ირინა ლობჟანიძე
ილიას სახელმწიფო უნივერსიტეტი, ლინგვისტურ კვლევათა ცენტრი

თანამედროვე ქართული ენის მორფოლოგიური ანალიზატორი და გენერატორი

რეზიუმე
წინამდებარე სტატიაში წარმოდგენილია თანამედროვე ქართული ენის მორფოლოგიური ანალიზატორი, რომელიც შეიქმნა სასრული პოზიციის ავტომატების გამოყენებით. სასრული პოზიციის ტექნიკური საშუალებები გამოიყენება სხვადასხვა ენის ფონოლოგიისა და მორფოლოგიის კომპიუტერული აღწერის დროს. სისტემა მოიცავს თანამედროვე ქართული ენის მეტყველების ნაწილების მორფოლოგიურ თვისებებს. მორფოტაქტიკა კოდირებულია ლექსიკონების, ხოლო ცვლილებები - რეგულარული გამოსახულებების სახით.

შესავალი

თანამედროვე ქართული ენა განეკუთვნება მორფოლოგიური კატეგორიებით მდიდარ ენებს. არსებობს ბუნებრივი ენის დამუშავების უამრავი სისტემა, რომელიც ემსახურება სიტყვის გარჩევას და მორფოლოგიური ტაგების მინიჭებას. აღნიშნული მოხსენება ეხება მსგავსი სისტემის შექმნასა და მისი განვითარების სამომავლო პერსპექტივას. შემუშავებული სისტემა მორფოლოგიურად აანალიზებს ტექსტს და თითოეულ სიტყვას ანიჭებს კონკრეტულ მორფოლოგიურ კატეგორიებს.
მოხსენების სტრუქტურა შემდეგია: 1. თანამედროვე ქართული ენის მეტყველების ნაწილებისა და ფლექსიური პარადიგმების მიმოხილვა; 2. მორფოლოგიური ანალიზატორის მოდელი და სტრუქტურა; 3. მორფოლოგიური ანალიზატორის ტესტირების შედეგები, შეცდომების ანალიზი და მისი განვითარების სამომავლო პერსპექტივა

მეთოდოლოგია

სასრული პოზიციის ტექნიკური საშუალებების გათვალისწინებით (Beesley K.R., Kartunnen L. 2003, Koskenniemi, K. 1983 და ა.შ.), თანამედროვე ქართული ენის მორფოლოგიური ანალიზატორი შეიქმნა xfst-სა და lexc-ის გამოყენებით.

პროექტის მოკლე აღწერა

წინამდებარე სტატიაში წარმოდგენილია *თანამედროვე ქართული ენის მორფოლოგიური ანალიზატორი,* რომელიც შეიქმნა სასრული პოზიციის ავტომატების გამოყენებით. სასრული პოზიციის ტექნიკური საშუალებები გამოიყენება სხვადასხვა ენის ფონოლოგიისა და მორფოლოგიის კომპიუტერული აღწერის დროს. ანალიზატორი შემუშავდა შოთა რუსთაველის ეროვნული სამეცნიერო ფონდის მიერ დაფინანსებული პროექტის (AR/320/4-105/11) ფარგლებში. ბუნებრივია, რომ მორფოლოგიურად ანოტირებული კორპუსის შექმნა მჭიდროდ უკავშირდება მორფოლოგიურად ანოტირებული ფაილების არსებობას. ამიტომაც, საჭირო გახდა თანამედროვე ქართული ენის მორფოლოგიური ანალიზატორის შექმნა. სისტემა მოიცავს თანამედროვე ქართული ენის მეტყველების ნაწილების მორფოლოგიურ თვისებებს. მორფოტაქტიკა კოდირებულია ლექსიკონების, ხოლო ცვლილებები - რეგულარული გამოსახულებების

სახით. სასრული პოზიციის ავტომატების, კერძოდ, ქსეროქსის სასრული პოზიციის ტექნოლოგიის (Xfst)[1] გამოყენებით შექმნილ მორფოლოგიურ ანალიზატორს შემდეგი სტრუქტურა ახასიათებს:

აღნიშნული სტრუქტურა მოიცავს მეტყველების ნაწილების სალექსიკონო ერთეულების შემდეგ ქვესიმრავლეებს:

LEXICON Nouns
LEXICON Adjectives
LEXICON Numerals
LEXICON Pronouns
LEXICON Conjunctions
LEXICON Particles
LEXICON Adverbs
LEXICON Postpositions
LEXICON Verbs
LEXICON Verbal Nouns
LEXICON Participles
LEXICON Punctuations
LEXICON Abbreviations

რომლებსაც, ესადაგება შენაცვლების წესები. აღნიშნული ლექსიკონებისა და შენაცვლების წესებზე დაყრდნობით იქმნება მორფოლოგიური ანალიზატორის მოდული. სხვადასხვა ტექსტზე გადამოწმებული რესურსი გამოიყენება ტოკენიზაციის, ლემატიზაციისა და ტაგირებისათვის.

დასკვნები

მოხსენებაში წარმოდგენილია ანალიზატორის სტრუქტურა, სამუშაო მოდული და განხილულია მისი შემდგომი განვითარების საფეხურები.

[1] http://www.cis.upenn.edu/~cis639/docs/xfst.html

Irina Lobzhanidze
irina_lobzhanidze@iliauni.edu.ge
Ilia State University, Centre of Linguistic Research

Morphological Analyzer and Generator of Modern Georgian Language

Summary

In this paper we present the *Morphological Analyzer of Modern Georgian Language* developed using finite-state automata. Finite State Techniques have been applied successfully in computational phonology and morphology of the world's major and minor languages. The system encodes the morphology of all inflected parts-of-speech of Modern Georgian. The morphotactics is encoded in the lexicons and alternation rules are encoded in regular expressions.

Background (Introduction)

Modern Georgian Language belongs to a morphologically rich languages. There are a lot of applications of Natural Language Processing, which give appropriate morphological tags to a given word and provide its analysis. This paper deals with the design of such system and the perspective of its further development. The system provides analysis of text and gives appropriate morphological tags to words.

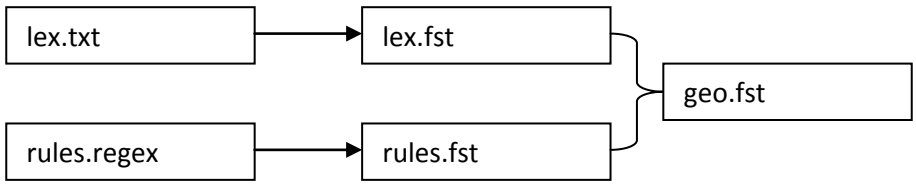
The paper is divided into the following parts: 1. Brief overview of Georgian PoS and their inflections; 2. Model and structure of morphological analyzer; 3. Testing and error analysis of morphological analyzer and the perspective of its development.

Methods

Following approaches of finite state techniques (Beesley K.R., Kartunnen L. 2003, Koskenniemi, K. 1983 etc.), a morphological analyzer of Modern Georgian has been created using xfst and lexc tools.

Brief Description of the Project

In this paper we present the *Morphological Analyzer of Modern Georgian Language* developed using finite-state automata. Finite state techniques have been applied successfully in computational phonology and morphology of the world's major and minor languages. The Analyzer was developed within the framework of the project (AR/320/4-105/11) financed by the Shota Rustaveli National Science Foundation. The development of morphologically annotated corpus is closely connected with the existence of morphologically annotated files. Thus, it was important to create a morphological analyzer for Modern Georgian Language. The system encodes the morphology of all inflected parts-of-speech of Modern Georgian. The morphotactics is encoded in the lexicons and alternation rules are encoded in regular expressions. The morphological transducer developed on the basis of Xerox Finite State Tools (Xfst)1 has the following structure:



The mentioned structure includes the following PoS Lexicons:
LEXICON Nouns
LEXICON Adjectives
LEXICON Numerals
LEXICON Pronouns
LEXICON Conjunctions
LEXICON Particles
LEXICON Adverbs
LEXICON Postpositions
LEXICON Verbs
LEXICON Verbal Nouns
LEXICON Participles
LEXICON Punctuations
LEXICON Abbreviations

The lexicon data are processed in accordance with the appropriate alternation rules. The morphological analyzer consists of the mentioned lexicons and alternation rules. It allows us to distinguish the appropriate lemma and morphological categories. This resource evaluated against different texts is used for tokenizing, lemmatising and tagging.

Conclusions

We present the structure and working module of the analyzer and describe the further stages of its development.

References

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

Gurevich, O. "A Finite-State Model of Georgian Verbal Morphology." *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL.* NY: Association for Computational Linguistics, 2006. 45-48.

Jurafsky, D., Martin, H. J. *Speech and Language Processing.* New Jersey: Pearson Education International, 2009.

Kapanadze, O. "Describing Georgian Morphology with a Finite-State System." *Lecture Notes in Computer Science*, 2010: 114-122.

Meurer, P. "A Computational Grammar for Georgian." *Lecture Notes in Computer Science*, 2009: 1-15.

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

მელიქიშვილი, დ. *ქართული ზმნის უღლების სისტემა.* თბილისი: ლოგოს პრესი, 2001.

შანიძე, ა. *ქართული ენის გრამატიკის საფუძვლები.* თბილისი: თბილისის უნივერსიტეტის გამომცემლობა, 1973.

Software

Xerox Finite-State Tools (tools: lexc, xfst, lookup; operating system: Windows).