ტექსტური ინფორმაციის ავტომატური დამუშავების მოდული ქართული ენის GeWordNet თესაურუსისთვის

ლიანა ლორთქიფანიძე

liana.lortkipanidze@tsu.ge

პრაქტიკული ინფორმატიკა,

კომპიუტერული მეცნიერების დეპარტამენტი

ივ. ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი

უნივერსიტეტის ქ 13, თბილისი 0186, საქართველო

საკვანძო სიტყვები: ელექტრონულ სიტყვათა ქსელი - WordNet, ტექსტის ავტომატური დამუშავება, GeWordNet თესაურუსი.

დღეს ინტერნეტის საძიებო სისტემები სულ უფრო ხშირად იყენებენ პრინსტონის WordNet თესაურუსის სტრუქტურის მსგავს ლექსიკო-სემანტიკურ მონაცემთა ბაზებს. ზოგადად WordNet (http://wordnet.princeton.edu/) ოთხი ნაწილისაგან შედგება, რომლებშიც შესულია არსებითი სახელები, ზმნები, ზედსართავები და ზმნიზედები. თითოეული მათგანი წარმოადგენს სემანტიკურ ქსელს, რომლის კვანძებიც შესაბამისი მეტყველების ნაწილების სინონიმური მწკრივებია.

GeWordNet პროექტის ფარგლებში მონაცემთა ბაზების ფორმირება ძირითადად ხელით წარმოებს. ქართული ენის სპეციფიკიდან გამომდინარე ამგვარად უფრო ხარისხიანი თეზაურუსის მიღებაა შესაძლებელი.

წარმოდგენილ სამუშაოში განვიხილავთ ქართული GeWordNet-ის მონაცემთა ბაზების ნაწილობრივ ავტომატურად ფორმირების მიდგომას. ძირითადი იდეა მდგომარეობს ინგლისური სინსეტების შესაბამისი ქართული შესატყვისით ჩანაცვლებაში. ამავდროულად არ უნდა დაირღვეს სემანტიკური ქსელის სტრუქტურა. ზოგ შემთხვევაში ასეთი ჩანაცვლების განხორციელება შესაძლებელია ავტომატურად, ვინაიდან ინგლისური სიტყვებისა და კოლოკაციების უბრალო ჩანაცვლება კი არ უნდა მოხდეს ქართულით, არამედ მხოლოდ ისეთი სიტყვების, რომლებიც სინონიმურად არიან ერთმანეთთან დაკავშირებული.

საწყის მონაცემთა ბაზად გამოიყენება English Princeton WordNet. მოხსენებაში განხილული და გაანალიზებული იქნება ასეთი მიდგომის დადებითი და უარყოფითი მხარეები.