

# Machine Translation in the Czech Republic: History, Methods, Systems

Jan Hajič

*Institute of Formal and Applied Linguistics  
Faculty of Mathematics and Physics, Charles University  
Malostranské nám. 25, 118 00 Prague 1  
Czech Republic*

Jan.Hajic@ff.cuni.cz

## Abstract

First experiments with machine translation were carried out in Prague in the late fifties, even before a research group of mathematical linguistics had been established. Larger scale research was started in the late sixties: a series of English to Czech translation experiments in the domain of technical abstracts then dominated this field throughout the seventies and the first half of the eighties. During the “perestrojka” years the group started large collaborative project on Czech to Russian machine translation in the domain of software manuals. After 1989, this project was abandoned; however, in 1993, the English/Czech project was revived (albeit on a smaller scale) using a partly different approach.

The emphasis in all machine translation systems has always been put on the analysis side, as expected. However, the point has not been the parsing procedure as such, but rather the representation theory having been developed in the group since its beginnings. Based on the Functional Generative Description, with a dependency-based (linguistic) meaning representation theory, the machine translation systems served as a “real-life” testbed for its various aspects. Nevertheless, several interesting NLP MT-related techniques have been discovered: so called fail-soft techniques for unparsed input, making the systems more robust; the transducing dictionary, for dictionary-less translation of words of Latin origin; the idea of keeping ambiguities across languages; and several simplifications made possible by the closeness of Czech and Russian (in the Czech to Russian system).

Currently, a new set of experiments is being carried out to explore the possibilities of combining linguistic and stochastic methods in the area of machine translation (between Czech and English). These experiments try to find out whether an automatic extraction of correspondences on parsed data is possible, and to what extent. The underlying parsers, albeit more shallow than the previous ones and quite differently implemented, again use the aforementioned dependency representation, which is believed to localize relations among sentence elements which, in turn, makes a more extensive use of statistical methods possible.