Project Description:
Designing and Constructing a Typologically Balanced Ukrainian Text Database

Emmerich Kelih (Graz, Austria),
Solomija Buk (Lviv, Ukraine),
Peter Grzybek (Graz, Austria),
Andrij Rovenchak (Lviv, Ukraine)

0. INTRODUCTION

The present contribution reports about a bilateral Austrian-Ukrainian research project. Starting with 2009, this project is financially supported by the Austrian Agency for International Cooperation in Education and Research (ÖAD), in the framework program “Scientific and Technological Co-Operation” and the Ministry of Education and Science of Ukraine (Project No. M/6-2009). Project partners are the Institute for Slavic Studies at Graz University (Emmerich Kelih, Peter Grzybek), the Institute for Theoretical Physics at Lviv University (Andrij Rovenchak), and the Institute for General Linguistics at Lviv University (Solomija Buk).

The major objective of this project is the design and construction of a typologically balanced text database for Ukrainian. With regard to future linguistic analyses, directed to the quantitative study of Ukrainian texts and language, special attention is paid to the inclusion of individual texts, rather than building a comprehensive corpus of heterogeneous language material. Particular attention is paid to the balanced inclusion of texts of various functional styles and text types, in order to cover the broad spectrum of text genres.

1. SHORT OVERVIEW: OBJECTIVES AND AIMS

For any empirical and/or statistical analysis of language and text the availability of digitized text material (language corpora) is an indispensable precondition. Especially in the field of corpus linguistics great efforts have been made to develop and build text archives, text collections and
general language corpora. For Ukrainian\(^1\), however, there are only a few projects pursuing this direction, and they shall be initially described here.

Two projects of general language corpora of Ukrainian are built independently at the Institute of Ukrainian Language (National Corpus of the Ukrainian Language) and at the Ukrainian Language-Information Fund (Ukrainian National Linguistic Corpus); both Institutes are parts of the National Academy of Sciences of Ukraine. Unfortunately, none of these two projects is available via Internet. National Corpus of the Ukrainian Language is planned as a balanced corpus, aiming to cover all text genres, and containing one million word occurrences (Dems’ka-Kul’čyc’ka 2005). The Ukrainian National Linguistic Corpus is designed as a non-balanced corpus, constantly being enlarged by texts from various domains (Šyrokov et al. 2005), and presently containing ca. 60 million word occurrences.

Some parallel corpora of Ukrainian are represented in the Internet. The project of the Ukrainian-Polish Corpus\(^2\) is carried out at the Institute of Slavic Studies of the Polish Academy of Sciences (Warsaw). It covers 70 \((2 \times 35)\) texts, which are aligned at the paragraph level. The Ukrainian-Russian parallel corpus\(^3\) of web publications uses automatic alignment by keywords.

Several projects from the domain of author lexicography are known for specific Ukrainian writers. The text corpus of Taras Shevchenko, a great Ukrainian poet and artist, was created in Canada for the compilation of the concordance of Shevchenko’s poetry (Ilnytzkyj/Hawrysch 2001). The same team under the supervision of O. Ilnytzkyj is presently working on the text corpus of H. Skovoroda, a Ukrainian poet and philosopher. The corpus of works of V. Shevchuk, a Ukrainian writer, was recently designed as a research tool for the study of key concepts in his works (Monakhova 2006).

---


\(^{2}\) http://corpus.domeczek.pl.

\(^{3}\) http://ling.infostream.ua.
The corpus of Ivan Franko’s texts is planned to cover all his works (estimated size: ca. seven million word occurrences), representing the Western variant of the Ukrainian language from the turn of the 20th century (Buk 2007). The morphological tagging of some works has already been accomplished; thus far, this work has provided an online concordance4 of Perekhresni stezhky (The Cross-Paths), a novel by Ivan Franko; on this basis, some statistic features of Franko’s text have successfully been studied (Buk/Rovenchak 2007a; 2007b).

What is problematic with such corpora is the fact that analyses done on their basis (collocation analysis, word form frequency, etc.), mostly ignore the specifics of individual texts, since the latter are not accessible as such.

From a quantitative linguistics perspective, however, the availability of semantically coherent and closed texts is of high importance and represents a primary interest. The analysis of individual texts can be seen as an attempt to minimize “internal” textual heterogeneity, characteristic of corpus analyses (cf. Altmann 1982; Kelih 2009, this volume). As Ju.K. Orlov has convincingly shown, particular quantitative regularities (e.g., the Zipf Law), hold true only for complete, “closed”, semantically coherent texts; as compared to this, text mixtures (and any corpus is such a mixture!), represent some kind of a quasi text, for which a harmonious behavior of frequency and length behavior cannot be expected.

With regard to the fact that for many Slavic languages there are no systematically constructed database, from which individual texts would be available, work in this direction has started in 2002 in the framework of the Graz project on Quantitative Text Analysis (QuanTA), to fill this gap. Up to now a text database, containing individual texts form different text types and functional styles (for details see below) for Slovene, Croatian, Serbian, Russian and Slovak with over 6000 individual texts, has been collected and pre-processed. With the present project, concentrating on Ukrainian only, another East Slavic language shall be added. The extension to Ukrainian is important for various reasons: irrespective of the fact that such a desirable text database has been generally missing, far-reaching options and possibilities will be provided for literary, linguistic, stylistic etc. analyses, particularly in a comparative perspective, with regard to other Slavic languages.

4 Available at: http://www.ktf.franko.lviv.ua/~andrij/science/Franko/concordance.html.
1.1. Detailed Project Description

The project’s program includes two major tasks:
1. Construction of a web based database of Ukrainian texts, on the basis of a typologically structured and balanced design;
2. Annotation and Tagging of the chosen texts, partly supported by relevant software.

In this context, it should be emphasized that the planned text database does not include the naïve claim for representativeness of Ukrainian as a literary or standard language. Representativeness is one of the greatest illusions in linguistics, not only in corpus linguistics. Rather the intention is to represent the broad spectrum of different text types and functional styles typical for a given language. Given the experience from intensive work on word length frequencies in various Slavic languages – see, among others, Grzybek/Kelih/Stadlober 2005, Grzybek/Stadlober/Kelih/Antić 2005, Kelih/Antić/Grzybek/Stadlober 2005) – it has turned out to be efficient, to motivate the selection of texts for the database with regard to their attribution to a particular text type and functional style.

1.2. Text Type and Functional Style

From a theoretical point of view, the concepts of functional style and text type are of crucial relevance with regard to the structure of the planned text database:

a. With reference to the achievements of the Prague School and other research in this direction, qualitatively different functional styles can be distinguished. The distinctions are made on the basis of general communicative (socially defined) functions of language and texts. Usually, a diverging number from five to eight functional styles are postulated, depending on the concrete theoretical position taken.

b. With regard to empirical research in the domain of text types, the whole spectrum of text types can be understood as a maximum differentiation of communicative utterances. This involves an enormous amount of pragmatic-communicative text situations, and results in an inventory of ca. 4000 text types which are assumed to play a crucial role in a given culture or language (cf. the bibliographical data given by Adamzik 1995).

It is hardly possible, of course, to take into consideration all text types when constructing a balanced text database. Having in mind sys-
tematic statistical analyses, however, this is not even necessary; rather, it must be guaranteed that the whole spectrum of different stylistic shapes is adequately represented. In this respect, a combined selection of different text types from all functional styles seems to be a pragmatic solution. Table 1 represents a schema with all functional styles and specific characteristic text types within each of them; this schema may serve as a starting point for the intended construction of a typologically balanced text database.

It seems reasonable that the schema represented in Table 1 is an adequate and sufficient orientation, and can serve as a basis for quantitative and statistical analyses. As a *minimal condition*, two different text types (boldfaced in Table 1) will be chosen for each of the seven functional styles, each of them represented by 30 individual texts. This results in a minimal number of 420 texts \((2 \times 7 \times 30)\) as a lower limit for systematic analyses.

### 1.3. Project Management and Work-Flows

The course of the project includes two phases, each of them planned to cover one year.

**Phase I:** The first steps are directed at text acquisition and selection. These steps include the analysis and documentation of available sources (internet, CD collections, etc.). In particular, usage of online versions of newspapers\(^5\) and of electronic news portals\(^6\) is planned, as well as of official web-ages of the Ukrainian Parliament\(^7\), some Ukrainian Churches\(^8\), freely available academic journals, electronic libraries of the Ukrainian literature\(^9\), etc. For some text types, it may be additionally necessary to process electronic texts by way of OCR software.

The next steps include the unification of different character encoding, the unified attribution of meta data (author, time, title, text type, functional style) for each single text.

**Phase II** primarily includes the preparation of tagging procedures, i.e. the text-specific treatment and/or annotation of headlines, chapter

---

\(^5\) [http://www.day.kiev.ua](http://www.day.kiev.ua), [http://www.zn.kiev.ua](http://www.zn.kiev.ua), [http://www.wz.lviv.ua](http://www.wz.lviv.ua), etc.


\(^7\) [http://www.rada.kiev.ua](http://www.rada.kiev.ua).


headings, abbreviations, numbers (years), foreign words, etc. which might influence text analyses unless carefully integrated. Additionally, a unified treatment and annotation of sentence boundaries, of direct and indirect speech, of quotations, etc. is necessary here. Finally, all the text have to be migrated to and incorporated into a central database structure; so they can be made publicly available to the academic community (given there are no copyright conflicts).

Table 1: Functional Styles and Text Types

<table>
<thead>
<tr>
<th>everyday style</th>
<th>scientific</th>
<th>administrative</th>
<th>journalistic</th>
<th>prose</th>
<th>poetry</th>
<th>drama</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td><strong>Private Letter</strong></td>
<td>Abstract</td>
<td>Instruction</td>
<td>News Agency communication</td>
<td>Autobiography</td>
<td>Elegy</td>
<td>Drama</td>
</tr>
<tr>
<td><strong>Diary Entry</strong></td>
<td><strong>Articles Social sciences</strong></td>
<td><strong>Business Letter</strong></td>
<td>Report</td>
<td>Biography</td>
<td>Epos</td>
<td>Comedy</td>
</tr>
<tr>
<td><strong>Natural sciences</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Joke</strong></td>
<td>Legal Text</td>
<td>Professional Article</td>
<td>Epistolary Novel</td>
<td>Poem</td>
<td>Tragedy</td>
<td></td>
</tr>
<tr>
<td><strong>Cooking Recipe</strong></td>
<td>Diploma Thesis</td>
<td>Expertise</td>
<td>Feuilleton</td>
<td>Epilogue</td>
<td>Ode</td>
<td>Drama in Verse</td>
</tr>
<tr>
<td><strong>...</strong></td>
<td>Ph.D. Thesis</td>
<td>Political Convention</td>
<td>Comment</td>
<td>Memoirs</td>
<td>Sonnet</td>
<td><strong>...</strong></td>
</tr>
<tr>
<td><strong>Presentation</strong></td>
<td>Sermon</td>
<td>Column</td>
<td>Short Story</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Review</strong></td>
<td>Contract</td>
<td>Reader’s Letter</td>
<td>Fable</td>
<td>Novel in Verse</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conference Report</strong></td>
<td><strong>Open Letter</strong></td>
<td><strong>Sport Reportage</strong></td>
<td>Parable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>...</strong></td>
<td>Resolution</td>
<td>Weather Report</td>
<td>Fairy Tale</td>
<td><strong>...</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Parliament speech</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>Novel</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>...</strong></td>
<td></td>
<td></td>
<td>Legend</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>...</strong></td>
<td></td>
<td>Myth</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Diary Novel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>...</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. RESULTS (EXPECTED)

The first concrete project result will be the text database itself, providing an adequate material basis for systematic stylistic, linguistic, statistical, etc. text analyses. The typologically balanced design, the pre-processing of all texts, and the specific annotations are essential characteristics of this database. In fact, a systematically constructed and publicly available collection of Ukrainian texts will be available in electronic form, representing the necessary pre-condition for automatic, typologically specific analyses of Ukrainian texts. In the long run, these analyses will provide theoretical insight into the synergetic functioning of texts and languages, and with its focus on Ukrainian, it will contribute to a deeper understanding of language-specific factors.

As to future perspectives, one should first think, by way of an example, of analyses of the Ukrainian writing system (cf. Grzybek/Kelih 2005, Buk/Mačutek/Rovenchak 2008); in case an automatic grapheme-phoneme (letter-allophone) conversion will be available, a reasonable next step would be a (comparative) analysis of the Ukrainian phonological system and phoneme frequencies. Another reasonable field of application can be seen in lexical and syntactic studies, including word length, sentence length, and the relation between sentence length and word length (Arens Law, Menzerath-Altmann Law) in various discourse styles (cf. Buk/Rovenchak 2008, Grzybek/Kelih/Stadlober 2008). When then the technical conditions will be fulfilled, and the sources will be available, as has been described above, it will be easy to systematically pursue these questions and, to be sure, quite a few others …

REFERENCES


Методи аналізу тексту

Methods of Text Analysis

Чернівці
Чернівецький національний університет
2009
METHODS OF TEXT ANALYSIS:

The omnibus volume provides results of various approaches to text analysis and related issues in Western and Eastern Europe, USA and China. Special attention is paid to quantitative methods in text analysis.

Editors:
Emmerich Kelih (Graz, Austria)
Viktor Levickij (Černivci, Ukraine)
Gabriel Altmann (Lüdenscheid, Germany)

Gedruckt mit der Unterstützung der Karl-Franzens-Universität Graz.

ЗМІСТ/CONTENTS

Художественное слово и поэтическая картина мира

Николай Алефиренко 3

Weight Factor Formalisms in the Study of Lexical Growth: The Case
of Textually Modelled Strings of English Verbs

Michael Bilynsky, Andriy Pereymybida, Gabriel Altmann 13

Дискурсивные и текстовые категории: к разграничению понятий

Лилия Безуглая 40

Rhythmishe Einheiten in Hülsen, Natur-Betrachtungen (1800)

Karl-Heinz Best 53

Hapax Legomena and Language Typology, a Case Study

Fan Fengxiang 63

Zur Homogenität von Graphemhäufigkeiten in Texten: Evidenz aus
dem Russischen

Emmerich Kelih 85

Slawisches Parellel-Textkorpus: Projektvorstellung von „Kak zakalja-
las’ stal” (KZS)“

Emmerich Kelih 106

Project Description: Designing and Constructing a Typologically Bal-
anced Ukrainian Text Database

Emmerich Kelih, Solomija Buk, Peter Grzybek, Andrij Rovenchak 125

A Densitometric Classification of Web Template Content

Christian Kohlschütter 133

Перлокутивний ефект висловлення-звинувачення

Антоніна Король 156

Слова-консоциации и побочный смысл: забытые идеи Г. Шпербе-
ра и К.О. Эрдмана

Виктор Левицкий 165

Arc length development and the highest word frequency

Ján Mačutek 182
Narrative psychological study of self and object representation with young deviant people
*Bernadette Péley, János László*

A modified text indicator
*Ioan-Iovitz Popescu, Gabriel Altmann*

Фоносемантический анализ текста (на материале русских и английских звуко-цветовых ассоциаций)
*Лариса Прокофьева*

Фоносемантический анализ текста
*Виктор Левицкий*

Теоретико-методологические предпосылки анализа текста англоязычной шутки
*Виктория Самохина (Дмитренко)*

Length and Style in Slovak
*Zuzana Venitová and Emília Nemcová*

The Well-Formedness of Two Psychoanalytic Word Categories in Portuguese Texts
*Andrew Wilson*

An integral qualitative-quantitative approach to the study of concept realization in the text
*Nadia Yesypenko*

Linguistics as a diagnostic tool for textual ineptness: Narrative perspective in Zamjatin's *Navodnenie*
*Olga T. Yokoyama*
Наукове видання

Методи аналізу тексту: Збірник наукових праць

Літературні редактори Колодій О.В.
Комп’ютерний набір та верстка Мацкуляк Ю.Й.


Реєстраційне свідоцтво ДК №891 від 08.04.2002 р.
Підписано до друку . Формат 60 х 84/16.
Папір офсетний. Друк офсетний. Ум. друк. арк.
Обл.-вид. арк. . Зам. . Наклад

Друкарня Чернівецького національного університету
58012, Чернівці, вул. Коцюбинського, 2.