დავით მიშელაშვილი

ინსტრუმენტული საშუალებები ბუნებრივი ენის ტექსტების სინტაქსური და მორფოლგიური ანალიზატორის შესადგენად

05.13.11 - გამოთვლითი მანქანების, სისტემების, კომპლექსებისა და ქსელების მათემატიკური და პროგრამული უზრუნველყოფა

ფიზიკა-მათემატიკის მეცნიერებათა კანდიდატის სამეცნიერო ხარისხის მოსაპოვებლად წარმოდგენილი დისერტაციის

ა ვ ტ ო რ ე ფ ე რ ა ტ ი

თბილისი, 2006

სადისერტაციო ნაშრომი შესრულებულია ივ. ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტის გამოყენებითი მათემატიკისა და კომპიუტერულ მეცნიერებათა ფაკულტეტის კომპიუტერების მათემატიკური უზრუნველყოფისა და ინფორმაციული ტექნოლოგიების კათედრაზე

სამეცნიერო ხელმძღვანელი: ჯემალ ანთიძე
ფიზიკა-მათემატიკის მეცნიერებათა კანდიდატი,
დოცენტი, 05.13.11

თემის აქტუალურობა

ბუნებრივი ენის ტექსტების კომპიუტერული დამუშავება თანამედროვე საინფორმაციო ტექნოლოგიების ერთერთ მნიშვნელოვან მიმართულებას წარმოადგენს. მასში ჩვენ ვხვდებით ისეთი სახის ამოცანებს, როგორიცაა: მართლწერის ავტომატური შემოწმება; მანქანური თარგმანი; საძიებო სისტემები, რომელთა მეშვეობითაც შესაძლებელია სიტყვის სხვადასხვა ფორმების მოძებნა ბაზაში; ბუნებრივ ენაზე დიალოგურ რეჟიმში მომუშავე პროგრამები; ტექსტების მრავალმიზნობრივი მორფოლოგიური, სინტაქსური და სემანტიკური ანალიზი; ენის სწავლებისათვის განკუთვნილი სავარჯიშო ტრენაჟორები და სხვა. გამოთვლითი ტექნიკის დღევანდელი მასშტაბებით განვითარების პირობებში, ძალზედ მნიშვნელოვანია შეგვეძლოს ამგვარი ამოცანების სწორად და ეფექტურად გადაჭრა.

თავდაპირველად, სანამ მოხდებოდა ახალი ფორმალიზმების შემუშავება და პროგრამული სისტემის რეალიზაცია, დაწვრილებით იქნა შესწავლილი თანამედროვე მიდგომები ბუნებრივ ენათა დამუშავების თეორიაში. განხილული იქნა სხვა მსგავსი სისტემების მუშაობის ძირითადი პრინციპები. პრაქტიკული ექსპერიმენტების შედეგად, აშკარა გახდა რომ ქართული ენისათვის და მრავალი სხვა ენებისათვის რომლებსაც ახასიათებთ მდიდარი მორფოლოგია, წინადადებაში სიტყვათა თავისუფალი წყობა და სინტაქსის სირთულე, გამოყენებული უნდა ყოფილიყო განსხვავებული მიდგომები. საჭირო იყო უფრო ზოგადი, მოქნილი და მძლავრი პროგრამული სისტემის შექმნა. პირველ რიგში კი აუცილებელი იყო შეგვემუშავებინა ახალი ფორმალიზმები, რომელთა ფარგლებშიც მოხერხდებოდა სხვადასხვა ბუნებრივი ენების გრამატიკის ჩაწერა ადამინისათვის ადვილად გასაგები და ლაკონური სახით.

ნაშრომის მიზანი

ნაშრომის ძირითადი მიზანი გახლავთ პროგრამული სისტემის შექმნა, რომლის გამოყენებაც შესაძლებელი იქნება ბუნებრივი ენაზე ჩაწერილი ტექსტების სინტაქსური და მორფოლოგიური, კომპიუტერული ანალიზისათვის. ბუნებრივი ენის მაგალითად აღებულ იქნა ქართული ენა, მისი გრამატიკის წესები და ლექსიკური მარაგი. თავისთავად, ასეთი პროგრამული სისტემის შექმნის პროცესი შემდეგ ცალკეულ მიზნებად შეგვიძლია დავყოთ:

1. თეორიული ბაზის მომზადება. ახალი ალგორითმული მიდგომების შემუშავება.
2. ბუნებრივი ენის გრამატიკული წესების ჩასაწერი ფორმალიზმების შექმნა.
3. პროგრამული სისტემის რეალიზაცია, რომელიც ეფუძნება 1 და 2 პუნქტში მიღებულ შედეგებს.
4. რომელიმე ბუნებრივი ენისათვის ამ სისტემის პრაქტიკული გამოყენება.

პროგრამული სისტემა ორი ნაწილისაგან შედგება, სინტაქსური და მორფოლოგიური ანალიზატორებისაგან. სინტაქსური ანალიზატორის მუშაობის მიზანია ბუნებრივ ენაზე ჩაწერილი წინადადების ანალიზი, მისი სტრუქტურის დადგენა (გარჩევის ხის აგება), ურთიერთკავშირის დადგენა მის შემადგენელ წევრებს შორის და წინადადების შესახებ ძირითადი გრამატიკული ინფორმაციის მოპოვება (მაგალითად: ინორმაცია სუბიექტის, ობიექტების და პრედიკატის შესახებ). მორფოლოგიური ანალიზატორის დანიშნულება გახლავთ სიტყვაფორმის დაშლა მის შემადგენელ მორფემებად და ამ დაშლის საფუძველზე სიტყვაფორმის ძირითადი გრამატიკული კატეგორიების დადგენა.

პროგრამული სისტემის შექმნის ძირითადი მოტივებია:

- იყოს იგი მაქსიმალურად გამოსადეგი ისეთი ბუნებრივი ენისათვის რომლებსაც გააჩნიათ მდიდარი მორფოლოგია, აქვთ რთული სინტაქსი, წინადადებაში დაშვებულია სიტყვათა თავისუფალი წყობა და სხვა
- აღმოიფხვრას ზოგიერთი ნაკლოვანება, რაც გააჩნდათ აქამდე არსებულ მსგავს პროგრამულ სისტემებს
- რაც შეიძლება გამარტივდეს ბუნებრივი ენის გრამატიკის წესების ჩაწერა

სიახლე და ძირითადი შედეგები

წარმოდგენილ ნაშრომში სიახლე და ძირითადი შედეგი არის პროგრამული სისტემა, ინსტრუმენტარი, რომელიც საშუალებას გვაძლევს სპეციალური ფორმალიზმების მეშვეობით მივაწოდოთ მას ბუნებრივი ენის გრამატიკის წესები, ლექსიკონი და გამოვიყენოთ იგი შესაბამისი ბუნებრივი ენის ტექსტების სინტაქსური და მორფოლოგიური ანალიზისათვის.

დეტალური სახით, ნაშრომში მიღებული სიახლეები და ძირითადი შედეგებია:

- სინტაქსური და მორფოლოგიური ანალიზისათვის შექმნილია ახალი, კომპლექსური ალგორითმები.
- შემუშავებულ იქნა ახალი ფორმალიზმები, რომელთა ფარგლებშიც მარტივად შეგვიძლია ჩავწეროთ ბუნებრივი ენის გრამატიკის შემადგენელი წესები.
- ამ ფორმალიზმების რეალიზაციის შედეგად, დაპროგრამების ენა C++_ზე დაიწერა პროგრამული სისტემა, რომელიც განკუთვნილია ბუნებრივ ენაზე ჩაწერილი ტექსტების სინტაქსური და მორფოლოგიური ანალიზისათვის.
- მიღებული სისტემა გამოიცადა ქართული ენისათვის. შემუშავებული ფორმალიზმების ფარგლებში ჩაიწერა ქართული ენის სინტაქსის და მორფოლოგიის წესების გარკვეული ნაწილი. შედგა სიტყვის ძირების ლექსიკონი (მცირე მოცულობით).

ნაშრომის დაწვრლებითი გაცნობის შემდეგ, შეგვიძლია რამოდენიმე კონკრეტული სიახლის გამოყოფა რომელიც აქამდე არ გვხვდებოდა სხვა ნაშრომებში. ეს სიახლეებია:

- კონტექსტისაგან თავისუფალი გრამატიკის წესების ჩაწერის გაფართოებული ვარიანტი, რომელიც ითვალისწინებს წესში არსებულ სიმბოლოთა თავისუფალ, ან ნაწილობრივ თავისუფალ წყობას.
- შეზღუდვების მუშაობის უფრო ზოგადი პრინციპი. წესზე დადებული შეზღუდვები წარმოიდგინება როგორც ლოგიკური გამოსახულებები,

რაც უფრო რთული გრამატიკული წესების ჩაწერის საშუალებას გვაძლევს.

- სინტაქსური და მორფლოგიური ანალიზატორების ფორმალიზმში შემოღებულია მრავალი მოსახერხებელი კონსტრუქცია: ცვლადები, კონსტანტები, თვისებათა სტრუქტურის ინიციალიზაციის ნაწილი, მრავალარგუმენტიანი ოპერაციები თვისებათა სტრუქტურებზე და სხვა.
- მორფოლოგიური ანალიზის პრინციპულად ახალი ალგორითმი, წესის შიგნით შეზღუდვების გადანაწილების შესაძლებლობით.
- პრეპროცესორი მორფოლოგიური ანალიზატორის გრამატიკის ფაილისათვის. მისი მეშვეობით შესაძლებელია პარამეტრიანი მაკრო ჩასმების გამოყენება. რაც გვეხმარება გრამატიკის ფაილში გამეორებადი (ან ერთმანეთის მსგავსი) ტექსტური ფრაგმენტების მოკლედ ჩაწერაში.

პრაქტიკული მნიშვნელობა

ნაშრომს გააჩნია პრაქტიკული მნიშვნელობა, ვინაიდან მასში წარმოდგენილია პროგრამული სისტემა რომელიც განკუთვნილია ბუნებრივ ენაზე ჩაწერილი ტექსტების სინტაქსური და მორფოლოგიური ანალიზისათვის. ამ სისტემის პრაქტიკაში გამოყენება მოხდა ქართული ენის მაგალითზე. ტესტირების მიზნით, სისტემაში გამოყენებული ფორმალიზმის ფარგლებში ჩაიწერა ქართული ენის გრამატიკის ნაწილი და მიღებულ იქნა წინადადებების და სიტყვაფორმების ანალიზი.

შეგვიძლია გამოვყოთ რამოდენიმე ძირითადი მომენტი ნაშრომის პრაქტიკული მნიშვნელობის შესახებ:

- მიღებული პროგრამულ სისტემას თავისთავად გააჩნია პრაქტიკული მნიშვნელობა ვინაიდან იგი არის ინსტრუმენტარი, რომელის გაშვებაც შეგვიძლია კომპიუტერზე და საშუალება გვეძლევა ჩავატაროთ ბუნებრივი ენის ტექსტების სინტაქსური და მორფოლოგიური ანალიზი.
- შემუშავებული ფორმალიზმები, რომელებიც საფუძვლად უდევს სინტაქსურ და მორფოლოგიურ ანალიზატორებს, ასევე პრაქტიკული მნიშვნელობის მატარებელნი არიან, ვინაიდან მათი მეშვეობით

სპეციალისტი (ლინგვისტი) ჩაწერს ბუნებრივი ენის გრამატიკას ფორმალური, კომპიუტერისათვის მისაღები სახით.

- პროგრამული სისტემა დაწერილია დაპროგრამების ენა C++_ის სტანდარტზე, მოდულების სახით, ობიექტზე ორიენტირებული დაპროგრამების სტილის გამოყენებით. ის ასევე შეგვიძლია განვიხილოთ როგორც კლასების ბიბლიოთეკა და გამოვიყენოთ იგი სხვადასხვა პროგრამების შიგნით, სადაც საჭიროა ბუნებრივ ენაზე ჩაწერილი ტექსტების ანალიზი.

ნაშრომის აპრობაცია

დისერტაციის შედეგები გამოქვეყნებულია 10 სამეცნიერო ნაშრომში და მოხსენებულია სხვადასხვა სამეცნიერო შეკრებებსა და სემინარებზე:

- საქართველოს მეცნიერებათა აკადემიის ენათმეცნიერების ინსტიტუტის კონფერენციებზე (2003, 2004 წწ.) [2, 4, 5]
- ვეკუას სახელობის გამოყენებითი მათემატიკის ინსტიტუტის სემინარები (2004, 2005 წწ.) [19, 20, 21]
- ენა, ლოგიკა, გამოთვლები - მეოთხე ინტერნაციონალური სიმფოზიუმი (2001 წ.) [17]
- ქართველ მათემატიკოსთა კონგრესი (2001 წ.) [18]

ძირითადი შედეგები მოხსენებული იქნა ვეკუას სახელობის გამოყენებითი მათემატიკის ინსტიტუტის კონფერენციაზე (2005 წ.)

დისერტაციის მოცულობა და სტრუქტურა

სადისერტაციო ნაშრომი შედგება შესავლის, 3 თავის, დანართისა და ციტირებული ლიტერატურისაგან. მოიცავს 117 ნაბეჭდ გვერდს. ასევე ცალკე არის აკინძული დანართი, რომელშიც გაფორმებულია პროგრამის და მასში გამოყენებული მოდულების კოდი, ასევე დამატებითი ფაილები Bison, Lex და Coco/R ინსტრუმენტებისათვის.

დისერტაციის მოკლე შინაარსი თავების მიხედვით

შესავალი შეიცავს სადისერტაციო თემასთან დაკავშირებული ლიტერატურის მიმოხილვას და დისერტაციის მოკლე შინაარსს.

დისერტაციის პირველი თავში - ”ბუნებრივი ენის კომპიუტერული ანალიზი” - განხილულია ბუნებრივი ენის კომპიუტერული ანალიზის საკითხები. დასმულია ბუნებრივი ენის ტექსტების კომპიუტერული ანალიზის პრობლემა და შემადგენელი ნაწილები: მორფოლოგიური, სინტაქსური და სემანტიკური ანალიზი. მორფოლოგიური და სინტაქსური კომპიუტერული ანალიზისათვის განხილულია პრობლემის გადაწყვეტისადმი ძირითადი თეორიული და პრაქტიკული მიდგომები.

პირველ პარაგრაფში დასმულია ბუნებრივ ენათა კომპიუტერული ანალიზის პრობლემა. ბუნებრივი ენის ტექსტების კომპიუტერული ანალიზის პროცესი და მისი შედეგის ფორმალიზაცია. ჩამოთვლილია ის ძირითადი პრაქტიკული ამოცანები, რომელთა გადაჭრისათვისაც აუცილებელია ბუნებრივ ენათა კომპიუტერული ანალიზის პრობლემის გადაწყვეტა. მაგ: მართლწერის ავტომატური შემოწმება, მანქანური თარგმანი და სხვა.

მეორე პარაგრაფში ჩამოთვლილია ბუნებრივი ენის კომპიუტერული ანალიზის შემადგენელი ნაწილები (მორფოლოგიური, სინტაქსური და სემანტიკური ანალიზი) და დაწვრილებით არის განხილული თითოეული მათგანი: დასმულია პრობლემა, მოყვანილია მარტივი მაგალითები და მითითებულია შესაძლო გადაჭრის გზები. აღნიშნულია რომ სადისერტაციო ნაშრომში განიხილება ბუნებრივი ენის მორფოლოგიური და სინტაქსური ანალიზის პრობლემა. სქემის სახით ნაჩვენებია ბუნებრივი ენის მორფოლოგიური და სინტაქსური ანალიზატორებისადან შემდგარი ერთიანი პროგრამული სისტემა. განხილულია მისი მუშაობის ზოგადი პრინციპი.

მესამე პარაგრაფში განხილულია ბუნებრივი ენის კომპიუტერული ანალიზისათვის გამოყენებული ძირითადი მიდგომები:

- თვისებათა სტრუქტურები და მათზე განსაზღვრული ოპერაციები
- ზოგადი, ქვემოდან-ზემოთ სინტაქსური გარჩევის ალგორითმი
- შეზღუდვები

- არადეტერმინისტული ძიების ალგორითმი (მორფოლოგიური ანალიზისათვის)

თითოეული მათგანი სათითაოდაა განხილული. აგრეთვე განხილულია უნიფიკაციის ოპერაციის მუშაობის პრინციპი. შეფასებულია სინტაქსური და მორფოლოგიური ანალიზის ალგორითმების სისწრაფე, როგორც განყენებულად, ასევე შეზღუდვების შემოწმების ალგორითმთან კომბინირებული სახით. ჩამოთვლილია მათი ძირითადი თვისებები. აღნიშნულია შეზღუდვებისადმი განზოგადოებული მიდგომა, ისინი წარმოიდგინება როგორც ლოგიკური გამოსახულებანი.

მეოთხე პარაგრაფი დათმობილი აქვს თვისებათა სტრუქტურების და მათზე განსაზღვრული ოპერაციების დაწვრილებით განხილვას. ამის მიზეზი არის ის, რომ ნაშრომში თვისებათა სტრუქტურები ძალზედ ფართოდაა გამოყენებული. ისინი გამოიყენება ინფორმაციის შესანახად და მისი მანიპულაციისათვის, როგორც სპეციფიკური მონაცემთა სტრუქტურები. თვისებათა სტრუქტურა წარმოადგენს თვისებების სიმრავლეს. თითოეულ თვისება არის ”სახელი” – ”მნიშვნელობა” ტიპის კონსტრუქცია. რაც მთავარია, თვისების მნიშვნელობა შეიძლება თავად იყოს თვისებათა სტრუქტურა. შესაბამისად, თვისებათა სტრუქტურა წარმოადგენს რეკურსიული ტიპის მონაცემთა სტრუქტურას, რაც მას უნივერსალურობას და მოქნილობას მატებს. წარმოდგენილია თვისებათა სტრუქტურების მაგალითები, მათ შორის ერთმანეთში ჩალაგებულ თვისებათა სტრუქტურების მაგალითებიც. ნაჩვენებია თუ როგორ ჩაიწერება თვისებათა სტრუქტურები და მათზე განსაზღვრული ოპერაციები ფორმალური სახით. დაწვრილებით არიან განხილულნი თვისებათა სტრუტურებზე განსაზღვრული ოპერაციები:

- მინიჭება
- ტოლობაზე შემოწმება
- უნიფიკაცია
- უნიფიკაციაზე შემოწმება
- შემოწმების მრავალარგუმენტიანი ოპერაციები

განსაკუთრებული ყურადღება აქვს დათმობილი უნიფიკაციის ოპერაციის განხილვას. ორი თვისებათა სტრუქტურა ერთმანეთთან უნიფიცირდება მაშინ და მხოლოდ მაშინ თუკი მათი ერთნაირი სახელის მქონე თვისებები არ ეწინააღმდეგებიან ერთმანეთს.

დისერტაციის მეორე თავში - "მორფოლოგიური გამრჩევი" - განხილულია მორფოლოგიური გამრჩევი, მისი მუშაობის პრინციპი და პროგრამული რეალიზაცია. დაწვრილებით არის განხილული მორფოლოგიური გამრჩევისათვის შემუშავებული ფორმალიზმი. განხილულია მაგალითები და ნაჩვენებია ქართული ენის მორფოლოგიის ფრაგმენტის ჩაწერა შემუშავებული ფორმალიზმის მეშვეობით.

პირველ პარაგრაფში დასმულია მორფოლოგიური ანალიზის ამოცანა. სიტყვაფორმის მორფემებად დაშლა განხილულია როგორც კომბინატორული ამოცანა, რომელიც შემდეგნაირად გამოიყურება:

მოცემული გვაქვს არაცარიელი ანბანი Σ.

ასევე მოცემული გვაქვს N რაოდენობა მორფემათა სიმრავლე M_1 M_2 M_3 ... M_N.

მორფემათა სიმრავლე M_i განისაზღვრება, როგორც Σ ანბანით ჩაწერილ სტრიქონთა სიმრავლე. $M_i = \{ m_1^i, m_2^i, \ldots m_{ki}^i \}$, $k_i > 0$, სადაც მორფემა m_j^i ეკუთვნის Σ^* -ს (შესაძლებელია, რომ მორფემა ცარიელი სტრიქონი იყოს).

მოცემული გვაქვს Σ ანბანით განსაზღვრული შემავალი სიტყვა (სტრიქონი) α, რომელიც უნდა დაიშალოს მორფემებად.

საჭიროა ვიპოვოთ მორფემათა ყველა ისეთი კომბინაცია m_{i1}, m_{i2}, ... m_{iN} (სადაც m_{ik} ეკუთვნის M_K -ს), რომლებიც კონკატენაციის შედეგად გვაძლევს α საწყის სიტყვას.

განხილულია არადეტერმინისტრული ძიების ალგორითმი, რომლის მეშვეობითაც ხდება დასმული პრობლემის გადაწყვეტა. ამ ალგორითმის მუშაობა ნაჩვენებია კონკრეტულ მაგალითზე. საწყის მათემატიკურ მოდელს ვაფართოებთ შეზღუდვების დამატებით, რათა მისი მეშვეობით შესაძლებელი გახდეს ბუნებრივი ენის სიტყვაფორმების სწორ კომბინაციებად დაშლა. გაფართოებული ამოცანის ამოსახსნელად გამოყენებულია იგივე ალგორითმი ოღონდ კომბინირებული სახით შეზღუდვების შემოწმების ალგორითმთან ერთად. ფორმალურად მორფოლოგიური წესის ჩაწერა შემდეგი სახით ხდება:

$W \rightarrow M_1 \{ C_1 \} M_2 \{C_2\} \ldots M_N \{ C_N \}$;

სადაც C_i წარმოადგენს შეზღუდვებს. შესაძლებელია შეზღუდვების გადანაწილება მორფემებთან მიმართებაში სხვადასხვა ადგილას. რაც უფრო მარცხნივ არის შეზღუდვა მით უფრო ადრეულ ეტაპზე ხდება მისი შემოწმება. თუმცა შეზღუდვას შეუძლია ოპერირება მოახდინოს მხოლოდ იმ მორფემათა თვისებებზე, რომლებიც მის აღწერამდე არიან მითითებულნი. ამ მეთოდში საინტერესო გახლავთ ის, რომ პრაქტიკულად ჩვენ ვინარჩუნებთ თავდაპირველად აღწერილ მათემატიკურ მოდელს. ძიების ალგორითმი მიმდინარეობს იგივენაირად, მაგრამ დამატებით ჩვენ წესში შეგვიძლია ჩავრთოთ შემოწმებები, რომელთა მეშვეობითაც ვზღუდავთ ამონახსნების სიმრავლეს და უკუვაგდებთ ძიების პროცესში არასასურველ განშტოებებს. ამასთანავე შეზღუდვების შემოწმება ხდება დინამიურად, სიტყვის დაშლის პროცესში. შესაბამისად უფრო მარცხნივ მდგარი შეზღუდვები უფრო ადრე შემოწმდებიან, რაც ალგორითმის ეფექტურობას მკვეთრად ამაღლებს.

მეორე პარაგრაფში განხილულია მორფოლოგიური გამრჩევისათვის შემუშავებული ფორმალიზმი. აღწერილია თუ როგორ განისაზღვრება მორფემათა კლასები, წესები და შეზღუდვები. დაწვრილებით არის განხილული შეზღუდვების მუშაობის პრინციპი. შეზღუდვების წესის შიგნით გადანაწილების შესაძლებლობა. შეზღუდვები წარმოადგენენ ლოგიკურ გამოსახულებებს. განხილულია მათი გამოთვლის პროცესი, ოპერაციათა პრიორიტეტები, დეკლარაციული და იმპერატიული ბუნება. განხილულია ფომრლიზმში შემოღებული შემდეგი საშუალებები: ცვლადებისა და კონსტანტების განსაზღვრა, ფაილის ჩართვის დირექტივა, კომენტარები.

მორფოლოგიის ფაილი პირობითად ორ ნაწილად შეგვიძლია გავყოთ: მორფემების კლასების აღწერად და წესების აღწერად.

მორფემების კლასის აღწერა იწყება '@' სიმბოლოთი, რომელსაც მოსდევს მორფემების კლასი სახელი, ტოლობის ნიშანი '=' და ფიგურულ ფრჩხილებში ჩამოწერილი მორფემების კლასის შესაძლო მნიშვნელობები. მორფემების კლასის შესაძლო წარმომადგენლები იწერება ბრჭყალეში. მას მოსდევს თვისებათა სტრუქტურა (შესაძლოა ცარიელი). განიხილოთ შემდეგი მორფემების კლასის აღწერა:

```
@ბრუნვა =
  {
  "ი"   [],
```

```
”მა” [],
”ი ს ” [],
”ს ”  [],
”ი თ” [],
.....
””  []
}
```

ამ ჩანაწერით განისაზღვრება მორფემათა კლასი ”ბრუნვა”. მისი შესაძლო ლექსიკური მნიშვნელობებია ”ი”, ”მა”, ”ის” ... და ა.შ. ანუ ქართულ ენაში არსებული ყველა შესაძლო ბრუნვის ნიშანი. აგრეთვე ცარიელი სტრიქონი ””, რომელიც გამოიყენება მორფემის ცარიელი მნიშვნელობის აღსანიშნავად.

განვიხილოთ წესის აღწერა მორფოლოგიის ფაილში. უმარტივეს შემთხვევაში თუკი წესს შეზღუდვები არ გააჩნია მისი აღწერა შემდეგნაირად გამოიყურება:

სიტყვა -> მორფემა_1 მორფემა_2 ... მორფემა_N ;

წესის მარცხენა მხარეში დგას ამ წესის სახელი, ხოლო მარჯვენა მხარეში ჩამოთვლილია მისი შემადგენელი მორფემები. ეს მორფემები მორფოლოგიის ფაილში უნდა განსაზღვრულნი იყვნენ ამ წესის აღწერამდე. მორფოლოგიის ფაილში შეიძლება მოცემული იყოს მრავალი ამგვარი წესი, თითოეული მათგანი წარმოადგენს სიტყვის დაშლის შესაძლო ვარიანტს. ანალიზატორი სიტყვის გარჩევის დროს განიხილავს ყველა შესაძლო წესს და შედეგად მოგვცემს სიტყვის ყველა დასაშვებ დაშლას.

წესი -> მორფემა_1 { შეზღუდვები_1} მორფემა_2 { შეზღუდვები_2 } ... მორფემა_N { შეზღუდვები_N } ;

შეზღუდვები მოსდევს მორფემას და იწერება ფიგურულ ფრჩხილებში. ყოველი მორფემის შემდეგ შეგვიძლია ჩავურთოთ შეზღუდვა, და მისი შემოწმება მოხდება მაშინ, როდესაც გამრჩევი ამ კონკრეტულ მორფემას შეუსაბამებს გარკვეულ დასაშვებ მნიშვნელობას. სიტყვის მორფემებად დაშლა მიმდინარეობს მარცხნიდან მარჯვნივ, შესაბამისად უფრო მარცხნივ მდგომი მორფემები უფრო ადრე მიიღებენ მნიშვნელობებს და უფრო მარცხნივ მგომი შეზღუდვები უფრო ადრეულ ეტაპზე ამუშავდებიან, რაც საშუალებას

გვაძლევს არ დაველოდოთ სიტყვის ბოლომდე დაშლას და დროულად შევამოწმოთ შეზღუდვები.

მესამე პარაგრაფში აღწერილია მორფოლოგიური გარჩევის პროგრამა. მისი მუშაობის პრინციპი. მომხმარებლის ინტერფეისი (ტექსტური რეჟიმი). ნაჩვენებია როგორ ხდება პროგრამის გაშვება, პროგრამიდან გამოსვლა და სიტყვაფორმების გარჩევა. განხილულია პროგრამის სტრუქტურა.

მეოთხე პარაგრაფი დათმობილი აქვს პრეპროცესორის განხილვას. ნაჩვენებია ის მიზეზები თუ რატომ გახდა საჭირო პრეპროცესორის დამატება მორფოლოგიური ანალიზის პროგრამისათვის. მოყვანილია მაგალითები, სადაც ნაჩვენებია თუ როგორ ხდება მაკრო-ჩასმების დეკლარაცია და მათი გამოძახება. განხილულია პარამეტრიანი მაკრო-ჩასმები. განხილულია რეკურსიული და ურთიერთრეკურსიული მაკრო-ჩასმების პრობლემა და მისი გადაჭრის ერთერთი მეთოდი.

მეხუთე პარაგრაფში განხილულია სხვადასხვა სირთულის მაგალითები, რომლებიც გვაცნობენ მორფოლოგიური ანალიზის პროგრამის ძირითად ასპექტებს. განხილულია, თუ როგორ ხდება გრამატიკის ფაილის შედგენა მორფოლოგიური გამრჩევისათვის, როგორ გამოიძახება პროგრამა და როგორ მიეწოდება გასარჩევი სიტყვაფორმები.

მეექვსე პარაგრაფში ზუსტი, ფორმალური სახით არის აღწერილი მორფოლოგიის ფაილის სტრუქტურა. სინტაქსური კონსტრუქციების აღსაწერად გამოყენებულია ბეკუს-ნაურის გაფართოებული ფორმალიზმი. მისი მეშვეობით აღწერილია მორფოლოგიის ფაილის ყველა კონსტრუქცია: მორფემათა კლასების განსაზღვრა, წესების განსაზღვრა, შეზღუდვები, თვისებათა სტრუქტურები და სხვა.

მეშვიდე პარაგრაფში მოყვანილია ქართული ენის მორფოლოგიის ფრაგმენტი რომელიც ჩაწერილია შემუშავებული ფორმალიზმის მეშვეობით და რომლის საფუძველზეც მორფოლოგიური გარჩევის პროგრამის გამოყენებით მიღებულ იქნა ექსპერიმენტული შედეგები. ქართული ენის მორფოლოგიის ფრაგმენტი 4 ფაილად არის დაყოფილი: არსებითი სახელის და ზედსართავი სახელის შემადგენელი მორფემათა კლასები, არსებითი სახელის და ზედსართავი სახელის განმსაზღვრელი წესები, ზმნების შემადგენელი მორფემათა კლასები, ზმნების განმსაზღვრელი წესები.

დისერტაციის მესამე თავში - ”სინტაქსური გამრჩევი” - განხილულია სინტაქსური გამრჩევი, მისი მუშაობის პრინციპი და პროგრამული

რეალიზაცია. დაწვრილებით არის განხილული სინტაქსური გამრჩევისათვის შემუშავებული ფორმალიზმი. განხილულია მაგალითები და ნაჩვენებია ქართული ენის სინტაქსის ფრაგმენტის ჩაწერა შემუშავებული ფორმალიზმის მეშვეობით.

პირველ პარაგრაფში განხილულია სინტაქსური ანალიზის პრობლემა და სინტაქსური გამრჩევისათვის შემუშავებული ფორმალიზმი. განხილულია კონტექსტისაგან-თავისუფალი გრამატიკის წესები, მათი გამოყენების შესაძლებლობა ბუნებრივი ენისათვის. დაწვრილებით არის განხილული წესში სიტყვათა თავისუფალი წყობის შემთხვევა და მის ჩასაწერად შემოტანილია სპეციალური კონსტრუქცია, სიმბოლოების ურთიერთმდებარეობათა მარეგულირებლები, რომელიც ბოლოში ემატება კონტექსტისაგან-თავისუფალი გრამატიკის წესს. ნაჩვენებია ბუნებრივი ენისთვის სტანდარტული კონტექსტისაგან-თავისუფალი წესების გამოყენებისას მათი არასაკმარისობა. რაც გამოწვეულია წინადადების წევრებს შორის შეთანხმებების პრობლემით. განხილულია ამ პრობლემის მოგვარების შესაძლებლობა შეზღუდვების გამოყენებით. წარმოდგენილია ფორმალიზმში გამოყენებული ძირითადი კონსტრუქციები.

მეორე პარაგრაფში აღწერილია სინტაქსური გარჩევის პროგრამა. მისი მუშაობის პრინციპი. მომხმარებლის ინტერფეისი (ტექსტური რეჟიმი). ნაჩვენებია, თუ როგორ ხდება გრამატიკის ფაილის მომზადება, პროგრამის გაშვება და წინადადებების გარჩევა. განხილულია პროგრამის სტრუქტურა. გარჩეულია პროგრამის ყველა ბრძანება და მოყვანილია მათი გამოყენების მაგალითები. ნაჩვენებია როგორ ვიღებთ ანალიზის შედეგს, გარჩევის ხეები და ინფორმაცია მათ შესახებ. განხილულია პროგრამის რეალიზაციის ზოგიერთი დეტალი.

მესამე პარაგრაფში განხილულია სხვადასხვა სირთულის მაგალითები, რომლებიც გვაცნობენ სინტაქსური ანალიზის პროგრამის ძირითად ასპექტებს. განხილულია, თუ როგორ ხდება გრამატიკის და ლექსიკონის ფაილის შედგენა სინტაქსური გამრჩევისათვის, როგორ გამოიძახება პროგრამა და როგორ მიეწოდება გასარჩევი წინადადებები.

მეოთხე პარაგრაფში აღწერილია სინტაქსური გამრჩევის ლექსიკონის ფაილის სტრუქტურა, მასში ინფორმაციის წარმოდგენა თვისებათა სტრუქტურების მეშვეობით. განხილულია სხვადასხვა მაგალითები.

მეხუთე პარაგრაფში ზუსტი, ფორმალური სახით არის აღწერილი სინტაქსური გამრჩევის გრამატიკის ფაილის სტრუქტურა. სინტაქსური კონსტრუქციების აღსაწერად გამოყენებულია ბეკუს-ნაურის ფორმალიზმი. მისი მეშვეობით აღწერილია სინტაქსური გამრჩევის გრამატიკის ფაილის ყველა კონსტრუქცია: წესების განსაზღვრა, შეზღუდვები, კონსტანტები და სხვა.

მეექვსე პარაგრაფში მოყვანილია ქართული ენის სინტაქსის ფრაგმენტი (სამპირიანი ზმნებით განსაზღვრული წინადადებები) რომელიც ჩაწერილია შემუშავებული ფორმალიზმის მეშვეობით და რომლის საფუძველზეც სინტაქსური გარჩევის პროგრამის გამოყენებით მიღებულ იქნა ექსპერიმენტული შედეგები.

დასკვნა და ძირითადი შედეგები

წარმოდგენილ სადისერტაციო ნაშრომში განხილულია ბუნებრივ ენაზე ჩაწერილი ტექსტების კომპიუტერული დამუშავების საკითხები, შემუშავებული ახალი მიდგომები და მათი რეალიზაციის საუძველზე მიღებულია სინტაქსური და მორფოლოგიური ანალიზისათვის განკუთვნილი პროგრამული სისტემა.

იმისათვის რომ პროგრამას შეეძლოს ტექსტების ანალიზი, სპეციალურად ამ პროგრამული სისტემისათვის შემუშავებული ფორმალიზმების ფარგლებში, უნდა ჩავწეროთ კონკრეტული ენის გრამატიკული წესები (სინტაქსური და მორფოლოგიური) და შევქმნათ სიტყვის ძირების ლექსიკონი. პროგრამული სისტემის შემუშავებისას ბუნებრივი ენის მაგალითად აღებულ იქნა ქართული ენა, თუმცა სისტემა სხვა მრავალი ბუნებრივი ენისთვისაც შეგვიძლია გამოვიყენოთ.

პროგრამისათვის აუცილებელი საწყისი ინფორმაცია, გრამატიკის და ლექსიკონის ფაილი, უნდა მოამზადოს შესაბამისი ენის სპეციალისტმა (ლინგვისტმა), რომელიც ღრმად ერკვევა ენის სინტაქსურ და მორფოლოგიურ სტრუქტურაში. მას შემდეგ რაც სპეციალისტი გაეცნობა ფორმალიზმს, რომელიც საკმარისად მარტივი ასათვისებელი და გამოყენებისათვის მოხერხებულია, იგი მზადაა ენის სინტაქსური და მორფოლოგიური წესები ფორმალური სახით ჩაწეროს შესაბამის გრამატიკის ფაილებში. ამის შემდეგ

უკვე შეგვეძლება სისტემის გამოყენება კონკრეტულ ენაზე ჩაწერილი ტექსტების ანალიზისათვის. გრამატიკის ფაილების შექმნა, მათი ტესტირება და კორექტირება შეიძლება გარკვეული პერიოდი გაგრძელდეს, ვიდრე არ მივიღებთ მაქსიმალურად ზუსტ შედეგს. წარმატებული ანალიზისათვის საჭიროა, რომ ფორმალური სახით ჩაწერილი გრამატიკა დიდი სიზუსტით ასახავდეს ბუნებრივი ენის გრამატიკას.

პროგრამული სისტემა გამოიცადა ქართული ენისათვის. ქართული ენის გარკვეული ნაწილისათვის, შემუშავებული იქნა გრამატიკის ფაილები სინტაქსური და მორფოლოგიური ანალიზისათვის. გრამატიკული წესები ძირითადად ჩაიწერა ა. შანიძის გრამატიკის მიხედვით, ხოლო ზმნების მორფოლოგიური ანალიზისათვის გამოყენებული იქნა დ. მელიქიშვილის მიერ შედგენილი ქართული ზმნების კლასიფიკაცია. მიღებულმა შედეგებმა ცხადყო რომ შექმნილი პროგრამული სისტემის გამოყენება საკმარისად ეფექტურია პრაქტიკული ამოცანების გადასაწყვეტად.