NomVallex: Valency Patterns of Semantically Classified Czech Nouns

Veronika Kolářová, Jana Klímová and Anna Vernerová
(Charles University, Faculty of Mathematics and Physics, Czech Republic)

1 Introduction

Corpus-based models for describing the syntactic environments of individual lexical items play an important role in corpus lexicography, as demonstrated for example by the corpus-driven approach to the lexical grammar of English called Pattern grammar (Hunston & Francis, 2000)) or the Corpus Pattern Analysis of Patrick Hanks (2013). While many valency lexicons are primarily intended for non-native speakers (e.g., Herbst et al. 2004), nouns are also covered in lexicons created mainly with NLP applications in mind, such as FrameNet[1] and NomBank 1.0[2]. Corpus-driven approach to valency of Czech nouns was applied by Čermáková (2009). Currently, valency patterns of Czech nouns are in focus of a new lexicographic project titled Corpus-based Valency Lexicon of Czech Nouns (Klímová, Kolářová, & Vernerová, 2016), using an acronym NomVallex[3].

2 The development of NomVallex

NomVallex is a project building upon the theory of valency developed within Functional Generative Description (Sgall, Hajičová, & Panevová, 1986) and extending two existing valency lexicons developed within this tradition, Vallex (a valency lexicon of Czech verbs; Lopatková et al., 2015, 2016) and PDT-Vallex[4] (containing valency patterns of verbs, nouns, adjectives and adverbs as they occurred in the Prague Dependency Treebank, Prague Czech-English Dependency Treebank and Prague Treebank of Spoken Czech; Hajič et al., 2003). Nouns to be included in NomVallex are selected based on the complexity of their valency patterns, special valency behaviour (e.g., special forms of participants, cf. Kolářová, 2014) and semantic class membership. Valency properties are captured in the form of valency frames for each meaning (lexical unit) of nouns included, and an enumeration of combinations of adnominal participants representing various valency patterns, as extracted from Czech corpora.

Currently, NomVallex follows the Vallex annotation scheme. Vallex was chosen as the base for the NomVallex project because it provides semantic class membership (Kettnerová, Lopatková, & Hrstková, 2008) and valency patterns for all meanings (i.e. lexical units)[5] of verbs included, while PDT-Vallex covers only the lexical units that were encountered in the data of the treebanks in the Prague Dependency Family. Where possible, NomVallex maps nominal lexical units to their source verbal lexical units contained in Vallex. Technically, adopting Vallex software, NomVallex can serve as a supplement to Vallex, providing not only nominal entries but also links between the two parts of speech. Links to PDT-Vallex data are also recorded wherever they can be established.

---

[1] https://framenet.icsi.berkeley.edu
[2] http://nlp.cs.nyu.edu/meyers/NomBank.html
[3] https://ufal.mff.cuni.cz/grants/nomvallex
[4] http://hdl.handle.net/11858/00-097C-0000-0023-4338-F
[5] Aspectual pairs (e.g., perfective *odpovědět* 'to answer'and imperfective *odpovídat* 'to answer') are usually treated within one lexeme (headword) in Vallex.

3 Extraction of valency patterns from Czech corpora

Czech is a highly inflectional language; valency participants of a word are primarily distinguished by their morphological category of case while the word order is very flexible, especially concerning valency participants expressed by prepositional phrases (PPs). Typically, an adnominal participant can be expressed by at least two forms (variants); in general, almost no combination of variants can be excluded. Therefore, searching for valency patterns of Czech nouns usually means searching for many various combinations of forms, including word order variants.

The following Czech lemmatized and morphologically annotated corpora are used: the synchronic part of the Czech National Corpus (CNC)[6], the web corpus Araneum Bohemicum Maximum[7] and the Prague Dependency Treebank (PDT 3.0)[8]. The PDT 3.0 is a small but manually syntactically annotated corpus, providing also semantic roles assigned to particular nodes (Mikulová et al., 2006). Using the CNC and the Araneum corpus, valency patterns of Czech nouns are being extracted either with the help of Sketch Engine's Word Sketches (Kilgarriff & Tugwell, 2001), or by sophisticated CQL queries specified in the KonText application[9]. Searching through the PDT is carried out by the tool called PML-TQ (Štěpánek & Pajas, 2010).

4 Semantic classes in NomVallex and a preliminary list of entries

Nouns representing five semantic classes are included in NomVallex, namely Communication (e.g. *odpověď* 'answer'), Exchange (e.g. *dodávka* 'delivery'), Contact (e.g. *dotyk* 'touch'), Mental action (e.g. *dojem* 'impression'), and Psychological nouns (e.g. *obava* 'fear'). The assignment of semantic class is carried over from Vallex: a noun is supposed to be assigned the same semantic class as its source verb in Vallex, with the exception of nouns that undergo a change in meaning. On the basis of the list of verbs in Vallex (see Table 1 for numbers of lexical units representing particular semantic classes), a preliminary list of noun entries was created. We aim to provide valency patterns of all types of Czech nouns with a meaning denoting an action or an abstract result of an action. These nouns are either derived from verbs by productive means (suffixes *-(e)ní/tí*, as in *vykládání* 'explaining // unloading' or *pojetí* 'conception') or by non-productive means including the zero suffix (such as *vykládka* 'unloading', *výklad* 'explanation / interpretation'). The preliminary list of candidate entries to be included in NomVallex currently contains 1230 lemmas, cf. Table 2.

	Communication	Exchange	Contact	Mental action	Psychological verbs	Total
Verbs in Vallex	428	182	125	338	143	1216

Table 1: Number of verbal lexical units in Vallex

[6] http://korpus.cz/
[7] http://ucts.uniba.sk/aranea_about/index.html
[8] http://ufal.mff.cuni.cz/pdt3.0
[9] http://wiki.korpus.cz/doku.php/en:manualy:kontext:index

	Communication	Exchange	Contact	Mental action	Psychological nouns	Total
Productively derived nouns	335	171	117	257	104	984
Non-productively derived nouns	110	38	14	56	28	246
Total	445	209	131	313	132	1230

Table 2: Number of lemmas of nouns included in the NomVallex preliminary list of entries

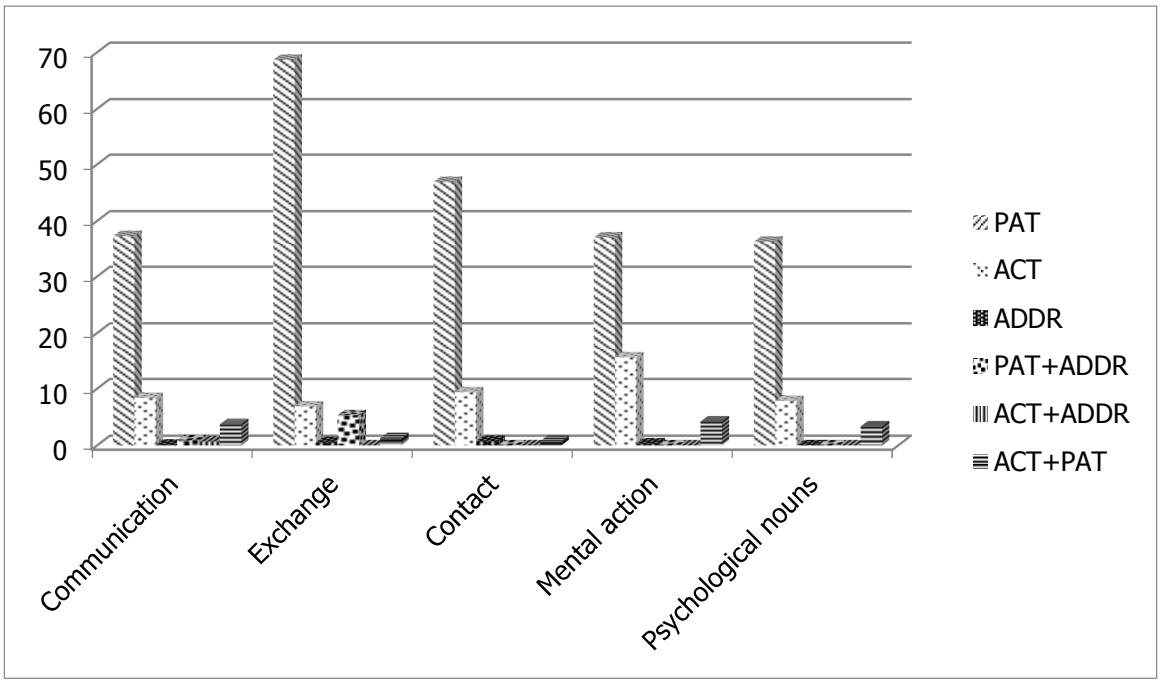
5 A quantitative analysis of combinations of participants in the PDT

As the first step, we carried out a quantitative analysis focusing on relative frequencies of combinations of participants modifying nouns representing the five selected semantic classes in the PDT 3.0. 623 such lemmas occurred in the PDT 3.0 in a total of 8273 occurrences (see Table 3).

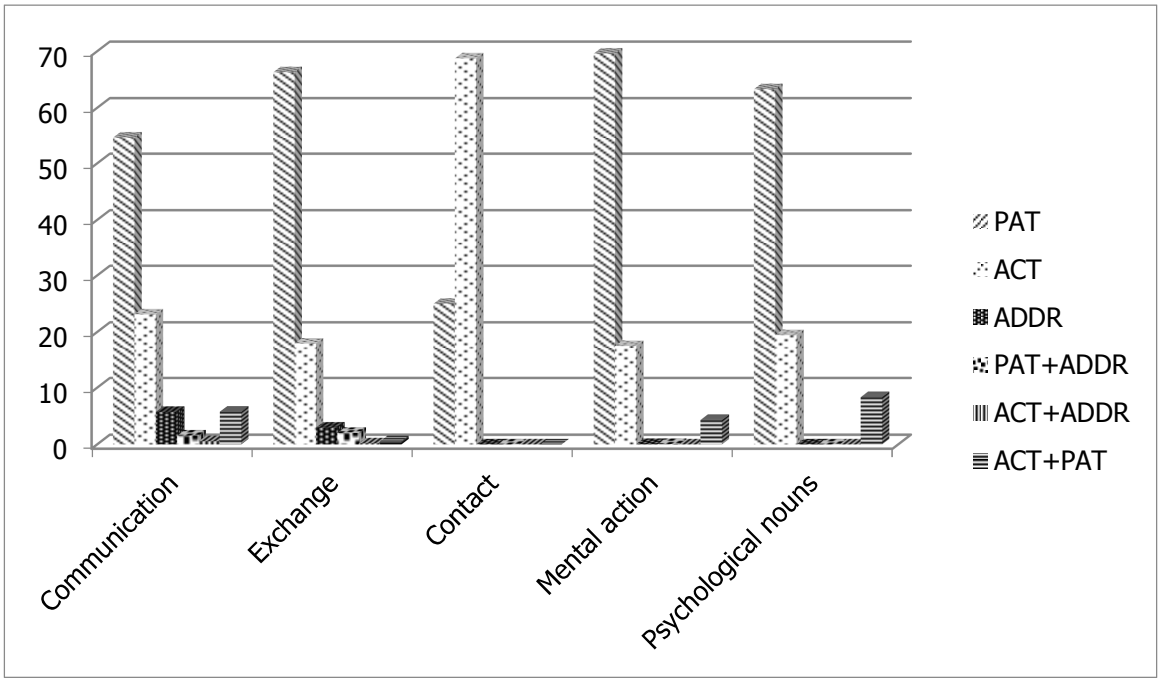
		Communication	Exchange	Contact	Mental action	Psychological nouns	Total
Productively derived nouns	Lemmas	145	94	30	107	29	405
	Occurrences	1552	699	128	1236	179	3794
Non-productively derived nouns	Lemmas	102	34	10	54	18	218
	Occurrences	2163	540	16	1256	504	4479
Total	Lemmas	247	128	40	161	47	623
	Occurrences	3715	1239	144	2492	683	8273

Table 3: Number of lemmas and occurrences of nouns found in the PDT 3.0

The Graphs 1 and 2 show that the most frequent combination is the case when only Patient (PAT) is expressed (with the exception of non-productively derived nouns of Contact which represent the least frequent class and so the numbers may be influenced by their rare occurrence). The case when only Actor (ACT) is expressed is the second most frequent combination, followed by the combinations Actor + Patient or Patient + Addressee, the latter of which is applicable only in the case of nouns that have an Addressee (ADDR) in their valency frame. Interestingly, relative frequencies of the combination Actor + Patient are very low with nouns of Exchange and nouns of Contact. Relative frequencies of combinations of three participants – no more than 0,13% – are not shown in the Graphs.



Graph 1: Relative frequencies of selected combinations of participants modifying productively derived nouns in the PDT 3.0



Graph 2: Relative frequencies of selected combinations of participants modifying non-productively derived nouns in the PDT 3.0

6 Valency patterns of Czech nouns based on the CNC and the Araneum corpus

The current work focuses on the extraction of all possible forms of noun participants and their combinations from the CNC and the Araneum corpus, concentrating on nouns of Communication. The extracted corpus data give evidence about the following phenomena:

(i) Nouns derived from perfective verbs by productive means show slightly different valency behaviour than nouns derived from the corresponding imperfective verbs, even when both types of nouns denote an action.

(ii) There is a strong tendency for non-productively derived nouns for an increased number of possible expressions of their participants in comparison with their source verbs, cf. valency frames for the verb *apelovat* 'to appeal' and the noun *apel* 'appeal' illustrated in (1) and (2). Interestingly, despite the higher number, some of the forms are shared between particular participants, such as the prepositional phrase *k* 'to'+DAT shared between ADDR and PAT in (3) and (4).

(1) apelovat 'to appeal'
ACT(NOM;obl) ADDR(*na* 'at'+ACC;obl) PAT(content_clause;obl)

(2) apel 'appeal'
ACT(GEN, possessive;obl) ADDR(DAT, *k* 'to'+DAT, *na* 'at'+ACC;obl) PAT(*k* 'to'+DAT, *proti* 'against'+DAT, content_clause,inf;obl)

(3) apel k lidem.ADDR 'an appeal to people'

(4) apel k ukončení násilí.PAT 'an appeal to end the violence'

7 Conclusion

We believe our description of valency patterns of Czech deverbal nouns representing the five semantic classes will result in a valuable source of information, facilitating a detailed comparison of valency patterns of Czech nominal and verbal lexical units, providing also information about word-formation relations and semantics, including shifts in meaning.

Acknowledgements

The research reported in the paper was supported by the Czech Science Foundation under the project GA16-02196S. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

Čermáková, A. (2009). *Valence českých substantiv*. Praha: Lidové noviny.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., & Pajas, P. (2003). PDT-VALLEX: Creating a Largecoverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*. Vaxjo University Press, 57-68.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press.

Herbst, T., Heath, D., Roe, I. F., & Götz, D. (2004). *A valency dictionary of English: a corpus-based analysis of the complementation patterns of English verbs, nouns, and adjectives*. Berlin: Walter de Gruyter.

Hunston, S., & Francis, G. (2000). *Pattern grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins Publishing.

Kettnerová, V., Lopatková, M., & Hrstková, K. (2008). Semantic Classes in Czech Valency Lexicon: Verbs of Communication and Verbs of Exchange. In *Lecture Notes in Computer Science, Vol. 5246, Proceedings of the 11th International Conference, TSD 2008*, 109-116. Berlin / Heidelberg: Springer.

Kilgarriff, A., & Tugwell, D. (2001). WORD SKETCH: Extraction and display of significant collocations for lexicography. In *Proc Collocations workshop*, ACL2001, Toulouse, France, 32-38.

Klímová, J., Kolářová, V., & Vernerová, A. (2016). Towards a Corpus-based Valency Lexicon of Czech Nouns. In I. Kernerman et al. (Ed.), *Globalex 2016, Lexicographic Resources for Human Language Technology*, 1-7. Available at: http://ailab.ijs.si/globalex/files/2016/06/LREC2016Workshop-GLOBALEX_Proceedings-v2.pdf.

Kolářová, V. (2014). Special valency behavior of Czech deverbal nouns. In O. Spevak (Ed.) *Noun Valency*, Amsterdam: John Benjamins, 19-60.

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., & Žabokrtský, Z. (2015). *VALLEX 3.0 – Valenční slovník českých sloves*. Charles University in Prague, [online] http://ufal.mff.cuni.cz/vallex/3.0/.

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., & Žabokrtský, Z. (2016). *Valenční slovník českých sloves VALLEX.* Praha: Karolinum.

Mikulová, M. et al. (2006). *Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual*. Technical Report TR-2006-30, Praha: ÚFAL MFF UK.

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

Štěpánek, J., & Pajas, P. (2010). Querying Diverse Treebanks in a Uniform Way. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA), 1828-1835. Malta: Valletta.