Computational Model of Modern Georgian Language and Searching Patterns for On-line Dictionary of Idioms

Irina Lobzhanidze
Ilia State University
irina_lobzhanidze@iliauni.edu.ge

Introduction

Any kind of electronic dictionary can be considered as a database; generally, its purpose is to provide adequate explanation or translation of separate words or multi-word expressions (MWE), to store information and to allow user to find appropriate language units. Following Gibbon (2000), there are four major prerequisites to the design of any lexicographic database, i.e. dictionaries:

1. Linguistic specification (of macrostructure and microstructure);
2. Database management system (DBMS) specification;
3. Specification of phases of lexicographic database construction: input, verification and modification;
4. Presentation of and access to lexical information: access, re-formatting, dissemination.

In case of Modern Georgian language, the main problems are associated from one point with linguistic specification, which corresponds to types of lexical information involving linguistic analysis for Modern Georgian Language and from another point – with access to lexical information stored in the database (DB) by end-user. The Modern Georgian language belongs to morphologically rich languages. Descriptions of Georgian morphological structure emphasize large number of inflectional categories; the large number of elements that verb or noun paradigms can contain; the interdependence in the occurrence of various elements and the large number of regular, semi-regular and irregular patterns. It means that the morphologically rich nature of Georgian expresses different levels of information at the word level and affects a compilation of dictionaries, i.e. lexicographic databases for Georgian language. Thus, the main issues, which are worth of mentioning are as follows:

a) representing of verbal forms in dictionary entries caused by the absence of infinitive (verbal noun vs verb in the third person singular);
b) polypersonalism of Georgian verb, which causes inclusion of different verbal patterns in the majority of Georgian printed or electronic dictionaries;
c) 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.

Present paper answers the above mentioned issues describing for instance the On-line Dictionary of Idioms prepared under the financial support of the Shota Rustaveli Science Foundation (Projects No Y-04-10, No LE/17/1-30/13) and the morphological analyzer of Modern Georgian Language (Project No AR/320/4-105/11) used for advanced search.

The Dictionary of Idioms is bilingual in Modern Georgian and Modern Greek, includes approximately 12000 entries and reflects the function and meaning of idioms. It combines features of translational and learner's dictionaries. The dictionary is available at http://idioms.iliauni.edu.ge/.

Section 1 Morphological Analyzer of Modern Georgian and Problems of Georgian Lexicography

The Morphological analyzer is developed as bi-directional finite state transducer by means of Xerox Finite State Tools (xfst and lexc), which is sufficient for capturing morphological structure of Modern Georgian Language. The system includes 13 blocks of the existing Part of Speech (PoS) of Modern Georgian language as well as separate blocks for Punctuation and Abbreviations, while the pattern for Verbal Paradigm is subdivided into additional 66 groups as described by Melikishvili (2001) and an additional group for irregular verbs.

The morphotactics of language is encoded in PoS lexicons and alternation rules are encoded in regular expressions.

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

The lexicon data are processed in accordance with the appropriate 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.

Section 2 Brief Descriptions of Dictionaries

The majority of Georgian electronic dictionaries (monolingual and bilingual, e.g. http://translate.ge/, http://ena.ge/ etc.) share similar lexicographic problems caused by the following:

a) Absence of an infinitive form of the verb, which affects dictionary entries and causes use of different patterns (some dictionaries include entries represented in the form of Verbal nouns so called masdars, e.g. Oniani (1966) etc., others – in the form of Verbs in the third-person singular of the present tense, e.g. Sakhokia (1979) etc., also, there are dictionaries possessing both of the above mentioned forms);

b) Georgian verb template consisting at least of twelve constituents implies existence of preverbs, person and version markers before the root. It makes impossible to find appropriate verb in dictionaries by initial letters of verbal noun or the third-person singular of the present tense i.e. in alphabetic order. And it forces Native and Non-Native speakers of Georgian to acquire grammatical information on Georgian verbal patterns with purpose of searching and translating, which have a negative impact on language acquisition.

Each dictionary stated above selects its own types of access to the data, generally, by special filters, which allow user to look for a word not only in the headword lists, but also in the whole database. This possibility is rather difficult to acquire taking into account that the end-user, generally, has a possibility to find some words without their meaning and cannot use them for his/her purposes.

Section 3 Methods

During the compilation of Morphological Analyzer of Modern Georgian language and the compilation of Online Dictionary of idioms, I have used different kind of approaches:

a) Finite state techniques, especially, xfst and lexc (as described by Beesley, Kartunnen 2003, Koskenniemi 1983 etc.) used for the compilation of the morphological analyzer of Modern Georgian;

b) 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.

Section 4. Findings and Hypothesis

The compilation of any dictionary includes the sequence of stages. 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[1] and additional one created in TLex system[2], add revised and new entries to TLex system, converted the prepared dictionary to .xml format and launched an on-line version of dictionary. At the same time we had to find solution for the problems described previously. So, special attention was paid to

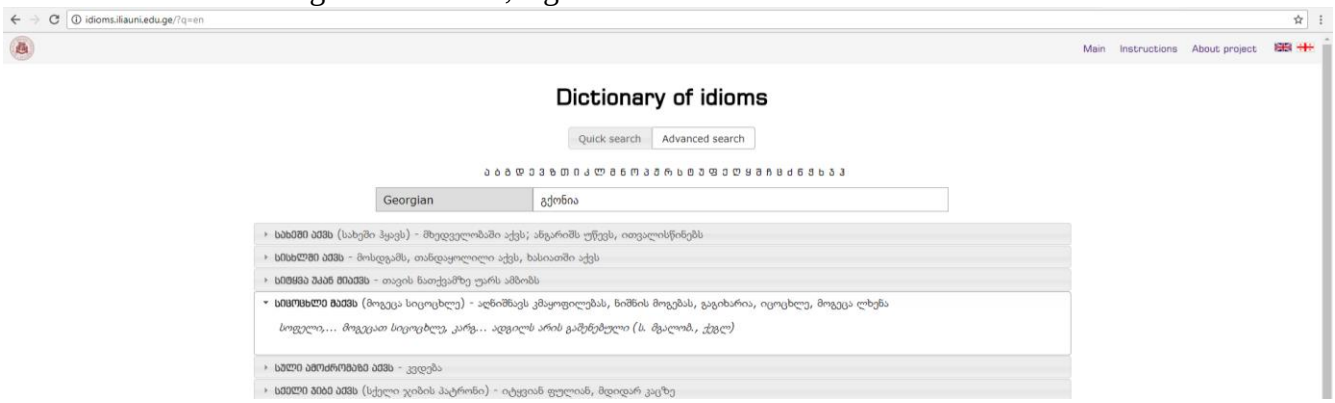
[1] http://corpora.iliauni.edu.ge/

[2] http://tshwanedje.com/

1. The Dictionary entries, which differ from the viewpoint of the elements represented in monolingual and bilingual parts. Most entries include information on lemma sign, derivational variants of use, etymological notes for some entries, definitions, literary citation with indication of literary source;
2. Ordering of entries that is closely connected to the following types 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 represents the occurrence of any number of characters

   Such kind of ordering means that if a user wants to find any word or constituent of multiword expression, it is allowed directly from the web.

At the same time the so called Advanced Search is a decision to the issues mentioned above, especially, the absence of infinitive form and the impossibility to find appropriate verbal MWE by initial letters of headwords. 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.



Modern Georgian allows forms like: *სიცოცხლე გქონია* 'you are happy', *სიცოცხლე მქონია* 'I am happy' etc. Such kind of forms can be seen in literary sources as well and the user whose knowledge of Modern Georgian is not very high will not be able to find them taking into account that the headwords in the dictionary entries for a verbal form *გქონია* 'you have' are *აქვს* 'has' or *ქონა* 'possession'. The system available online determines the lemma sign for a verb *გქონია:*

გქონია:აქვს+V+Intr+Res1+<DatSubj>+<NomObj>+Subj2Sg+Obj3

And then based on the lemma *აქვს* 'has' provides search in the database and returns all verbal MWE associated with the above-mentioned verb. As a result, the system gives the end-user access to the headword of the appropriate word.

Section 5 Conclusions

The compilation of on-line dictionary is useful for the further development of computational approaches to Georgian language. Taking into account that the compilation of monolingual and bidirectional bilingual dictionary of idioms is over, there is a possibility to represent results of our research and describe the further stages of its development. The dictionary is available at http://idioms.iliauni.edu.ge/.

Keywords: on-line dictionary of idioms, morphological analyzer and generator of Modern Georgian Language, multi-word expressions, verbal patterns

References

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

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

Carter, J. M. (1996). Corpus to Corpus: A Study of Translation Equivalence. *International Journal of Lexicography* , 171-178.

Van Eynde, F., Gibbon, D. (2000). *Lexicon Developmen for Speech and Language Processing.* London: Kluwer Academic Publishers.

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. 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*, 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,* 1-15.

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

Lobzhanidze, I. (2016). Online Dictionary of Idioms, Proceedings of the XVII EURALEX International Congress, Ivane Javakhishvili Tbilisi University Press, 710-717.

Lobzhanidze, I. (2015). Finite State Morphological Approach to Georgian Verbal Paradigm, TICCSAM, 79-85.

Lobzhanidze, I. (2013). Morphological Analyzer and Generator of Modern Georgian Language, GMLT, 82-83.

Lobzhanidze, I. (2012). For the Compilation of Modern Georgian – Modern Greek Dictionary of Idioms. *10th International Conference of Greek Linguistics.* Komotini: The Democritus University of Thrace, 899-904

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

Ooi, V. B. (1998). *Computer Corpus Lexicography.* Edinburgh: Edinburgh University Press.

Rundell, B. S. (2008). *The Oxford Guide to Practical Lexicography.* New York: Oxford University 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.*

Lexicographic Sources

Rayfield, D., Apridonze, Sh., Margalitadze T. etc. (2006). *A Comprehensive Georgian-English Dictionary.* Garnett Press.

Βλαχόπουλος, Σ. (2007). *Λεξικό των ιδιωτισμών της νέας ελληνικής.* Αθήνα: Κλειδάριθμος.

Δημητρίου, Α. (1995). *Λεξικό νεοελληνισμών: Ιδιωτισμοί, στερεότυπες μεταφορές και παρομοιώσεις, λέξεις και φράσεις από την καθαρεύουσα.* Αθήνα: Γρηγόρη .

Μπαμπινιώτης, Γ. (1998). *Λεξικό της Νέας Ελληνικής γλώσσας, Κέντρο λεξικολογίας.* Αθήνα.

ონიანი, ა. (1966). *ქართული იდიომები.* თბილისი: ნაკადული.

სახოკია, თ. (1979). *ქართული ხატოვანი სიტყვათქმანი.* თბილისი: მერანი.

ჩიქობავა, ა. (1950-1964; 2008). *ქართული ენის განმარტებითი ლექსიკონი.* თბილისი: საქართველოს სსრ მეცნიერებათა აკადემიის გამომცემლობა