4FX: Automatic Detection of Light Verb Constructions in a Multilingual Corpus

Anita Rácz[1], István Nagy T[1], Veronika Vincze[2]

[1]University of Szeged, Department of Informatics
raczanita89@gmail.com, nistvan@inf.u-szeged.hu
[2]Hungarian Academy of Sciences, Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu

In this paper we describe the 4FX corpus, the first English, Hungarian, Spanish and German parallel corpus, which is manually annotated for light verb constructions (LVCs). For corpus construction, legal texts from the JRC-Acquis legal parallel corpus were selected. Annotation principles and statistical data on the corpus are also provided, and data for the four different languages are contrasted. We also present the results of a machine learning-based approach that allows us to identify light verb constructions in free texts. The tool was originally implemented to automatically detect Hungarian and English LVCs. However, we were able to easily adapt this data-driven machine learning-based approach to the other languages, since manually annotated corpora are also available in Spanish and German in the 4FX corpus. Moreover, we were able to define language-specific features, like the gender of the noun in Spanish and German, for the machine learning-based method to detect LVCs in free texts in these different languages. Our applied method proved to be sufficiently robust, since it outperformed our dictionary labeling baseline method in the case of all the four different languages.