Mapping Czech and English Valency Lexicons: Preliminary Report

Vít Baisa, Karel Pala, Zdeňka Sitová, and Jakub Vonšovský

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xbaisa,pala,xsitova,xvonsovs}@fi.muni.cz

Abstract. We describe here a very first attempt to connect two valency lexicons: Pattern Dictionary of English Verbs (PDEV) and VerbaLex. Both lexicons contain verbs together with their syntactic structure (arguments of the verbal predicate) and semantic restrictions (semantic types typical for a given verb argument). The lexicons are similar in overall but differ in details since their formalisms are tailored for the respective languages. They also differ in a way they have been built: whilst the former resource has been built using Corpus Pattern Analysis methodology the latter has been built upon previous datasets Brief, Vallex and Czech WordNet. We present a preliminary work on linking English patterns in PDEV with their Czech equivalents: frames in VerbaLex.

Keywords: valency, lexicon, PDEV, CPA, VerbaLex, ontology, WordNet

1 Introduction

Valency lexicons are lexical resources containing valency frames (patterns) of individual verbs. The frames contain information about verb arguments (such as direct and indirect object, subject), their morphosyntactic properties (such as cases in Czech) and their semantic roles (such as agens, patient, instrument).

For Czech there are two valency lexicons, one is Vallex [1] by Žabokrtský (approximately 6,000 Czech verbs) based on formalism of Functional Generative Description (FGD) and VerbaLex [2] developed by Hlaváčková et al. (approximately 10,500 Czech verbs). For English we use PDEV by Hanks et al. [3].

The mentioned valency lexicons for Czech basically share the morphosyntactic information (about cases and adverbial phrases) but they differ in their inventories of the semantic roles: Vallex uses about 40 roles, VerbaLex uses complex roles consisting of the main roles (48) and selectional restrictions (900).

In the PDEV, the description of the morphosyntactic properties of the verb arguments is different from the Czech lexicons as English displays the fixed word order (SVOMPT). The semantic roles and types are based on Pustejovsky's shallow ontology containing 228 items.

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2014, pp. 139–145, 2014.

Our goal is to exploit the overall similarity (structure of frames and patterns) and propose possible equivalents of English patterns and Czech frames. In this paper we present a preliminary analysis of correspondences between the two lexical resources. We believe that the resulting translation valency dictionary would be very useful resource for natural language processing tasks, mainly for machine translation.

2 Pattern Dictionary of English Verbs

The PDEV[1] [4] is a result of a long-term work by Patrick Hanks and his colleagues. Currently, it is being developed within project *Disambiguation of Verbs by Collocation*[2] (DVC) at University of Wolverhampton.

The method of building the lexicon is based on finding corpus evidence: the English verb patterns are created only when observed in a sample of corpus examples for a given English verb. This technique of *Corpus Pattern Analysis* (CPA) was invented by Patrick Hanks [4]. The corpus used in CPA is the written part of British National Corpus.

The focus of CPA is on the prototypical syntagmatic patterns with which verbs in use are associated. Verb patterns in PDEV consist not only of the basic *argument structure* or *valency structure* of each verb (typically with semantic values stated for each of the elements), but also of subvalency features, where relevant, such as the presence or absence of a determiner in noun phrases constituting a direct object. For example, the meaning of *take place* is quite different from the meaning of *take his place*. The possessive determiner makes all the difference to the meaning in this case.

3 VerbaLex

The Czech lexicon[3] [2] has been initially based on the following resources:

1. the starting repertoire of the verbs has been taken from syntactic lexicon of verb valencies called BRIEF by Pala and Ševeček [5],
2. Czech WordNet valency lexicon developed within the Balkanet project.
3. The tool for handling the structure of the lexicon has been partially inspired by the editor developed for the above-mentioned Vallex. A new editor has been developed and is used for editing and browsing VerbaLex.

The verbs in VerbaLex are grouped into synsets in the same way as in Princeton WordNet [6]. Approximately 8,000 of them are linked to the equivalent English WordNet synsets.

[1] `http://www.pdev.org.uk`

[2] `http://clg.wlv.ac.uk/projects/DVC/`

[3] `http://nlp.fi.muni.cz/verbalex/html2/generated/alphabet/`

4 Related work

We have already mentioned Vallex as a similar resource for Czech. Framenet for English [7] should be mentioned as well though it is not only a verb lexicon. Valency lexicons are available for a number of languages: German [8], French, Italian, Russian [9], Polish [10] and others.

There have been some attempts at linking valency lexicons, e.g. [11] describes their ongoing efforts in aligning two valency lexicons PDT-VALLEX and EngValLex on the basis of a parallel treebank. The token alignment is done manually by annotators whose task is to go through the verb occurrences in the treebank, collect a typical representative of a frame mapping and control and decide potential conflicting cases. Once collected, the frame mapping is automatically applied to all its other potential representantives.

Related to our effort is also EngValLex [12]—transformation of the PropBank [13] lexicon to the structure of Vallex. After linguistic comparison of PropBank and Vallex, PropBank was automatically converted to FGD-compliant form which was later manually refined. The method is as follows: first, all slots have been renamed using functors, second, the non-obligatory free modifiers have been deleted and optional elements marked. Third, frames corresponding to the same verb sense have been merged. Fourth, the lexicon has been refined in the process of treebank annotation by addition of other frames, whole verb lemmas, and also, the PropBank adapted frames were corrected manually with respect to the language data available in the English part of parallel treebank. [11]

5 Analysis of differences and similarities

For this study, the verbs have been selected in the way that there was only one pattern in PDEV which helps the translation into VerbaLex and avoids ambiguity. There are 313 single-pattern verbs in PDEV. For some of them it is not possible to find Czech translation equivalents,[4] thus they have been left out from further analysis.

There are some features (grammatical categories) in Czech that do not have their respective counterparts in English. One of them is category of aspect: in the regular cases in which the members of an aspect pair preserve the same meaning, the category of aspect can remain in the frames, as in pair like *zrychlit, zrychlovat* (to accelerate). Similar category that should be preserved is category of case (7 grammatical cases in Czech).

Some verbs in PDEV which simply do not have direct translation equivalents in VerbaLex (*calcify, demystify, ignore,...*) are excluded from further considerations. On the other hand, there are many verbs in VerbaLex for which we cannot find the translational equivalents in PDEV because it is too small so far[5]

[4] This is caused by special terminology from very limited domains in BNC.

[5] There are roughly 1,100 completed verbs in PDEV.

or because many complex Czech verbs can not be translated on lexical level, for example: *povytáhnout* (to pull something out a bit), *poposednout si* (to move on a bit) etc.

If we look at PDEV and VerbaLex we can observe that their ontological structures are considerably different which complicates the mapping. The ontology in VerbaLex is partly based on the Top Ontology used in EuroWordNet [14] and on selected literals from Princeton WordNet. In PDEV *Shallow Ontology* by Pustejovsky [15] is used. For example, *Group* class in PDEV contains subclasses *Human Group, Vehicle Group, Animal Group, Physical Object Group* which have their own respective categories in VerbaLex. Only very few classes inherit their mapping such as PDEV *Machine* → `<artifact:1>` in VerbaLex. This means we have to uncover relations between every single class by analysing more and more words. Nevertheless, so far it seems we can go up in the classes to find a match such as for *water* which corresponds to `SUBS<liquid substance:1>` in VerbaLex and has its own class in PDEV. In one of our analyses it maps `SUBS<liquid substance:1>` to *Entity* having *Water* class as one of its descendants.

From 21 analyses, 9 patterns were mapped without any problems from one lexicon to another. 10 patterns were mapped with some imperfections such as missing frames in PDEV (for example out of 5 frames in Verbalex only 2 had a match in PDEV) or small mismatches in frames (obligatory requirement in Verbalex). Those small mismatches in frames which happened in 2 cases could be somehow penalized in an automatic tool. Only one record was unmappable (*burrow*) because frames were mismatched (Verbalex did not cover case of burrowing animals) and one record was not present at all in Verbalex (*disregard*). For some examples, see Table 1.

Table 1: Some mapping examples, PDEV on left, VerbaLex on right

Animate physical object	
Human Group | *Human*	`AG<person|child|...>`
Animal | *Bird* (all animals)	`AG<animal>`
Institution	`GROUP<institution>|AG<person>`
Precise mappings	
Machine	`ART<artifact>|INS<device>`
Body Part	`PART`
Artwork	`COM<written communication>`
Fluid | *Beverage*	`SUBS<liquid substance>`
Imprecise mappings	(one of possible mappings)
Action	`MAN,how`
Activity	`ACT<act>`
Eventuality	`Event`
Entity	`GROUP<institution>`
Physical Object	`OBJ`
Anything causes *Anything*	`REAS<reason>,díky,kvůli`

In the mapping, AG (agens) can also be PAT (patient), ENT (entity) or SOC (associate) thus *Human* would map to AG<person> as well as to PAT<person>.

Record analysis example

VerbaLex frames (VN) and PDEV patterns (PN) are in typewriter typeface. PDEV pattern implicatures are in italics.

rozmazlovat (cosset)
[V1] AG <person:1> V PAT<person:1>
[V2] AG <person:1> V PAT<person:1> (ACT<act:2>)
[P1] [[Human 1]] cosset [[Human 2]]
[[Human 1]] cares for [[Human 2]] in an excessively protective and fussy way
Comment: exact match.

zakrýt (blanket)
[V1] (překrýt) OBJ <object> V OBJ<object>
[V2] (překrýt) OBJ <object> V OBJ<object> PART<part>
[V3] AG<person> V PAT<person> (ART<covering>)
[P1] [[Stuff|{PhysicalObject1=PLURAL}]] blanket
[[Location|PhysicalObject2]]
[[Location|Physical Object 2]] becomes covered by a layer of
[[Stuff|Physical Object 1 = PLURAL]]
Comment: *Physical Object* is mapped to OBJ<object> here, but in fact it is still quite general class.

zbankrotovat (bankrupt)
[V1,2] ENT<person|instit> V (REAS<reason>díky,kvůli)
[P1] [[Human1|Instit1|Event]] bankrupt [[Human2|Instit2]]
[[Human1|Instit1|Event]] causes [[Human2|Instit2]] to not have enough money to pay his or her or its debts
Comment: *Human 2* | *Institution 2* → to <person>|<institution>. REAS<reason> maps to *A* causes *B*.

6 Discussion & conclusions

So far we have uncovered several promising relations between the two lexicons. Unfortunately, as their ontology structures are completely different, we would need to analyse tens or hundreds possible pairs to get a more complex image of possible mappings. So far from about 20 records we are already able to map animate physical objects and some of inanimate objects which together form the biggest group in PDEV. This is a basis for further investigation and for a rule-based approach to proposing and linking possible equivalents from PDEV and VerbaLex. The main problems consist of the following:

1. the ontologies used in PDEV and VerbaLex are structured differently, there is a shallow ontology in PDEV and a sort of the Aristotelian ontology based on Top Ontology from EuroWordNet in VerbaLex.
2. Basic items in VerbaLex are synsets containing usually more than one verb lemma, whereas in PDEV the basic items are the individual verb lemmas. This, however, can be handled by obtaining appropriate lists from VerbaLex (by expanding and filtering the verb list).

The comparison of the two ontologies is a separate task that should be further investigated and deserves a separate paper.

We hope that in the near future we will be able to propose and implement an automatic tool with high accuracy of PDEV patterns translations of VerbaLex frames and vice versa.

Acknowledgement This work was partially supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013 and by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.

References

1. Žabokrtský, Z., Lopatková, M.: Valency information in vallex 2.0. The Prague Bulletin of Mathematical Linguistics (87) (2007) 41–60
2. Hlavácková, D., Horák, A.: Verbalex–new comprehensive lexicon of verb valencies for czech. In: Proceedings of the Slovko Conference, Citeseer (2005)
3. Hanks, P., Pustejovsky, J.: A pattern dictionary for natural language processing. Revue française de linguistique appliquée **10**(2) (2005) 63–82
4. Hanks, P.: Lexical Analysis: Norms and Exploitations. Mit Press (2013)
5. Pala, K., Ševeček, P., et al.: Valence českých sloves. (1997)
6. Fellbaum, C.: WordNet. Wiley Online Library (1998)
7. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics (1998) 86–90
8. Hinrichs, E.W., Telljohann, H.: Constructing a valence lexicon for a treebank of german. In: Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories. (2009) 41–52
9. Lyashevskaya, O.: Bank of russian constructions and valencies. In: LREC. (2010)
10. Przepiórkowski, A.: Towards the design of a syntactico-semantic lexicon for polish. In: Intelligent Information Processing and Web Mining. Springer (2004) 237–246
11. Šindlerová, J., Bojar, O.: Building a bilingual vallex using treebank token alignment: First observations. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). (2010)
12. Cinková, S.: From propbank to engvallex: Adapting the propbank-lexicon to the valency theory of the functional generative description. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'06). (2006)
13. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: An annotated corpus of semantic roles. Computational Linguistics **31**(1) (2005) 71–106

14. Vossen, P.: A multilingual database with lexical semantic networks. Springer (1998)
15. Rumshisky, A., Hanks, P., Havasi, C., Pustejovsky, J.: Constructing a corpus-based ontology using model bias. In: FLAIRS Conference. (2006) 327–332