Czech-Russian and Czech-Slovak Parallel Corpora via a Simple Web Interface

Radovan Garabík, Natalia Klyueva, Petra Galuščaková, Ondřej Bojar and Miroslav Týnovský

We describe Czech-Slovak and Czech-Russian parallel corpora that were initially created as training data for Machine Translation systems. Those corpora have been available in a format suitable for the computer processing, but theoretical linguists interested in the resources would not benefit much from them. So we decided to put the corpora into a user-friendly environment.
They are now accessible via a simple www interface, based on the Manatee backend. Both parts of each corpus can be queried using full CQL syntax with regular expression based search of wordforms, lemmas and POS/morphological tags. A simplified search is provided if a non-CQL entry is detected, providing a simple intuitive regular expression search of selected attributes (wordform and lemma). The results are displayed in parallel side-by-side sentence aligned tabular format.

The texts for the Czech-Russian-English parallel corpus were downloaded from the Project Syndicate page. The corpus contains above 100000 sentences. We are currently enlarging the corpus with the books from online libraries.

Czech-Slovak parallel corpus was compiled from several freely available data sources. More than 5,5 million sentences were collected and morphologically annotated. Similarly to Czech-Slovak corpus, English-Slovak corpus was compiled using the same sources.