Irina LOBJANIDZE
Associate Professor
Ilia State University
Tbilisi, Georgia

For the annotation of Titlo Diacritic

Abstract: The paper describes different levels of annotation used in the Corpus of Modern, Middle and Old Georgian Texts. Aiming at building a new, extensive and representative tool for Georgian language the Corpus was compiled under the financial support of the Shota Rustaveli National Science Foundation and the Ilia State University (AR/266/1-31/13). In particular, the Corpus of Georgian language is envisaged as collecting a substantial amount of data needed for research. The scope and representativeness of texts included as well as free accessibility to it makes the corpus one of the most necessary tools for the study of different texts in Modern, Middle and Old Georgian (see, http://corpora.iliauni.edu.ge/). The corpus consists of different kind of texts, mainly: a) Manuscript-based publications; b) Reprints; c) Previously unpublished manuscripts and; d) Previously published manuscripts and covers Modern, Middle and Old Georgian.

The paper presents the research area, the design and structure and applications related to the compilation of the corpus, in particular, different levels of annotation as meta-data, structural mark-up and linguistic annotation at word-level, especially, from the viewpoint of Titlo Diacritic.

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 refers generally to the differences between the annotation of Modern and Old Georgian texts; and Section 4 presents the answers to the research questions.

Keywords: Digitalization of Manuscripts, Morphological Analyzer, Meta-Data, Structural Mark-up, Titlo Diacritic, Linguistic Annotation

1. Background and Research Auestions

Aiming at building a new, extensive and representative tool for Georgian language the Corpus was compiled under the financial support of the Shota Rustaveli National Science Foundation and the Ilia State University (AR/266/1-31/13). In particular, the Corpus of Georgian language is envisaged as collecting a substantial amount of data needed for research. The scope and representativeness of texts included as well as free accessibility to it makes the corpus one of the most necessary tools for the study of different texts in Modern, Middle and Old Georgian (see, http://corpora.iliauni.edu.ge/).

This corpus introduces a perspective to analyze Georgian language from diachronic point of view providing scholars with flexible tool to analyze linguistic phenomena over a long period of time and to create a base for historical study of language. The source (http://corpora.iliauni.edu.ge/) includes two types of freely-available corpora: monolingual and bilingual. The monolingual part of corpus is subdivided into two parts: Modern Georgian, Middle Georgian and Old Georgian Languages. The Sub-corpus of Modern Georgian Language is equipped with linguistic annotation and covers period from 1832 to 2012. The linguistic annotation was provided by the Morphological analyzer of Modern Georgian Language developed at the Ilia State University. The Sub-corpus of Middle and Old Georgian language includes a) Manuscript-based publications; b) Reprints; c) Previously unpublished manuscripts and; d) Previously published manuscripts of Athonite and Pre-Athonite Periods and *the Georgian Chronicles*. The Bilingual Corpus also available online, generally, consists of Georgian text aligned against its translation into Armenian (for instance see, *The Georgian Chronicles* by Abuladze, Ilia, 1953). A query program allows user to retrieve data from a single text or from the whole corpus. Every word is accompanied by a brief context for every occurrence.

The compilation of Georgian language corpus, which includes Modern Georgian, Middle and Old Georgian texts, has shifted the focus from creating accurate digital texts to the possibility of their annotation i.e. the possibility to provide presentational markup to text, image or other

data. Thus, the annotation is subdivided into Header and Body levels in accordance with TEI standard.

In linguistics annotation stage includes both extra-linguistic and intra-linguistic data. Extra-linguistic data is subdivided in two parts and from one point of view refers to the text itself and includes information on author, title, edition etc. and from another – to structural annotation of texts including chapter, paragraph, sentence etc. Intra-linguistic data refers to lemmas and Parts of Speech (PoS) markers and is closely connected to the structural characteristics of language under investigation.

The Corpus of Modern Georgian texts is annotated not only at the level of structural data, but also at the level of linguistic data including lemmas and PoS markers for different words. The Corpus of Modern Georgian texts is equipped with query set based on the above-mentioned markers. Such kind of annotation was provided by means of the morphological analyzer for Modern Georgian language developed within the framework of project AR/320/4-105/11 financed by the Shota Rustaveli National Science Foundation. The analyzer created by means of xfst includes appropriate lexicons and rules existing in Modern Georgian Language. It can be used for the annotation of Modern Georgian texts, but it does not describe the structure and changes in Georgian language, which took place since the VIth century and can't be used for the annotation (lemma and grammar forms) of Old Georgian manuscripts, especially, with respect to the follows:

1. The script: a) the Georgian scripts are subdivided into three writing systems, especially, Asomtavruli, Nuskhuri and Mkhedruli used in Georgian manuscripts of different centuries; b) in 1879 five letters (especially, ჱ (he), ჲ (hie), ჳ (vie), ჴ (qari), ჵ (hoe) from Georgian alphabet have been discarded, but they were used in manuscripts and played an important role from the linguistic point of view, especially, in the formation of some morphological structures e.g. ჱ (he), ჲ (hie) were markers of nominative case in vowel-ended nouns etc., c) Georgian manuscripts also use non-syllabic ჷ (u), which was a part of so called ascending diphthongs and participated in some phonetic processes;
2. The tokenization: POS tagging of texts are closely connected to the possibility to separate the text into tokens; these tokens may be words, punctuation markers, multi-word expressions etc. The main problem is that some Georgian manuscripts were not segmented according to white space between words or were partially segmented according to

white space and special marker like :·, others have not punctuation markers at all;

3. The titlo diacritic: Georgian manuscripts used an extended diacritic symbol drawn as a zigzag over words sometimes between words; this symbol has different meanings depending on the context or grammatical form e.g. postpositions or case markers were, generally, put under the titlo diacritic, and;
4. The potential ambiguity at the morphology level.

Therefore, in the paper I would like to pay special attention to the differences between annotation levels in Modern and Old Georgia, for instance, there will be considered some peculiarities as they are represented in the sub-corpora of *the Corpus of Modern, Middle and Old Georgian*, especially, the problems associated with presentation of Diacritics in Unicode and their processing for the future morphological analysis.

2. Methods and Theoretical Prerequisites

2.1. Standards

The Methodological background of *the Corpus of Modern, Middle and Old Georgian* is closely connected with the following: a) the taxonomy of corpus design; b) the difference between published texts, published and unpublished manuscripts (the majority of manuscripts represented in the corpus are kept at the National Centre of Manuscripts) and reprints; c) the standards, which allow us to process the existing textual heritage.

In our case, the following goals were set addressed for *the Corpus of Modern, Middle and Old Georgian*:

1. **The interdisciplinary approach to the text.** The main goal was to provide the compilation, systematization and on-line accessibility of printed texts and manuscript versions of narrative in order to facilitate a corpus approach to the study of the above-mentioned texts;
2. **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;
3. **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.

The annotation of corpus at the linguistic level needs the compilation of analyzer for Old Georgian language and, correspondingly, the decisions to the problems of tokenization and usage of titlo diacritic. Thus, additional approaches are associated with finite state techniques (Beesley 2003, Koskenniemi 1983 etc.) especially, with xfst and lexc tools.

2.2 Diacritics in Unicode

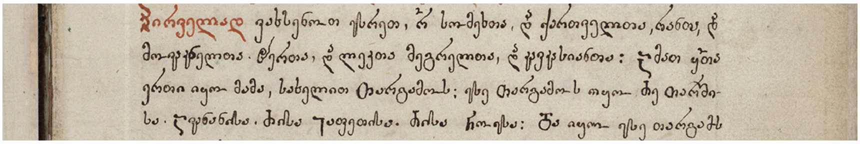
Unicode is a character encoding standard that treats alphabetic characters and symbols by means of numeric value and attribute. Before a Unicode, a very large number of Georgian fonts such as including AcadNusx, LitNusx etc. used an ASCII-based mapping. Georgian alphabet in Unicode is represented in the following standard ranges:

- 10A0–10FF – for *Mkhedruli* and *Asomtavruli* Scripts, which include a paragraph separator (⁖) and some other symbols;
- 2D00–2D2F – for *Nuskhuri Script*

Unicode use Combining Diacritical Marks (range: 0300–036F) for the representation of diacritics. A Georgian diacritic, especially, titlo diacritic is not represented at all in the ranges associated with Georgian language, but is represented in the range associated with diacritical marks. But for our case we have used a diacritic represented by a Greek Perispomeni of 1FC0[1] range so called combining tilde chosen because of its availability in text documents. But such kind of symbol cannot be considered as an extended diacritic drawn over a text; it looks like a character placed between letters, but not over them, e.g.

1. see, https://unicode.org/charts/PDF/U1F00.pdf (access 26.12.2017)

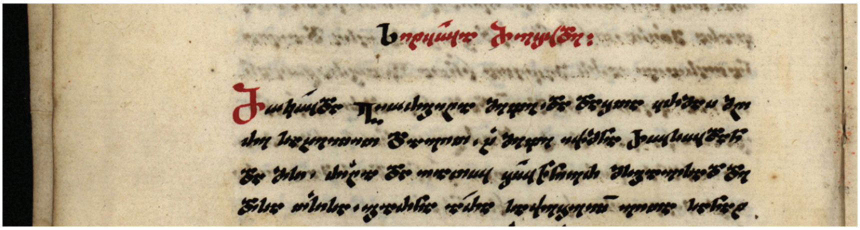
პირველად. ვახსენოთ ესრეთ, რ˜ სომეხთა, და ქართველთა რანთა, და მოვაკანელთა. ჵერთა, და ლეკთა მეგრელთა, და კავკასიანთა: ამათ ყ˜თა ერთი იყო მამა, სახელით თარგამოს იყო ძე თარშისა. ავანანისა, ძისა იაფეთისა. ძისა ნოესა: და იყო ესე თარგამოს...	Let us initially say the fact that the Armenians, Georgians, Rans, Movaknels, Herethians, Leks, Megrels and Caucasians had one father named Thargamos, son of Tharshi, son of Avanani, son of Japheth, son of Noah.



Mkhedruli Script in S-5316, 1r

Ⴑⴈⴒⴗⴓⴀⴢ Ⴋⴄⴘⴣⴃⴄ: Ⴋⴍⴉⴓⴃⴀ Ⴂⴈⴍⴐⴂⴈ ⴋⴄⴔⴄ, ⴃⴀ ⴃⴀⴗⴇⴀ ⴗⴐⴋⴀⴢ ⴋⴚⴄ ⴛⴄ ⴑⴀⴞⴄⴊⴈⴇ Ⴃⴀⴅⴈⴇ. ⴗ˜ ⴋⴄⴔⴄ ⴈⴕⴋⴌⴀ Ⴐⴓⴑⴓⴃⴀⴌ ⴃⴀ ⴋⴈⴑⴈ. ⴗ˜ⴊⴀ ⴃⴀ ⴈⴀⴇⴗ ⴗ ⴗⴍⴚⴞⴈⴄⴊⴈ ⴋⴑⴂⴀⴅⴑⴀⴃ ⴃⴄ ⴃⴈⴑⴀ ⴇ˜ⴑⴈⴑⴀ.	Chapter Seven: And the king George died and left a child David. And his sister Rusudan became a Queen. And she was beautiful like her mother.

vs



Nuskhuri Script in S-2518, 16v

As it can be seen from the examples, the titlo diacritics in Old Georgian Manuscripts, especially, in *the Georgian Chronicles* are not similar according to their forms and depend on different factors including century of compilation and hand of transcriber. Sometimes titlo diacritic is just a horizontal zigzag line over words as in Rt-XV-N-3, Rt-XII-N-I, sometimes just a vertical zigzag line as in Q-383, 162r, sometimes just a line over words as in S-354, 245 and sometimes just a point between or after a word as in H-2303, 32v etc.

3. Finding and Hypothesis

3.1. Corpus Access and Meta-Data information

The texts included in the corpus have been enriched with meta-data information. Metadata is defined as 'data about data'[2] 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 the 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 allows the researcher to perform complex queries on annotated data and to manage structured documents.

The Corpus of the Modern, Middle and Old Georgian was designed to contain printed as well as manuscript versions, which, naturally, required different metadata schemes to apply to:

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

Thus, every text included in the corpus has been annotated according to the appropriate metadata schemes. Additional web-interface has been built to upload texts (see figure 1), which allowed to distinguish the format of annotation for different kind of texts, especially, for published and unpublished manuscripts.

Naturally, the metadata information depends on the type of text, but all corpus files are furnished with the following basic information:

1. **Project description**: Funding institution, Leading institution, Responsible person: first name, last name, Responsible person's obligations, Responsible person, institution, Project name;
2. **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.;

2. Bournard 2005: 30.

3. **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
4. **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;
5. **Manuscript language and script,** Manuscript language and script (Asomtavruli – Majuscule, Nuskhuri – Minuscule/Cursive, Mkhedruli – Civil;
6. **Physical condition of the manuscript,** Form of the object, Material, paper, Number of papers, 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);
7. **Formal description of the manuscript,** Description of handwriting, Script description, Description of miniatures and decorations, Metatexts;
8. **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

Collection and processing of this information requires significant time and human resources, especially in case of unpublished manuscripts. Therefore, the employees of the National Center of Manuscripts, T. Khakhviashvili, N. Bilanishvili, G. Shubitidze, Sh. Tumanishvili and N. Datashvili were involved in manuscript digitalization. They provided the digitalization of manuscripts and the filling of meta-data about available manuscripts.

Two different querying systems have been applied to the retrieval of information. 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. *The Sub-Corpora of the Modern Georgian* is additionally equipped with linguistic querying system, which allows user to find words in accordance with their

PoS and appropriate morphological categories, the similar querying system for *the Sub-Corpora of the Middle and Old Georgian* will be launched after the implementation of morphological analyzer of Old Georgian Language.

It is well known that the manuscripts were written in Asomtavruli, Nuskhuri and Mkhedruli. Some texts were mixed by Asomtavruli and Mkhedruli, others were completely 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 manuscript. 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[3]. The button Mkhedruli allows the user to switch between Nuskhuri and Mkhedruli script of Georgian text.

3.2. Titlo Diacritic in the Corpus of Middle and Old Georgian Language

Taking into account that the texts of *the Georgian Chronicles* were compiled in different centuries, they cover different linguistic data and represent changes which took place in Georgian language since 973. At the same time, the majority of published manuscripts has been changed and adapted to the modern Georgian language with purpose to provide understanding of the text by the final reader. Thus, the use of titlo diacritics for instance in *the Georgian Chronicles* was generally omitted in published texts and substituted by standard words. Some publications preserved titlo diacritics, but diminished their number, e.g.

1. *The Georgian Chronicles*, *Queen Mariam*'s version, published in 1906 by E. Takaishvili; words in this publication are separated by colons and the text includes 845 titlo diacritics used for sacred names so called Nomina Sacra. In this text we can see 514 abbreviated forms of "the Lord" in different cases, e.g. ღ̃თნ [g̃t'n, Lord:ERG.SG], ღ̃თსა [g̃t'sa, Lord:DAT.SG], ღ̃თისა [g̃t'isa, Lord:GEN.SG] etc.,
2. History and life of Saint Nino, published in 1946 by S. Kubaneishvili; the text includes only two titlo diacritics used in conjunction "as about, but" e.g. ხ̃ [x̃, but].

3. Mkhedruli script is used for Modern Georgian Language.

The manuscripts of *the Corpus of Middle and Old Georgian Language* include a large number of titlo diacritics used for different purposes including preservation of writing space and presentation of different contexts. The number of scribal abbreviations available in the manuscript block of the corpus is equal to 250424. Thus, the processing of titlo diacritic should be considered as a part of morphological analysis by means of computer, which requires implementation of morphological knowledge and accompanying phonological processes. The identification of titlo diacritic needs two stages generation and analysis; it means that all rules should operate on two levels: lexical and surface.

At the same time, some of Old Georgian Manuscripts were translated from Greek language; that's why for instance the processing of titlo diacritic in *the Georgian Chronicles* (non-translated narrative) is closely connected to the general rules of scribal abbreviations. There are well-known types of abbreviations, which can be met not only in Georgian Manuscripts, but also in the majority of Medieval Manuscripts worldwide:

1. suspension i.e. only the first part of a word is written, the last part consists of diacritic mark. According to K. Danelia (1997) suspension in its pure form can't be met in Old Georgian Manuscripts (the inscriptions of Kala-Bolnisi can be considered as an exception), *the Georgian Chronicles or other texts represented in the Corpus of Middle and Old Georgian Texts* can't be considered as an exception to that rule;
2. contraction i.e. middle part of a word is omitted; in its pure form a word has only the first and the last letters, otherwise an impure contraction has one or more letters in the middle part. In *the Corpus of Middle and Old Georgian* we can see both of these types implemented in different ways depending on the context, e.g. რ̃ი [r̃i, which:NOM.SG], ყ̃ი [q̃i, every:NOM.SG], ქ̃ყ̃ნსა [k`̃q̃nsa, country:DAT.SG], ს̃ფლ̃ვი [s̃p`l̃vi, grave:NOM.SG] etc. In Greek Manuscripts such kind of abbreviations are generally met in Nomina Sacra.
3. truncation i.e. only the first letter of a word is written, while other letters are substituted by a titlo diacritic. In *the Georgian Chronicles* such kind of abbreviation is commonly used e.g. რ̃ [r̃, for, because] used 11229 times, ხ̃ [x̃, but] used 16050 etc.

Also, in *the Corpus of Middle and Old Georgian* can be seen the following types of abbreviations:

a) abbreviated of phrases or sometimes sentences, e.g. კ̃ხრ [k̃xr, blessed:NOM.SG=be:2SG], აღზრდილ̃რს [ag̃zrdil̃rs, grown:NOM.SG=be:3SG] etc.;

b) missing vowels, e.g. ჩ̃მ [č̀̃m, me:DAT.1SG], ჯ̃რნი [j̃̌rni, donkey:NOM.PL], and vowels placed over the letters as can be seen in H-1067 (2r), H-1082 (15r) etc. The last type of abbreviation will not be treated taking into account that digitized versions of manuscripts don't show the difference between words and words with vowels placed over other letters.

So, all the above-mentioned variations can be encoded during the processing of old Georgian manuscripts.

3.3. Linguistic Level Annotation by means of the Morphological Analyzer and Titlo Diacritic

Modern Georgian language belongs to a morphologically rich languages. Descriptions of Georgian morphological structure emphasize the large number of inflectional categories; the large number of elements that a verb or a noun 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 Georgian morphology a rather difficult task.

The Morphological analyzer of Modern Georgian language has been 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 by means of Xerox Calculus (especially, xfst and lexc). 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 language can't be considered as completely agglutinating. According to the existing definitions, the main peculiarity of agglutinating languages is that the root of a word doesn't change and each affix added directly to the root has its own grammatical function. From this point of view 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.

At the same time, the decisions to the computer processing and annotation of words for *the Corpus of Middle and Old Georgian* is closely connected to the similar techniques used for the processing of Old Georgian texts, but the special attention is paid to the above-mentioned types of abbreviations existing in *the Corpus of Middle and Old Georgian*. For instance the quantity of words with titlo diacritic in *the Corpus of the Georgian Chronicles* is equal to 251318 including those of pure/impure contraction (218405) including missing vowels and abbreviated phrases (we couldn't find any sentence level abbreviations in *the Georgian Chronicles*) and truncation (13 percents 32913 words). It should be mentioned that contraction at the level of phrases can be predicted for the samples with auxiliary verbs at the lexical levels, but other cases are more difficult and need additional rules for tokenization of phrase. It is known that some of Georgian manuscripts partially belong to a style of writing without spaces or other marks between words (Scriptio continua); sometimes there are used special marks for separation of words as :· like H-1067, Q-795, sometimes the format is mixed as in S-1444 etc. but, generally, old Georgian manuscripts doesn't use punctuation marks.

In finite state transducers, the lexical (grammar) level is considered as an upper state and the surface (word) level – as a lower state. Thus, the mapping between these two states, e.g.

Upper Level → რამეთუ+Abbr

Lower Level → რ̃

Can be described by a regular expressions compiled into a single finite state transducer for morphological analyzis and generation. The lexical side of the transducer includes an fs tag +Abbr (abbreviation). The surface state contains all valid forms of abbreviated word. One of the possible scripts helps us to represent each word containing ̃ titlo sign as an Abbreviation unit.

But we can see that titlo diacritic is used not only for the cases of pure abbreviation of words, but is a part of words with concrete lexical structure expressing PoS and their categories; it means that the possibility explained above can't be completely adopted. It can be used only for the cases of Truncation, but it can't be used for the cases of contraction or abbreviation of phrases or sentences, e.g. Let's consider a proper noun ალექს̃ნდრ

(Alek`s˜ndr) used in different manuscripts of *the Georgian Chronicles*, especially, H-988, M-13 etc.

ალექს˜ნდრ˜მ (alek's˜ndr˜m)	ალექსანდრ+Abbr	ალექსანდრ+N+Prop+Name+Sg+Erg	1
ალექს˜ნდრე (alek's˜ndre)	ალექსანდრ+Abbr	ალექსანდრ+N+Prop+Name+Sg+Nom	17
ალექს˜ნდრემ (alek's˜ndrem)	ალექსანდრ+Abbr	ალექსანდრ+N+Prop+Name+Sg+Erg	2
ალექს˜ნდრეს (alek's˜ndres)	ალექსანდრ+Abbr	ალექსანდრ+N+Prop+Name+Sg+Dat	5
ალექს˜ნდრესავით (alek's˜ndresavit')	ალექსანდრ+Abbr	ალექსანდრ+N+Prop+Name+Sg+Dat+ Emp+Post(like)	1
ალექს˜ნდრესგ˜ნ (alek's˜ndresg˜n)	ალექსანდრ+Abbr	ალექსანდრ+N+Prop+Name+Sg+Gen +Post(from1)	4
ალექს˜ნდრესი (alek's˜ndresi)	ალექსანდრ+Abbr	ალექსანდრ+N+Prop+Name+Sg+Gen	1
ალექს˜ნდრესსა (alek's˜ndressa)	ალექსანდრ+Abbr	ალექსანდრ+N+Prop+Name+Sg+Dat +Dat+Emp	3
ალექს˜ნდრესცა (alek's˜ndresc'a)	ალექსანდრ+Abbr	ალექსანდრ+N+Prop+Name+Sg+Dat +Ptcl+Emp	1

We can see that these forms can't be annotated only at the level of Abbreviation; they need additional marks including PoS and other morphological categories. And according to the structure of Nouns in Old Georgian language the lemma sign for such kind of words should be Absolute case as opposite to Nominative case used in Modern Georgian.

Thus, the most important is to distinguish a sample lemma for a set of lexical forms, which should be treated as a whole. Such kind of processing can be completed only after the compilation of the morphological analyzer for Old Georgian language, but at this stage we can predict the most frequent units of truncation and some of units for contraction, which have titlo diacritic used in the words similar to the forms existing in Modern Georgian. It means that we can't predict forms with double cases, but forms with similar structure as e.g. ალექს˜ნდრესავით are predictable.

4. Conclusions

The goal of this paper has been to demonstrate the computational approach to the series of complex morphological problems with regards to *the Middle and Old Georgian Language*. The approach to the processing of titlo diacritic has been tested on several texts containing titlo diacritics. Thus, a word containing titlo diacritic can be considered as a structural pattern with Zero unit, which participates in the formation of its meaning and structure; sometimes it is used as a morpheme, sometimes as a whole word or expression, but it needs further generalization.

Bibliography

Beesley, Kenneth, Karttunen, Lauri, *Finite State Morphology,* Stanford, CSLI Publications, 2003.

Cappelli, Adriano, *The Elements of Abbreviation in Medieval Latin Paleography,* Kansas, University of Kansas, 1982.

Chankieva, Tsatsa, "Norms of abbreviation in old Georgian Manuscripts (V-X cc.)", in *Paleographic Researches*, Tbilisi, Georgian Adacemy of Sciences, 1965, p. 57-94.

Chankieva, Tsatsa, "Norms of abbreviations in Old Georgian Manuscripts", in *Problems of Paleography and codicology in the USSR* (Paleographic), 1974, p. 434-439.

Danelia, Korneli, Sarjveladze, Zurab, *Georgian Paleography.*, Tbilisi, Nekeri, 1997.

Gurevich, Olga, "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, 2006, p. 45-48.

Javakhishvili, Ivane, *Georgian Paleography,* Tbilisi, Tbilisi State University, 1949.

Jurafsky, Daniel, Martin, James, *Speech and Language Processing,* New Jersey, Pearson Education International, 2009.

Kapanadze, Oleg, "Describing Georgian Morphology with a Finite-State System", in *Lecture Notes in Computer Science*, South Africa, Springer, 2010, p. 114-122.

Koskenniemi, Kimmo, *Two-level morphology: A general computational model for word-form recognition and production,* Helsinki, University of Helsinki, 1983.

Lobzhanidze, Irina, "Realization of Verbal Models in the Corpus of Modern Georgian Language", in *Language, Logic and Computation*, Tbilisi, TSU, Nekeri, 2015, pp. 79-85.

Lobzhanidze, Irina, "Morphological Analyzer and Generator of Modern Georgian Language", in *GMLT*, Tbilisi, TSU, 2013, p. 82-83.

McEnery, Tony, Wilson, Andrew, *Corpus Linguistics,* Edinburgh, Edinburgh University Press, 2011.

Melikishvili, Damana, *Conjugation System of Georgian Verb*, Tbilisi, Logos Press, 2001.

Meurer, Paul, "A Computational Grammar for Georgian", *Lecture Notes in Computer Science,* Germany, Springer, 2009, p. 1-15.

Shanidze, Akaki,*The Basics of the Georgian language grammar, Tbilisi, TSU, 1973.*

Sinclair, John, *Corpus, concordance, collocation: Describing English Language,* Oxford, Oxford University Press, 1991.

Stump, Gregory, *Inflectional Morphology: a Theory of Paradigm Structure,* NY, Cambridge University Press, 2001.