

# Corpora for Lithuanian Language Technologies

Rūta Marcinkevičienė, Agnė Bielinskienė, Vidas Daudaravičius, Erika Rimkutė

Vytautas Magnus University  
Kaunas, Lithuania  
ruta@hmf.vdu.lt

## Abstract

The paper overviews the development and the present state of the Lithuanian language corpora during the last decade and their application for a number of fields of language research and product development. The main focus is on the problematic issues of compilation, annotation and presentation of the large general corpus, nevertheless, smaller specific corpora, such as morphologically tagged and parallel corpora are discussed.

## 1. Introduction

A corpus is conceived here as a large collection of electronic texts compiled with the general purpose of representing a sub-language at a certain period of time or the language itself. Any corpus comprises two equally important parts: collection of texts and tools for their processing. The types of corpora, defined on the bases of several oppositions (such as large vs. small, balanced or representative vs. unbalanced, dynamic vs. static, unannotated vs. annotated) reveal a wide range of possible corpora and their applications. The paper presents several types of corpora of the Lithuanian language widely used for HLT.

## 2. The Corpus of the Contemporary Written Lithuanian Language

The Corpus of the Contemporary (comprising the period since 1990) Written Lithuanian Language compiled at Vytautas Magnus University can be described as a large (more than 100 million running words in its

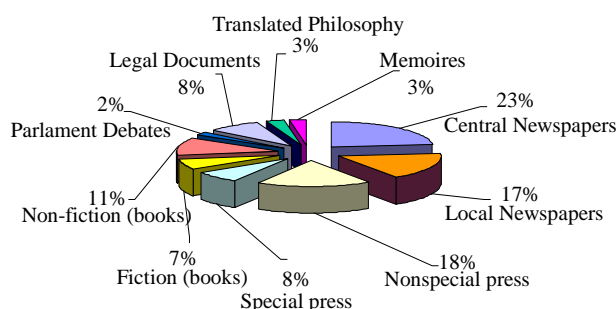
Internet version), balanced, dynamic, i.e. constantly being enlarged and updated, unannotated corpus of full texts. A yearly increase in size is ca 10 million running words.

### 2.1. The Structure of the Corpus

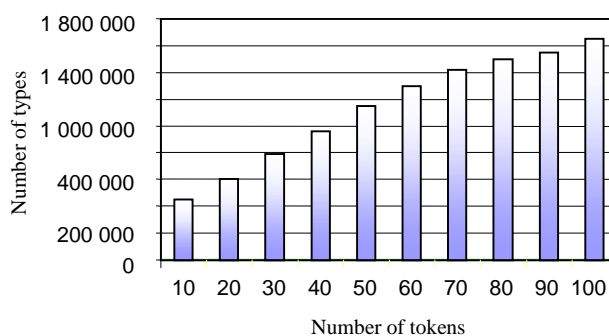
The corpus is balanced from the perspective of the production and readability of texts. It is designed as a general purpose monitor corpus aimed at various applications.

The overall structure of the corpus is reflected in picture 1. A considerable part of the corpus is made up by press materials since they are abundantly produced and consumed. Fiction and non-fiction, legal documents and memoirs are major groups of different genres. All the texts are original writings except for translated philosophy and Parliament debates. The latter component comes from the spoken variety, however, is considerably transformed while editing. Therefore it is included into the corpus of the written language.

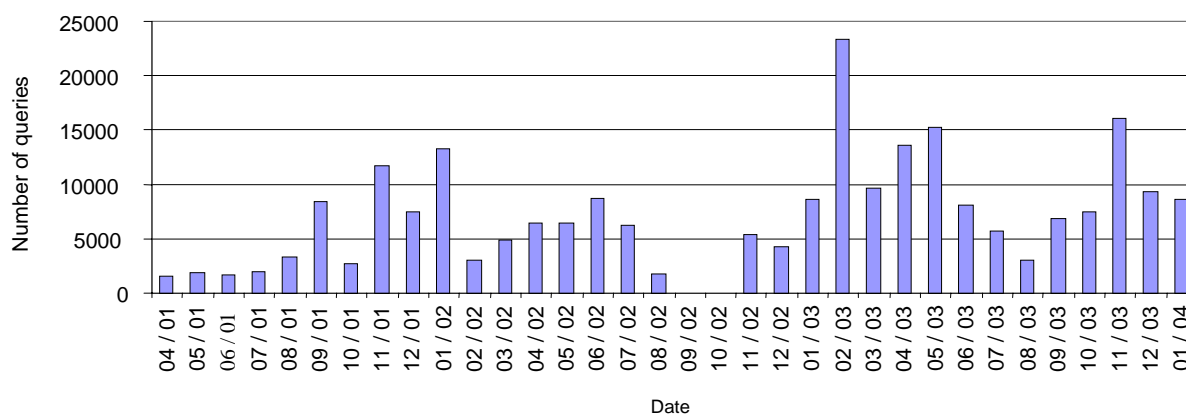
The present design of the corpus represents the chosen variety of the contemporary Lithuanian language. This fact is proved by the saturation curve (see picture 2), which shows



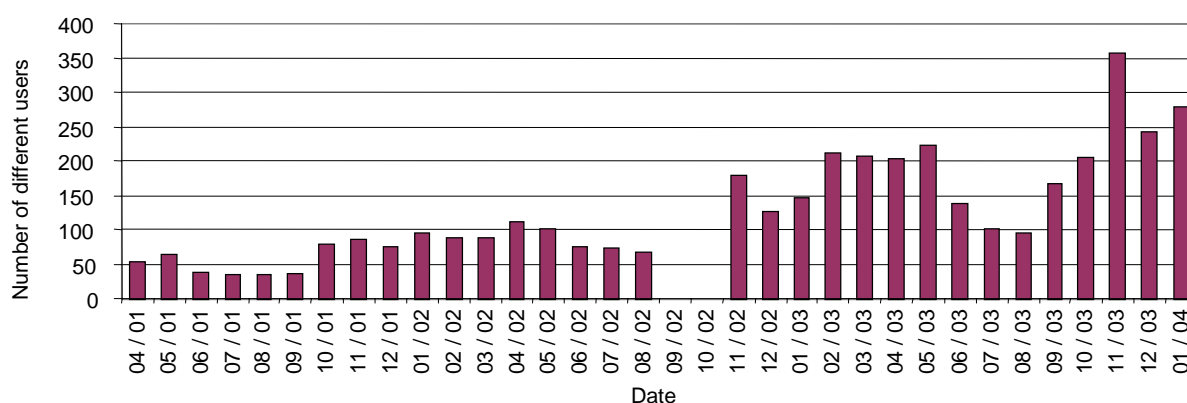
**Picture 1.** The structure of the Corpus (tokens – 102,909,906; types – 1,699,560)



**Picture 2.** Saturation of the Corpus



Picture 3. The usage of the Corpus on WWW



Picture 4. The users of the Corpus on WWW

that the increase of tokens does not influence markedly the variety of types any longer. Only changes in design, e.g. addition of the spoken component, would influence the number of types.

## 2.2. Tools and Sub-products

The tools used for the management and query of this corpus are both universal and language specific. The latter includes tools made at VMU. The corpus is compiled with the help of Collector, the tool for extracting newspaper articles from www (Daudaravičius, 2000). The Internet version of the corpus is supplied by Internet Corpus Administration Kit, which comprises registry of new texts, search-indexes, and corpus exploration tools. Subsequent analysis can be performed with the help of Lexicographer, a tool for classification and statistical analysis of concordances. The most recent tool of the Kit is meant for the detection of collocations and specification of their boundaries (Daudaravičius, Marcinkevičienė 2004).

The sub-products directly obtained from the corpus include regularly updated word frequency lists, concordances and the list of collocations for every query. The last two are available on Internet (<http://donelaitis.vdu.lt>). The corpus is frequently queried from Lithuania and abroad. Pictures 3 and 4 reflect the number of queries and the number of different users, respectively, per month. In average every user presents 40 queries. There are ca 250 users of the Corpus at present thus ca 9000 queries per month. The users inside the country are the Parliament, research institutes and universities, and some private companies.

The next stage in the development of the corpus beside its increase in size and variety would be automatic morphological and syntactic annotation as well as more sophisticated query system.

## 3. The Annotated Corpus

Morphologically annotated corpus of more than 1 million running words is compiled using a pattern of design similar to the large

corpus. It is made up exceptionally of full texts.

### 3.1. The Type and Tool of the Annotation

The annotated corpus was tagged automatically with the help of morphological analyzer, lemmatizer and tagger Lemuoklis (Zinkevičius, 2000). A word or word form is characterized grammatically by a combination of properties with respect to 13 categories: part of speech, aspect, reflexiveness, voice, mood, tense, group, degree, definiteness, gender, number, case and person, e.g. see the annotation of most frequent notional word in the corpus *Lithuania*:

```
<word="LIETUVOS" lemma="Lietuva" type="tikr dktv  
mot.gim vnsk K">
```

The tool processes over 30,000 tokens per minute on ATX Intel Celeron (433 MHz, 64 MB RAM). It uses morphological rules expressed in digital tables and word-root lexicons that enable us to analyse milliards of theoretically available words forms. However, theoretically possible and actually used numbers of word forms differ considerably. Great number of theoretically possible readings increases ambiguity and demands careful manual processing.

Thus automatic analysis of the corpus was followed by manual processing in order to disambiguate homoforms. The amount of ambiguous cases in the corpus was 47%, the amount of the cases unidentified by the tool was 11 % (Rimkutė 2003).

### 3.2. Feedback from the Corpora for Tools

Lemuoklis, as it was originally designed, was mostly based on existing grammars and dictionaries and did not take into account either context information and word frequencies or their semantics. The corpus applications of the tool and its feedback enable creation of algorithms for the reduction of ambiguity. Moreover, inadequacies of morphological interpretation and classification, inherited from the traditional grammars and dictionaries, are solved applying the corpus data. Finally, systemic knowledge about the correlation between word groups and their relevant morphological features is added to Lemuoklis. Extended and improved, Lemuoklis forms the bases for the

Lithuanian spell checker, produced by Fotonija.

Another tool that benefits from both annotated and unannotated corpora is morphological analyzer and synthesizer MorfoLema. It is based on the authentic data, e.g. the list of adjectives that can be derived by adding suffixes to certain nouns. The tool was incorporated in the machine translation system *Česilko*, designed at the institute of Formal and Applied linguistics (ÚFAL) of Charles University in Prague and applied among others for the Czech-Lithuanian languages.

Last but not least, the data from the corpora is used for the currently designed parser. The dictionary of syntactic combinability or valency is mostly based on the corpora since currently available dictionaries provide its compilers with severely limited data (e.g. the dictionary of verbal valency includes only 2000 verbs out of 40,000).

## 4. The Parallel Corpus

The Parallel Corpus of the Lithuanian and some other languages (Czech, English, and German) consists of two parts (each ca 1 million running words in size): Czech – Lithuanian and Lithuanian – Czech corpus of predominantly literary prose and English – Lithuanian and English – German texts of EU legislation.

The source language and translation texts were aligned at the level of sense units, usually equivalent to one sentence. Deviation from a sentence-by-sentence alignment occurs where the unit of translation involves a one-to-many or many-to-one relationship. The alignment was carried out with the help of ParaConcord for the Czech literary prose and Vanilla aligner for EU legislation. The latter tool was adapted to our specific needs changing its interface and adding a more detailed level of alignment, i.e. texts are aligned on the level of sentence segments.

The Parallel Corpus is used for the search of translation equivalents (words and collocations). It is also a platform for the design of lexicographic and automated translation tools as well as foreign language teaching aids (Skoumalová, 2000, Marcinkevičienė 1998).

## 5. Afterword

Corpora in general and corpora of lesser used languages in particular serve multiple purposes. They form a part of cultural heritage, they bridge languages, they serve as encyclopedia of the present time. Last but not least, corpora serve as a source of statistical and linguistic knowledge. The case of the Lithuanian language corpora exemplify their application for HLT.

## 6. Bibliographical References

Vidas Daudaravičius 2001 *Interneto teksto paruošimas automatinei analizei*, In Proceedings of 6<sup>th</sup> conference of master and doctoral students, Vytautas Magnus University, 2001 04 26.

Vidas Daudaravičius and Rūta Marcinkevičienė. 2004. Gravity Counts for the Boundaries of Collocations. In

*International Journal of Corpus Linguistics*, John Benjamins Publishing Company, Amsterdam/Philadelphia, forthcoming.

Rūta Marcinkevičienė. 1998. Parallel Corpora and Bilingual Lexicography. In *Germanic and Baltic Linguistic Studies and Translation*, Homo Liber, Vilnius, pp.40-47.

Hana Skoumalová. 2000. Bridge dictionaries. In *The Proceedings of N ninth EURALEX International Congress*, Universität Stuttgart, pp. 799-803.

Erika Rimkutė. 2003. Morfologinio daugiareikšmiškumo tipologija. In *Lituanistica*, No 4 (56), pp. 60–78.

Vytautas Zinkevičius. 2000. Lemuoklis – morfologinei analizei. In *Darbai ir Dienos*, Vytauto Didžiojo universiteto leidykla, Kaunas, pp. 245-273.