Van's upon a Time: Copulas in Dependency Grammars

Katalin Ilona Simkó[1], Veronika Vincze[1,2]

[1]University of Szeged, Department of Informatics
Szeged, Árpád tér 2.
kata.simko@gmail.com
[2]MTA-SZTE Research Group on Artificial Intelligence
Szeged, Tisza Lajos körút 103.
vinczev@inf.u-szeged.hu

Most linguistic phenomena do not have one uniform analysis that describes them perfectly, rather there are multiple frameworks offering a number of different approaches each. This paper aims to investigate the different ways the Hungarian copula *van* is described in dependency syntax.

The *van* copula is not only analysed multiple ways in theoretical syntax, but also in computational syntax [1]. The Hungarian verb *van* – as well as many other equivalent verbs in the languages of the world – has an existential as well as a copular use. The existential is used just like any other main verb, expressing being somewhere or in a certain state. The copular *van* makes up the predicate of the sentence together with a nominal predicate, but in Hungarian, the nominal predicate alone is present in the surface structure in present tense, third person sentences, the verb is absent, which is problematic for the syntactic analysis.

There are three different approaches in computational dependency syntax to describe Hungarian *van*. The function head analysis is the original annotation of the Szeged Dependency Treebank [2]; it treats all different types of *van* the same way: it is always the head of the sentence, when it is not present, a virtual node is inserted into the structure manually to take its place. The content head approach distinguishes the existential and the copular uses of *van*; the existential *van* is the head just like all other main verbs, while the copular verb is never the head: the nominal predicate is. The complex label analysis does not use virtual nodes either, but instead it marks the missing verb on the label of the nominal predicate head.

As all three annotations are available on the same text, the Népszava part of Szeged Corpus [3], we could see how each of them perform under the same conditions. We used the Bohnet parser [4] with all three analyses to measure their performance in ULA and LAS, as well as manual error analysis focusing on the errors related to the copular structure.

We found that the content head analysis works best for parsing copular structures, as it does not require manual insertion of virtual nodes, like the function head analysis and does not make the analysis overly complicated, like the complex label approach.

References

1. Simkó, K.I.: Magyar kopulás szerkezetek az elméleti és a számítógépes szintaxisban. Master's thesis, Szegedi Tudományegyetem (2015)
2. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010, Valletta, Malta, ELRA (2010)
3. Vincze, V., Varga, V., Simkó, K.I., Zsibrita, J., Nagy, Á., Farkas, R., Csirik, J.: Szeged Corpus 2.5: Morphological Modifications in a Manually POS-tagged Hungarian Corpus. In: Proceedings of LREC 2014, Reykjavik, Iceland, ELRA (2014) 1074–1078 ACL Anthology Identifier: L14-1241.
4. Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). (2010) 89–97