Automatic extraction of data: Slovenian case revisited

Iztok Kosem, Simon Krek, Polona Gantar

At SKEW3, we presented a new Sketch Grammar and GDEX configurations for Slovene, designed specifically for automatic extraction of corpus data (from the 1,18-billion-word Gigafida corpus), which were to be used for the purposes of devising entries for the Slovene Lexical Database. A key part of the automatic extraction was an API script, written in the months following SKEW3 in collaboration with the Sketch Engine team, which was later used not only for the purposes of Slovene Lexical Database but also in a terminological project called TERMIS, the latter also enabled us to test the API script on a smaller and specialized corpus and to test automatic extraction of data for multi-word items.

In our presentation, we will focus on the process of automatic extraction used in the Slovene Lexical Database project and to some extent in the TERMIS project, and its evaluation. This will include a closer look at the API script, its versions and development according to the needs of different projects, as well as the role of sketch grammar and GDEX in this process. A part of presentation will be dedicated to a more detailed overview of how we determined the settings for extraction, e.g. thresholds for the extraction of grammatical relations and collocates.

The evaluation of the extracted data consisted of two parts: firstly, a selection of automatic entries were cleaned and completed with missing information (e.g. sign posts, meaning frames), as to get a comparison between the time spent writing a manual entry and a (semi)-automatic entry. Secondly, the automatic extraction was taken under scrutiny, in order to identify potential improvements in the API script and its components and in the presentation of data, and to document the processes involved in the clean-up of data.

We will present the findings of the evaluation, paying particular attention to the potential improvements to word sketch data (and relatedly, sketch grammar). For example, our analysis showed that some collocates with high frequency were sometimes not exported because of their low salience, which was either below the threshold limit or was so low that the collocates were not found among the first X collocates that were exported per relation (in our case, the default setting was usually 25). Consequently, for collocates in certain grammatical relations a score that would take into account collocate position on both raw frequency and salience lists may be the best way to identify, and order, the relevant candidates.

We will conclude the presentation by outlining future work, including improvements to the API script and the plan to use automatic extraction of corpus data in the making of a new dictionary of contemporary Slovene.