PRACE FILOLOGICZNE

Warszawa 2012 TOM LXIII

FRANTIŠEK MARTÍNEK
Institute of Czech Language and Theory of Communication
Charles University in Prague, Faculty of Arts
KATEŘINA RYSOVÁ
Institute of Formal and Applied Linguistics
Charles University in Prague, Faculty of Mathematics and Physics

ON A CORPUS OF OLDER CZECH TEXTS AND ITS USAGE[1]

SŁOWA KLUCZOWE: korpus diachroniczny, język czeski XVI wieku, transliteracja, szyk wyrazów, swobodne modyfikacje wyrazowe

KEYWORDS: diachronic corpus, humanistic Czech, transliteration, word order, free verbal modifications

1. Introduction

In the first part of the paper, the principles of the Corpus of Humanistic Czech[2] are presented, with the focus on the linguistic aspects of the building of the corpus and searching within the corpus. The selected method of transcription is discussed and the theoretical reasons for it are given. Furthermore, the selection of the texts included is explained.

[1] This paper was supported by the doctoral grant project No. 16 809 *Linguistic Analysis of Czech Humanistic Texts (Lingvistická analýza českých humanistických textů)*, financed by the Grant Agency of Charles University, by the grant projects No. 405/09/0729 *From the Structure of a Sentence to Textual Relationships (Od struktury věty k textovým vztahům)*, No. P406/12/0658 Coreference, discourse relations and information structure in a contrastive perspective (*Koreference, diskurs a aktuální členění v kontrastivním pohledu*), financed by the Grant Agency of the Czech Republic, and by the project LINDAT-Clarin LM2010013.

[2] The adjective *humanistic* does not refer to any special text types but to the first period of Middle Czech; the second one is called Baroque Czech (1620–1780).

In the second part, the results of an analysis of word order in humanistic Czech, based on the Corpus of Humanistic Czech, are presented, and in this way, exploration possibilities of this corpus are shown with regard to syntactic phenomena. In particular, we explored the position of obligatory verbal modifications, which are dependent on verbs of the type *zacházeti, nakládati s něčím* ('to deal with something').

2. Corpus of Humanistic Czech

2.1. General principles

The Corpus of Humanistic Czech was created as the material base of the *Linguistic Analysis of Czech Humanistic Texts* project in 2011. It will be included in the diachronic part of the Czech National Corpus as a guest corpus.

It is a balanced (see 2.2) electronic corpus of texts printed from 1500 to 1620. It includes more than half a million word forms and consists of approximately 50 texts and extracts of longer texts. In March 2011, all the texts were digitised (by students) and approximately two thirds have been already checked. The purpose of the balanced corpus is first of all didactic: it is intended for morphological, syntactical and lexical analyses for use on university bachelors' and masters' courses of the Czech language.

The texts included are transcribed, i.e. transformed into the Modern Czech orthographical system with respect to phonetical and phonological peculiarities of the older language.

Any problematic or questionable phenomena in texts where the script enables two (or more) interpretations are recorded in transliteration. The main reason for this method is that transliteration does not enable any effective searching in electronic texts (cf. Kučera 1998: 306f.). Moreover, the Czech 16th century orthography – the so-called *bratrský pravopis*, "Czech Brethren orthography" – is heterogeneous, and using transliteration, one has to solve analogous problems as with transcription.

Annotation tags are embedded in the text under similar principles which are used in the Czech diachronic corpus DIAKORP. The only difference is caused by the fact that lemmatization of the corpus is not planned. Irregular word forms are thus marked and complemented by the default form, which enables the user