საგრანტო პროექტის სახელწოდებაქართულ სიტყვათა ქსელის კომპაილერი – GeWordNet, **საგრანტო ხელშეკრულება №** AR/84/4-105-14

საგრანტო პროექტის სამეცნიერო ხელმძღვანელი ლიანა ლორთქიფანიძე

პროექტის ხანგრძლივობა 28.04.2015-28.04.2017

რეზიუმე

ინტერნეტ-სივრცეში საძიებო სისტემების ინტელექტუალიზაცია მნიშვნელოვნად ზრდის ძიების სიჩქარესა და ხარისხს. დოკუმენტებში ძიებისას ბუნებრივ წინააღმდეგობას ქმნის სინონიმია (სხვადასხვა ცნების აღნიშვნა ერთი და იგივე სიტყვით ან ტერმინით) და პოლისემია (საერთო სემანტიკის მქონე ცნებების აღნიშვნა სხვადასხვა სიტყვით ან ტერმინით). ბოლო წლებში ეს პრობლემა ბევრი ენისთვის დაძლეულია სპეციალური ელექტრონული WordNet ტიპის თესაურუსების გამოყენებით.

ჩვენ მიერ შექმნილი GeWordNet-ი არის ერთგვარი ლექსიკური ონტოლოგია კომპიუტერულ მეცნიერებაში. მისი მიზანია ერთდროულად შექმნას ლექსიკონისა და თესაურუსის კომბინაცია, რომელიც ხელს შეუწყობს ტექსტის ავტომატური ანალიზის განხორციელებას და ხელოვნური ინტელექტის ამოცანების შესრულებას.

ლექსიკონი წარმოდგენილია ოთხი ქსელისაგან, რომლებშიც გაერთიანებულია ძირითადი მეტყველების ნაწილები: არსებითი სახელები, ზმნები, ზედსართავი სახელები და ზმნიზედები.

GeWordNet-ის უშუალო პოტენციური მომხმარებლები იქნებიან სხვადასხვა ინტერნეტ-საძიებო სისტემები (Google, Yandex, Yahoo და სხვ.), ქართული ენით დაინტერესებული სხვადასხვა დისციპლინარული სპექტრის მეცნიერები (ეთნოლოგები, ანთროპოლოგები, სოციოლინგვისტები, ლექსიკოგრაფები....). GeWordNet-ი, ისევე როგორც WordNet-ი სხვა ენებისთვის, პოპულარული იქნება საქართველოს სხვადასხვა უნივერსიტეტის ჰუმანიტარული ფაკულტეტების სტუდენტებში, საზღვარგარეთ მცხოვრები ქართველებისთვის და ქართული ენის შესწავლით დაინტერესებული ნებისმიერი პირისთვის.

WordNet თესაურუსი გამოიყენება:

- ინფორმაციის ძიებისას მომხმარებლის მოთხოვნის გასაფართოებლად პარადიგმატულად და სინტაგმატურად დაკავშირებული სიტყვების მეშვეობით. ასეთი სიტყვებია, მაგალითად, სინსეტის (SynSet – სინონიმური მწკრივი, რომელშიც გაერთიანებულია მსგავსი მნიშვნელობის მქონე სიტყვები) კომპონენტები, ან კონტექსტური ძიებისათვის საჭირო „ზმნა-აქტანტი“-ს ტიპის კავშირები;
- ფორმალური გრამატიკების ლექსიკონად, განსაკუთრებით ზმნების ვალენტობისა და არსებითი და ზედსართავი სახელების ამომწურავად აღწერისას;
- სპეციალიზებული ლექსიკონების (სამედიცინო, ეკონომიკური, გეოგრაფიული, ბიოლოგიური და სხვ.) შესადგენად;
- ენის სხვადასხვა ქვესისტემების (მაგალითად, დიალექტური ლექსიკონი) შესადგენად;
- სიტყვათა სინტაგმატური მიმართებების საშუალებით სიტყვების არაერთმნიშვნელოვნობის მოსახსნელად;
- ტექსტის ავტომატური დამუშავებისა და ინფორმაციული ძიების პროგრამულ დანართებში დოკუმენტების ფილტრაციისა და რუბრიკაციის ხარისხის გასაზრდელად;
- ჰიპერონიმული მიმართებების საფუძველზე აზრობრივად ახლო მდგომი ტექსტების განსაზღვრისთვის.

მომხმარებელს GeWordNet-ის გამოყენება შეუძლია თესაურუსის მართვისთვის შემუშავებული Web-სერვისის საშუალებით, რომელიც განთავსებულია მისამართზე: GeWordNet.gtu.ge.

საბოლოო ანგარიში

GeWordNet დიდი სემანტიკური ქსელია. მისთვის დამახასიათებელია ორი სახის სემანტიკური მიმართებები: ლექსიკური (სიტყვა-სიტყვა) და კონცეპტუალური (კონცეპტი-კონცეპტი). ყველაზე მნიშვნელოვანი ლექსიკური მიმართება არის სინონიმია. GeWordNet თესაურუსის საბაზისო სტრუქტურული ერთეული არის არა ცალკეული სიტყვა, არამედ სინონიმური მწკრივი ე. წ. სინსეტი, რომელიც აერთიანებს მსგავსი მნიშვნელობის სიტყვებს და ცნებებს. სინსეტებს შორის დამყარებულია სასრული რაოდენობის ასოციატიურ-სემანტიკური მიმართებები, როგორიცაა ჰიპონიმია (სახე-გვარეობითი), მერონიმია (ნაწილი-მთელი), ლექსიკური მიმართება (კაუზაცია, პრესუპოზიცია) და სხვ.; მათ შორის ძირითად როლს ასრულებს ჰიპონიმია, რომელიც სინსეტების იერარქიული (ხისებრი) სტრუქტურის ორგანიზების შესაძლებლობას იძლევა.

GeWordNet თესაურუსის გამოყენება შესაძლებელია:

- ინფორმაციის ძიებისას მომხმარებლის მოთხოვნის გასაფართოებლად პარადიგმატულად და სინტაგმატურად დაკავშირებული სიტყვების მეშვეობით. ასეთი სიტყვებია, მაგალითად, სინსეტის (SynSet) კომპონენტები, ან „ზმნა-აქტანტი“-ს ტიპის კავშირები, რომლებიც კონტექსტური ძიებისათვის არის საჭირო;

- ფორმალური გრამატიკების ლექსიკონად, განსაკუთრებით ზმნების ვალენტობის, არსებითი და ზედსართავი სახელების ამომწურავი აღწერისას;

- სპეციალიზებული ლექსიკონების (მაგალითად, სამედიცინო, ეკონომიკური, გეოგრაფიული, ბიოლოგიური და სხვ.) შესადგენად;

- სხვადასხვა დიალექტებისა და ენების ლექსიკონების შესადგენად;

- სიტყვათა სინტაგმატური მიმართებების საშუალებით კლასიკური ამოცანის - სიტყვების არაერთმნიშვნელოვნების მოსახსნელად;

- ტექსტის ავტომატური დამუშავებისა და ინფორმაციული ძიების პროგრამულ დანართებში დოკუმენტების ფილტრაციისა და რუბრიკაციის ხარისხის გასაზრდელად;

- ჰიპერონიმული მიმართებების საფუძველზე აზრობრივად ახლო მდგომი ტექსტების განსაზღვრისთვის.

პროექტი განხორციელდა საქართველოს ტექნიკური უნივერსიტეტის არჩილ ელიაშვილის მართვის სისტემების ინსტიტუტის ენობრივი და სამეტყველო სისტემების განყოფილების ბაზაზე.

პროექტის პირველ ამოცანას წარმოადგენდა ტექსტური ინფორმაციის დამუშავების ვექტორული სივრცის მოდელების ალგორითმიზაცია, პროგრამული რეალიზაცია და WordNet თესაურუსის აგებისას მათი ეფექტურობის შეფასება. ამოცანის შესრულების სავარაუდო დრო თვეების მიხედვით – 1 - 23-ე თვე. საანგარიშო პერიოდში ტექსტური ინფორმაციის ვექტორული წარმოდგენის თანამედროვე მეთოდების გამოყენებით შესრულდა ტექსტური ინფორმაციის დამუშავების პროგრამული მოდული. აღნიშნულ თემატიკაზე გაკეთდა მოხსენება კონფერენციაზე WordNet თესაურუსის ტექნოლოგიის სტანდარტების შესახებ http://ict-mc.gtu.ge/conference.pdf; გაკეთდა მოხსენება კონფერენციაზე: ტექსტური ინფორმაციის ავტომატური დამუშავების მოდული ქართული ენის GeWordNet თესაურუსისთვის http://conference.ens-2017.tsu.ge/uploads/5892067e85b23Lortkipanidze_geo.pdf .

პროექტის მეორე ამოცანას წარმოადგენდა ქართული ენის განმარტებითი, ქართულ სინონიმთა, ქართული იდიომების, თანამედროვე ქართული ენის იდეოგრაფიული და უცხო სიტყვათა ლექსიკონების ლექსიკოგრაფიულ მონაცემთა ბაზების შევსება და რედაქტირება. ამოცანის შესრულების სავარაუდო დრო თვეების მიხედვით – 1 – 23-ე თვე. საანგარიშო პერიოდში შეიქმნა ქართული ენის განმარტებითი, ქართულ სინონიმთა, ქართული იდიომების, თანამედროვე

ქართული ენის იდეოგრაფიული და უცხო სიტყვათა ლექსიკონის ლექსიკოგრაფიულ მონაცემთა ბაზა.

პროექტის მესამე ამოცანას წარმოადგენდა ქართული ენის განმარტებითი, ქართულ სინონიმთა, ქართული იდიომების, თანამედროვე ქართული ენის იდეოგრაფიული და უცხო სიტყვათა ლექსიკონების ლექსიკოგრაფიულ მონაცემთა ბაზების ფორმატირება სალექსიკონო ერთეულის ინფორმაციული ველების მიხედვით. ამოცანის შესრულების სავარაუდო დრო თვეების მიხედვით – 7 – 23-ე თვე. საანგარიშო პერიოდში ჩატარდა ფორმატიზაცია სალექსიკონო ერთეულის ინფორმაციული ველების მიხედვით ქართული ენის განმარტებითი, ქართულ სინონიმთა, ქართული იდიომების, თანამედროვე ქართული ენის იდეოგრაფიული და უცხო სიტყვათა ლექსიკონის ლექსიკოგრაფიულ მონაცემთა ბაზებში.

პროექტის მეოთხე ამოცანას წარმოადგენდა ლექსიკოგრაფიულ მონაცემთა ბაზების ლექსიკონების მორფოლოგიური, სინტაქსური და სემანტიკური ანოტირება. ამოცანის შესრულების სავარაუდო დრო თვეების მიხედვით – მე-7 – 23-ე თვე. საანგარიშო პერიოდში ჩატარდა ქართული ენის განმარტებითი, ქართულ სინონიმთა, ქართული იდიომების, თანამედროვე ქართული ენის იდეოგრაფიული და უცხო სიტყვათა ლექსიკონის მორფოლოგიური, სინტაქსური და სემანტიკური ანოტირება.

პროექტის მეხუთე ამოცანას წარმოადგენდა თესაურუსის სტრუქტურის მიხედვით ჰიპონიმური ხის ავტომატური ფორმირების ალგორითმის შემუშავება. ამოცანის შესრულების სავარაუდო დრო თვეების მიხედვით – მე-7 – 23-ე თვე. საანგარიშო პერიოდში დასრულდა თესაურუსის სტრუქტურის მიხედვით ჰიპონიმური ხის ავტომატური ფორმირების ალგორითმის შემუშავება. აღნიშნულ თემატიკაზე გაკეთდა მოხსენებები ტექსტური ინფორმაციის დამუშავების ვექტორული სივრცის მოდელების ალგორითმიზაციის შესახებ http://ict-mc.gtu.ge/conference.pdf; თესაურუსის სტრუქტურის მიხედვით ჰიპონიმური ხის ავტომატური ფორმირების ალგორითმის შესახებ http://www.ice.ge/new/pages/inst/confer/Chiqobavas%20sakitkhavebi/XXVI.pdf; თესაურუსის სტრუქტურის მიხედვით ჰიპონიმური ხის ავტომატური ფორმირების ალგორითმის პროგრამული რეალიზაციის შესახებ https://iliauni.edu.ge/uploads/other/38/38129.pdf; WordNet თესაურუსის აგებისას ტექსტური ინფორმაციის დამუშავების ვექტორული სივრცის მოდელების ეფექტურობის შესახებ http://conference.ens-2017.tsu.ge/lecture/view/743. გამოიცა სამეცნიერო სტატიები ტექსტური ინფორმაციის დამუშავების ვექტორული სივრცის მოდელების ალგორითმიზაციის შესახებ; ვექტორული სივრცის მოდელები და ქართულენოვანი ტექსტების დამუშავება; თესაურუსის სტრუქტურის მიხედვით ჰიპონიმური ხის ავტომატური ფორმირების პროგრამული რეალიზაცია (დანართი 2).

პროექტის მეექვსე ამოცანას წარმოადგენდა თესაურუსის სტრუქტურის მიხედვით ჰიპონიმური ხის ავტომატური ფორმირების ალგორითმის პროგრამული რეალიზაცია. ამოცანის შესრულების სავარაუდო დრო თვეების მიხედვით – მე-7 – 23-ე თვე. საანგარიშო პერიოდში შემუშავდა თესაურუსის სტრუქტურის მიხედვით ჰიპონიმური ხის ავტომატური ფორმირების ალგორითმი და დასრულდა GeWordNet თესაურუსის სამომხმარებლო ინტერფეისის პროგრამული რეალიზაცია. აღნიშნულ თემატიკაზე საქართველოს ტექნიკური უნივერსიტეტის არჩილ ელიაშვილის მართვის სისტემების ინსტიტუტის შრომათა კრებულში დაიბეჭდა სტატია (დანართი 2).

პროექტის მეშვიდე ამოცანას წარმოადგენდა GeWordNet – ქართული ენის ლექსიკური სისტემის ორგანიზება პრინსტონის WordNet თესაურუსის შესაბამისად. რადგან GeWordNet-ის ლექსიკამ უნდა ასახოს ლექსიკონის ყველაზე მნიშვნელოვანი სტრუქტურული მიმართებები და მოიცვას თანამედროვე ქართული ენის ძირითადი ბირთვი, ამიტომ ლინგვისტურ რესურსში გაერთიანდება რამდენიმე სხვადასხვა პლანის აღწერა: ტრადიციული ლექსიკოგრაფიული, ენობრივი ცნობიერების მოდელი და მონაცემთა წარმოდგენა კომპიუტერული ფორმით. ყოველივე ეს იმედს გვაძლევს, რომ GeWordNet-ის გამოყენება შესაძლებელი იქნება სხვადასხვა საინფორმაციო სისტემებში. აღნიშნულ

თემატიკაზე სამეცნიერო კონფერენციაზე გაკეთდა მოხსენებები: პრინსტონის WordNet თესაურუსის შესაბამისად ორგანიზებული GeWordNet ქართული ენის ლექსიკური სისტემის მოდელის შესახებ http://gtu.ge/pdf/konf/verbaluri_komunikacia_ge.pdf; „ქართული ენის ლექსიკური სისტემა GEWORDNET თესაურუსში“. https://www.tsu.ge/data/image_db_innova/shanidze.programa.Tezisebi.pdf.

პროექტის მერვე ამოცანას წარმოადგენდა GeWordNet თესაურუსის ინტერნეტში განთავსება. ამოცანის შესრულების სავარაუდო დრო თვეების მიხედვით: მე-19 – 23-ე თვე.

შემუშავდა ქართულ სიტყვათა ქსელის ვებსაიტის Front-end-ის მოდულები: საიტის დიზაინ-მაკეტი, საიტის განლაგება, CMS შაბლონები, მომხმარებლის ინტერფეისის ვიზუალიზაცია, ვებ-ანიმაცია.

შემუშავდა ქართულ სიტყვათა ქსელის ვებსაიტის Back-end-ის მოდულები: საიტის სერვერული ნაწილის რეალიზაცია, მონაცემთა ბაზების ინტეგრაცია და დაკავშირება მომხმარებლის (Front-end) მხარესთან, სერვერზე საიტის პროგრამული უზრუნველყოფის მართვა. ვებგვერდის მისამართია www.GeWordNet.gtu.ge .

პროექტის მეცხრე ამოცანას წარმოადგენდა GeWordNet თესაურუსის ტესტირება. ამოცანის შესრულების სავარაუდო დრო თვეების მიხედვით: 24-ე თვე.

2017 წლის 3 მაისს ტექნიკური უნივერსიტეტის ადმინისტრაციულ კორპუსში შედგა GeWordNet თესაურუსის პრეზენტაცია და ტესტირება. პრეზენტაციას ესწრებოდნენ ტექნიკური უნივერსიტეტის პრორექტორი - ზურაბ გასიტაშვილი, ტექნიკური უნივერსიტეტის არჩილ ელიაშვილი მართვის სიტემების ინსტიტუტის დირექტორი - ნუგზარ ყავლაშვილი, შოთა რუსთაველის ეროვნული სამეცნიერო ფონდის საზოგადოებასთან ურთიერთობის სამსახურის უფროსი - თეა მჟავანაძე, უნივერსიტეტისა და ინსტიტუტის თანამშრომლები.

ტესტირებამ წარმატებით ჩაიარა და დამსწრეთა მოწონება დაიმსახურა.