An interface between the Czech valency lexicon PDT-Vallex and corpus manager KonText

Abstract

We present a user interface between the Czech valency lexicon, PDT-Vallex [1], and KonText[1] – a web application for querying corpora available within the LINDAT/CLARIN project. KonText allows evaluation of simple and complex queries, displaying their results as concordance lines, computing frequency distribution, calculating association measures for collocations and further work with language data. For every verb in a concordance line, our interface allows to display information concerning its valency frame in a separate window if corresponding entries exist in PDT-Vallex, as well as a list of possible valency frames for that particular verb. Information concerning verb frame comprises verb lemma, frame elements with semantic roles, vocabulary-style description and examples from PDT-Vallex, Prague Dependency Treebank [2] and Prague Czech-English Dependency Treebank [3]. The information is requested by REST-API from the valency lexicon PDT-Vallex that contains over 11000 valency frames for more than 7000 verbs which occurred in Prague Dependency Treebank, Prague Czech-English Dependency Treebank or Prague Dependency Treebank of Spoken Czech. We use a fork of KonText application (developed by the Institute of the Czech National Corpus) that has been further extended by the Institute of Formal and Applied Linguistics to suit the needs of LINDAT/CLARIN project. The plugin we present provides a unique solution for Czech language that integrates valency information from the Czech valency lexicon with the means of querying Prague Dependency Treebank.

References

[1] Z. Urešová, J. Štěpánek, J. Hajič, J. Panevova, and M. Mikulová, "PDT-vallex: Czech valency lexicon linked to treebanks." LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

[2] E. Bejček, E. Hajičová, J. Hajič, P. Jínová, V. Kettnerová, V. Kolářová, M. Mikulová, J. Mírovský, A. Nedoluzhko, J. Panevová, L. Poláková, M. Ševčíková, J. Štěpánek, and Š. Zikánová, "Prague dependency treebank 3.0," 2013. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[3] J. Hajič, E. Hajičová, J. Panevová, P. Sgall, O. Bojar, S. Cinková, E. Fučíková, M. Mikulová, P. Pajas, J. Popelka, J. Semecký, J. Šindlerová, J. Štěpánek, J. Toman, Z. Urešová, and Z. Žabokrtský, "Announcing prague czech-english dependency treebank 2.0," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, (İstanbul, Turkey), pp. 3153–3160, ELRA, European Language Resources Association, 2012.

[1] `https://ufal.mff.cuni.cz/lindat-kontext`