Title: **Evaluation of Error Mark-Up in a Learner Corpus of Czech**

Author: Barbora Štindlová

Department: Institute of Czech Language and Theory of Communication, Faculty of Arts, Charles University in Prague

Supervisor: prof. PhDr. Karel Šebesta, CSc.

Abstract:

The thesis deals with the topic of Czech as a second language, while introducing methods of corpus linguistics as applied to texts produced by language learners. The context is the process of building and exploiting a learner corpus, with a focus on its error mark-up and options for evaluating the annotation scheme.

Learner corpora have become a major resource for investigating a learner interlanguage and a significant incentive for many different types of research and teaching of second/foreign languages. They are used mainly for contrastive studies of native and non-native speakers, i.e. for contrastive interlanguage analysis, and for computer-aided error analysis of the learner language. This kind of analysis is crucially dependent on the type and quality of the error mark-up. In every error-annotated corpus the error annotation is based on an error typology, which is necessarily problematic from a number of theoretical aspects. Evaluation of the reliability and validity of the annotation scheme design is therefore an important step in the build-up of a learner corpus.

The thesis is concerned primarily with the technical aspects and specific issues involved in the digitization of hand-written texts, with options for the error annotation of non-native speakers' language, and with the issues of its evaluation. At the same time, a significant amount of space is devoted to the questions of methodology, architecture and purpose of the compilation of learner corpora, because the topic of a non-native speakers' corpus and its exploitation in the Czech environment is quite recent and thus a more detailed introduction is justified.

In the first part (A), several major approaches to the issues of foreign/second language acquisition are briefly summarized and the developments in the theory of error in non-native speakers' language are presented in more detail. In part B, a summary of the current state of the field is presented together with an overview of existing corpora of non-native speakers' language, the result of a questionaire-based research and a detailed analysis of available learner corpora. The third part (C) presents a learner corpus of non-native speakers' Czech (CzeSL), focusing on the issues of text transcription. In the fourth part (D), the error annotation scheme proposed for CzeSL is subjected to evaluation. To assess the reliability of the annotation scheme a measure of inter-annotator agreement – the coefficient kappa – is used. The measured results of the inter-annotator agreement, the analysis of the problematic points in the annotation scheme, and the evaluation of the scheme, including the error taxonomy, represent some of the main assets of the present thesis.

Keywords: learner corpus, error annotation, text transcription, inter-annotator agreement, interlanguage