns show the ratio of CPHR and DPHR nodes to the total (tectogrammatical) nodes of the sample. Given the rather small sample size, the number of these nodes is small. The precision and recall figures should therefore be considered with this in mind.

The manual proofreading of the sample of sentences showed that only 46 % of all MWEs were identified automatically. This is caused by the fact that only MWEs listed in the Slovak Valency Lexicon (SVL) are detected. As a translation of the original PDT-Vallex dictionary, which only includes MWEs present in the PDT data, SVL currently has a limited coverage of MWEs. As soon as more MWEs are added into SVL, the recall of our method will improve.

type	number CPHR	number DPHR	ratio CPHR [%]	ratio DPHR [%]	precision CPHR [%]	recall CPHR [%]	precision DPHR [%]	recall DPHR [%]
newspaper	14	15	0.36	0.39	89	53	100	33
professional	28	7	0.38	0.09	95	72	100	57
fiction	24	31	0.64	0.83	91	42	88	23
overall	66	53	0.44	0.35	93	57	94	30

Table 1: Precision and recall of automatic annotation of MWEs.

6. Conclusions and Future Work

We presented a work-in-progress report of the creation of the Slovak Valency Lexicon (SVL) interlinked with the Slovak Dependency Treebank (SDT), aimed at annotating multiword entities (MWEs).

The Slovak Valency Lexicon, created by a translation of the Czech PDT-Vallex lexicon and subsequent post-processing of multiword expression entries, is considered the first successful outcome of our experiments. It contains 10 038 verbs and 741 MWE entries (261 valency frames for light verbs and 480 frames for phrasemes).

The lexicon can be further used for the purpose of contrastive analysis of syntactic and semantic properties of Slovak and Czech. The list of multiword expressions can be used to examine syntactic patterns of multiword expressions and will be used for automatic verification of the forthcoming Lexicon of Slovak Verbal Collocations.

The other outcome of this paper is the method for automatic conversion of the SDT to a deep-syntactic/semantic representation following the annotation schema of the Prague Dependency Treebank, which is specifically aimed at annotating MWEs – light verb constructions and phrasemes – using a list

[7] Our subset preserved the genre balance described in Section 2.

of MWE candidates. Our results show that with this method we can identify MWEs with very good precision.

Our further immediate plans include work on improving MWE coverage in the SVL; in particular, extending the list of MWEs and adding further features that would help for their automatic identification in the syntactic treebank. A broader aim of our research is to create a full Slovak valency dictionary with links to the Czech PDT-Vallex lexicon and to use the extracted verbal frames in compiling a collocation dictionary of Slovak verbs.

Acknowledgements

We thank Mária Šimková and Katarína Gajdošová from the Slovak National Corpus for their work on Slovak language treebank. This paper was created within the research plan of the ICT COST Action IC1207 Parseme: Parsing and multi-word expressions. Towards linguistic precision and computational efficiency in natural language processing.

References

Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J., and Žabokrtský, Z. (2012). Prague Dependency Treebank 2.5 – a Revisited Version of PDT 2.0. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 231–246, Mumbai, India. Coling 2012 Organizing Committee.

Bojar, O., Žabokrtský, Z., Dušek, O., Galuščáková, P., Majliš, M., Mareček, D., Maršík, J., Novák, M., Popel, M., and Tamchyna, A. (2012). The Joy of Parallelism with CzEng 1.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC 2012)*, pages 3921–3928, Istanbul, Turkey, May. ELRA, European Language Resources Association.

Garabík, R. and Šimková, M. (2012). Slovak Morphosyntactic Tagset. *Journal of Language Modelling*, 0(1):41–63.

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z., and Bémová, A. (1999). *Anotace Pražského závislostního korpusu na analytické rovině: pokyny pro anotátory*. Technical Report 28, ÚFAL MFF UK, Prague, Czech Republic.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolárová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, pages 57–68, Växjö, Sweden.

Hajič, J. (2004). *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Karolinum, Praha.

Popel, M. and Žabokrtský, Z. (2010). TectoMT: Modular NLP Framework. In *Proceedings of IceTAL, 7th International Conference on Natural Language Processing*, pages 293–304, Berlin - Heidelberg. Springer-Verlag.

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Academia and Dordrecht: Reidel, Prague.

Sgall, P. (1967). *Generativní popis jazyka a česká deklinace*. Academia, Praha.

Šimková, M. and Garabík, R. (2006). Синтаксическая разметка в Словацком национальном корпусе. In *Труды международной конференции Корпусная лингвистика – 2006*, pages 389–394, Sankt-Petersburg. St. Petersburg University Press.

Urešová, Z. and Pajas, P. (2009). Diatheses in the Czech Valency Lexicon PDT-Vallex. In *Slovko 2009, NLP, Corpus Linguistics, Corpus Based Grammar Research*, pages 358–376, Brno. Tribun.

Urešová, Z. (2011a). *Valence sloves v Pražském závislostním korpusu. Studies in Computational and Theoretical Linguistics*. Ústav formální a aplikované lingvistiky, Praha.

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

Žabokrtský, Z., Ptáček, J., and Pajas, P. (2008). TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Ohio. Columbus.

Zeman, D. (2008). Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 213–218, Marrakech, Morocco.