საქართველოს მეცნიერებათა ეროვნული აკადემიის მოამბე, ტ. 19(193), №2, 2025
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 19(193), no. 2, 2025

Linguistics

Grammatical Dictionary Compiler as a System for Kartvelian Languages

Liana Lortkipanidze[*], Anna Chutkerashvili[**]

[*] *Faculty of Exact and Natural Sciences, Ivane Javakhishvili Tbilisi State University, Georgia*
[**] *Archil Eliashvili Institute of Control Systems, Georgian Technical University, Tbilisi, Georgia*

(Presented by Academy Member Avtandil Arabuli)

Abstract. This paper presents an automated system designed for compiling grammatical dictionaries of the Georgian language and its dialects. The system is based on a lexicon-oriented approach, which involves identifying words with common grammatical markers and integrating them into dialectal dictionaries to efficiently expand lexical resources. Additionally, the system employs an innovative lemmatization algorithm for processing unknown words, enabling the automatic generation of base forms and their paradigms. The developed system demonstrated high efficiency in the automatic creation of grammatical dictionaries for the Georgian language. Testing on literary corpora showed that only 2% of non-dictionary word forms required manual correction after lemmatization. The proposed affix-based algorithm significantly outperformed traditional methods relying solely on suffixes, confirming the system's effectiveness in lexical resource expansion and its adaptability to other Kartvelian languages. © *2025 Bull. Georg. Natl. Acad. Sci.*

Keywords: acquisition of lexicon, Georgian grammatical dictionary, lemmatization rules, morphological processor

Introduction

The difference between the grammatical dictionary and already existing ordinary dictionaries is that grammatical dictionary offers the user not only hear word forms but also the whole paradigm derived from this word i.e. it shows the whole set of language vocabulary. It contains the morphological and partial syntactic characteristics of the language units. The usual dictionary cannot help the user to understand the text, because of the agglutinative-inflection structure of many languages (partially Georgian), the word can change its form in the context, so that it can be difficult to find the basic form.

Grammatical dictionaries are used successfully for various languages for text annotation. As a rule, in the process the system of compiling and enrichment of grammatical dictionaries by words must be included. For example, a word that is not included in the dictionary may occur in the corpus. In addition, the subsystems in the language have different orthography, morphology and vocabulary. A dictionary compiler is needed to solve such tasks.

In some languages, grammatical analyzers are used for these purposes.

The paper proposes the description of grammatical dictionary compilers for Georgian language.

Methods

The paradigms used to construct the dictionary may be derived as follows:

- Automatically – unsupervised method. The program can define the word-form in the paradigm that combines the lexemes with one lemma and one part of speech in the boundaries of one paradigm.
- Manually – supervised method. Linguist decides how to form paradigms in order to unite the lexemes and their corresponding lexical forms.
- Mixed method – automatically and manually. It is sometimes necessary to use both methods to get the aimed result.

All words in the Georgian grammatical dictionary will be input as a headword. Noun – in nominative case. Verb – in present tense of the third person singular form of subject or object.

The lexical input of the dictionary will be compiled from the following information:

1. The lemma;

2. Grammatical features of a lexeme;

2.1. Part of speech;

2.2. For noun: animated or inanimate concrete or abstract; for adjective: basic and derived; for numeral: cardinal, ordinal and fractional; for pronouns: by lexical content; for verbs: transitivity, voice, version, causation, aspect, mood; for adverb: according to the lexical content; In the case of the unchangeable words will be indicated its type: postposition, conjunction, particle or interjection;

3. For title forms the type of the root will be indicated: whether the root is ended with vowel or consonant will be defined for nouns; as for verbs, the stem type will be indicated according to the thematic marker;

4. Types of declination or conjugation.

We use automatic enlargement of a dictionary from the Georgian corpus by means of a set of lemmatized lists of unknown word forms. The tag set we use for Georgian is the one developed for the Georgian corpus [1]. It consists of about 100 different tags, where each character in the tag string corresponds to a single morpho-syntactic category.

Georgian needs more advanced methods than suffix replacement for receiving a lemma from so-called OOV (out-of-vocabulary) words [2]. Process of affixing can include pre-fixing, and suffixing. Therefore, we have created a trainable lemmatizer that adds and removes both suffixes, prefixes and infixes as needed. As already mentioned contrary to English or French, Georgian is an inflected language (About Georgian morphology see [3]). This means that nouns, adjectives and numerals (among others) are inflected according to their inflectional class (or paradigm), and the stem itself can be affected. The latter occurs in particular for some nouns and adjectives in their genitive singular form. For example, *mamali* ‘rooster’, in which the root is *mamal-*, has the genitive form *mamlis*. Here is an example to achieve with the training algorithm. Suppose we have the following pairs of the Georgian full forms and lemmas: *saxlši* → *saxli* (Translation: “in the house” → house).

If this were the sole input given to the training program, it should produce a transformation rule for noun suffix like this: *-*ši* → *-*i.*

With this rule, a lemmatizer would be able to construct the correct lemma for some words that had not been used during the training, such as the words: *vašlši*, *jamši*, *k’aradaši*.

However, for the most words, the lemmatizer would simply fail to produce correct output, because not all words do contain the literal strings only -*ši*, but -*ebši* also. In this case a transformation rule must be like this: * -*ebši* → *-*i.*

We correct it, having added the rule: to begin search with the longest suffix string.

Letting the program construct lemmatization rules requires an extended list of full form – lemma

pairs the program can exercise on – at least tens of thousands and possibly over a million entries [4,5], what does a task impracticable. In the current work, we want to present an idea of the advantages of a mixed algorithm of an affix-based for verbs and only a suffix-based for other parts of speech.

The acquisition of lexicon from the corpus can be achieved by the iteration of a four-step loop:

1. Building a rule set using the morphological processor.

2. Building all possible lemmas for OOV words found in the corpus.

3. Ranking these possible lemmas according to their likelihood given the corpus.

4. Validation best ranked lemmas.

Building a rule set and possible lemmas. The training algorithm generates a data structure consisting of rules that a lemmatizer must use to build a lemma from the full form.

If the elected rule produces the only one lemma from the full form, nothing needs to be done. Otherwise, it is necessary to choose the correct rule. The training process terminates when the full forms in all pairs in the training set are transformed to their corresponding lemmas. After training, the data structure becomes permanent and a lemmatizer can use rules. The lemmatizer must select and remove rules in the same way as the training algorithm. All words from the training set and without it has to be lemmatized correctly. However, it may fail to produce the correct lemmas for verbs that were not in the training set.

During training, the morphological generator GeoTrans [4,5] is used. With its help, it is possible to generate paradigms 35000 verbs and 65000 words from other parts of the speech. The system covers all templates of paradigms of word change.

For training a lemmatizer, the GeoTrans system generates word-forms according to a paradigm pattern for a given lemma. After that, the algorithm for compiling the rules of lemmatization is used to find the longest non-overlapping similar parts in a given full form – lemma pair. For example, in the pairs: *saxlis → saxli; saxlebi → saxli; saxlni → saxli; saxlma → saxli; saxlebma → saxli; saxlsa → saxli; saxlze → saxli; saxlebze → saxli; saxlši → saxli; saxlebši → saxli* – the longest common substring is *saxl*. These similar parts are replaced with wildcards and the corresponding lemmatization rules can be created for each full form:

Table 1. Examples of lemmatization rules

Saxlis	**is → *i*	*saxlsa*	**sa → *i*
Saxlebi	**ebi → *i*	*saxlze*	**ze → *i*
Saxlni	**ni → *i*	*saxlebze*	**ebze → *i*
Saxlma	**ma →* i*	*saxlši*	**ši → *i*
Saxlebma	**ma → *i*	*saxlebši*	**ebši → *i*

We suppose that no more than three kinds of rules are applied to the lemma during the lemmatization process: the suffix and/or the prefix rule. As already noted, the prefix rule is used only for the lemmatization of verbs. In view of the fact that the system covers all the main verbs of the Georgian language, we assume that in the corpus new verbs can be found only in derived forms.

In modern Georgian, for the formation of new verbs, only 12 percent of all possible prefixes are the most productive. There is also no need to consider the process of infixation as it is absent from the verb's derivatives. Given these factors, the lemmatization process is significantly accelerated.

For the training of a lemmatizer, a list of verbs is compiled, the lemmas of which have the same prefixes as in the aforementioned 12 percent (one copy for each). From these verbs, the GeoTrans system generates all possible forms. After that, the rule generation algorithm displays the rules of lemmatization.

Procedure of Lexical acquisition "is based on the idea that open-class lemmata are likely to occur in more than one form" [6]. The program of corpus lemmatization should isolate the groups belonging to the same lemma in an ordered array. We believe that the group with identical lemmas should include at least two members. The result of lemmatization

mainly depends on the members of the selected groups. The prefixing process is very active in the Georgian language and the full forms obtained from different lemmas can be placed one after another in alphabetically ordered list. Consequently, to count the number of occurrences of lemmas, the ranking algorithm will receive from the corpus a list of words ordered on lemmas.

Algorithm of Ranking. At the first stage, lists of rules are compiled according to the method described in the 3rd paragraph. Each rule is assigned the probability PRi of the fact that the Ri rule generates the correct lemma using the following equation:

$$PRi=1/NRi, \text{ where } NRi \quad (1)$$

Ri denotes the number of occurrences of rules that generate different lemmas using identical word form affixes. For example, if the full form of the word *kalis* is given, then the lemmatizer may apply two rules for it:

1) *is → **a* and 2) *is→**i*.

Using them, we will have:

1) *kalis → kali* (translate: 'woman' in genetive case → 'woman' in nominative case)

2) *kalis →kala* (translate: 'skull' in genetive case → 'skull' in nominative case)

Thus, for given suffixes, two rules are used and both are assigned a probability equal to ½ = 0.5.

Then, the lemmatizer program creates an ordered list of words from the corpus, where to each word (Wi) corresponds the number of occurrences of a word in the corpus (OCi); hypothetical lemma (Li) of this word; the number of the rule (NRi) and probability (PRi) of the rule of lemmatization.

Our algorithm is based on the assumption that the probability (Pi) of the form (Wi) comes from the form of Lemma (Li) satisfies the following equation:

$$P_i = \frac{\sum_{i=1}^{n} PR_i * OC_i}{\sum_{i=1}^{m} PR_i * OC_i}.$$

Here *n* is the number of lines of the record with identical lemmas. *m* is the total number of recording lines with all possible lemmas of this word-forms.

Thus, it is possible to resolve ambiguity that happens when words have more than one lemma. The problem can arise in case presumable lemmas of new words are created from the only forms. In such cases, the program separately outputs the list of lemmas to remove ambiguity. Cycle completes with manual validation.

The Extension of lexicon. At the output, the lemmatization algorithm issues list of words with variety of lemmas with corresponding probabilities. After this, the algorithm chooses from the list with identical forms the one for which the corresponding lemma has the highest probability. For the example described in the previous section, this list will look like in Table 2.

Table 2. Fragments of the lemmatization data matrixes

W	OC	L	RN	PR	PR*OC	P
kala	1	kala	1	1	1	0.16
kalad	8	kali	5	0.5	4	0.84
kalebi	109	kali	7	0.5	54.5	0.84
kalebis	31	kali	11	0.5	15.5	0.84
kalebistvis	5	kali	3	0.5	2.5	0.84
kalebit	3	kali	9	0.5	1.5	0.84
kali	426	kali	15	1	426	0.84
kalis	200	kali	17	0.5	100	0.84
kalma	115	kali	18	1	115	0.84
kalo	36	kali	19	1	36	0.84
kals	197	kali	20	1	197	0.84
kalta	1	kali	13	1	1	0.84
kaltan	1	kali	14	1	1	0.84
kalze	9	kali	12	1	9	0.84

At the next step the sampling algorithm removes duplicates of the lemmas and at the end, a ready list is issuing for expanding the dictionary (Table 3).

Table 3. Fragments of the lemmatization data matrixes

L	P
kala	0.156743
kali	0.843257

Conclusion

The system was tested for lemmatization of non-dictionary forms from the corpus of novels of Otar Tchiladze. In total, a vocabulary was compiled from the corpus with 95224 word-forms. Of these, 74900 were lemmas. After lemmatization of the non-dictionary words, only 2% of them needed manual disambiguation.

The obtained results with the new affix algorithm are better than those of only suffix lemmatizer. This means that the new algorithm is good at generalizing over small groups of words with complex morphology, like Georgian language.

Acknowledgements

This paper is published with the support of the Shota Rustaveli National Science Foundation of Georgia (SRNSFG) under the grant FR-21-3509.

ენათმეცნიერება

გრამატიკული ლექსიკონის კომპილატორი – სისტემა ქართველური ენებისათვის

ლ. ლორთქიფანიძე*, ა. ჩუტკერაშვილი**

** ივანე ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი, ზუსტ და საბუნებისმეტყველო მეცნიერებათა ფაკულტეტი, საქართველო*
*** საქართველოს ტექნიკური უნივერსიტეტი, არჩილ ელიაშვილის სახ. მართვის სისტემების ინსტიტუტი, თბილისი, საქართველო*

(წარმოდგენილია აკადემიის წევრის ა. არაბულის მიერ)

ნაშრომში წარმოდგენილია ავტომატიზებული სისტემა, რომელიც განკუთვნილია ქართული ენისა და მისი დიალექტების გრამატიკული ლექსიკონების შესადგენად. სისტემა ეფუძნება ლექსიკურ რესურსებზე ორიენტირებულ მიდგომას, რაც გულისხმობს გრამატიკული მარკერების მქონე სიტყვების იდენტიფიკაციას და მათ ინტეგრირებას დიალექტურ ლექსიკონში, ლექსიკური ბაზის ეფექტური გაფართოების მიზნით. გარდა ამისა, სისტემა იყენებს ინოვაციურ ლემატიზაციის ალგორითმს უცნობი სიტყვების დასამუშავებლად, რაც უზრუნველყოფს საბაზისო ფორმებისა და მათი პარადიგმების ავტომატურ გენერირებას. შემუშავებულმა სისტემამ მაღალი ეფექტურობა აჩვენა ქართული ენის გრამატიკული ლექსიკონების ავტომატური შექმნის პროცესში. ლიტერატურულ კორპუსებზე ჩატარებულმა ტესტირებამ აჩვენა, რომ არალექსიკონური სიტყვა-ფორმების მხოლოდ 2%-ს დასჭირდა ხელით კორექტირება ლემატიზაციის შემდეგ. აფიქსებზე დაფუძნებულმა შემოთავაზებულმა ალგორითმმა მნიშვნე-

ლოვნად აჯობა ტრადიციულ, მხოლოდ სუფიქსებზე დაფუძნებულ მეთოდებს, რაც ადასტურებს სისტემის ეფექტურობას ლექსიკური რესურსების გაფართოებაში და მის ადაპტირების შესაძლებლობას სხვა ქართველურ ენებზე.

REFERENCES

1. Lortkipanidze L., Amirezashvili N., Chutkerashvili A., Javashvili N., Samsonadze L. (2017) Syntax annotation of the Georgian literary corpus. *Theoretical Computer Science and General Issues. Proceedings of the 11th International Tbilisi Symposium on Logic, Language, and Computation, TbiLLC 2015*, *Revised Selected Papers:* 89-97, Berlin/Heidelberg.
2. Kanis J., Müller L. (2015) Automatic lemmatizer construction with focus on OOV words lemmatization in text, speech and dialogue, *Lecture Notes in Computer Science*, 132-139. Berlin/Heidelberg.
3. Harris A. (1981) Georgian syntax: a study in relational grammar. New York: Cambridge University Press.
4. Beridze M., Lortkipanidze L., Nadaraia D. (2015) Dialect Dictionaries with the Functions of representativeness and morphological annotation in Georgian dialect corpus. *Theoretical Computer Science and General Issues. 10th International Tbilisi Symposium on Logic, Language, and Computation, TbiLLC 2013*, *Revised Selected Papers*: 82-96. Berlin Heidelberg.
5. Lortkipanidze L. (2006) Application of GeoTrans System in the Georgian "Spell Checker". *Proceedings of the LEPL Archil Eliashvili Institute of Control Systems*: 187-192. Tbilisi (in Georgian).
6. Hana J., Feldman A. (2004) Portable language technology: Russian via Czech *Proceedings of the Midwest Computational Linguistics Colloquium*, Bloomington, Indiana.

Received January, 2025