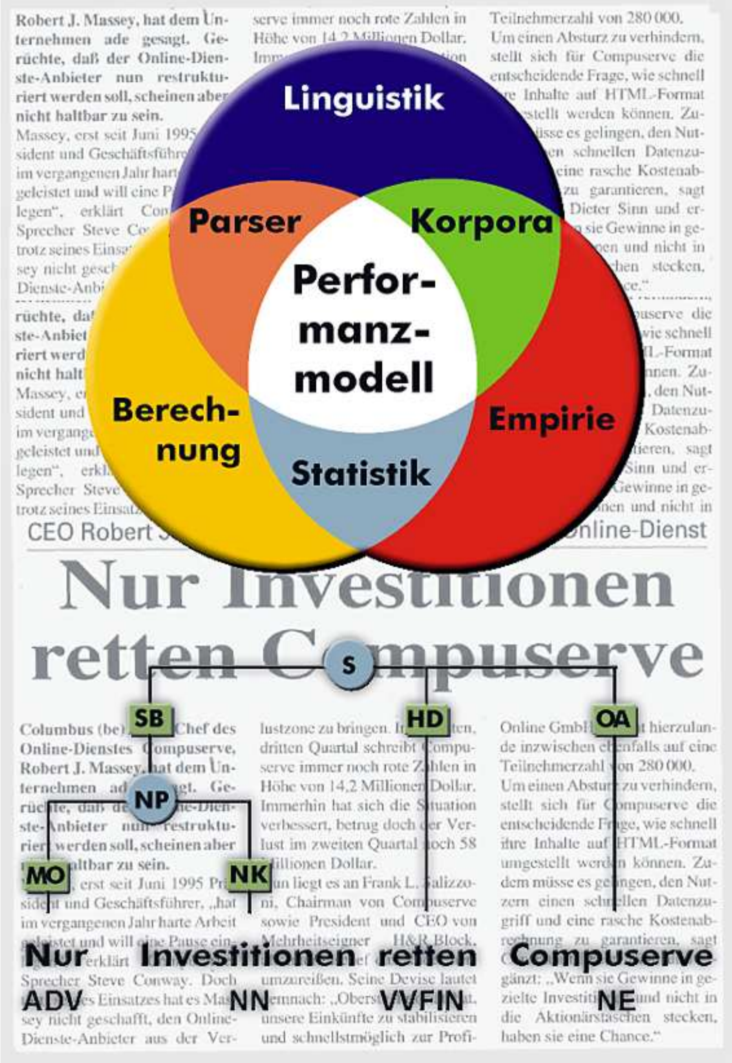
NEGRA: Concurrent Grammar Processing



NEGRA is a collaborative project involving researchers in computational linguistics and computer science. The project aims to develop hybrid technologies for modeling human language processing. Research builds on methodologies previously investigated independently, with the goal of developing new techniques which combine the advantages of these approaches.

NEGRA combines modern linguistic theory, large amounts of real linguistic data, and a range of computational methods. Large collections of natural language text (corpora) are combined with their linguistic interpretations to provide an empirical foundation for research. The corpora provide rich analyses of everyday language, and also supply information about the frequency with which various linguistic phenomena occur. Modern statistical language processing technologies exploit such frequency information to automatically learn language in terms of statistical regularities. Once trained, these systems can deal accurately and robustly with ambiguous and previously unseen sentences. The richness and complexity of human language also requires the use of sophisticated constraint-based parsing systems, which exploit linguistic knowledge, derived from current linguistic theories. To enable this, we are investigating concurrent processing techniques which permit efficient understanding of language via quasi- parallel processing.

The key question in our research is how to combine rich constraint-based systems and robust statistical processing techniques in a way which best capitalizes on the strengths of each. The integration of these paradigms promises to form the basis for the next genaration of speech and language processing technologies, and is therefore the central focus of NEGRA.

To support the combination of linguistic, constraint-based and statistical approaches, an important result was the development of the first German linguistically analysed corpus. The NEGRA Corpus currently consists of approximately 20,000 newspaper sentences taken from the Frankfurter Rundschau, and it continues to grow. The linguistic analysis of the corpus was generated semi-automatically using techniques developed within the project. They are part of a boot-strapping process, enabling our research on automatic learning, the development of robust statistical parsing techniques, and models of human language use, in the SFB and many other projects.

Prof. Dr. Werner H. Tack
SFB 378 – der Sprecher
Universität des Saarlandes
Postfach 151150
D-66041 Saarbrücken

URL: http://www.coli.uni-sb.de/info/projects/negra.html
Contact: Prof. Dr. Hans Uszkoreit
Telephone: +49-681-302-4115
Telefax: +49-681-302-4700
email: hansu@coli.uni-sb.de