# The Statistical Approach to Natural Language Processing: Achievements and Open Problems

**Hermann Ney**
RWTH Aachen University, Aachen
DIGITEO Chair, LIMSI-CNRS, Paris

## Abstract

When IBM research presented a statistical approach to French-English machine translation in 1988, the linguistic community was shocked because this approach was a hit in the face of the then received machine translation theories. Since then we have seen a dramatic progress in statistical methods for speech recognition, machine translation and other tasks in natural language processing. This talk gives an overview of the underlying statistical methods. In particular, the talk will focus on the remarkable fact that, for all these tasks, the statistical approach makes use of the same four principles:

- Bayes decision rule for minimum error rate,
- probabilistic models, e.g. Hidden Markov models or conditional random fields, for handling strings of observations (like acoustic vectors for speech recognition and written words for language translation),
- training criteria and algorithms for estimating the free model parameters from large amounts of data,
- the generation or search process that generates the recognition or translation result.

Most of these methods had originally been designed for speech recognition. However, it has turned out that, with suitable modifications, the same concepts carry over machine translation and other tasks in natural language processing. This lecture will summarize the achievements and the open problems in this area of statistical modelling.