

ანოტაცია

ქართულ ენაში ზმნას რთული სისტემას აქვს. უცხოელებისთვის ქართულის სწავლაში დასახმარებლად ჩვენ ვქმნით ქართული ზმნების ნაუღლები ფორმების დაკავშირებულ მონაცემთა ბაზას - KARTUVERBS. ის ამჟამად შეიცავს 5 მილიონამდე ნაუღლებ ფორმას, რომლებიც დაკავშირებულია 16 000-ზე მეტ ზმნასთან. ბაზაში 80 მილიონზე მეტი კავშირია. ბაზიდან ინფორმაციის გამოთხოვა/მიღების დრო მისაღებია მაშინაც კი, როდესაც ეს სისტემა გაშვებულია პირად კომპიუტერზე. ბაზის მონაცემები დაგენერირებულია გარე ტექსტური წყაროდან. ამ პროექტში დეტალურადაა წარმოდგენილი ყველა იმ პროცესი, რომლითაც ხდება ტექსტური სტრუქტურირებული მონაცემების ამოღება, მათი გაუმჯობესება, ვალიდაცია და შემდეგ, სემანტიკურად დაკავშირებულ მონაცემებად გარდაქმნა. პროცესში თანმიმდევრულადაა გამოყენებული მონაცემების გაუმჯობესების სხვადასხვა ინსტრუმენტები. მასწავლებლიანი მანქანური სწავლების ალგორითმით - გადაწყვეტილების ხე, გაიწვრთნა კლასიფიკაციის მოდელი სახელზმნების გამოსაცნობად. სახელზმნა, როგორც წესი, წარმოადგენს ლექსიკონების ლემა. მოდელმა აჩვენა 98%-99% სიზუსტე. შედეგი დასტურდება ქართული ეროვნული ქართული კორპუსის ლემებთან შედარებითაც. როდესაც კორპუსში წარმოდგენილია ზმნის ლემა, ჩვენს სისტემის ლემა ემთხვევა მას. დანარჩენი შემთხვევებისთვის ვიყენებთ აკადემიურ ქრაუნდსორსინგ (crowdsourcing) პლატფორმას - HEADWORK, სადაც დარგის ექსპერტები და სხვა მონაწილეები შემოთავაზებული ფორმების სახელზმნის სისწორეს ადასტურებენ ან უარყოფენ ან/და ტოვებენ შენიშვნას/კომენტარს/წინადადებას.

ქართული ენის შემსწავლელთათვის KARTUVERBS არის ციფრული ლექსიკოგრაფიული ინსტრუმენტი, რომელიც მათ ეხმარება ნებისმიერი ნაუღლები ფორმიდან სახელზმნის იდენტიფიცირებაში და სთავაზობს API-ს თარგმნისთვის. ქართველი ლექსიკოგრაფებისთვის KARTUVERBS არის სისტემა, რომელიც ციფრულად შეიცავს თითქმის ყველა ქართულ ზმნას და მათ ნაუღლებ ფორმას და ემსახურება როგორც პედაგოგიურ, ასევე სამეცნიერო მიზნებს ქართული ზმნის უღლების რთული და გამონაკლისებით დატვირთული ბუნებიდან გამომდინარე. სხვა ენების ლექსიკოგრაფებისთვის KARTUVERBS წარმოადგენს სემანტიკური ვებ სტანდარტებისა და ტექნოლოგიების საფუძველზე აგებულ ხელსაწყოს, რაც მონაცემებს მანქანის მიერ წაკითხვადს და თავსებადს ხდის ინტერნეტში. პროექტის ფარგლებში შექმნილი პროგრამები/სკრიპტები თავისუფლად არის ხელმისაწვდომი.

საკვანძო სიტყვები: დაკავშირებული მონაცემები, მანქანური სწავლება, გადაწყვეტილების ხე, ქართული ზმნა, ლექსიკონის ლემა