Correcting and Enlarging the Slovak UD Treebank

Vladimír Benko

Comenius University in Bratislava, UNESCO Chair in Plurilingual and Multicultural Communication (Bratislava, Slovakia)
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics (Bratislava, Slovakia)

Key-words: Slovak Dependency Treebank, annotation errors, data-type balance

Introduction

Slovak belongs to languages with a large morphologically tagged corpus[1] usable as data for training taggers, referred to as 'Manually Morphologically Annotated Corpus' (MMAC). This resource is continually being maintained and corrected. Its current size is almost 1.2 million tokens, and its latest Version 6.0 has been completed in 2017.

There are, however, several issues related to this dataset.

(1) The texts to be processed were extracted from the Slovak National Corpus (SNC) without taking care about the user license, and several parts of the data seem to be covered by the copyright. This resulted to situation that, as a whole, the dataset cannot be published under any type of open license.
(2) As the respective metadata related to the user license has been deleted after extracting the respective texts from the SNC, a special procedure of aligning them is to be written to find out what is in fact the proportion of the copyrighted material.
(3) Some linguistic decisions related to capitalization of lemmas, lemmatization of collective nouns, tagging of proper nouns, etc., are incompatible with those taken within a parallel project of the Slovak morphological database, which poses a problem if a language model for a tagger considering both training corpus and morphological database is to be produced.
(4) A considerable amount of inconsistencies still exists in the data, especially in lemmatization and PoS assignments of rare phenomena (typos, foreign words, symbols, etc.)

There is, unfortunately, almost no documentation available about the process of creation and/or current state of this resource.

The Slovak Syntactic Treebank

Somewhere in the middle of the MMAC Project an offspring initiative has been launched targeted at creation a syntactically annotated subset of MMAC according to Prague Dependency Treebank (PDT) methodology and tools (Šimková & Garabík, 2006).

Slightly more than 50 % of the MMAC contents (35,000 sentences, 570,000 tokens) have been annotated by two annotators according the guidelines from the PDT Manual (Hajičová, E. & Sgall, P., 1999) but, for various reasons, the process was not completed, and the resulting treebank was not published.

---

[1] https://korpus.sk/en/corpora-and-databases/snc-corpora/r-mak-versions/

About ten years later, the data in a form 'as is' has been sent to Dan Zeman in Prague, who converted it to the UD format and performed a validation. Part of the data with 100% identical annotations has been extracted and published in 2016 as Slovak SNK UD at the UD Portal. Since then, only minor maintenance operations have been performed, mostly as a result of improvements of the UD validation procedures.

What is the problem

The resulting Treebank currently cosists of 10,604 sentences containing 106,043 tokens, which is, at least at a first sight not bad. There are, however, several issues related to it.

(1) The decision that identically annotated sentences should be considered resulted in having mostly very short sentences present in the Treebank only, which limits is usability as a training resource.
(2) The data is rather unbalanced from the text type perspective, containing almost 80% of fiction texts.
(3) The data still contain many PoS tag and lemmatization errors, that have been already fixed in MMAC
(4) The copyright status of some texts has to be checked.

What We Want to Do

Having very good experience with using work of graduate students of linguistics in crowdsourcing (Benko, 2018), that resulted in significant improvements of the Slovak morphological lexicon by manually annotating corpus out-of-vocabulary (OOV) lexical items, we would also like to apply this method to addressing some of the issues mentioned above. In the first round, we want to correct annotation errors by concentrating on sentences with only minimal differences in annotations. Later, we want to improve the balance of the data by adding more sentences belonging to non-fiction and media data types extracted from our Araneum Slovacum (Benko, 2016) web-crawled corpus.

The crowdsourcing phase is to start at the beginning of the next school year.

What Do We Expect from the Training School

One of the decisions to be made is what tools should we use for correcting the existing Slovak treebank. While using the PDT tools (such as TrEd[2]) could save data in the original format, using newer tools might be more user-friendly, i.e., needing less training for the annotators. We hope to by able to take the decision after getting acquainted with the tools presented in the framework of the School.

A Final Remark

I understand that (at least 'technically') I am not eligible for funding my presence at the School. On the other hand, as the acquired information is to be passed to our students, I believe that the 'eligibility spirit' could be observed 😊

[2] https://ufal.mff.cuni.cz/tred/

References

Benko, V. (2016). Feeding the 'Brno Pipeline'. In Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016, pp. 19–27, 2016. © Tribun EU 2016

Benko, V. (2018). Crowdsourcing for the Slovak Morphological Lexicon. In S. Krajci (ed.): ITAT 2018 Proceedings, pp. 126–129, CEUR Workshop Proceedings Vol. 2203, ISSN 1613-0073.

Gajdošová, K., Šimková, M., et al. (2016). Slovak dependency treebank. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University. Accessible at: http://hdl.handle.net/11234/1-1822.

Hajičová E. & Sgall P. (eds.) (1999). Anotace na analytické rovině. Návod pro anotátory. Praha, 1999. Accessible at: http://ufal.mff.cuni.cz/pdt

Šimková, M. and Garabík, R. (2006). Синтаксическая разметка в Словацком национальном корпусе. In Труды международной конференции Корпусная лингвистика – 2006, pp. 389–394, St. Petersburg University Press, Russia.

Zeman, D. (2017). Slovak Dependency Treebank in Universal Dependencies. In Journal of Linguistics/Jazykovedný časopis, 68(2), DOI: 10.1515/jazcas-2017-0048.