Development of large-scale lexicons and lemmatizers for English and German using finite-state and constraint-grammar technology

The Department of Computational Linguistics at the HHU Düsseldorf, in collaboration with ComNet GmbH, Würselen (leading partner), TU Dresden (Institute of Acoustics and Voice Communication) and Cognesys GmbH, Aachen, develop software for automatic filtering of text and video content on the Internet. The aim is to provide an application that goes beyond building blacklists or identification of single keywords and understands and evaluates the complete content of a webpage. The project *Jugendschutz im Internet. Development of Innovative Technologies for the Identification of Internet Content that is Harmful to Adolescents* is funded by the Bundesministerium für Wirtschaft und Technologie (BMWi).

The main module of the application will be based on *Cognitive Ergonomic Solution* (CES) of Cognesys GmbH, a universal speech and text interface for automatic acquisition und further processing of the meaning of spoken and written information.

The performance of CES can be significantly enhanced if the input is annotated with lemmata and relevant morphosyntactic information. Our contribution to the project will be to develop broad-coverage lemmatizers for German and English.

Each parser will consist of the following language-specific modules: tokenizer, sentence splitter, normalizer, morphological analyzer (lexicon) and simple phrase tagger. The last module is needed for the disambiguation of homonymous forms belonging to different part-of-speech classes (e.g., *books* (noun plural) and *books* (verb, 3rd person singular, present tense)) and for linking split verb forms in German and phrasal verbs in English in order to provide the correct lemma (e.g., *aufhören* for *Hör auf!*).

In our talk we will present not only the lemmatizers that are being developed for the project but also how they can be extended to include grammatically relevant semantic information.