IDIOMS: PRINTED OR DIGITAL?

I. Lobzhanidze, S. Berikashvili
17.12.2019

OUTLINE

- Introduction
- Problems of Georgian Lexicography
  - Lemmatization Problems
  - Access to Headwords in Alphabetical Order
- Findings
  - The dictionary System: constituents and type of search
  - The dictionary System: advanced search option
- Conclusions

FOUR MAJOR PREREQUISITES TO THE DESIGN OF ANY LEXICOGRAPHIC DATABASE

Linguistic specification (of macrostructure and microstructure);

Database management system (DBMS) specification;

Specification of phases of lexicographic database construction: input, verification and modification;

Presentation of and access to lexical information: access, re-formatting, dissemination.

MAIN PROBLEMS OF GEORGIAN LEXICOGRAPHY

Representation of verbal forms in dictionary entries caused by the absence of infinitive (verbal noun so call *masdars* vs verb in the third person singular);

Polypersonalism of Georgian verb, which causes inclusion of different verbal patterns in the majority of Georgian printed or electronic dictionaries, e.g. Chikobava's Explanatory Dictionary etc.;

Searching patterns for verbal forms in electronic and online dictionaries having in mind that it is completely impossible to focus on a lemma for verbal forms and to provide their setting in alphabet order.

LEMMATIZATION PROBLEM

According to the Morpho-syntactic Annotation Framework(MAF):

a LEMMA *is a lemmatized form class of inflected forms differing only by inflectional morphology. In European languages, the lemma is usually the /singular/ if there is a variation in /number/, the /masculine/ form if there is a variation in /gender/ and the /infinitive/ for all verbs. In some languages, certain nouns are defective in the singular form, in which case the /plural/ is chosen. In Arabic, for a verb, the lemma is usually considered to be the third person singular with the accomplished aspect.*

The lemma for the nominal paradigm is represented with the nominative singular, but the Georgian verb does not have an infinitive. Thus there is no clear rule with regards to the representation of a lemma for the verbal paradigm.

LEMMATIZATION PROBLEM
VERBAL IN GEORGIAN DICTIONARIES

In the majority of Modern Georgian dictionaries include the following approaches to the base form of dictionary entries:

- Verbal noun, the so-called masdar form;
- Root-based form, the so-called abstract root;
- The third person singular in the present or future indicative.

LEMMATIZATION PROBLEM
VERBAL NOUN APPROACH

This approach considers a verbal noun to be a base form for a verbal entry and sometimes considers it to be an infinitive of the verbal paradigm, keeping in mind that the extraction of an abstract root from given forms is a simpler task than the other way around etc. This approach can be seen in the dictionaries of MWEs as well (see Sakhokia, T.: Georgian Figurative Expressions, (1950-1955) etc.), where in headwords verbal constituents of idioms or other compounds are represented in the form of verbal nouns in spite of the fact that a verbal constituent is used in the citation form. Such a type of headword representation has a negative influence on the meaning of idiomatic expressions as a whole and the majority of idioms require the fixed grammatical structures of concrete verbs, but not verbal nouns.

LEMMATIZATION PROBLEM
ROOT BASED APPROACH

The second approach differs from the previous one by representing the headwords in the form of an abstract verbal root with appropriate paradigms. For dictionary users it is rather difficult to find the appropriate meaning of words by trying to determine their possible verbal roots and to derive them from the existing structures without basic knowledge of Georgian grammar, especially the rules for the formation of verbal paradigms.

This principle of dictionary entries cannot be shared for the dictionaries of MWEs because idioms or any other kind of phraseological units are groups of words with a fixed lexical composition and grammatical structure, but not a word taken separately.

LEMMATIZATION PROBLEM
THE THIRD PERSON SINGULAR APPROACH

This approach considers a verb in the third singular subject form in the present or future indicative, which includes forms indicating grammatical categories like version, causation etc. as well. This kind of approach is shared by the dictionaries of idioms (see Oniani, A.: Georgian Idioms, (1966) etc.), where in headwords verbal constituents are represented in two ways:

- as required by the fixed grammatical structure of the MWE, and
- as a verb in the third singular subject form in the present or future indicative if a fixed grammatical structure of the MWE can be violated.

PROBLEMS OF GEORGIAN LEXICOGRAPHY
ACCESS TO HEADWORDS IN ALPHABETIC ORDER

There are two options of alphabetization, namely,

- word-by-word, where spaces between words take precedence, or
- letter-by-letter, where spaces and hyphens are disregarded.

Both of these options are easily adopted for the Georgian language, but at the same time the position of the verbal root in the verbal template, which occupies the fourth slot in the nine-slot template, makes it impossible to arrange verbs in alphabetical order without paying attention to preverbs, prefixal person markers and version markers.

PROBLEMS OF GEORGIAN LEXICOGRAPHY
ACCESS TO HEADWORDS IN ALPHABETIC ORDER

For example, if we consider the third person singular in the present or future indicative of verbs: *u-vl-is* 'looks after, tends to smb.',*a-u-vl-is*'will make the rounds of smth.',*cha-u-vl-is* 'will go down smth.', *she-u-vl-is* 'will drop in to see smth.' etc. as a main form, the quantity of dictionary entries with affixes attached to the same stem and reflecting different meanings will be more than enough.

Otherwise, if we consider the appropriate verbal noun *-svla* 'going, walking' to be the headword of the dictionary entry, different meanings of verbal constituents like those mentioned above will not be represented at the appropriate level. We can put MWEs with verbal nouns like *gverd-is avla* 'avoiding, ignoring', *gverd-is chavla* 'ignoring', but there do not exist forms like **gverd-is uvla* or **gverd-is shevla*.

PROBLEMS OF GEORGIAN LEXICOGRAPHY
ACCESS TO HEADWORDS IN ALPHABETIC ORDER

Thus, the dictionary of idioms should contain MWEs with verbal constituents to preserve all possible meanings (1.a, 1.b, 1.c, 1.d):

1.a *gverd-s* *a-u-vl-is* ‘He/she will overlook/bypass smb./smth.’

side-DATPV-prv-go-FUT.3SGSBJ

1.b *gverd-s* *u-vl-is* ‘He/she disregards, avoids smb./smth.’

side-DATprv-go-PRES.3SGSBJ

1.c *gverd-s* *cha-u-vl-is* ‘He/she will ignore smb./smth.’

side-DATPV-prv-go-FUT.3SGSBJ

1.d *gverd-s* *she-u-vl-is* ‘He/she will drop in to see smb./smth.’

side-DATPV-prv-go-FUT.3SGSBJ

PROBLEMS OF GEORGIAN LEXICOGRAPHY
ACCESS TO HEADWORDS IN ALPHABETIC ORDER

There are two types of verbal idioms which should be described separately, specifically

1. those which do not undergo grammatical transformations as given in example (2.a) for the subject and (2.b) for the object verbal paradigms, and,
2. those which show morphosyntactic flexibility with regards to the perfective/imperfective aspects (3.a) and tenses (3.b)and allow inflections etc.

PROBLEMS OF GEORGIAN LEXICOGRAPHY
ACCESS TO HEADWORDS IN ALPHABETIC ORDER

tavze buzs ar isvams, but not **tavze buzs isvams*; *tavi ara makvs*, but not **tavi makvs* etc.

2.a	*tav-ze*	*buz-s*	*ar*	*i-sv-am-s*
	head-on	fly-DAT	NEG	prv-seat-TS-PRES.3SGSBJ

'He/she doesnot let a fly land on one's head; ~He/she is haughty'

2.b	*tav-i*	*ara*	*m-akv-s*
	head-NOM NEG	1SGOBJ-have-PRES.3SGSBJ	

'I have no head; ~I am not able to do smth.'

PROBLEMS OF GEORGIAN LEXICOGRAPHY
ACCESS TO HEADWORDS IN ALPHABETIC ORDER

3.a tav-s	*i-gd-eb-s* 'He/she is throwing a head; ~ He/she is insolent'
head-DAT	prv-throw-TS-PRES.3SGSBJ
tav-i	*ča-i-gd-o* 'He/she was throwing a head; ~He/she was insolent'
head-NOM	PV-prv-throw-TS-AOR.3SGSBJ
3.b qu̇r-s	*u-gd-eb-s* 'He/she is throwing an ear; ~He/she listens to smb.'
ear-DAT	prv-throw-TS-PRES.3SGSBJ
qu̇r-i	*v-u-gd-e* 'He/she was throwing an ear; ~He/she listened to smb.'
ear-NOM	1SGSBJ-prv-throw-AOR

PROBLEMS OF GEORGIAN LEXICOGRAPHY
ACCESS TO HEADWORDS IN ALPHABETIC ORDER

Also, there are idioms with non-fixed wording,e.g. *dushash-i mo-s-d-is* vs *mo-s-d-is dushash-i* (4.a, 4.b), non-fixed canonical forms, e.g. *dana-s ɣor-is ḳud-ze ga-ṭex-s* vs *dana-s ɣor-is ḳud-ze gada-ṭex-s* (5.a, 5.b), or with the possibility of substitution, e.g. *cecxl-s u-ḳid-eb-s* vs *al-cecxl-s u-ḳid-eb-s* (6.a, 6.b)

4.a dushash-i *mo-s-d-is*

double_sixes-NOM PV-3SGOBJ-come-TS-PRES.3SGSBJ

4.b mo-s-d-i-s *dushash-i*

PV-3SGOBJ-come-TS-3SGSBJ.PRES double_sixes-NOM

'He/she gets double sixes; ~ He/she is very lucky'

PROBLEMS OF GEORGIAN LEXICOGRAPHY
ACCESS TO HEADWORDS IN ALPHABETIC ORDER

5.a dana-s	*ɣor-is*	*ḳud-ze*	*ga-ṭex-s*
knife-DAT	pig-GEN	tail-on	PV-break-FUT.3SGSBJ
5.b dana-s	*ɣor-is*	*ḳud-ze*	*gada-ṭex-s*
knife-DAT	pig-GEN	tail-on	PV-break-FUT.3SGSBJ

‘He/she will break knife on tail; ~He/she will leave the job half-done’

PROBLEMS OF GEORGIAN LEXICOGRAPHY
ACCESS TO HEADWORDS IN ALPHABETIC ORDER

6.a cecxl-s	*u-ḳid-eb-s*
fire-DAT	prv-light-TS-PRES.3SGSBJ
6.b al-cecxl-s	*u-ḳid-eb-s*
flame-fire-DAT	prv-light-TS-PRES.3SGSBJ

‘He/she makes smb. burn with love or sells smth. for a fortune’

PROBLEMS OF GEORGIAN LEXICOGRAPHY
ACCESS TO HEADWORDS IN ALPHABETIC ORDER

As it can be seen a word-by-word approach with space precedence does not work correctly for these more flexible examples of MWEs and affects the compilation of dictionaries as well.

So, to overcome the problems of lemmatization and alphabetization the main focus of our research was to compile an Online Dictionary of Idioms and to provide user-friendly access to lexical information stored in the DB on the basis of the appropriate linguistic specification and the morphological analyzer of the Georgian language.

FINDINGS

Finite state techniques, especially, xfst and lexc (as described by Beesley, Kartunnen 2003, Koskenniemi1983 etc.) used for the compilation of the morphological analyzer of Modern Georgian;

Approaches of modern corpus based lexicography (as described by Atkins 2008, Sinclair 1996, Ooi 1988 etc.) used for the compilation of the On-line dictionary of idioms by means of TLex system.

FINDINGS

In the case of the Online Dictionary of Idioms, we determined the form of the on-line dictionary and the structure of entries, revised the existing units using the concordance from the corpus of Modern Georgian Language available at http://corpora.iliauni.edu.ge/ and additional one created in TLex system, add revised and new entries to TLex system, converted the prepared dictionary to .xml format and launched the on-line version of dictionary

The On-Line Dictionary of Idioms (MWE) available at http://idioms.iliauni.edu.ge/

FINDINGS
CONSTITUENTS AND TYPE OF SEARCH

Monolingual Dictionary of Idioms (Modern Georgian MWE)

Bi-directional Bilingual Dictionary of Idioms (Modern Greek – Modern Georgian and vice versa)

Type of Search:

- **Quick Search:** Type in keyword or phrase that you are looking for, then press ENTER;
- **Advanced Search:** Perform a more extensive search associated with grammatical structure;
- **Alphabetic Search:** Browse the dictionary from ა (a) to ჰ (h);
- **Wild Card:** * can represent the occurrence of any number of characters

FINDINGS
ADVANCED SEARCH

This option performs search for any kind of word as it is met in the raw text and gives users possibility to see direct translation of its initial form in our case it is the third person singular for verbs and nominative case singular for nouns, e.g.

სიცოცხლე გქონია 'you are happy', *სიცოცხლე მქონია* 'I am happy' etc.

verbal forms *მქონია, გქონია* 'I, you have' etc. are connected to *აქვს* 'has' or *ქონა* 'possession'

FINDINGS
ADVANCED SEARCH

The system available online determines the lemma sign for a verb ‘გქონია’ by means of the morphological analyzer and then, the morphological analyzer returns to the system the initial lemma of a word, in our case, the lemma ჰქვს ‘has’ and the system provides search in the database and returns all verbal MWE-s associated with the above-mentioned lemma form.

FINDINGS
CONCLUSIONS

As shown in this presentation, the use of two-level morphology and finite-state technology is both theoretically and technologically suitable for the Modern Georgian language if we keep in mind that in spite of long distance dependencies within words the concatenative structure of Modern Georgian can be implemented without difficulty using finite state transducers like *lexc* and *xfst*.

And, the morphological transducer can easily be adapted for different purposes, especially for solving the problems of lemmatization and alphabetization noticed in Georgian dictionaries (in monolingual and bilingual ones).

THANK YOU

irina_lobzhanidze@iliauni.edu.ge

svetlana.berikashvili@iliauni.edu.ge