The Linguistic Model of the Georgian Language (Synthesis of Nominal word-forms)

Ketevan Datukishvili, Nana Loladze, Merab Zakalashvili (Tbilisi)

DOI: https://doi.org/10.62235/dk.3.2024.8512

ketevan.datukishvili@tsu.ge || ORCID: 0009-0009-2104-556X
nana.loladze@tsu.ge || ORCID: 0009-0005-1817-9963
merabza@gmail.com || ORCID: 0009-0006-0757-6856

Abstract: In this paper we present a unique linguistic model designed for the morphological synthesis and analysis of the Georgian language. The morphological structure of the Georgian language is fully covered by this model. Based on it, we have created a software with a special tool, namely, a processor enabling the generation and analysis of word forms. We would like to emphasise the fact that no particular grammatical theory is used in our linguistic model. Rather, the language data is provided in a different format and in a structured way, taking into account the theories currently in use. Our model as presented in this paper is a collection of morphological equations (often around 3,000 units) required to generate every word form from a single stem. Currently, more than 266 million word-forms can be synthesised utilising the Georgian language's morphological processor. While not all of these forms can be found in electronic texts or extensive corpora, they are all viable. These words have been referred to as "potential forms". The possible forms are crucial for studying a variety of topics pertaining to natural language processing. Additionally, they aid in the resolution of specific artificial intelligence challenges. Beyond, we intend to pinpoint the exact frequencies of the data produced by the processor and specify the domains in which they are utilised.

Keywords: Linguistic model, Synthesis of word-forms, Morphological processor

With the aim of morphological synthesis and analysis of the Georgian language, we have created a special linguistic model. This model is a complete representation of the morphological structure of the Georgian language. Based on the given model, we have created a software with a special tool, namely, a processor enabling the generation and analysis of word forms.

In our linguistic model, the language is not described based on a certain grammatical theory. Instead, taking into account the existing theories, the language data are presented in a different format in a formalised way. The paper analyses the model and offers a corresponding mechanism of synthesis.

Each morpheme in this model has a symbolic marker. The unity of these markers yields a morphological formula of a word. Our method is based on the expression of words by means of symbolic formulae. This enables the representation of all types of word-forms in a formalised way.

The model reflects all the nominal word-forms derived from the same root, in particular, case-forms like მთას ('to the mountain' (dative), მთებისა ('of the mountains' (genitive), etc.; forms with postpositions like მთაზე ('on the mountain'), მთისთვის ('for the mountain'); forms with particles like მთასა-ც ('also the mountain'), მთაში-ვე ('in the mountain itself'); forms with auxiliary verbs like მთა-ა ('is a mountain'), მთაში-ა ('is in the mountain'); forms with indirect speech particles like მთამ-ო ('he/she said that the mountain'), მთაში-მეთქი ('I said, in the mountain'); and other word-forms.

Based on the table of morphemes, each affix is expressed by means of a special symbol, for instance: ებ – A_1; ნ – A_2; თ – A_3; ი – B_2; მა – B_3 etc. Any word-form can be expressed by means of these symbols. The unity of the symbols yields the morphological formula of a word. For example: ბავშვ-ებ-მა ('children', ergative) – R + A_1 + B_3; ბავშვ-ის-ა-თვის ('for the child') – R + B_6 + C_1 + D_7. R is the root, whereas other symbols denote affixes (markers of number, case, postposition and so on).

The expression of words by means of morphological formulae is our method to enable the presentation of all kinds of words in a formalised way. The morphological formula is a pattern, based on which the program can make a synthesis of word-forms from the root.

Certain types of formulae are united in a model. A model is a unity of morphological formulae necessary for the generation of all word-forms from one stem (on the average, about 3,000 units). The phonological part of the model describes the changes related to the nominal inflections and their positions. The program also offers a list of phonological types.

Based on the above-mentioned linguistic model, we have created a special software tool: a morphological processor of the Georgian language. It consists of two components: a database and the program part. Based on the model and the phonological type, the synthesiser generates inflectional forms of the nominal parts of speech. As a result, the program yields a complete list of the nominal forms, including case-forms, nominal forms with postpositions, forms with particles and so on. This list can be arbitrarily referred to as a complete paradigm of the noun which, as we have mentioned, embraces approximately 3,000 units. Currently, the statistical data of the morphological processor are as follows:

- lemmas – 113,860
- formulae generated by the synthesiser – 916,893
- word-forms generated by the synthesiser – 266,478,980

Currently, using the morphological processor of the Georgian language, it is possible to implement the synthesis of over 266 million word-forms. Not all of these forms are found in electronic texts or ample corpuses, yet, they are potentially possible. We have termed these words as "potential forms".

The "potential forms" are important for the research of diverse issues related to the processing of natural languages. In addition, they help solve certain tasks of artificial intellingence. We also plan to identify precise frequencies of the data generated by means of the processor and define the fields of their usage.

ქართული ენის ლინგვისტური მოდელი
(სახელური სიტყვაფორმების სინთეზი)

ქეთევან დათუკიშვილი, ნანა ლოლაძე, მერაბ ზაკალაშვილი
(ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი)

DOI: https://doi.org/10.62235/dk.3.2024.8512
ketevan.datukishvili@tsu.ge || ORCID: 0009-0009-2104-556X
nana.loladze@tsu.ge || ORCID: 0009-0005-1817-9963
merabza@gmail.com || ORCID: 0009-0006-0757-6856

1. შესავალი

ბუნებრივ ენათა დამუშავების ერთ-ერთ მნიშვნელოვან ამოცანას წარმოადგენს მორფოლოგიური ანალიზი და სინთეზი, რისთვისაც სხვადასხვა მეთოდი გამოიყენება. ენის დამუშავებისას უპირატესად ჯერ ხორციელდება მონაცემთა ანალიზი და მასზე დაყრდნობით – სინთეზი. ჩვენ თავდაპირველად შევასრულეთ სინთეზის ამოცანა და შემდეგ, მისი ალგორითმის საფუძველზე – ანალიზი. ჩვენი სინთეზის პროგრამა ეფუძნება ქართული ენის ლინგვისტურ მოდელს. ფლექსიის მოდელის ავტორია ქეთევან დათუკიშვილი, დერივაციის მოდელისა – ნანა ლოლაძე. პროცესორით სინთეზირებული ფორმების ვერიფიცირება ხორციელდება ტექსტების დიდ მასივებში (კორპუსებსა და სხვა ელექტრონულ წყაროებში).

სტატიაში განვიხილავთ ქართული ენის სახელის ლინგვისტურ მოდელს, რომლის მიხედვითაც სრულდება სახელურ სიტყვაფორმათა სინთეზი. მოდელი უზრუნველყოფს ერთი ფუძისგან ყველა იმ ფორმის გენერირებას, რომლებიც თეორიულად (ლინგვისტური წესების შესაბამისად) დასაშვებია თანამედროვე ქართულ ენაში.

ქართული, როგორც აგლუტინაციური ენა, აფიქსთა დიდი რაოდენობითა და ფორმაწარმოების მრავალფეროვნებით გამოირჩევა, ამიტომ განსაკუთრებით მნიშვნელოვანია იმგვარი ინსტრუმენტების შექმნა, რომლებიც უზრუნველყოფს ქართული ენის სინთეზსა და ანალიზს მორფოლოგიის დონეზე. ამ ამოცანის გადაწყვეტის არაერთი ცდა არსებობს: მოირერი (2007), მარგველანი (2008), ანთიძე (2009), ლობჟანიძე (2021)... ავტორები უპირატესად იყენებენ ენის მორფოლოგიური ანალიზის სტანდარტულ მეთოდებს, კერძოდ, სასრული პოზიციის ავტომატებს. ეს არის საყოველთაოდ ცნობილი მეთოდი ბუნებრივ ენათა მორფოლოგიური ანალიზისათვის,

მაგრამ გარკვეულ მოდიფიკაციას მოითხოვს იმ ენათა სისტემების აღწერისას, რომლებიც ფორმაწარმოების მრავალფეროვნებით ხასიათდება.

ჩვენ საკითხის გადაწყვეტის სხვა გზა ავირჩიეთ: ქართული ენისათვის შევქმენით ლინგვისტური მოდელი. მასში აღწერილია ენის ფლექსიური და დერივაციული სისტემები, აგრეთვე მორფონოლოგიური და მორფოსინტაქსური პროცესები. ამ ინფორმაციის საფუძველზე განხორციელდა პროგრამული უზრუნველყოფა: შეიქმნა ინსტრუმენტი – მორფოლოგიური პროცესორი, რომლითაც შესაძლებელია ერთი ძირიდან სხვადასხვა ფუძის წარმოება დერივაციისა და კომპოზიციის საშუალებით. მაგალითად: ძირიდან „ძმა“ გენერირდება შემდეგი ფუძეები: დერივაციით – მოძმე, საძმო, უძმო, ძმობილ, ძმურ და მისთ. კომპოზიციით – და-ძმა, დედისძმა, ნახევარძმა, ცოლისძმა, ძმაბიჭ, ძმადნაფიც, ძმაკაც, ძმისშვილ და მისთ. თითოეული ფუძისაგან კი გენერირდება სახელური და ზმნური ფორმები. ზემოთ წარმოდგენილ ყველა ფუძეს აქვს სახელის პარადიგმა, ზოგი ფუძისაგან კი (ძმ, ძმობილ. ძმაბიჭ...) იწარმოება აგრეთვე ზმნური პარადიგმები. ამგვარად წარმოდგენილი მონაცემები საშუალებას იძლევა განხორციელდეს ქართული ენის სრული მორფოლოგიური სინთეზი – აღიწეროს როგორც ფლექსიური, ისე დერივაციული სისტემები.

სტატიაში წარმოვადგენთ ლინგვისტური მოდელის შექმნის ჩვენეულ მეთოდს, კერძოდ, განვიხილავთ სახელის მოდელსა და მასზე დაფუძნებული სინთეზის მექანიზმს. აღვწერთ პროცესორის სტრუქტურას.

2. ქართული ენის სახელის ლინგვისტური მოდელი.

ჩვენს ლინგვისტურ მოდელში ენა არ არის აღწერილი რომელიმე კონკრეტული გრამატიკული თეორიის მიხედვით, არამედ არსებული თეორიების გათვალისწინებით ენობრივი მონაცემები წარმოვადგინეთ განსხვავებული ფორმატით – ფორმალიზებული სახით. მასში სრულად არის ასახული ენის მორფოლოგიური სისტემა. ფორმალური წესების საშუალებით აღიწერა ქართული ენის ფლექსიური და დერივაციული შესაძლებლობანი, ფონოლოგიური პროცესები, მორფოსინტაქსური თავისებურებანი და სხვ.

მანქანური დამუშავებისათვის შექმნილ ლინგვისტურ მოდელში გარკვეულ შემთხვევებში, კომპიუტერული დამუშავების სპეციფიკის გამო, ტრადიციული გრამატიკებისაგან განსხვავებულად აღვწერთ ენობრივ მოვლენებს. მაგალითად, მო-

დელ'ში სახელურ ფორმათა რიგ'ში განიხილება შედგენილი შემასმენლები, რომლებიც წარმოადგენს ერთ სიტყვას, მაგრამ მოიცავენ ორ კომპონენტს: სახელურ ნაწილსა და დამხმარე ზმნის ხმოვან ელემენტს, მაგალითად:

ა) სახლია = სახლი + ა

სახლ-ი – სახ. ბრუნვის ფორმა

ა (<- არის) – მეშველი ზმნა

ბ) სახლისაა = სახლისა + ა

სახლ-ის-ა – ნათ. ბრუნვის ფორმა, გავრცობილი

ა (<- არის) – მეშველი ზმნა

გ) სახლ'შია = სახლ'ში + ა

სახლ-'ში – თანდებულიანი სახელი

ა (<- არის) – მეშველი ზმნა და ა. შ.

ტრადიციულ გრამატიკებ'ში, ბუნებრივია, ეს სიტყვები სახელებ'ში არ განიხილება, მაგრამ ჩვენ მათ სახელურ ფორმათა გვერდით წარმოვადგენთ მანქანური დამუშავების სპეციფიკის გამო: ანალიზატორმა უნდა იცნოს ერთ სიტყვად გაფორმებული ნებისმიერი ერთეული, ამისათვის კომპიუტერულ მოდელ'ში ასახული უნდა იყოს ყველა ამგვარი მონაცემი. ზემოთ დასახელებულ შედგენილ შემასმენლებს (სახლი-ა, სახლისა-ა, სახლ'ში-ა...) ზმნურ პარადიგმებ'ში ვერ განვიხილავთ, რადგან ზმნური ელემენტი აქ არის მხოლოდ -ა, სიტყვის მთავარი ნაწილი კი სახელური ფორმაა (ბრუნვის ფორმა, თანდებულიანი ან ნაწილაკიანი სახელი). ამიტომ ამგვარ ერთეულებს ჩვენ სახელური პარადიგმების წევრებად წარმოვადგენთ, მიუხედავად იმისა, რომ ისინი ერთგვარი შესიტყვებებია (უფრო სწორად, მომდინარეობს შესიტყვებისაგან) და არა სახელური სიტყვაფორმები.

ტრადიციული გრამატიკებისაგან განსხვავებული აღწერაა წარმოდგენილი ჩვენს მოდელ'ში სხვა შემთხვევებ'შიც.

2.1. მოდელ'ში ასახული სახელური ფორმები

მოდელ'ში ასახულია ყველა სახელური ფორმა, რომლებიც ერთი ფუძიდან შეიძლება იწარმოოს, კერძოდ, ესენია:

1. ბრუნვის ფორმები (გაუვრცობელი):

მხ. რ.	**მრ. რ.** (ებიანი)	**მრ. რ.** (ნართანიანი)
სახ. მთა	მთ-**ებ-ი**	მთა-**ნ-ი**
მოთხ. მთა-**მ**	მთ-**ებ-მა**	მთა-**თა**
მიც. მთა-**ს**	მთ-**ებ-ს**	მთა-**თა**
ნათ. მთ-**ის**	მთ-**ებ-ის**	მთა-**თა**

და ა. შ.

2. გავრცობილი ბრუნვის ფორმები:

მხ. რ.	**მრ. რ.**
მიც. მთას-**ა**	მთებს-**ა**
ნათ. მთის-**ა**	მთების-**ა**

და ა. შ.

3. თანდებულიანი ფორმები:

მხ. რ.	**მრ. რ.**
მთა-**ზე**	მთებ-**ზე**
მთის-**თვის**	მთების-**თვის**
მთ-**ამდე**	მთებ-**ამდე**

და ა. შ.

4. გავრცობილი თანდებულიანი ფორმები:

მხ. რ.	**მრ. რ.**
მთი**სთვის**-ა	მთე**ბისთვის**-ა
მთა**მდის**-ა	მთე**ბამდის**-ა

და ა.შ.

5. ნაწილაკდართული ფორმები:

- **ნაწილაკდართული ბრუნვის ფორმები:**
  მთამა-**ც**, მთასა-**ც**, მთას-**ღა**... მთებსა-**ც**...
- **ნაწილაკდართული თანდებულიანი ფორმები:**
  მთაზე-**ც**, მთისთვისა-**ც**, მთისგან-**ვე**, მთებში-**ღა**...

6. გავრცობილი ნაწილაკდართული ფორმები:

მთაც-**ა**, მთისაც-**ა**, მთაზეც-**ა**, მთამდეც-**ა**...

7. **მეშველზმნიანი ფორმები**:
   - **მეშველზმნიანი ბრუნვის ფორმები:**
     მთა-**ა**, მთასა-**ა**, მთითა-**ა**...
   - **მეშველზმნიანი თანდებულიანი ფორმები:**
     მთაში-**ა**, მთაზე-**ა**, მთასთანა-**ა**...
   - **მეშველზმნიანი ნაწილაკიანი ფორმები:**
     მთაცა-**ა**, მთებიღა-**ა**, მთაშიცა-**ა**...
8. **სხვათა სიტყვის ნაწილაკის დართვით მიღებული ფორმები**:
   - **სხვათა სიტყვის ნაწილაკის დართვით მიღებული ბრუნვის ფორმები:**
     მთამ-**ო**, მთის-**მეთქი**, მთებით-**თქო**...
   - **სხვათა სიტყვის ნაწილაკის დართვით მიღებული თანდებულიანი ფორმები:**
     მთაში-**ო**, მთაზე-**მეთქი**, მთასთან-**თქო**...
   - **სხვათა სიტყვის ნაწილაკის დართვით მიღებული ნაწილაკიანი ფორმები:**
     მთაცა-**ო**, მთებიღა-**მეთქი**, მთაშიც-**თქო**...
   - **სხვათა სიტყვის ნაწილაკის დართვით მიღებული მიღებული მეშველზმნიანი ფორმები:** მთაა-**ო**, მთებია-**მეთქი**, მთაშია-**თქო**...

როგორც წარმოდგენილი ნიმუშებიდან ჩანს, ქართულ ენაში სახელები მრავალფეროვანი ფორმაწარმოებით გამოირჩევა. ერთ სიტყვაფორმაში შეიძლება შეგვხვდეს ერთი, ორი, სამი და ა. შ. მორფემა. მორფემათა მაქსიმალური რაოდენობა არის 10. წარმოვადგენთ ნიმუშებს სხვადასხვა რაოდენობის მორფემათა შემცველი სახელებისთვის:

1. **ერთი მორფემა** – მთა
2. **ორი მორფემა** – მთა-ს
   ფუძე (**მთა**) + ბრუნვის ნიშანი (**ს**)
3. **სამი მორფემა** – მთ-ის-გან
   ფუძე (**მთ**) + ბრუნვის ნიშანი (**ის**) + თანდებული (**გან**)
4. **ოთხი მორფემა** – მთ-ის-ა-გან
   ფუძე (**მთ**) + ბრუნვის ნიშანი (**ის**) + საგრძობი (**ა**) + თანდებული (**გან**)
5. **ხუთი მორფემა** – მთ-ებ-ის-ა-გან

ფუძე **(მთ)** + რიცხვის ნიშანი **(ებ)** + ბრუნვის ნიშანი **(ის)** + სავრცობი **(ა)** + თანდებული **(გან)**

6. **ექვსი მორფემა** – მთ-ის-ა-გან-ა-ც
   ფუძე **(მთ)** + ბრუნვის ნიშანი **(ის)** + სავრცობი **(ა)** + თანდებული **(გან)** + სავრცობი **(ა)** + ნაწილაკი **(ც)**
7. **შვიდი მორფემა** – მთ-ებ-ის-ა-გან-ა-ც
   ფუძე (მთ) + რიცხვის ნიშანი (ებ) + ბრუნვის ნიშანი (ის) + სავრცობი (ა) + თანდებული (გან)+ სავრცობი **(ა)** + ნაწილაკი **(ც)**
8. **რვა მორფემა** – მთ-ებ-ის-ა-გან-ა-ც-ო
   ფუძე **(მთ)** + რიცხვის ნიშანი **(ებ)** + ბრუნვის ნიშანი **(ის)** + სავრცობი **(ა)** + თანდებული **(გან)**+ სავრცობი **(ა)** + ნაწილაკი **(ც)** + სხვათა სიტყვის ნაწილაკი **(ო)**
9. **ცხრა მორფემა** – მთ-ებ-ის-ა-გან-ა-ც-ა-ა
   ფუძე **(მთ)** + რიცხვის ნიშანი **(ებ)** + ბრუნვის ნიშანი **(ის)** + სავრცობი **(ა)** + თანდებული **(გან)**+ სავრცობი **(ა)** + ნაწილაკი **(ც)**+ სავრცობი **(ა)** + მეშვეული ზმნა **(ა)**
10. **ათი მორფემა** – მთ-ებ-ის-ა-გან-ა-ც-ა-ა-ო
    ფუძე **(მთ)** + რიცხვის ნიშანი **(ებ)** + ბრუნვის ნიშანი **(ის)** + სავრცობი **(ა)** + თანდებული **(გან)**+ სავრცობი **(ა)** + ნაწილაკი **(ც)**+ სავრცობი **(ა)** + მეშვეული ზმნა **(ა)** + სხვათა სიტყვის ნაწილაკი **(ო)**

ჩვენს ლინგვისტურ მოდელში ასახულია სხვადასხვა მორფემის კომბინაციით მიღებული ყველა სახელი. როგორც ცნობილია, ფუძეებს აფიქსური მორფემები დაერთვის განსაზღვრული წესების მიხედვით. ამასთან, ზოგჯერ სახელურ ფორმებში ხდება ფონოლოგიური ცვლილებები. ეს კანონზომიერებანი აღწერილია შესაბამისი წესებისა და მონაცემების სახით, რომლებიც განთავსებულია მოდელის სხვადასხვა კომპონენტში.

2.2. სახელის მოდელის კომპონენტები.

სახელის მოდელი შედგება შემდეგი კომპონენტებისაგან:

- მორფემათა რანგობრივი ცხრილები
- სიტყვაფორმათა შაბლონები
- ფონოლოგიური ტიპების სია

განვიხილავთ ამ კომპონენტებს.

2.2.1. მორფემათა რანგობრივი ცხრილი.

მართალია, ქართულ ენაში სახელის ფლექსიურ აფიქსთა რაოდენობა მნიშვნელოვნად დიდია, მაგრამ ისინი განაწილებულია მკაცრი რანგობრივი პრინციპით, კერძოდ:

- სიტყვაფორმებში მორფემათა მიმდევრობა წარმოდგენილია რანგების მიხედვით;
- სიტყვაფორმაში დასტურდება ამა თუ იმ რანგის მხოლოდ ერთი ელემენტი.

ნიმუშად წარმოვადგენთ სახელურ მორფემათა ცხრილის ფრაგმენტს.

N	R	A	B	C	D	E	F	G	H	I
1		-ებ	-0	-ა	-ში	-ა	-ც	-ა	-ა	-მეთქი
2		-ნ	-ი	-ი	-ზე		-და			-თქო
3		-თ	-მა		-დან		-ვე			-ო
4			-მ		-იდან					
5			-ს		-მდე					
6			-ის		-ამდე					
7			-ს		-თვის					

ცხრილი 1. სახელურ მორფემათა ცხრილის ფრაგმენტი
R-ფუძე, A-რიცხვის ნიშანი, B-ბრუნვის ნიშანი, C-სავრცობი1, D-თანდებული,
E-სავრცობი2, F-ნაწილაკი, G-სავრცობი3, H-მეშველი ზმნა, I-სიტყვა-სიტყვითი ნაწილაკი

ცხრილი 1-ის შესაბამისად, მოდელში თითოეული მორფემა გამოისახება სიმბოლური ნიშნით, მაგალითად:

ებ – A_1

ნ – A_2

თ – A_3

ი – B_2

მა – B_3

მ – B_4

ს – B_5 და ა. შ.

ცხრილში მორფემები წარმოდგენილია ენის მანქანური დამუშავების სპეციფიკის გათვალისწინებით და ზოგ შემთხვევაში აფიქსთა აღწერა არ ემთხვევა გრამატიკებში წარმოდგენილ მონაცემებს. განვიხილავთ ამგვარ შემთხვევებს.

როგორც ცნობილია, ქართული ენის სახელურ ფორმებში დასტურდება სავრცობი -ა. ტრადიციულ გრამატიკებში განიხილება ის აფიქსები, რომელთა გავრცობაც ხდება; კერძოდ, ცნობილია სავრცობის გამოყენების პოზიციები: ბრუნვის ნიშნის შემდეგ: სახლ-ს-ა, სახლ-ის-ა... თანდებულის წინ ან შემდეგ: სახლის-ა-თვის, სახლის-თვის-ა... ნაწილაკის წინ ან შემდეგ: სახლს-ა-ც, სახლში-ც-ა... და ა. შ.

მართალია, -ა ერთი მორფოლოგიური ოდენობაა (სავრცობი), მაგრამ მორფემათა რანგობრივ ცხრილში იგი წარმოდგენილია სამ სხვადასხვა სვეტში (C, E, G) იმის მიხედვით, თუ რომელი აფიქსია გავრცობილი, ე. ი. -ა რანგობრივად რომელ პოზიციას იკავებს სხვა მორფემებთან მიმართებით:

- თუ გავრცობილია ბრუნვის ნიშანი (B რანგის მორფემა), მაშინ ა = C_1 : ბავშვ-ს-ა;
- თუ გავრცობილია თანდებული (D რანგის მორფემა), მაშინ ა = E_1 : ბავშვ-თან-ა;
- თუ გავრცობილია ნაწილაკი (F რანგის მორფემა), მაშინ ა = G_1 : ბავშვი-ც-ა.

-ა სავრცობის ამგვარი დიფერენცირება მნიშვნელოვანია, რადგან ერთ ფორმაში შეიძლება მონაწილეობდეს ორი ან სამი -ა და თითოეულ მათგანს უნდა ჰქონდეს შესაბამისი სიმბოლური ნიშანი, რომ პროგრამამ შეძლოს ფორმის სინთეზი (ან ანალიზი). მაგალითად:

ბავშვ-ის-ა-თვის-ა-ც-ა – ამ ფორმაში პირველი ა = C_1 , მეორე ა = E_1 , და მესამე ა = G_1 .

უნდა აღინიშნოს ისიც, რომ ზოგჯერ სხვადასხვა სუფიქსი ფორმალურად ემთხვევა ერთმანეთს. ასეთ შემთხვევებში ისინი ორჯერ არის შეტანილი ცხრილში მათი ფუნქციური სხვაობის დასაფიქსირებლად. მაგალითად, -ს არის როგორც მიცემითი, ისე ნათესაობითი ბრუნვის ნიშანი, ცხრილში ისინი ამგვარად აისახება:

მიცემითი ბრუნვის ნიშანი: ს = B_5

ნათესაობითი ბრუნვის ნიშანი: ს = B_7

მაშასადამე, თითოეულ სახელურ მორფემას ცხრილში შეესაბამება თავისი სიმბოლური ნიშანი. ნებისმიერი სიტყვაფორმა შეიძლება გამოისახოს ამ ნიშნებით, რომელთა ერთობლიობა ქმნის სიტყვის მორფოლოგიურ ფორმულას. მაგალითად:

ბავშვ-ებ-მა – R + A_1 + B_3

ბავშვ-ის-ა-თვის – R + B_6 + C_1 + D_7 …

მორფოლოგიური ფორმულებით გამოსახვა არის ჩვენ მიერ შემოთავაზებული მეთოდი, რომლითაც შესაძლებელია ყველა ტიპის სიტყვაფორმის ფორმალიზებული სახით წარმოდგენა.

ფორმულით გამოისახება ნებისმიერი გრამატიკული ფორმა. მაგალითად:

ა) მოთხრობითი ბრუნვის ფორმის ფორმულა არის: R + B_3,

ბ) მრავლობით რიცხვში ნათესაობითი ბრუნვის ფორმის ფორმულა არის: R + A_1 + B_6

გ) -ზე თანდებულიანი ფორმის ფორმულა არის: R + D_2 და ა. შ.

მორფოლოგიურ ფორმულებში R (ფუძე) არის ცვლადი ერთეული, ფორმულის ყველა დანარჩენი კომპონენტი არის უცვლელი (განსაზღვრული) მონაცემი. ერთი და იმავე ფორმულით სხვადასხვა ფუძისგან შეიძლება მივიღოთ კონკრეტული გრამატიკული ფორმები. მაგალითად, თუ ავიღებთ მოთხრობითი ბრუნვის ფორმულას (R + B_3) და ვცვლით R-ს, მივიღებთ მოთხრობითი ბრუნვის ფორმებს.

მოთხრობითი ბრუნვის ფორმულა: **R + B_3**

B_3 = -მა

ფორმულაში **R**-ის ცვლით მიღებული სიტყვაფორმები:

R= კაც: კაც-მა

R= სახლ: სახლ-მა

R= ბავშვ: ბავშვ-მა...

თუ სახელი ხმოვანფუძიანია, მაშინ მოთხრობითი ბრუნვის ფორმულა იქნება ამგვარი: **R + B_4**

B_4 = -მ

ამ ფორმულით სხვადასხვა ფუძისგან მიიღება შემდეგი სიტყვაფორმები:

R= მთა: მთა-მ

R= დედა: დედა-მ

R= სპილო: სპილო-მ...

თუ ავიღებთ **-ზე** თანდებულიანი სახელის ფორმულას (R + D_2) და ვცვლით R-ს, მივიღებთ -ზე თანდებულიან ფორმებს:

-ზე თანდებულიანი სახელის ფორმულა: **R + D_2**

D_2 = -ზე

ფორმულაში R-ის ცვლით მიღებული სიტყვაფორმები:

R = კაც: კაც-ზე

R = სახლ: სახლ-ზე

R = ბავშვ: ბავშვ-ზე

R = მთა: მთა-ზე

R = დედა: დედა-ზე

R = სპილო: სპილო-ზე...

მაშასადამე, მორფოლოგიური ფორმულა არის გარკვეული ყალიბი, რომლის მიხედვით პროგრამას შეუძლია ამა თუ იმ ფუძიდან შესაბამისი სიტყვაფორმების სინთეზი.

2.2.2. სიტყვაფორმათა შაბლონები

შაბლონი წარმოადგენს იმ მორფოლოგიური ფორმულების ერთობლიობას, რომლებიც საჭიროა ერთი ფუძიდან ყველა სიტყვაფორმის დასაგენერირებლად, მაშასადამე, სრული პარადიგმის მისაღებად. შაბლონში სიტყვაფორმათა ნიმუშები ჩაწერილია ფორმულების სახით.

ცნობილია, რომ ქართულ ენაში სახელთა გარკვეული ჯგუფები ერთმანეთისაგან განსხვავდება ფორმაწარმოების თვალსაზრისით. შესაბამისად, გამოიყოფა შაბლონთა სხვადასხვა ტიპი. ისინი დამოკიდებულია სახელის ფუძეზე და პრაქტიკულად ემთხვევა ბრუნების ტიპებს, რომლებსაც ქართულ ენაში სხვადასხვა მეცნიერი განსხვავებულად გამოყოფს: შანიძე (1980), ჩიქობავა (1998), ონიანი (2003), უთურგაიძე (2009), არაბული (2016), ქირია (2022) და სხვ. ტიპებად დაყოფის პრინციპი განსხვავებულია: ფონოლოგიური, მორფოლოგიური, მორფონოლოგიური და ა. შ. მოდელისათვის რელევანტურია ფუძეზე დართულ მორფემათა აღწერა, რადგან მათი ერთობლიობა ქმნის ყალიბს (შაბლონს), რომლის საფუძველზეც უნდა განხორციელდეს სიტყვაფორმათა გენერაცია. ამიტომ ტიპიზაციისას ჩვენ ვითვალისწინებთ მორფოლოგიურ პრინციპს: შაბლონებს გამოვყოფთ იმის მიხედვით, თუ რომელ აფიქსებს (ბრუნვის ნიშნებს, თანდებულებს, ნაწილაკებს...) დაირთავს სახელი.

ტრადიციულ გრამატიკებში კუმშვად-კვეცადი და ი-ბოლოხმოვნიანი უკვეცელი სახელები ცალკე ტიპებად არ განიხილება. ჩვენ მათ გამოვყოფთ ცალკე შაბ-

ლონების სახით, რადგან ისინი განსხვავდება ყველა დანარჩენისგან ფორმაწარმოების (ფუძეზე დართული მორფემების) თვალსაზრისით. მართალია, სხვაობას ქმნის სულ რამდენიმე ფორმა, მაგრამ ამ უკანასკნელთა წარმოება შეუძლებელია სპეციალური შაბლონის გარეშე.

გვაქვს 5 ტიპის შაბლონი. წარმოვადგენთ მათ ცხრილის სახით. თითოეულ ტიპს ნიმუშად ვუწერთ ერთ-ერთ ამგვარ სახელს, რომელიც აზუსტებს, რა სახის სიტყვათა შაბლონია ის.

ბრუნვა	**I ტიპი „კაცი“**	**II ტიპი „დედა“**	**III ტიპი „ფანჯარა“**	**IV ტიპი „წყარო“**	**V ტიპი „ტრამვაი“**
სახელობითი	-ი	-Ø	-Ø	-Ø	-Ø
მოთხრობითი	-მა	-მ	-მ	-მ	-მ
მიცემითი	-ს	-ს	-ს	-ს	-ს
ნათესაობითი	-ის	-ის	-ის	-ს	-ს
მოქმედებითი	-ით	-ით	-ით	-თი	-თ
ვითარებითი	-ად	-დ	-ად	-დ	-დ
წოდებითი	-ო	**-Ø /-ვ / -ო**	**-Ø /-ვ**	**-Ø /-ვ / -ო**	**-Ø /-ვ / -ო**

ცხრილი 2. სახელის შაბლონთა ტიპები

წარმოდგენილი 5 შაბლონი აღწერს სახელთა ბრუნების ძირითად (რეგულარულ) ტიპებს; ამის გარდა არსებობს არარეგულარული ბრუნების ტიპები, რომლებსაც მიეკუთვნება ნაცვალსახელები; ამგვარი შემთხვევებისთვის ცალკე შაბლონები გამოიყოფა.

არარეგულარულ შაბლონებს ქმნიან აგრეთვე ე. წ. ფორმაუცვლელი სიტყვები (ზმნიზედები, თანდებულები, ნაწილაკები და შორისდებულები). როგორც ცნობილია, ეს მეტყველების ნაწილები ფორმაუცვლელ სიტყვათა ჯგუფს მიეკუთვნება, შესაბამისად, ტრადიციულ გრამატიკაში ისინი სახელურ ფლექსიაში არ განიხილება. აღნიშნული სიტყვები, მართალია, არ იბრუნვის, მაგრამ ქართულ ენაში ფორმაწარმოება გარკვეული სახით მათაც ახასიათებს, რადგან დაირთავენ ფლექსიურ აფიქსებს, კერძოდ, სავრცობს, თანდებულებსა და ნაწილაკებს. მაგალითად, ზმნიზედა **აქ** დაირთავს თანდებულებს: აქე-**დან**, აქა-**მდე**... ნაწილაკებს: აქა-**ც**, აქ-**ვე**, აქ-**ღა**... ამიტომ ამ ტიპის სიტყვებიც მოდელში სახელური ფორმატით აღიწერება. მათთვის ცალკე შაბლონებია შექმნილი.

მოდელში სახელთათვის 20-ზე მეტი შაბლონია (რეგულარული თუ არარეგულარული) წარმოდგენილი. თითოეულ მათგანში ფორმულების სახით აღწერილია: ბრუნვის ფორმები, თანდებულიანი ფორმები, ნაწილაკდართული ფორმები,

მეშველზმნიანი ფორმები და სხვათა სიტყვის ნაწილაკების დართვით მიღებული ფორმები.

ნიმუშად წარმოვადგენთ პირველი შაბლონის ფრაგმენტებს; მოდელში შაბლონი მოიცავს მხოლოდ ფორმულებს, აქ კი, ფორმულების უკეთ გააზრების მიზნით, ფრჩხილებში მითითებულია მათი შესაბამისი ნიმუშები (სიტყვაფორმები) ფუძისათვის **სახლ**.

შაბლონი 1

ბრუნვის ფორმები

მხ. რ.		მრ. რ.		
სახ.	$R+B_2$	(სახლ-ი)	$R+A_1+B_2$	(სახლ-ებ-ი)
მოთხ.	$R+B_3$	(სახლ-მა)	$R+A_1+B_3$	(სახლ-ებ-მა)
მიც.	$R+B_5$	(სახლ-ს)	$R+A_1+B_5$	(სახლ-ებ-ს)
ნათ.	$R+B_6$	(სახლ-ის)	$R+A_1+B_6$	(სახლ-ებ-ის)
მოქ.	$R+B_7$	(სახლ-ით)	$R+A_1+B_7$	(სახლ-ებ-ით)
და ა. შ.				

თანდებულიანი ფორმები (გაუვრცობელი და გავრცობილი)

$R+D_1$	(სახლ-ში)
$R+D_2$	(სახლ-ზე)
$R+D_9$	(სახლ-თან)
$R+D_9+E_1$	(სახლ-თან-ა)
$R+B_2+D_{10}$	(სახლ-ი-ვით)
$R+B_2+D_{10}+E_1$	(სახლ-ი-ვით-ა)
$R+D_4$	(სახლ-იდან)
$R+D_4+E_1$	(სახლ-იდან-ა)
და ა. შ.	

ნაწილაკიანი ფორმები (გაუვრცობელი და გავრცობილი)

$R+B_2+F_1$	(სახლ-ი-ც)
$R+B_2+F_1+G_1$	(სახლ-ი-ც-ა)
$R+B_3+F_1$	(სახლ-მა-ც)
$R+B_3+F_1+G_1$	(სახლ-მა-ც-ა)
$R+B_5+C_1+F_1$	(სახლ-ს-ა-ც)
$R+B_5+C_1+F_1+G_1$	(სახლ-ს-ა-ც-ა)
$R+D_1+F_1$	(სახლ-ში-ც)
$R+D_1+F_1+G_1$	(სახლ-ში-ც-ა)
$R+D_2+F_1$	(სახლ-ზე-ც)

$R+D_2+F_1+G_1$	(სახლ-ზე-ც-ა)
$R+D_6+F_1$	(სახლ-ამდე-ც)
$R+D_6+F_1+G_1$	(სახლ-ამდე-ც-ა)

და ა. შ.

მეშველზმნიანი ფორმები

$R+B_2+H_2$	(სახლ-ი-ა)
$R+B_5+C_1+H_2$	(სახლ-ს-ა-ა)
$R+B_7+C_1+H_2$	(სახლ-ით-ა-ა)
$R+A_1+B_2+H_2$	(სახლ-ებ-ი-ა)
$R+A_1+B_5+C_1+H_2$	(სახლ-ებ-ს-ა-ა)
$R+A_1+B_7+C_1+H_2$	(სახლ-ებ-ით-ა-ა)
$R+D_1+H_2$	(სახლ-ში-ა)
$R+D_2+H_2$	(სახლ-ზე-ა)
$R+D_4+E_1+H_2$	(სახლ-იდან-ა-ა)
$R+D_6+H_2$	(სახლ-ამდე-ა)
$R+D_9+E_1+H_2$	(სახლ-თან-ა-ა)

და ა. შ.

ამგვარი ფორმით (ფორმულების სახით) არის წარმოდგენილი შაბლონებში სახელური სიტყვაფორმები. რეგულარულ შაბლონებში ფორმათა რაოდენობა მერყეობს 736-იდან 756-ამდე. საშუალოდ – 746. ეს არის სახელები სხვათა სიტყვის ნაწილაკების (-მეთქი, -თქო, -ო) გარეშე. ყველა სიტყვას დაერთვის სამივე ნაწილაკი. საბოლოოდ, შაბლონში სიტყვაფორმათა სრული რაოდენობა არის 4-ჯერ მეტი, ე. ი. საშუალოდ 3 000 ერთეული.

მაშასადამე, ქართულ ენაში ერთი ფუძიდან დასაშვებია დაახლოებით 3 000 სახელური სიტყვის გენერაცია. რასაკვირველია, აქედან ყველა სიტყვაფორმა არ დასტურდება ხელმისაწვდომი ელექტრონული რესურსების ტექსტებში. ამის შესახებ ქვემოთ გვექნება მსჯელობა (იხ. 5. პოტენციური ფორმები).

2.2.3. ფონოლოგიური ტიპები.

მოდელის კიდევ ერთი კომპონენტია ფონოლოგიური ტიპების ჩამონათვალი. სახელთა ფორმაწარმოებისას მნიშვნელოვანია სახელურ ფუძეთა ფონოლოგიური მახასიათებელი, რომელიც განსაზღვრავს, იცვლება თუ არა ფუძე ფორმაწარმოების დროს და თუ იცვლება, რა ტიპის ცვლილებები ახასიათებს მას.

აქაც გვაქვს გარკვეული სხვაობა ტრადიციულ გრამატიკებში მოცემულ ფონოლოგიურ ტიპებსა და ლინგვისტურ მოდელში დაფიქსირებულ მონაცემებს შორის. მაგალითად, როგორც ცნობილია, არსებობს ხმოვანფუძიანი კვეცადი სახელები, რომელთა ბოლო ხმოვანი იკარგება (იკვეცება) მხოლობითი რიცხვის ორ ბრუნვასა (ნათესაობითი და მოქმედებითი) და ებ-იან მრავლობითში. მაგრამ ეს პროცესი ყველგან ერთგვაროვანი არ არის, კერძოდ: ხმოვანფუძიან სახელთა ნაწილი იკვეცება სამივე პოზიციაში (**დედა**: **დედ**-ის, **დედ**-ით, **დედ**-ებ-ი...), ნაწილი – მარტო მხოლობით რიცხვში (**სარკე**: **სარკ**-ის, **სარკ**-ით, მაგრამ მრ. რ. – **სარკე**-ებ-ი...), ნაწილი კი – მხოლოდ ებიან მრავლობითში (**გოგონა**: **გოგონა**-სი, **გოგონა**-თი, მაგრამ მრ. რ. – **გოგონ**-ებ-ი...). ფონოლოგიური ცვლილების მხოლოდ ერთი ტიპი (კვეცადი) ვერ ასახავს ამ სხვაობას, ამიტომ ზემოთ დასახელებული ვარიანტებისთვის ჩვენ გამოვყოფთ სამ სახეობას: კვეცადი მხოლობითში, კვეცადი მრავლობითში და კვეცადი ორივე რიცხვში. ამავე მიზეზით ცალკე ტიპებად არის გამოყოფილი ორი სახეობა კუმშვად სახელებშიც: კუმშვადი ორივე რიცხვში (**სოფელი**: **სოფლ**-ის, **სოფლ**-ებ-ი...) და კუმშვადი მხოლობითში (**სასმელი**: **სასმლ**-ის, მაგრამ მრ. რ. – **სასმელ**-ებ-ი...). ფონოლოგიური ტიპების ამგვარი დანაწევრება საშუალებას იძლევა, ზუსტად განხორციელდეს აღნიშნული ცვლილებები ფორმათა სინთეზის დროს.

წარმოვადგენთ ფონოლოგიურ ტიპებს:

- უცვლელი (კაცი, წყარო)
- კუმშვადი ორივე რიცხვში (სოფელი)
- კუმშვადი მხოლობითში (სასმელი)
- ნაწილობრივ კუმშვადი (მინდორი)
- კუმშვად-თანხმოვანდამკარგველი (ამბავი)
- კვეცადი ორივე რიცხვში (დედა)
- კვეცადი მხოლობითში (სარკე)
- კვეცადი მრავლობითში (გოგონა)
- კუმშვად-კვეცადი (ფანჯარა)

მოდელში აღწერილია, რა ცვლილებები ხორციელდება თითოეული ფონოლოგიური ტიპის შემთხვევაში და რომელ პოზიციებში. ამჯერად შემოვიფარგლებით მხოლოდ ამ ტიპების დასახელებით, ფონოლოგიური ცვლილებების მოდელს სამომავლოდ ცალკე უფრო დეტალურად განვიხილავთ.

3. ქართული ენის მორფოლოგიური პროცესორი.

ზემოთ განხილულ ლინგვისტურ მოდელზე დაფუძნებით შევქმენით პროგრამული ინსტრუმენტი – ქართული ენის მორფოლოგიური პროცესორი. იგი შედგება ორი კომპონენტისგან: მონაცემთა ბაზისა და პროგრამული ნაწილისაგან.

3.1. მონაცემთა ბაზა. ბაზა მოიცავს ლინგვისტურ მოდელსა და ძირების ლექსიკონს.

ლინგვისტურ მოდელში, როგორც ზემოთ აღვნიშნეთ, ცხრილებისა და სიების სახით არის წარმოდგენილი სიტყვაფორმათა სინთეზისა და ანალიზისათვის საჭირო სხვადასხვაგვარი ენობრივი მონაცემები: მორფემათა რანგობრივი ცხრილები, ფონეტიკური ტიპების ჩამონათვალი, სახელისა და ზმნის პარადიგმათა შაბლონები, პირისა და რიცხვის ნიშნები, პირთა კომბინაციები და ა. შ.

მოდელის განთავსება მონაცემთა ბაზაში საშუალებას იძლევა, საჭიროების შემთხვევაში მოდელში განხორციელდეს ცვლილებები პროგრამული ჩარევის გარეშე. გარდა ამისა, ამგვარად ორგანიზებული მოდელის საფუძველზე სამომავლოდ ვფიქრობთ სპეციალური ინსტრუმენტის შექმნას, რომელიც გამოიყენება სხვადასხვა ენის მორფოლოგიური სინთეზისა და ანალიზის შესასრულებლად. იგი განსაკუთრებით მორგებული ინსტრუმენტი იქნება აგლუტინაციურ ენათა დამუშავებისას.

ძირების ლექსიკონი მოიცავს „ქართული ენის განმარტებითი ლექსიკონის“ რვატომეულს, აგრეთვე სხვადასხვა წყაროდან (ლექსიკონებიდან, კორპუსებიდან...) დამატებულ მასალას. ამ ეტაპზე ლექსიკონში დაფიქსირებულია 24 000-ამდე ძირი და მათგან ნაწარმოები 90 000-ზე მეტი ფუძე.

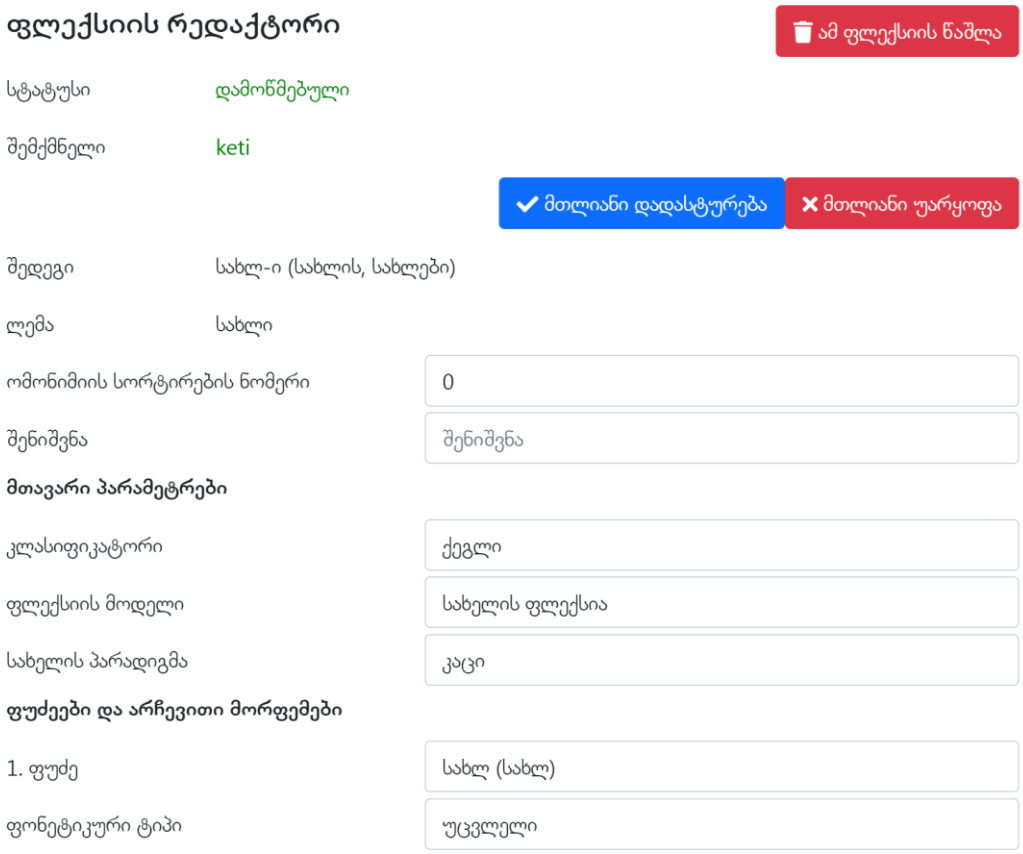
ლექსიკონში ძირები შეყვანილია შესაბამისი პარამეტრებით, კერძოდ, მიეთითება ძირის ფონეტიკური ტიპი და მისგან დერივაციული მოდელებით წარმოქმნილი ახალი ფუძეები, თითოეულ ფუძესთან დაფიქსირებულია მისგან მიღებულ სახელური და / ან ზმნური პარადიგმის შაბლონთა ტიპები, ფუძის ფონეტიკური ტიპი, ზმნური პარადიგმის შესაქმნელად საჭირო მონაცემები: ზმნის პირიანობა, გარდამავლობა და ა. შ.

სახელური ფორმების გენერაციისათვის ფუძეს მიეთითება ორი ინდექსი: შაბლონის ტიპი და ფონოლოგიური ტიპი. მაგალითად:

სახელის ფუძე	შაბლონის ტიპი	ფონოლოგიური ტიპი
მზე	„დედა“	კვეცადი მხოლობით'ში
ფურცელ	„კაცი“	კუმშვადი ორივე რიცხვ'ში
სახლ	„კაცი“	უცვლელი
დაფა	„დედა“	კვეცადი ორივე რიცხვ'ში
პეპელა	„ფანჯარა“	კუმშვად-კვეცადი
ენძელა	„წყარო“	კვეცადი მრავლობით'ში
კინო	„წყარო“	უცვლელი
ბგერა	„დედა“	კვეცადი ორივე რიცხვ'ში

ცხრილი 3. ლექსიკონ'ში განთავსებულ მონაცემთა ნიმუშები სახელის ფლექსიისათვის.

მონაცემთა ბაზის რედაქტორ'ში ეს ინფორმაცია დაფიქსირებულია შემდეგნაირად:



სურათი 1. პროცესორ'ში სახელის ფლექსიურ მონაცემთა შეყვანის ფორმატი

3.2. პროგრამული ნაწილი

პროგრამული ნაწილის კომპონენტებია: პროგრამა-რედაქტორი, სინთეზის პროგრამა (სინთეზატორი) და ანალიზის პროგრამა (ანალიზატორი).

პროგრამა-რედაქტორი არის მონაცემთა ბაზის მართვის სისტემა. ეს არის პროცესორის ის ნაწილი, რომელთანაც უშუალოდ უხდებათ მუშაობა ლინგვისტებს. იგი მოიცავს მოდელის რედაქტორსა და ძირების რედაქტორს. მათი საშუალებით შესაძლებელია ძირების ლექსიკონისა და მოდელის პარამეტრების შეყვანა, აგრეთვე მათი რედაქტირება: პარამეტრების შეცვლა ან ახლის დამატება.

მოდელსა და ძირების ლექსიკონში წარმოდგენილი მონაცემების საშუალებით სინთეზის პროგრამით (მორფოლოგიური სინთეზატორით) ხორციელდება სიტყვაფორმათა სინთეზი; სინთეზის ალგორითმის გამოყენებით ანალიზის პროგრამა (მორფოლოგიური ანალიზატორი) ახორციელებს უკუპროცესს: სიტყვაფორმათა ანალიზს.

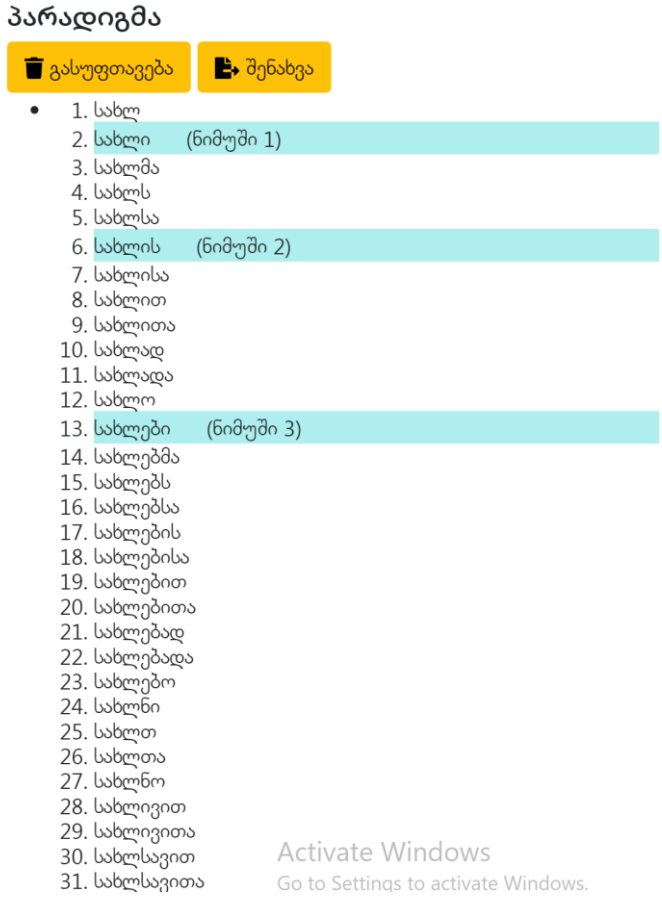
4. სახელის სინთეზით მიღებულ სიტყვაფორმათა სიები (პარადიგმები).

შაბლონისა და ფონოლოგიური ტიპის მიხედვით სინთეზატორში ხდება სახელთა ფლექსიური ფორმების გენერაცია, რომლის შედეგად პროგრამა გვაძლევს სახელური ფორმების სრულ სიას; სია მოიცავს ბრუნების პარადიგმას, თანდებულიან სახელებს, ნაწილაკიან სახელებს და ა. შ. პირობითად ამ სიას შეიძლება სახელის სრული პარადიგმა ვუწოდოთ. პარადიგმის ფორმებიდან ცალკე გამოიყოფა და სინთეზირებულ ფორმათა ჩამონათვალში ფიქსირდება ლემა (მხოლობითი რიცხვის სახელობითი ბრუნვის ფორმა), მოინიშნება და ლემასთან ერთად მიეთითება აგრეთვე მხოლობითი რიცხვის ნათესაობითი ბრუნვისა და ებიანი მრავლობითის სახელობითი ბრუნვის ფორმები, რომლებიც მნიშვნელოვანია სახელის ფონოლოგიური ტიპის საილუსტრაციოდ.

ქართულ ლექსიკოგრაფიაში ამგვარი პრაქტიკა არსებობს: ქართული ენის განმარტებით ლექსიკონში სახელის ფონოლოგიური ტიპის საილუსტრაციოდ ფრჩხილებში მითითებულია ნათესაობითი ბრუნვის ფორმა, მაგრამ ხშირ შემთხვევაში (როდესაც მხოლობითი და მრავლობითი რიცხვის ფორმები განსხვავებულია კუმშვის ან კვეცის თვალსაზრისით) ეს ინფორმაცია საკმარისი არ არის, ამიტომ ჩვენ ვუთითებთ ორ მონაცემს: როგორც მხოლობითი რიცხვის ნათესაობითი ბრუნვის, ისე ებ-იანი მრავლობითის ფორმას. მაგალითად:

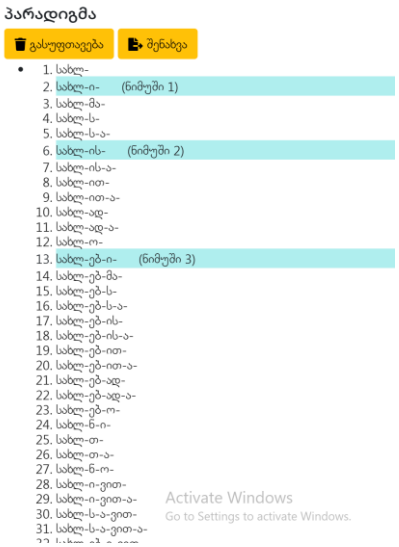
სარკე – სარკე (სარკის, სარკეები)

სიტყვაფორმათა პარადიგმას ამგვარი სახე აქვს:



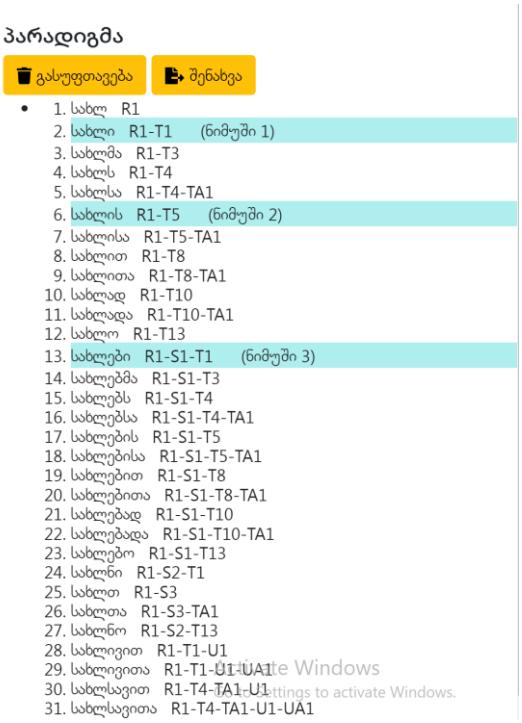
სურათი 2. სახელის სრული პარადიგმის ფრაგმენტი

პარადიგმაში შესაძლებელია სიტყვაფორმათა დაშლა აფიქსებად. ამისთვის არის სპეციალური ღილაკი: „დეფისებით“:



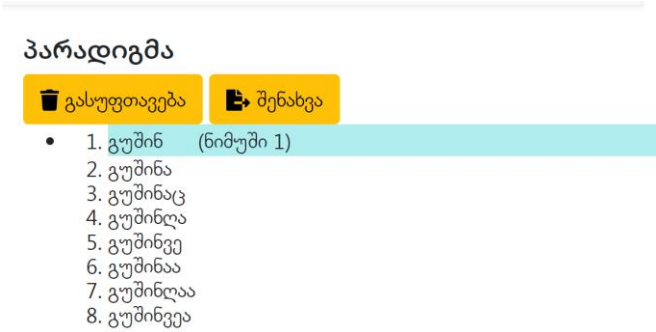
სურათი 3. მორფემებად დაშლილ სახელთა სრული პარადიგმის ფრაგმენტი

შესაძლებელია აგრეთვე სიტყვაფორმათა წარმოდგენა ფორმულების სახით:



სურათი 4. სახელებისა და მათი შესაბამისი ფორმულების სრული პარადიგმის ფრაგმენტი

არარეგულარული შაბლონები, ბუნებრივია, ფორმათა მცირე რაოდენობას შეიცავს. წარმოვადგენთ მათ ნიმუშსაც:



სურათი 5. არარეგულარული პარადიგმის ფრაგმენტი

5. პოტენციური ფორმები

როგორც ზემოთ აღვნიშნეთ, ჩვენს მიზანს წარმოადგენდა ისეთი მოდელის შექმნა, რომელიც უზრუნველყოფს თანამედროვე ქართულ ენაში არსებული და პოტენციურად დასაშვები ნებისმიერი სიტყვაფორმის სინთეზს.

დღეისათვის მორფოლოგიური პროცესორის მონაცემთა სტატისტიკა ამგვარია:

- ლემები – 113 860
- სინთეზატორით დაგენერირებული ფორმულები – 916 893
- სინთეზატორით დაგენერირებული სიტყვაფორმები – 266 478 980

მაშასადამე, ბოლო მონაცემებით, ქართული ენის მორფოლოგიური პროცესორი აგენერირებს 266 მილიონზე მეტ სიტყვაფორმას. ყველა ეს ფორმა არ არის დადასტურებული ელექტრონულ ტექსტებში ან დიდი მოცულობის კორპუსებში, მაგრამ პოტენციურად დასაშვებია (ლინგვისტური მოდელი განსაზღვრავს მათ გენერაციას). ამგვარ სიტყვებს პირობითად „პოტენციურ ფორმებს“ ვუწოდებთ.

პოტენციური ფორმების დაფიქსირება პროცესორში ჩვენ საჭიროდ მიგვაჩნია. მოსალოდნელია, რომ დღეისათვის „პოტენციურად“ მიჩნეული ფორმები შეგვხვდეს რომელიმე ახალ (წერილობით ან ზეპირმეტყველების) ტექსტში. პროცესორმა უნდა შეძლოს მათი ამოცნობა. გარდა ამისა, საინტერესოა განისაზღვროს, მოდელით ნაგულისხმევი სიტყვაფორმებიდან რომელი ვარიანტებია რეალიზებული ენაში (ამასთან – რა სიხშირით) და რომელი – არა. ეს საკითხი ცალკე კვლევის საგანია. საინტერესოა აგრეთვე სხვადასხვა ენის მონაცემთა შედარება ამ თვალსაზრისით.

კვლევის შემდგომ ეტაპზე დიდი მოცულობის ტექსტური მონაცემებიდან ვგეგმავთ მასალის მოძიებას, რის საფუძველზეც განისაზღვრება სიტყვაფორმათა სიხშირული მახასიათებლები: აქტიურია ის, პასიური თუ პოტენციური. პოტენციური ფორმები ზოგიერთი ინსტრუმენტისათვის (მართლწერის შემმოწმებელი პროგრამები, სათარგმნი სისტემები და სხვ.) რელევანტური არ არის. მაგრამ ისინი მნიშვნელოვანია ბუნებრივ ენათა დამუშავების სხვადასხვა საკითხის კვლევისას, აგრეთვე ხელოვნური ინტელექტის ცალკეული ამოცანების გადასაწყვეტად. სამომავლოდ დაზუსტდება პროცესორით გენერირებულ მონაცემთა სიხშირული მახასიათებლები და განისაზღვრება მათი გამოყენების სფეროები.

გამოყენებული ლიტერატურა

Meurer, P. (2007). A Computational Grammar for Georgian. Lecture Notes in Computer Science. Berlin: Springer.

ანთიძე, ჯ. (2009). ფორმალურ ენათა და გრამატიკათა თეორია, ბუნებრივი ენების კომპიუტერული მოდელირება, თბილისი: გამომცემლობა „ნეკერი“. ISBN 978-9941-404-92-4.

არაბული, ა. (2016). ახალი ქართული ენა, წიგნი I, სალიტერატურო ენის მორფოლოგია, გ. გოგოლაშვილისა და ა. არაბულის რედაქციით, თბილისი: „თბილისის სახელმწიფო უნივერსიტეტის გამომცემლობა“.

გოგოლაშვილი, გ. კვანტალიანი, ც. შენგელია, დ. (1989). ქართული ენის ზმნური ფუძეების ლექსიკონი, თბილისი: გამომცემლობა „მეცნიერება“.

გოგოლაშვილი, გ. კვანტალიანი, ც. შენგელია, დ. (1991). ქართული ენის სახელზმნური ფუძეების ლექსიკონი, თბილისი: გამომცემლობა „მეცნიერება“.

დათუკიშვილი ქ. (2020). ბრუნების ტიპებისათვის თანამედროვე ქართულში, იკე, XLVIII, თბილისი: „ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტის გამომცემლობა“ – https://ice.tsu.ge/wp-content/uploads/2021/04/IKE-48-1-265-%E2%80%93-2020.pdf

დათუკიშვილი, ქ. ლოლაძე, ნ. ზაკალაშვილი, მ. (2024). ქართული ენის ლინგვისტური მოდელი სიტყვაფორმათა სინთეზისა და ანალიზისთვის, მესამე საერთაშორისო კონფერენცია „საქართველო და კავკასია – წარსული, აწმყო, მომავალი“, თბილისი: „საქართველოს უნივერსიტეტის გამომცემლობა“.

დათუკიშვილი, ქ. ლოლაძე, ნ. ზაკალაშვილი, მ. (2024). დერივაციული მოდელები ქართული ენის მორფოლოგიურ პროცესორში, ენათმეცნიერ-კავკასიოლოგთა VII საერთაშორისო სიმპოზიუმის მასალები, თბილისი: „ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტის გამომცემლობა“. ISBN 978-9941-36-275-0

დათუკიშვილი, ქ. ლოლაძე, ნ. ზაკალაშვილი, მ. (2013). ქართული ენის მორფოლოგიური პროცესორი – მანქანური ანალიზისა და სინთეზის ინსტრუმენტი, ჟურნალი „ქართველოლოგია“, N4, თბილისი: „ილიას სახელმწიფო უნივერსიტეტის გამომცემლობა“.

ლობჟანიძე, ი. (2021). „ქართული ენის მორფოსინტაქსური ანოტირებისა და სასრული პოზიციის მორფოლოგიური ანალიზის პრინციპები“, თბილისი: „ილიას სახელმწიფო უნივერსიტეტის გამომცემლობა“. ISBN 978-9941-18-366-9

მარგველანი, ლ. (2008). ქართული ენის კომპიუტერული მოდელები, თბილისი: გამომცემლობა „ინტელექტი“. ISBN 5-7859-0078-5

ონიანი, ალ. (2003). თანამედროვე ქართული სალიტერატურო ენა, თბილისი: „სულხან-საბა ორბელიანის სახელობის თბილისის სახელმწიფო პედაგოგიური უნივერსიტეტის გამომცემლობა“.

უთურგაიძე, თ. (2009). ქართული ენის დონეთა ძირითადი მახასიათებლების ურთიერთზემოქმედებისათვის გლობალურ ენობრივ სისტემაში, თბილისი: გამომცემლობა „მერიდიანი“. ISBN 978-9941-10-194-6

ქართული ენის განმარტებითი ლექსიკონი. (1950 – 1964). ტ. I – VIII, თბილისი: „საქართველოს სსრ მეცნიერებათა აკადემიის გამომცემლობა“. – https://ice.tsu.ge/liv/ganmartebiti.php

ქართული ენის განმარტებითი ლექსიკონი. (2008 – 2019). ტ. I – IV, თბილისი: გამომცემლობა „მერიდიანი“ – https://ice.tsu.ge/liv/ganmartebiti.php

ქართული ენის ეროვნული კორპუსი – http://gnc.gov.ge/gnc

ქართული ლექსიკონი – https://www.ganmarteba.ge/

ქირია, ჭ. (2022). არსებითი სახელის ბრუნების პარადიგმული მოდელები თანამედროვე ქართულში, „საენათმეცნიერო ძიებანი“, XIII, თბილისი. ISSN 1987-6653

შანიძე, ა. (1980). ქართული ენის გრამატიკის საფუძვლები, თხზულებანი 12 ტომად, ტ. III, თბილისი: „თბილისის უნივერსიტეტის გამომცემლობა“.

ჩიქობავა, არნ. (1998). ქართული ენის ზოგადი დახასიათება, თბილისი: გამომცემლობა „ქართული ენა“.

ჯორბენაძე, ბ. ლოლაძე, ნ. კიკონიშვილი, მ. (2014). ქართული ენის სახელური ფუძეების ლექსიკონი, თბილისი: „თბილისის უნივერსიტეტის გამომცემლობა“. ISBN 978-9941-13-326-8