Corpus of the Georgian Language

Nino Doborjginidze, Irina Lobzhanidze
Ilia State University
e-mail: nino_doborjginidze@iliauni.edu.ge, irina_lobzhanidze@iliauni.edu.ge

Abstract

The paper is the public presentation of the research program the Corpus of Modern, Middle and Old Georgian Texts including the Knight in the Panther's Skin and the Georgian Chronicles. The Corpus was compiled with the support of Ilia State University and Shota Rustaveli National Science Foundation aiming at building a new, extensive and representative corpus of the Georgian language. This paper is structured as follows: Section 1 includes background and research questions; Section 2 presents a methodological approach and briefly summarizes its theoretical prerequisites; Section 3 includes the findings and hypothesis, which refer to a compilation of corpus including morphological analyzer, corpus access and meta-data Information; and Section 4 presents the answers to the research questions and corpus itself.

Keywords: corpus of the Georgian language; morphological analyzer of the Georgian Language

Introduction

The corpus of the Georgian language is subdivided into several units such as:

- Text Block consisting of published texts and manuscripts in digital format. The search can be carried out by means of different filters;
- Visual Block consisting of illustrations from different manuscripts represented in the corpus, including, samples of calligraphy and bindings.

The paper presents the research area, the design and structure and applications related to the compilation of the corpus. At this moment, the beta version of Corpus is available at http://corpora.iliauni.edu.ge/.

1 Background and Research Questions

The compilation of the corpus of the Georgian language was carried out at the Centre of Linguistic Research at Ilia State University. This corpus introduces a perspective to analyze the Georgian language from the diachronic point of view providing scholars with flexible tool to analyze linguistic phenomena over a long period of time and to create a basis for the historical study of the language. The source (http://corpora.iliauni.edu.ge/) includes two types of freely-available corpora: monolingual and bilingual. The paper describes both of them, especially, their balance, size, designs, representativeness and types of annotation. The monolingual part of the corpus is subdivided into three parts: Modern Georgian, Middle Georgian and Old Georgian. The Sub-corpus of the Modern Georgian language is equipped with linguistic annotation and covers the period from 1832 to 2012. The linguistic annotation was provided by morphological analyzer of Modern Georgian developed at Ilia State University. The bilingual part of the corpus includes texts of different dimensions and design, especially, the Corpus of the Knight in the Panther's Skin and the Corpus of the Georgian

Chronicles. The first one is the database of the Georgian epic, including all published and unpublished Georgian texts aligned against their English translations; the collection is big and includes about 5 million words. The second one is the database of the Georgian Chronicles, which includes all published and unpublished texts and is subdivided into monolingual and bilingual parts. The bilingual part includes the Georgian text aligned against its translation into the Armenian language. The corpus has collected about 45 unpublished manuscripts and 10 published works. A query program allows users to retrieve data from a single text or from the whole corpus. Every word is accompanied with a brief context for every occurrence.

2 Methodological Approach and Theoretical Prerequisites

According to the general definition applied in corpus linguistics, a corpus is a collection of texts 'bound' together through specific parameters and principles. More broadly, a corpus can be defined through the following quotes:

1. A corpus is a collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis (Francis 1992: 17).

2. A corpus is understood to be a collection of samples of running text. The texts may be in spoken, written or intermediate forms, and the samples may be of any length (Aarts 1991: 45).

There are numerous corpora that contain samples of texts representing different periods of language development. Such kind of corpora are useful for the study of language changes and development. Thus, a corpus is one of the outputs of textual data processing. At the same time, it is a tool enabling the study of language data and the application of these data for the further production of dictionaries and grammars of language. But the main purpose of the Corpus is the desire to obtain and make available a great variety of texts, which can be used to represent the development of the Georgian language.

The methodological background of the Corpus is closely connected with the following: a) the taxonomy of corpus design; b) the difference between published and unpublished manuscripts; c) the standards, which allow us to process the existing textual heritage.

In our case, the following goals were addressed for the corpus of the Georgian Chronicles:

- *The interdisciplinary approach to the text.* The main goal was to provide the compilation, systematization and online accessibility of printed and manuscript versions;
- *The corpus design and representativeness.* The main blocks of the corpus design were structured in the following way: a) monolingual block of narrative; b) bilingual block of narrative; c) visual heritage;
- *The machine-readable standard.* A number of standards were to be applied and taken into account including: 1. ISO standards for natural language processing: Word segmentation of written texts (ISO 24614), Morpho-syntactic annotation framework (MAF) (ISO 24611, 246121 and 24615), Feature structures (ISO 24610), Lexical markup framework (LMF) (ISO 24613); 2. Additional standards: Data Category Registry (ISO 12620), Language codes (ISO 639 or IETF BCP-47), Script-codes (ISO 15924), Country codes (ISO 3166), Date and Time Formats (ISO 8601) and Unicode (ISO 10646) [ISO-TC37] and TEI XML P5 recommendations.

3 Findings and Hypotheses

3.1 Corpus access and Meta-data Information

The texts included in the corpus have been enriched with meta-data information. Meta-data is defined as "data about data" (Bournard 2005: 30) and provides additional information about corpus texts. Most language corpora apply different metadata schemes ranging from simple to highly sophisticated. Here are the most general metadata types: a. Bibliographic information about corpus texts; b. Descriptive information about corpus components; c. Documentary information about the corpus itself. The TEI recommendations have widely been applied lately for metadata sampling and structuring. These annotation schemes allow the researcher to perform complex queries on annotated data and to manage structured documents.

The corpus was designed to contain printed as well as manuscript versions of texts, which, naturally, required different metadata schemes to apply to:

1. Manuscript-based publications;
2. Reprints;
3. Previously unpublished manuscripts;
4. Previously published manuscripts.

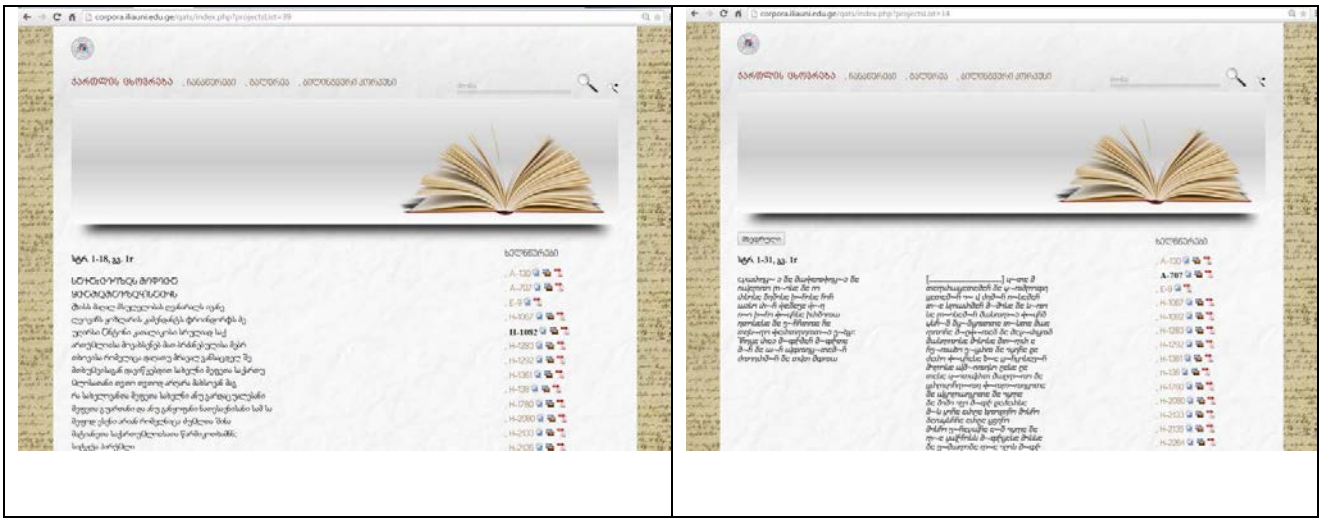
Thus, every text included in the corpus has been annotated according to the appropriate meta-data schemes. Additional web-interface has been built to upload texts, which allowed to distinguish the format of annotation for different kind of texts, especially for published and unpublished manuscripts. Naturally, the meta-data information depends on the type of text, but all corpus files are furnished with the following basic information:

- *Project description*: Funding institution, Leading institution, Responsible person: first name, last name, Responsible person's obligations, Responsible person, institution, Project name;
- *File description*: File author: first name, last name, File source, File language, File size, kB, Date of creation, Place of creation, Information about file revision, etc.;
- *Printed text description*: Text title, Author: first name, last name, Source language, Date of creation, Place of origin, Publisher, Place of publication, Date of publication, Editor, Translator, Illustrator, Number of volumes/issues, Number of pages, Text pages from … to …, ISBN/ISSN, Availability, Distributor, Authorized institution, Notes;
- *Manuscript description*: Location, Name of repository, Number of repository, Name of collection, Additional identification code, catalogue number, Manuscript author, copyist or compiler: first name, last name, The responsibilities of an editor etc., Manuscript title;
- *Manuscript language and script*: Manuscript language and script (Asomtavruli - Majuscule, Nuskhuri – Minuscule/Cursive, Mkhedruli – Civil);
- *Physical condition of the manuscript*: Form of the object, Material, Number of pages, Paper size type, Height, Width, Manuscript condition (description of revisions, damage), Foliation type (e.g. recto, verso, etc.), Paper collation type (e.g. mixed sequence);
- *Formal description of the manuscript*: Description of handwriting, Script description, Description of miniatures and decorations, Metatexts;
- *History of the manuscript*: Place of origin, Date of origin from (date), Date of origin to (date), Provenance from creation to archiving (if any), Information about manuscript purchase or donation.

Two different querying systems were applied for the retrieval of information. The first allows the

researcher to find a word in all data stored, to extract the context and all the document information and the second one allows the researcher to provide more complicated search based on the text description both for published and unpublished manuscripts.

It is well known that the Old Georgian manuscripts were written in Asomtavruli, Nuskhuri and Mkhedruli. Some texts were written in both: Asomtavruli and Mkhedruli, others were completely in Nuskhuri. The problem was how to represent such distinction between texts and, at the same time, how to make them easily acquired by the final reader. At the same time, the main goal of the project was to keep the digital format of manuscripts. Thus, all possible scripts are kept and query system can retrieve information from Old Georgian Texts in spite of the script used and, at the same time, the final reader has the opportunity to see the text written in Mkhedruli (see figures 1-2).



Figures 1-2: Corpus of the Georgian Chronicles

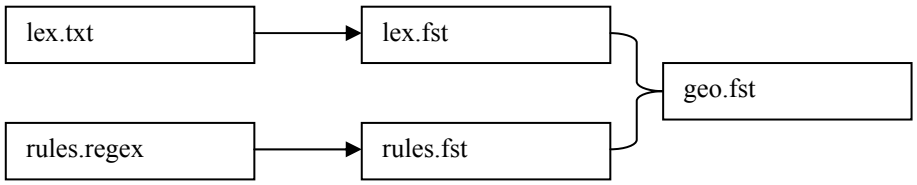
The button Mkhedruli allows the user to switch between Nuskhuri and Mkhedruli script of the Georgian text.

3.2 Morphological analyzer of the Modern Georgian language

Modern Georgian belongs to the morphologically rich languages. Descriptions of Georgian morphological structure emphasize the large number of inflectional categories; the large number of elements that verbal or nominal paradigms can contain; the interdependence in the occurrence of various elements and the large number of regular, semi-regular and irregular patterns. All the above-mentioned peculiarities make computational model of the Georgian morphology a rather difficult task.

The morphological analyzer of the Modern Georgian language was developed using finite state automata. Such kind of tools has been applied to the analysis of phonology and morphology in different languages. The analyzer was developed within the framework of the project AR/320/4-105/11 financed by the Shota Rustaveli National Science Foundation. The morphotactics is encoded in the lexicons and alternation rules are encoded in regular expressions. In addition to the above-mentioned peculiarities, we had to take into account the fact that Modern Georgian cannot be considered to be completely agglutinating. According to the existing definitions, the main peculiarity of agglutinating languages is that the root of a word does not change and each affix added directly to the root has its own grammatical function. From this point of view, the Georgian language is of mixed nature; especially, the paradigm of Georgian verb undergoes non-concatenative processes, which are more difficult from the viewpoint of computer generation.

The morphological transducer developed on the basis of Xerox Finite State Tools (Xfst)1 has the following structure:



The mentioned structure includes the following PoS Lexicons:
LEXICON Nouns
LEXICON Adjectives
LEXICON Numerals
LEXICON Pronouns
LEXICON Conjunctions
LEXICON Particles
LEXICON Adverbs
LEXICON Postpositions
LEXICON Verbs
LEXICON Verbal Nouns
LEXICON Participles
LEXICON Punctuations
LEXICON Abbreviations

The lexicon data are processed in accordance with the appropriate alternation rules. The morphological analyzer consists of the mentioned lexicons and alternation rules. It allows us to distinguish the appropriate lemma and morphological categories. This resource evaluated against different texts is used for tokenizing, lemmatizing and tagging.

Any kind of morphological analysis is based on the internal structure of a word. Words without internal changes like conjunctions and particles can be considered as an exception. Naturally, any kind of morpheme conveys additional information about the grammatical function of a word. Georgian verbs, generally, use bound morphemes to show its grammatical function. The main ways of use are as follows:

a) affixation, e.g.

Lexical Level: Ipfv+წანწალებ-s+V+RelStat+Intr+AutAct+FutCond+<NomSubj>+Subj3Sg

Surface Level: იწანწალებდა

b) root vowel alternation, e.g.

Lexical Level: Ipfv+დრეკ-ს+V+Din+Trans+Act+Pres+<NomSubj>+<DatObj>+Subj3Sg+Obj3

Surface Level: დრეკს

Lexical Level: Pfv+დრეკ-ს+V+Din+Trans+Act+Aor+<ErgSubj>+<NomObj>+Subj3Sg+Obj3

Surface Level: მოდრიკა

c) root alternation, e.g.

Lexical Level: Ipfv+ეუბნებ-ა+V+Trans+Act+Fut+<NomSubj>+<DatObjRec>+<DatObj>+

Subj3Sg+ObjRec3+Obj3

Surface Level: ეტყვის.

The main peculiarity of agglutinating language is that the root of a word does not change and each affix is characterized by a concrete grammatical function. Georgian language is of mixed nature; Georgian verb contains a lot of morphs, which can be described as agglutinating and at the same time as inflecting ones. Thus, the formation of Georgian verbal paradigm undergoes non-concatenative processes, which are more difficult from the viewpoint of computer generation.

The results of analyzer are used for the annotation of texts in the Corpus of Modern Georgian at the level of morphology and for the organization of query system, which, at this stage of development, is able to find about 250 grammatical categories.

The diversity of texts in the corpus requires permanent renovation of the analyzer and control for this renovation. Thus, to reach the above-mentioned task, the inner panel of the Corpus of Modern Georgian, at this stage, is equipped with the following blocks:

- Text Upload – the mentioned block provides the upload of texts of different genres to the Corpus of the Modern Georgian language and the annotation of text already uploaded;
- Texts – the mentioned block is subdivided into three parts: the first part provides the checking of meta-annotation and its correction (in case of need), the second part allows us to remove homonymy, to check the results of analyzer and to correct them (in case of need);
- Word Correction – in spite of correction of texts, we meet some words, which do not exist and can be considered as orthographic mistakes. The analyzer cannot provide their annotation. Thus, such kind of words can be corrected directly online;
- Unknown words – the texts of different genres have different type of words, e.g. terminological words, borrowings etc. These types of words are not represented in the lexicon of analyzer. Thus, the system collects such kind of words. This block allows the editor to assign appropriate class to the word. The above-mentioned information allows us to include the word directly in the lexicon.

4 Corpus Design and Taxonomy

The main function of a corpus is to display information to users. In this respect, it was very important to decide from the outset what type of information would be accessible online. However, the contents were revised during the project implementation. At present, the beta-version of the entire Corpus is accessible at http://corpora.iliauni.edu.ge.

References

Aarts, J. (1991). Intuition-based and observation-based grammars *English Corpus Linguistics. Studies in Honour of Jan Svartvik*, pp. 44-62.

Beesley, K., Karttunen, L. (2003). *Finite State Morphology.* Stanford: CSLI Publications.

Bournard, Lou (2005). Metadata for Corpus Work. In *Developing Linguistic Corpora: A Guide to Good Practice*, M. Wynne, Oxford: Oxford Books, pp. 30-46.

Francis, W. N., (1992). Language corpora BC. *Directions in corpus linguistics.* J. Svartvik, Berlin: Mouton de Gruyter, pp. 17-32.

Gurabanidze, M. (2004). *Computer Processing of Georgian Historical Sources.* Tbilisi: Tbilisi University Publishing House.

Gurevich, O. (2006). A Finite-State Model of Georgian Verbal Morphology. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL.* NY: Association for Computational Linguistics. pp. 45-48.

Jurafsky, D., Martin, H. J. (2009). *Speech and Language Processing.* New Jersey: Pearson Education International.

Kapanadze, O. (2010). Describing Georgian Morphology with a Finite-State System. In *Lecture Notes in Computer Science*, pp. 114-122.

McEnery, T., Wilson, A., (2011). *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Meurer, P., (2011). Constructing an annotated corpus for Georgian – Tools and resources. In *Symposium on Language, Logic and Computation*, Kutaisi.

Meurer, P. (2009). A Computational Grammar for Georgian. *Lecture Notes in Computer Science,* pp. 1-15.

Stump, G. T. (2001). *Inflectional Morphology: a Theory of Paradigm Structure.* NY: Cambridge University Press.

Melikishvili, D. (2001). *Conjugation System of Georgian Verb*. Tbilisi: Logos Press.

Sinclair, J., (1991). *Corpus, concordance, collocation: Describing English Language.* Oxford: Oxford University Press.

Shanidze, A., (1973). *The Basics of the Georgian language grammar. Tbilisi: TSU.*

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: Benjamins.