Corpus Architect
Abstract

Jan Pomikálek

Faculty of Informatics, Masaryk University, Brno, Czech Republic
xpomikal@fi.muni.cz

Abstract. In the Corpus Architect project we are developing a system for creating text corpora, which will be easy to use even for non-technical users. The system creates corpora from two data sources – users' own text documents and web pages. Users can upload their texts in various formats or ask the system for adding domain specific web texts into their corpora in an automated way. In the latter case the domain of the web texts is defined in terms of key words.

Once all the desired texts are collected for the corpus, with a single click users can have the texts part-of-speech tagged, lemmatized and loaded into the Sketch Engine corpus manager. With the corpus manager, users can instantly make use of fast searching in the corpus. They also immediately get access to important corpus derived statistical information, such as word sketches and statistical thesaurus.

The interface of the Corpus Architect is designed with an emphasis on simplicity and easiness of use. The user is asked for as little input as possible. No decisions are requested for options if the value can be detected in an automated way or a reasonable default can be used.

A live demo of the system will be presented. However, as long as this is a work in progress and many features are still unimplemented, only the basic functionality will be shown.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, pp. 10–10, 2008. © Masaryk University, Brno 2008