საგრანტო პროექტის სახელწოდება	ქართული ენის კორპუსის სრული (მორფოლოგიური, სინტაქსური, სემანტიკური) ანოტირების სისტემა
საგრანტო ხელშეკრულება №	№31/65
საგრანტო პროექტის სამეცნიერო ხელმძღვანელი	გიორგი ჩიკოიძე
პროექტის ხანგრძლივობა	25.04.2013 – 24.04.2016

რეზიუმე

პროექტის ფარგლებში შემუშავდა ქართული ენის მორფოლოგიური, სინტაქსური და სემანტიკური ანალიზატორი. სუბკორპუსად, რომელზედაც გამოვცადეთ ქართული ენის სრული ანოტირების სისტემა, შევარჩიეთ ჩვენი დროის გამოჩენილი ქართველი მწერლის ოთარ ჭილაძის პროზა. კორპუსში შესულია ოთარ ჭილაძის რომანები: გზაზე ერთი კაცი მიდიოდა, მარტის მამალი, რკინის თეატრი, ყოველმან ჩემმან მპოვნელმან, აველუმი, გოდორი.

პროექტის ფარგლებში ანოტირებული კორპუსის დახმარებით შესაძლებელია:

- კონკრეტული სიტყვაფორმის მოძიება და კონკორდანსის სახით გამოტანა;
- სიტყვაფორმის ძიება ლემის მიხედვით;
- წყვეტილი ან უწყვეტი სინტაგმის მიხედვით სიტყვაფორმათა ჯგუფის ძიება;
- სიტყვაფორმების ძიება მორფოლოგიური მახასიათებლების მიხედვით;
- სხვადასხვა ლექსიკო-გრამატიკული სტატისტიკური მონაცემების მოპოვება;
- კონკორდანსიდან შერჩეული სტრიქონების ცალკეულ ფაილში შენახვა.

კორპუსის ტექსტი ანოტირებულია მორფოლოგიური, სინტაქსური და სემანტიკური მარკერებით, რომლებშიც ასახულია მწერლის ენის მორფოლოგიური, სინტაქსური და სემანტიკური სტრუქტურა. კორპუსში გამოიყო 655,811 სიტყვაფორმა და 97,155 სიტყვათხმარება. ტექსტში ყველა დონეზე ნაწილობრივ მოხსნილია ომონიმია.

კორპუსი განთავსებულია საქართველოს ტექნიკური უნივერსიტეტის ვებგვერდზე http://geocorpora.gtu.ge/#/texts.

კომპიუტერული ლინგვისტიკის განვითარებისა და მშობლიური ენის სათანადო დონისა და გავრცელების არის შესანარჩუნებლად, თანამედროვე ელექტრონული ენობრივი კორპუსების არსებობა წარმოადგენს მეტად მნიშვნელოვან და პრიორიტეტულ საშუალებას. ის გვთავაზობს როგორც ენის სისტემურობის შემეცნებას (მოდელირება), ისე, მისი დღემდე შექმნილი კონკრეტული მასალის, კერძოდ, ლიტერატურული ძეგლების ასახვას, ფიქსირებას, შესწავლასა და მათ გამოყენებას ენობრივი სისტემის კვლევისა (ენის მოდელის აგება) და პრაქტიკული მიზნებისთვის (მთარგმნელობითი, დიალოგური, ენის მასწავლი კომპიუტერული სისტემები).

ზოგადად, რამდენადაც ანოტაცია მოიცავს ტექსტის ენის შესახებ ნებისმიერი სახის ანალიტიკურ ინფორმაციას, იმდენად წარმატებული ანოტირების შემდეგ ფასდაუდებელი მასალა გროვდება ენობრივი სისტემის კომპიუტერული მოდელების ასაგებად და სხვადასხვა ლინგვისტური ჰიპოთეზების შესამოწმებლად. ეს კი, ჩვენი აზრით, პროექტის ერთ-ერთ მნიშვნელოვანი შედეგია.

საბოლოო ანგარიში

პროექტის მიზანი იყო პროგრამული ინსტრუმენტის შექმნა, რომლის დახმარებითაც შესაძლებელი იქნებოდა ტექსტური კორპუსების ნახევრადავტომატური ანოტირება მორფოლოგიურ, სინტაქსურ და სემანტიკურ დონეებზე. ამისათვის სამუშაო თერთმეტ ძირითად ამოცანად დაიყო:

1) ქართული ენის კორპუსის ტექსტების მოპოვება და სტრუქტურირება

კორპუსისთვის შევარჩიეთ ჩვენი დროის გამოჩენილი ქართველი მწერლის, ოთარ ჭილაძის რომანები: გზაზე ერთი კაცი მიდიოდა, მარტის მამალი, რკინის თეატრი, ყოველმან ჩემმან მპოვნელმან, აველუმი, გოდორი. ოთარ ჭილაძის რომანების ელექტრონული ტექსტების მოპოვებაში დაგვეხმარა გამომცემლობა „არეტე“-ს დირექტორი ბატონი გია დარსალია, რისთვისაც მას დიდ მადლობას ვუხდით. მოხდა ყველა ტექსტის კონვერტირება ვორდის ფაილად და მისი გრაფომეტრიული დამუშავება. ტექსტებს მოსცილდა ზედმეტი ხარვეზები (ჰარები) და ცარიელი აბზაცები. ყველა ნაწარმოების მიხედვით შეიქმნა ერთიანი ტექსტური ბაზა. ტექსტები ყველგან ჩაიწერა Sylfaen შრიფტით. ერთიან ტექსტურ ბაზას შესაბამისი საიდენტიფიკაციო კოდით დაუკავშირდა კორპუსის სიტყვანი. თითოეული უნიკალური სიტყვა საიდენტიფიკაციო კოდით დაუკავშირდა ავტორის, ნაწარმოების, თავის/ქვეთავის, აბზაცისა და წინადადების საიდენტიფიკაციო კოდს.

2) ქართული ენის კორპუსის ტექსტების მეტაანოტირება

შესწავლილ იქნა მეტაანოტირების საერთაშორისო სტანდარტები: TEI (Text Encoding Initiative), CES (Corpus Encoding Standard), CDIF (Corpus Document Interchange Format), XCES (Corpus Encoding Standard for XML) და XML ტექნოლოგია. მათზე დაყრდნობით ტექსტურ კორპუსში განსათავსებლად დოკუმენტებისთვის შემუშავდა ე.წ. „ტექსტის პასპორტის“ შედგენის ტექნოლოგია, რაც გულისხმობს ტექსტური დოკუმენტების ფაილის დასაწყისში მეტამონაცემების პარამეტრების ჩაწერას. ყველა ტექსტს მიენიჭა შესაბამისი პასპორტი. მეტაანოტირების საერთაშორისო სტანდარტებზე დაყრდნობით ტექსტურ კორპუსში განსათავსებლად შემუშავდა ვებაპლიკაცია. შეიქმნა ვებსერვერი და მის ბაზაში განთავსდა ერთი ავტორის ლინგვისტური კორპუსის მასალა.

3) ზმნის რეგულარული სუპერპარადიგმების ჩამოყალიბება

ჩამოყალიბდა და შეირჩა ზმნური სუპერ-პარადიგმა როგორც ერთი და იმავე ლექსემისგან ნაწარმოები პარადიგმების ერთობლიობა, რომელსაც, სხვადასხვა საწყისი ფორმების შემთხვევაში, შეიძლება გააჩნდეს განსხვავებული სემანტიკური და გრამატიკული სტრუქტურა. გამოიყო რეგულარული და არარეგულარული სუპერპარადიგმები. მათზე დაყრდნობით ჩატარდა ზმნების კლასიფიკაცია.

4) მორფოლოგიური ანალიზატორის შემუშავება

შეიქმნა დესკტოპ აპლიკაცია, სადაც სალიტერატურო და ავტორისეული ენების მორფოლოგიური ანალიზის პროგრამები ცალკე პროგრამულ უტილიტებადაა რეალიზებული. მორფოლოგიურად გაანალიზდა კორპუსის სიტყვანი. კორპუსში გამოიყო 655,811 სიტყვაფორმა და 97,155 სიტყვათხმარება. გაირჩა, ციფრებით წარმოდგენილი 40 რიცხვითი სახელი, 38 სიტყვათხმარება ხარვეზითაა ჩაწერილი. დარჩენილი 97,076 სიტყვათხმარებიდან ომონიმია მოეხსნა 84,686 ერთეულს, ომონიმურია 12,390 სიტყვათხმარება, მიახლოებით გაანალიზდა 480 სიტყვათხმარება, სავარაუდო მარკერები მიეწერა 4,526 სიტყვათხმარებას.

5) სინტაქსური ანალიზატორის შემუშავება

სინტაქსურმა ანოტირებამ უნდა ასახოს ცალკეული წინადადების სინტაქსური სტრუქტურა, რომლის მიხედვითაც აიგება სიტყვათა კავშირების ხე. სინტაქსური გრაფით წარმოდგენილი წინადადების თითოეული სიტყვა დაწყვილებულია თავის დომინანტთან (ერთწევრიან წინადადებაში თვითონ ეს სიტყვაა დომინანტი). სახელდებითი წინადადების გარდა, რომელშიც

მთავარ დომინანტად საგნის, ან მოვლენის აღმნიშვნელი პირველივე სიტყვა ჩაითვლება, შემასმენელი არის სინტაქსური ხის მთავარი წვერო - დომინანტი. წინადადების მხოლოდ ერთ წევრს აქვს მიწერილი მთავარი დომინანტის ნიშანი - D. დანარჩენი წევრები დამოკიდებული წევრებია და ყველას მიწერილი აქვს თავის დომინანტ წყვილთან მიღებული კავშირის შესაბამისი ტიპის მარკერი. სინტაქსურად გაანალიზდა და სინტაქსური მახასიათებელი მიეწერა 97,155 ერთეულს. კორპუსში სინტაქსური ომონიმია მოხსნილი არაა.

6) სემანტიკური ანალიზატორის შემუშავება

ტექსტების სემანტიკური ტეგების მიწერის დროს ერთ-ერთი მოთხოვნა ასეთია: მნიშვნელოვანი ადგილი უნდა ეკავოს სიტყვაფორმის შესაბამისი ლემის (ამოსავალი სიტყვის) ხმარების არეს, ანუ მითითებას იმაზე, თუ რა სიტყვებს ეხამება მოცემული სიტყვა და რა აზრობრივ მიმართებაშია იგი სხვა სიტყვებთან. თავდაპირველად სიტყვის სემანტიკური ინფორმაცია როგორც ლექსიკონის ერთ-ერთი ზონა მოთავსებული იყო სიტყვათა ლემებთან ე. წ. "განმარტებით-კომბინატორულ ლექსიკონში". ხოლო შემდეგ, მასზე დაყრდნობით მოხდა ტექსტების სემანტიკური ტეგირება. ტექსტებში სიტყვისთვის მიწერილია ლექსიკო-სემანტიკური ინფორმაცია ჯგუფის შემდეგი მარკერებით: 1. ძირითადი მახასიათებელი - მეტყველების ნაწილი; 2. თვით ლექსიკო-სემანტიკური ინფორმაცია; 3. დერივაციული (სიტყვაწარმოქმნითი) მახასიათებლები.

7) ლექსიკურ-გრამატიკული შესაბამისობების დადგენა-შეთანადება თანამედროვე ნორმატიულ ლექსიკონებთან და მორფოლოგიური და სინტაქსური ლექსიკონების გამდიდრება

შესწავლილ იქნა საერთაშორისო ორგანიზაციების (EAGLES (Expert Advisory Group on Language Engineering Standards), ISLE (International Standards for Language Engineering), LGR (The Leipzig Glossing Rules), ISO (International Standardization Organization)) მიერ მიღებული მორფოლოგიური ანოტირების სტანდარტები. მათი გათვალისწინებით მორფოლოგიური პროცესორის ლექსიკონები შეივსო ახალი მარკერებით. მორფოლოგიური ანალიზატორის დესკტოპ აპლიკაციას დაემატა პროგრამული უტილიტა მორფოლოგიური მახასიათებლებისთვის სტანდარტული მარკერების ნახევრად-ავტომატურად მინიჭებისათვის.

8) ქართულ ზმნათა სუპერპარადიგმების სრული კლასიფიკაცია და სათანადო სემანტიკური ინფორმაციის ლექსიკონში ასახვა.

მორფოლოგიური ანალიზატორის ლექსიკონში ზმნების მეთაურ სიტყვებთან მიწერილია მითითება იმ კლასზე, რომელსაც სუპერ-პარადიგმა განეკუთვნება და იქვე ჩამოთვლილია ეტაპობრივი პარადიგმების გამომხატველი ერთეულების მისამართები, როგორიცაა კაუზირება - CAUS, პროცესი (მიმდინარეობა) - PROC და რეზულტატი (შედეგი) – RES.

9) კორპუსის ინტერაქტიული ანალიზატორის პროგრამული რეალიზაცია

შეიქმნა ვებ-აპლიკაცია eSketch-ის უტილიტა, რომლის დახმარებითაც შესაძლებელია თითოეული უნიკალური სიტყვის მოძებნა წინადადების კონტექსტში, რომლის წევრსაც ეს სიტყვა წარმოადგენს აგრეთვე, ანოტირებულ კორპუსში შესაძლებელია ომონიმიის ხელით მოხსნა ანალიზატორის ფანჯარაში, რომელშიც სიტყვაფორმის გარჩევის რამდენიმე პასუხიდან ოპერატორი მონიშნავს ერთ-ერთს. ვებაპლიკაციაში ცალკეული ნაწარმოების მიხედვით ფუნქციონირებს ომონიმურ სიტყვაფორმათა სიხშირის მთვლელი და ანოტირებულ სიტყვაფორმათა კონკორდანსული წარმოდგენა.

10) ანოტირებული კორპუსის ინტერნეტში განთავსება

ოთარ ჭილაძის ყველა რომანის ტექსტი განთავსებულია კორპუსში. ტექსტები ანოტირებულია მორფოლოგიურ, სინტაქსურ და სემანტიკურ დონეზე. გრძელდება ომონიმიის ხელით მოხსნა.

11) ტესტირება

ოთარ ჭილაძის რომანების მიხედვით შედგენილი ქართული ენის ლინგვისტური კორპუსი განთავსებულია საქართველოს ტექნიკური უნივერსიტეტის ვებგვერდზე http://geocorpora.gtu.ge/#/texts. ტესტირება წარმატებულადაა ჩატარებული.