მიგრაციული ქვეკორპუსი ქართული დიალექტების კორპუსში[1]

ლია ბაკურაძე, მარინა ბერიძე, დავით ნადარაია
თსუ არნოლდ ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი, საქართველო
l.bakuradze@gmail.com, marine.beridze@gmail.com, david.nadaraia@gmail.com

ეს მოხსენება არის გაგრძელება სამუშაოსი, რომელიც დაიწყო 2003 წლიდან და რომელიც ვითარდება ერთი დიდი პროექტის – „საქართველოს ლინგვისტური პორტრეტის“ ფარგლებში.

პროექტის სამუშაო პრიორიტეტები ისეთი თანმიმდევრობით დალაგდა, რომ ჯერ შექმნილიყო ქართული დიალექტური სივრცის რეპრეზენტაციული კორპუსული მოდელი და შემდგომ მომხდარიყო ამ მოდელის ცოდნის ინსტრუმენტად ქცევა სხვადასხვა ტექნოლოგიური კომპონენტის, მათ შორის, ლექსიკოგრაფიული რედაქტორისა და კარტოგრაფიული ვიზუალიზაციის საშუალებების დამატებით. სამუშაო პერსპექტივა ასეთი იყო: ქართული დიალექტური სივრცის ტექსტური პორტრეტის შექმნა; ქართული დიალექტური სივრცის ლექსიკოგრაფიული პორტრეტის შექმნა; მე-20 საუკუნის ქართული მიგრაციული მარშრუტების აღწერა (მიგრაციული პორტრეტის შექმნა); არსებული მონაცემების კარტოგრაფიული ვიზუალიზაცია. ამ სამუშაოს შედეგები განუწყვეტლივ ქვეყნდება და ახლდება ქდკ-ის სხვადასხვა კომპონენტის სახით და აისახება ჩვენს პუბლიკაციებში.

ამჯერად წარმოვადგენთ მიმდინარე პროექტს „ენა და კულტურული მეხსიერების დინამიკა XX საუკუნის საქართველოს შიდა მიგრაციების კონტექსტში“, რომელიც ერთ-ერთ აქტივობად გულისხმობს ქდკ-ის მიგრაციული ქვეკორპუსის შექმნას. მე-20 საუკუნის რამდენიმე მძლავრმა მიგრაციულმა ტალღამ მნიშვნელოვნად შეცვალა ქართული დიალექტების გავრცელების სურათი და საგრძნობლად ჩამოაშორა ის მანამდე ცნობილ ისტორიულ-ეთნოგრაფიულად სურათს.

წარმოდგენილი სამუშაოს სიახლე ისაა, რომ ის პირველად უყრის თავს და სამეცნიერო აპარატის დართვით აქვეყნებს მე-20 საუკუნის საქართველოში მიმდინარე შიდა მიგრაციების დოკუმენტურ მასალას, ჩაწერილს პირველი და მეორე თაობის მიგრანტებისგან (დამუშავდება დაახლოებით 50-70 საათი ვიდეოჩანაწერი). ასევე სიახლეა ის, რომ მასალა დამუშავდება კორპუსული სახით და ის ჩაერთვება ქვეკორპუსის სახით ქართული დიალექტების კორპუსში (ქდკ – http://corpora.co); განახლდება ქდკ-ს მიგრაციული ბაზა როგორც ტექნოლოგიურად, ისე ინფორმაციულად – ის მაქსიმალურად მოიცავს საქართველოში მე-20 საუკუნეში განხორციელებულ მიგრაციებს, შექმნის ამ მიგრაციებზე ფაქტობრივი მასალის მოპოვების დამატებით არხებსა და საშუალებებს, სრულყოფს მიგრაციების კარტოგრაფიულ სურათს, შექმნის ინფორმაციული უკუკავშირის ტექნოლოგიურ ინსტრუმენტს ქდკ-ს პლატფორმაზე.

[1] კვლევა განხორციელდა შოთა რუსთაველის საქართველოს ეროვნული სამეცნიერო ფონდის მხარდაჭერით [გრანტის ნომერი FR-21-336].

კორპუსის მეტატექსტური ანოტირების სისტემა და მიგრაციების კარტოგრაფიული ვიზუალიზაციის ბაზა, რომელიც მნიშვნელოვნად განახლდება ამ პროექტის ფარგლებში, დაფუძნებულია მიგრაციების დეტალური კლასიფიკაციის სისტემაზე. მიგრაციათა ძირითად საკლასიფიკაციო მახასიათებლებად გამოვყოფთ: მიგრაციის ტიპს (მასობრივი, ინდივიდუალური); მიგრაციის სახეს (იძულებითი, ეკოლოგიური, საყოფაცხოვრებო), ჩასახლების ტიპი (კომპაქტური, არაკომპაქტური), კომპაქტური ჩასახლების ტიპი (კონტაქტური, არაკონტაქტური), მიგრაციის დრო, მიგრაციის თაობა (თვითონ, მშობლები, მშობლების მშობლები, წინაპრები), მიგრაციის საწყისი და საბოლოო პუნქტები, მიგრირებულთა ეთნიკური და კუთხური წარმომავლობა, დამხვდური მოსახლეობის ეთნიკური და კუთხური წარმომავლობა, მიგრირებულთა სალაპარაკო ენა და დიალექტი, დამხვდური მოსახლეობის სალაპარაკო ენა და დიალექტი.... ამ მახასიათებლებით აღწერილი ტექსტები და მიგრაციული მარშრუტები დამატებით საკვლევ ღირებულებას ანიჭებს გამოქვეყნებულ მასალას როგორც ლინგვისტური, ისე ინტერდისციპლინური თვალსაზრისით.

მიგრაციების შესახებ ინფორმაცია კორპუსში ორი სხვადასხვა მონაცემის ირგვლივ არის სტრუქტურირებული, ესენია: ინფორმაცია მთქმელის შესახებ და ინფორმაცია მასალის ჩაწერის ადგილის შესახებ. ეს უკანასკნელი საფუძვლად უდევს მიგრაციების ბაზას და გულისხმობს გეოგრაფიულ წერტილებთან მრავალდონიანი მიგრაციული ინფორმაციის მიბმას.

პროექტის ფარგლებში იქმნება მიგრაციების ბაზის ახალი ვერსია, რომელიც ჩაეშენება ტექსტის დამატების რედაქტორში. გარდა მიგრაციის ნაკადის აღმწერი ტეგებისა, გეოგრაფიულ წერტილს ენიჭება სხვადასხვა სახის ინფორმაცია, მაგალითად, მიგრანტთა გვარები, დასახლებული პუნქტის მიკროტოპონიმები, ბიბლიოგრაფია კონკრეტული მიგრაციის შესახებ, წყაროები, ხალხური გადმოცემა და სხვ. მიგრაციების ბაზის სამუშაო ინტერფეისი ინფორმაციის შევსებისა და განახლების მუდმივ შესაძლებლობას იძლევა, ხოლო სამომხმარებლო ინტერფეისში შესაძლებელია სხვადასხვა მახასიათებლების კონფიგურირებით ძიება და საძიებო ერთეულების შესახებ ინფორმაციის შესაბამისი კარტოგრაფიული ვიზუალიზაცია.

A Migration Subcorpus within the Georgian Dialect Corpus[1]

Lia Bakuradze, Marina Beridze, Davit Nadaraia
TSU Arnold Chikobava Institute of Linguistics, Georgia
l.bakuradze@gmail.com, marine.beridze@gmail.com, david.nadaraia@gmail.com

This paper will proceed with the activities started in 2003 which have been developed within the framework of a large-scale project "The Georgian Linguistic Portrait".

The operation priorities of the project have been arranged in chronological order so as to elaborate a representative corpus model of the Georgian dialect space and then to transform this model into a knowledge instrument by adding technological components, among them a lexicological editor and cartographic visualization tools. The operation perspective was the following: to create a text portrait of the Georgian dialect space; to describe Georgian human migration routes of the 20th century (to set up a migration portrait); and to visualize available cartographic data. The results of this work have been continuously published and updated as certain components of the Georgian Dialect Corpus and have been reflected in our publications.

Currently, we present a project "Language and Dynamics of Cultural Memory under Migration Conditions in the 20th Century Georgia (LACONIC)". One of the activities of the project is to create a migration subcorpus of the Georgian Dialect Corpus. Several migration waves in the 20th century have substantially changed the overall picture of the spread of dialects and have divided it from the historical-ethnographic picture known up to now.

The novelty of the project is that for the first time by adding a scientific apparatus it collects and publishes document materials of the domestic migrations in Georgia in the 20th century recorded from first and second generation migrants (a 50-70-hour video recording will be examined). In addition, the material will be scrutinized in the form of a corpus and it will be included as a subcorpus within the Georgian Dialect Corpus (http://corpora.co); the migration base will be renewed technologically and with regard to information. It covers to the full the migrations performed in the 20th century, creates additional channels and means for obtaining factual materials on migrations, develops a cartograpic picture of migrations, produces a technological instrument for informative feedback on the platform of the Goergian Dialect Corpus.

The metatextual annotation system of the Corpus and the base of a cartographic vizualization of migrations, which will be substancially renewed within the framework of this project, is based on the detailed classification system of migration. We can offer main classification characteristics of migration: migration type (mass, individual), migration mode (forced, environmental, economic), type of settlement (compact, dispersed), type of compact settlement (contact, non-contact), migration time, migration

[1] This research was supported by Shota Rustaveli National Science Foundation of Georgia [grant number FR-21-336].

generation (themselves, parents, grandparents, ancestors), initial point and endpoint of migration, ethnic and regional origin of migrants, ethnic and regional original of inhabitants, spoken language and dialect of migrants, spoken language and dielect of inhabitants, etc. The texts and human migration routes described according to these features give additional research value to the published material with reference to linguistics and interdisciplinary approaches.

The information about migration is structured in the Corpus around two different data. They are the following: information about the teller and information about the place of recording. The latter is the basis for the migration database and implies the attachment of multistage migration information to certain geographical points.

Within the framework of the project a new version of the database for migrations will be elaborated which will be built in an additional text editor. Along with descriptive tags of migration flows, a geographical point provides different types of information, for example, family names of migrants, microtoponyms of build-up areas, bibliography about a certain migration process, sources, folk stories, etc. The working interface gives us an opportunity to constantly supplement and update information, while it is possible to search in the customer's interface according to different characteristics with a certain configuration. In addition, the customer's interface provided a cartographic visualization relevant to the information about search hits.

ქდკ-ის ონლაინლექსიკონები – საით მივდივართ?[1]

მარინა ბერიძე, ნინო შარაშენიძე, დავით ნადარაია
თსუ არნოლდ ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი, საქართველო
marine.beridze@gmail.com, nino.sharashenidze@tsu.ge, david.nadaraia@gmail.com

„ქართული დიალექტური კორპუსი“ არ არის კლასიკური კორპუსი, ის იქცა მრავალფუნქციურ საცნობარო-საინფორმაციო სისტემად, რომელიც რამდენიმე კომპონენტისგან შედგება. ესენია: საკუთრივ კორპუსი, ტექსტების ბიბლიოთეკა, ქდკ-ის ლექსიკოგრაფიული ბაზა/დიალექტური ონლაინლექსიკონები, მიგრაციების ბაზა. ამასთან, კორპუსში ინტეგრირებულია ტექსტური კორპუსისა და მიგრაციების ბაზის მონაცემების კარტოგრაფიული ვიზუალიზაციის ინტერაქციური სისტემა.

[1] კვლევა განხორციელდა შოთა რუსთაველის საქართველოს ეროვნული სამეცნიერო ფონდის მხარდაჭერით [გრანტის ნომერი HE-21-873].

ლექსიკოგრაფიული რედაქტორი ამ პროდუქტის ერთი მნიშვნელოვანი ტექნოლოგიაა. ეს არის უნივერსალური, ინტერაქტიული, მოქნილი ლექსიკოგრაფიული ინსტრუმენტი კორპუსზე დაყრდნობილი ახალი დიალექტური ლექსიკონების შესაქმნელად. ქდკ-ის ლექსიკოგრაფიული სისტემა წარმოადგენს კორპუსის არქიტექტურაშივე ჩაშენებული ლექსიკონების რედაქტორს – კომპილატორს – ერთგვარ „ლექსიკოგრაფიულ ლაბორატორიას“. ის პლატფორმის დამოუკიდებელი ნაწილია და წარმოადგენს ლექსიკოგრაფიული მასალის შეკრების, დამუშავების და ლექსიკონებად გარდაქმნის ინსტრუმენტს.

ლექსიკოგრაფიული ბაზის ინფორმაციული წყაროებია: გამოქვეყნებული ბეჭდური ლექსიკონების ელექტრონულ ვერსიები; არასალექსიკონო სამეცნიერო რესურსებიდან ამოკრებილი ლექსიკოგრაფიული მასალები; ქართული დიალექტური კორპუსის ტექსტური მასივის ლექსიკა.

ლექსიკოგრაფიულ რედაქტორში პარალელურად მიმდინარეობს სხვადასხვა ტიპის სამუშაო **ქდკ-ის დიდი დიალექტური ონლაინლექსიკონის** შესაქმნელად.

დღეისათვის ქდკ-ის ექვსი ონლაინლექსიკონია მომხმარებლისთვის სრულად ხელმისაწვდომი (თუშური, იმერული, ინგილოური, ფერეიდნული, ჩვენებურების ქართული, ლაზური). ეს ლექსიკონები გამოქვეყნებულია, თუმცა მათი შევსება და ცვლილება განუწყვეტლად მიმდინარეობს. გარდა ამისა, ხელმისაწვდომია სხვა ლექსიკონების სიტყვარი, ე.წ. „მარცხენა მხარე“. ეს ლექსიკონები შეადგენენ ერთიანი ქართველური ლექსიკის ბაზის ნაწილს. ამ ბაზიდან მათი ცალკე ლექსიკონებად „დაგენირირება“ საძიებო სისტემით ხდება.

ქდკ-ის ერთიანი ლექსიკოგრაფიული კონცეფციის მიხედვით მუშავდება და გამოსაქვეყნებლად მზადდება ახალი ონლაინლექსიკონები.

2018 წელს ფრანკფურტის წიგნის ბაზრობისთვის მომზადდა სამი კორპუსული ლექსიკონის ექსპერიმენტული გამოცემა, ესენია: თუშური, იმერული და ფერეიდნული დიალექტების კორპუსული ლექსიკონები (ბერიძე, ბაკურაძე, ნახუცრიშვილი 2018; ბერიძე, შარაშენიძე, კიკნაძე 2018; ცოცანიძე, ბერიძე 2018).

ექსპერიმენტული ლექსიკონების მომზადების მთლიანი ციკლი მიმდინარეობდა ქდკ-ის ლექსიკოგრაფიულ რედაქტორში. ეს ციკლი სხვადასხვა ტიპის სამუშაოს მოიცავდა იმის მიხედვით, თუ რა იყო სალექსიკონო სტატიის ამოსავალი ლექსიკოგრაფიული წყარო: ბეჭდური ლექსიკონი, ლექსიკოგრაფიული კითხვარით მოძიებული/გადამოწმებული მასალა, დიალექტური ტექსტი, სხვადასხვა სამეცნიერო წყარო. ზოგიერთი სალექსიკონო სტატია მხოლოდ ერთი წყაროს მონაცემებზე დაყრდნობით არის შექმნილი, ზოგი კი რამდენიმე წყაროდან გამოკრებილი ინფორმაციითაა გამდიდრებული.

მას შემდეგ, რაც შეიქმნა ჩვენთვის საინტერესო ლექსიკონების ონლაინვერსია, სპეციალურად ამ მიზნით დაწერილი პროგრამული ინსტრუმენტის საშუალებით შეიქმნა სალექსიკონო სტატიის მაკეტი – ლემის, ლემის ლათინური ტრანსკრიპციის, ლემის გრამატიკული მარკერის, ვარიაციებისა და მათი გრამატიკული მარკერების, სიტყვის წარმომავლობისა და უცხო ენაზე დაწერილობის, ილუსტრაციის, ენციკლოპედიური ინფორმაციის, აგრეთვე ლექსიკოგრაფიული ბმულების (იგივეა რაც, იხილე) გადმოცემის სასურველი გრაფიკული „ესკიზი“. ამ პროგრამული

ხელსაწყოს საშუალებით დაგენერირდა ლექსიკონების ბეჭდური პირველსახე, რომელსაც საბოლოო სასტამბო ვერსიის მისაღებად დასჭირდა მხოლოდ შრიფტის შეცვლა, სპეციფიკური ასონიშნების გადამოწმება და ზედაპირული კორექტურა. ამ ექსპერიმენტმა შექმნა ლექსიკოგრაფიული სამუშაოს დისტანციურად ორგანიზების პერსპექტივა.

სამივე ლექსიკონზე დაახლოებით 10 ადამიანი მუშაობდა სხვადასხვა სტატუსით (შემდგენლები, ოპერატორები, კორექტორები, რედაქტორები) და ისინი ხელს არ უშლიდნენ ერთმანეთს.

ქდკ-ის ლექსიკოგრაფიული რედაქტორი გამუდმებით იხვეწება ახალი ამოცანების შესაბამისად. ამჟამად მზადდება ინგილოური ლექსიკონის ბეჭდური ვერსია, რომელიც მოიცავს როგორც უკვე გამოცემული ბეჭდური ლექსიკონების შესახებ ინფორმაციას, ისე ლექსიკურ ერთეულებს ქდკ-ის ახალი ტექსტური კოლექციიდან, აგრეთვე ინგილოური დიალექტის ალიაბათური თქმის ახალ ლექსიკოგრაფიულ მონაცემებს, რომელიც აგრეთვე ჩვენს პროექტში მზადდება.

კუნძულურ დიალექტებს განსაკუთრებული ლექსიკოგრაფიული აღწერა სჭირდება, კერძოდ, დგება შიდადიფერენციაციისა და ქართული ენის სხვა ქვესისტემების (ძველი და ახალი ქართულის, სხვა ქართული დიალექტების) ლექსიკოგრაფიულ მონაცემებთან მიმართების ასახვის საკითხი, ამ ლექსიკონების სასწავლო ფუნქციის გაძლიერების საკითხი (კუნძულური თემის წარმომადგენელმა საკუთარი დიალექტის გამოყენება რომ შეძლოს სასწავლო მიზნებით) და სხვ. შესაბამისად, მიმდინარე პროექტში – „დიალექტური კუნძული ტრანსეთნიკურ არეალში - ქართული ენის ინგილოური დიალექტი აზერბაიჯანში“ – დაგეგმილია ლექსიკოგრაფიული რედაქტორის ახალი ფუნქციონალების შექმნა.

მოხსენებაში წარმოდგენილი იქნება ლექსიკოგრაფიული რედაქტორის ახალი ვერსია. განვიხილავთ მის ბაზაზე ღია და მისაწვდომი პროგრამული უზრუნველყოფის შექმნის საკითხს, რათა ქართულენოვან მომხმარებელს შეეძლოს მისი გამოყენებით ლექსიკოგრაფიული მუშაობის წარმოება. ამ პერსპექტიული პროდუქტის უპირატესობა ანალოგიურ უცხოენოვან და ადგილობრივ ანალოგებთან იქნება ის, რომ მისი სამუშაო ინტერფეისი იქნება ქართულენოვანი, პროდუქტი იქნება ღია და უფასო, სამუშაო ველების კონფიგურირება, კორექტირება იქნება შესაძლებელი და რაც მთავარია, მომხმარებელს მიეწოდება არა მხოლოდ ტექნოლოგიური ჩარჩო, არამედ, ტექსტური და ლექსიკოგრაფიული ბაზების მასალაც საკუთარი კვლევისა თუ ლექსიკოგრაფიული პროდუქტის შესაქმნელად.

Online Dictionaries of the Georgian Dialect Corpus – Where are we going to?[1]

Marina Beridze, Nino Sharashenidze, Davit Nadaraia
TSU Arnold Chikobava Institute of Linguistics, Georgia
marine.beridze@gmail.com, nino.sharashenidze@tsu.ge, david.nadaraia@gmail.com

The Georgian Dialect Corpus is a classical corpus which has been turned into a multifunctional reference and information system consisting of several components: the corpus itself, a library of texts, a lexicographic database/online dictionaries of dialects, and a migration database. Moreover, an interactive system of cartographic visualization of the data of text corpora and a migration base are integrated within the corpus.

One the most important technologies of this product is a lexicographic editor. This is a universal, interactive, and flexible lexicographic instrument based on the corpus to create new dialect dictionaries. The lexicographic system of the Georgian Dialect Corpus represents an editor built into the architecture of the corpus – a compiler – a kind of "lexicographic laboratory". It is an independent part of the platform and it functions as an instrument for the collection, processing and transformation of lexicographic materials into dictionaries.

Information sources for the lexicographic base are the following: electronic versions of printed dictionaries; lexicographic materials collected from non-lexical scientific recourses; textual mass lexes of the Georgian Dialect Corpus.

In parallel, in the lexicographic editor different types of activities are now underway to elaborate **a comprehensive dialect online dictionary**.

So far, six online dictionaries are available for customers (Tushetian, Imeretian, Fereydani, the Chveneburebi Georgian, Laz). These dictionaries are published; however, their amplification and modification are still an ongoing process. In addition, a glossary of dictionaries, so called "the left side", is accessible. The dictionaries are part of the overall Kartvelian lexis base. The search system "generates" them as separate dictionaries from the base.

New online dictionaries are being prepared and elaborated according to the overall lexicographic conception of the Georgian Dialect Corpus.

An experimental edition of three corpus dictionaries was compiled for the 2018 Frankfurt Book Fair: the corpus dictionaries of the Tushetian, Imeretian and Fereydani dialects (Beridze, Bakuradze, Nakhutsrishvili 2018; Beridze, Sharashenidze, Kiknadze 2018; Tsotsanidze, Beridze 2018).

The whole cycle of preparing the experimental dictionaries was implemented in the lexicographic editor of the Georgian Dialect Corpus. This cycle comprised different types of work depending on the starting lexicographic source of an entry: the printed dictionary, the material searched/checked by a

[1] This research was supported by Shota Rustaveli National Science Foundation of Georgia (SRNSFG) [grant number HE-21-873].

lexicographic questionnaire, dialect texts, different scientific sources, etc. Some entries are created on the basis of the data from only one source, while some of them are enriched by means of the information extracted from several sources.

Since the online version of dictionaries which is of great interest to us was elaborated, an entry model has been created by means of a programme instrument designed specifically for this purpose – the relevant graphic "lay-out" of representing lemmas, Latin transcription of lemmas, grammatical markers of lemmas, variations and their grammatical markers, etymology of a word and its spelling in a foreign language, illustrations, encyclopedic information, and lexicographic links (for example, see…). The printed prototype of the dictionaries has been generated by means of this programme instrument which only needed to change a font in order to get the final printing version, to double-check specific character-signs and proofread. This experiment has created the perspective of organizing the lexicographic work remotely.

Approximately 10 people worked on the three dictionaries having different responsibilities (collectors, operators, proofreaders, editors) without hindering each other.

The lexicographic editor of the Georgian Dialect Corpus has been continuously refining in compliance with new objectives. Currently, a printed version of the Ingiloian dictionary is being prepared, covering the information about the dictionaries that have been already printed as well as the lexical units from the textual base of the Georgian Dialect Corpus, and new lexicographic data of the Aliabati accent of the Ingiloian dialect which are also being prepared within our project.

Insular dialects require a particular lexicographic description, namely an issue of internal differentiation reflecting the Georgian language with respect to lexicographic data of other subsystems (Old Georgian, New Georgian, other Georgian dialects) and strengthening the educational function of these dictionaries (so that insular community would be able to use the vernacular for educational purposes), etc. Correspondingly, in the ongoing project – "Georgian Dialect Island in the Transethnic Area – Ingiloian Dialect in Azerbaijan" – we plan to create new functionals of the lexicographic editor which we present in this paper.

The paper will present a new version of the lexicographic editor. Moreover, it will examine an aspect of elaboration of open and accessible programme software created on the basis of the editor in order to enable Georgian speaking users to implement lexicographic activities. The advantage of this promising product compared to other foreign and local analogues will be its Georgian working interface. In addition, the product will be open and free; it will be possible to configure and alter work fields; users will be given not only the technological framework but also the materials from the textual and lexicographic databases in order to carry out their own researches and create lexicographic products.

მეგრული ენის ანოტირებული ზეპირი კორპუსის თეორიული და პრაქტიკული ჩარჩო

რუსუდან გერსამია, ირინა ლობჟანიძე, თამუნა სხულუხია, ნინო წულაია
ილიას სახელმწიფო უნივერსიტეტი, საქართველო
rgersamia@iliauni.edu.ge, irina_lobzhanidze@iliauni.edu.ge,
tamuna.skhulukhia.2@iliauni.edu.ge, nino.tsulaia.1@iliauni.edu.ge

1. შესავალი

მეგრული უმწერლობო ქართველური ენაა, რომელიც გავრცელებულია დასავლეთ საქართველოს კოლხეთის დაბლობზე. მეგრული ტექსტების დოკუმენტაცია მე-19 საუკუნის ბოლოდან დაიწყო და უშუალოდ ამ პერიოდიდან შესაძლებელი გახდა იმ ენობრივ და სოციოკულტურულ ცვლილებებზე დაკვირვება, რომლებსაც მეგრული ენა ორ საუკუნეზე მეტი ხნის განმავლობაში განიცდიდა. ბოლო ოცდაათწლიანი ცვლილებების საფუძველს გლობალიზაცია და მიგრაციული პროცესები წარმოადგენს. შედეგად ბევრი მეგრული ტრადიცია და რიტუალი შეიცვალა და მიეცა დავიწყებას, ქცევის ძველ წესებს საზოგადოება დღეს არ იზიარებს; თაობათა შორის ცოდნის გადაცემაც, გარკვეულწილად, პრობლემურია; ახალგაზრდა თაობის ლექსიკა გაღარიბებულია და ენობრივი კონსტრუქციები მიისწრაფვის გამარტივებისკენ.

შესაბამისად, მეგრული ტექსტების დიგიტალიზაცია და შესაბამისი ანოტაცია გადაუდებელ ამოცანას წარმოადგენს. ამავდროულად, მნიშვნელოვანია ხანდაზმულთა მეტყველების კონსერვაციის უზრუნველყოფა, ამასთან ინფორმაციის მოპოვება თანამედროვე ლინგვისტური სიტუაციისა და ენობრივი ცვლილებების შესახებ. მეგრული სამეტყველო კორპუსის შედგენის პროექტი მოიცავს შემდეგ ამოცანებს:

1. მეგრული ზეპირი ტექსტების კონსერვაცია და მორფოლოგიურად ანოტირებული კორპუსის შედგენა, რომელშიც წარმოდგენილი იქნება საველე ექსპედიციების დროს მოპოვებული ინფორმაცია მეგრელების ლინგვისტური და სოციოკულტურული სიტუაციის შესახებ;
2. აღწერითი გრამატიკის გენერირება და ლექსიკონის შედგენა Fieldwork Language Explorer-ის (FLEx) საშუალებით.

მეგრული ტექსტების ბეჭდურად გამოცემა იწყება 1880 წელს ალექსანდრე ცაგარლის „მეგრული ენის ეტიუდებით“, მას შემდეგ სხვადასხვა დროს გამოიცა არაერთი ტექსტი; კორპუსის კოლექციაში შევა დღემდე გამოცემული ტექსტები, მათ შორის მკვლევართათვის ცნობილი, თუმცა ნაკლებად გამოყენებული, ტექსტები წიგნებიდან: Мингрельская азбука (1899, Тифлис), რობერტ ბლაიხშტაინერი „Georgesche und Mingrelische Texte“ (ვენა, 1919). ცხადია, ტექსტებში აისახა

მეგრული ენის ფლობის დონე და საზოგადოების დამოკიდებულება თავიანთი ტრადიციების მიმართ. შესაბამისად, კორპუსში წარმოსადგენი მასალა, მათ შორის, საველე ექსპედიციების დროს მოპოვებული ახალი მასალა (ტექსტი, აუდიო/ვიდეო) სამეგრელოს ბოლოდროინდელი სოციო-კულტურული და სოციოლინგვისტური მდგომარეობის ასახვასაც ემსახურება.

2. საველე მუშაობა

საველე სამუშაოები მოიცავს მასალის მოპოვებას სამეგრელოს ექვსი ადმინისტრაციული ერთეულის სოფლებსა და ცენტრში, ასევე აფხაზეთიდან დევნილთა კომპაქტურ დასახლებებში. ჩაიწერება არა მხოლოდ ხანდაზმულთა, არამედ ახლაგაზრდების მეტყველებაც, თემატურად აღიწერება ის მასალა, რომელიც ასახავს ენობრივი და კულტურული ცოდნის გადაცემის საკითხებს.

ცნობილია, რომ დილმანი (Dillman 1978; 2000) განასხვავებს კითხვების ხუთ თემატურ ტიპს: *ქცევას, შეხედულებას, ცოდნას, დამოკიდებულებებს და ატრიბუტებს.* საველე სამუშაოების დროს მონაცემთა შეგროვების, ამოღების და დაკვირვების ეტაპებზე დაყრდნობით აღიწერება ზემოთ აღნიშნული ხუთი ტიპი:

- ქცევა: ენობრივ-კულტურული სიტუაცია (მაგ. ნათესაობის აღმნიშვნელი ფორმები, რიტუალები, მისალმება/გამომშვიდობება და სხვ.);
- შეხედულებები: მეგრელთა აზრი მათი ენობრივი სამყაროსა და სხვა ენების გავლენის შესახებ;
- ცოდნა: ენის ათვისება, რომელიც უკავშირდება უფროსებსა და ბავშვებს შორის ცოდნის გადაცემას;
- დამოკიდებულება: მეგრელთა დამოკიდებულება საკუთარი ენისა და მისი ენობრივი თავისებურებებისადმი;
- ატრიბუტები: ინფორმაცია ინფორმატორების შესახებ (მაგ. სქესი, ასაკი და ა.შ.).

3. წინასწარ შედეგი

სამუშაოს შედეგები განთავსდება ონლაინრესურსის სახით. ამ მომენტში საწყისი ეტაპი მოიცავს კონვერტორის მომზადებას ტექსტების ტრანსკრიფციისა და ტრანსლიტერაციის უზრუნველსაყოფად. უკვე შემუშავებულია კონვერტაციის სამი ძირითადი მიმართულება (გერსამია, ლობჟანიძე 2022). პირველი ტიპის მიმართულება გულისხმობს ქართულ დამწერლობაზე დაფუძნებული მეგრულიდან საერთაშორისო ფონეტიკურ ანბანზე (IPA) გადასვლას, მეორე – მეგრულიდან ISO 9984 რომანიზაციის სქემაზე და მესამე გამყრელიძე-მაჭავარიანის ტრანსლიტერაციის სისტემაზე (1965 წ.) და პირუკუ.

4. დასკვნები და მომავალი გამოწვევები

პროექტის შედეგები მოიცავს მეგრულის ანოტირებულ სამეტყველო / ზეპირ კორპუსს, რომელიც ხელმისაწვდომი იქნება ონლაინრეჟიმში. ამ პროექტისთვის სპეციალურად შემუშავებული მორფოსინტაქსური ანოტაცია დაედება საფუძვლად აღწერით გრამატიკისა და ონლაინლექსიკონს.

ანოტირება განხორციელდება ენის დოკუმენტირების ხელმისაწვდომი ინსტრუმენტების გამოყენებით, კერძოდ, FLEx გამოყენებული იქნება აღწერითი გრამატიკისა და ლექსიკონის გენერაციისათვის, ხოლო ELAN აუდიო და ვიდეოფაილების ანოტირებისათვის ლაიფციგის გლოსირების წესების და ევროტიპის სპეციფიკაციების შესაბამისად.

პროექტის შედეგები შეიძლება გამოიყენებოდეს როგორც სასწავლო რესურსი სხვადასხვა საგანმანათლებლო პროგრამაში, რომელიც მოემსახურება უმწერლობო ქართველური ენების სწავლებას უმაღლესი განათლების სხვადასხვა საფეხურზე, სხვადასხვა სახელმძღვანელოებისა და გრამატიკული წესების მოსამზადებლად. პროექტის შედეგები ასევე მნიშვნელოვანი იქნება უმწერლობო ქართველური ენების პოპულარიზაციისთვის.

ლიტერატურა

Bakker, Dik, König, Ekkehard, Dahl, Östen, Haspelmath, Martin, Koptjevskaja-Tamm, Maria, Lehmann, Christian, Siewierska, Anna. 1993. Eurotyp Guidelines. European Science Foundation in Language Typology

Bleichsteiner, Robert. 1919. *Kaukasische Forschungen: t. Georgische und mingrelische Texte*. Volume 1 of Osten und Orient Osten und Orient, 1. Reihe, 1. Bd. Wien: Forschungsinstitut für Osten und Orient

Comrie, Bernard, Haspelmath, Martin, Bickel, Balthasar. 2008. *The Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-morphene Glosses.* Max Planck Institute for Evolutionary Anthropology

Dillman, Don A. 1978. Mail and Telephone Surveys: The Total Design Method, 1978. John Wiley: New York

Dillman, Don A. 2000. Internet and Mail Surveys: The Tailored Design Method, 2000. John Wiley: New York

გამყრელიძე, თამაზი, მაჭავარიანი, გივი. 1965. *სონანტთა სისტემა და აბლაუტი ქართველურ ენებში. თბილისი: მეცნიერება*

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press

ლობჟანიძე, ირინა, გერსამია, რუსუდანი. 2022. *მეგრული ტექსტის კონვერტორი* https://irinalobzhanidze.com/megrelian/converter/converter.html . წვდომა 15 ივნისი, 2022

Standardization, ISO. 1996. *Information and documentation — Transliteration of Georgian characters into Latin characters, No 9984.* https://www.iso.org/standard/17892.html. Accessed 15 June, 2022

William Goodrich (1979) Dillman, Don. *Mail and Telephone Surveys-the Total Design Method*. New York: John Wiley & Sons, 1978, Journal of Advertising, 8:1, 52

Theoretical and practical framework of the Spoken Megrelian Corpus

Rusudan Gersamia, Irina Lobzhanidze, Tamuna Skhulukhia, Nino Tsulaia
Ilia State University, Georgia
rgersamia@iliauni.edu.ge, irina_lobzhanidze@iliauni.edu.ge,
tamuna.skhulukhia.2@iliauni.edu.ge, nino.tsulaia.1@iliauni.edu.ge

1. Introduction

Megrelian is an unwritten language of the Kartvelian group of languages, spoken in the Kolkheti lowlands of the western Georgia. The documentation of Megrelian texts dates back to the end of the 19th century, and since this period it is possible to follow the linguistic and sociocultural changes that the Megrelian language underwent for more than two centuries. The globalization and migration processes can be considered as the basis for the changes in the last 30 years. As a result, a lot of Megrelian traditions and rituals have been changed and forgotten, behavior rules shared by the society in the past are not used today; the generational knowledge transmission is somehow problematic; the vocabulary of the young generation is impoverished and strives to simplify the linguistic constructions.

As a result, the digitization of the Megrelian texts and their appropriate annotation can be considered an urgent task. At the same time, it is important to provide conservation of the speech elderly speakers and to get information on the modern linguistic situation paying special attention to language change. As a result, the project on the compilation of Megrelian spoken corpus includes the following:

1. Conservation of Megrelian spoken texts and compilation of a morphologically annotated corpus, which encompass material with regards to the recent and sociocultural situation of Megrelain obtained through the linguistic fieldwork;
2. Generation of sketch grammar and compilation of a dictionary by means of Fieldwork Language Explorer (FLEx).

The first Megrelian published text compiled by Alexander Tsagareli dates back to 1880. The collection includes all published texts, including those from the following books: Megrelian alphabet (Tiflis, 1899), Georgesche und Mingrelische Texte by Robert Bleichsteiner (Vienna, 1919) and authorized publications, which reflect the proficiency of the Megrelian language and the attitude of the Megrelian community to their own traditions. New material (text, audio / video) collected during the fieldwork will be added to the collection as well with the purpose to show the recent sociolinguistic situation of Samegrelo.

2. Field work

Fieldwork includes obtaining material in the villages and centres of six administrative units in Samegrelo and in the compact settlements of IDPs from Abkhazia. The data will be collected not only from

elderly, but also from young informants to deal with the language and cultural transmission issues.

Dillman (1978; 2000) differentiates between five types of question content: behavior, beliefs, knowledge, attitudes and attributes. The implementation of fieldworks, data collecting, elicitation and observation stages will rely on the description of the above-mentioned five types:

1. Behavior: Language-cultural situation (e.g., forms of kinship, rituals, greetings / farewells, etc.);
2. Beliefs: Opinion of the Mingrelian community about their linguistic world and influence of other languages;
3. Knowledge: Language acquisition related to the generational knowledge transmission between adults and children;
4. Attitudes: Attitude of the Megrelian community to their own language and its linguistic features;
5. Attributes: Information about the informants (e.g. gender, age etc.).

3. Preliminary results

The results of the work done will be implemented in the form of the online resource. At this moment the initial stage encompasses preparation of a converter to provide the transcription and transliteration of texts. Three main points of conversion have been already developed (Gersamia et al. 2022). The first type encompasses bi-directional conversion from Georgian-script based Megrelian to International Phonetic Alphabet (IPA) and vice versa, the second and the third types provide transliteration from Megrelian to ISO 9984 romanization scheme and to Gamkrelidze and Machavariani's system of transliteration (1965) and vice versa.

4. Conclusions and future challenges

The results of the project will encompass appropriately annotated spoken corpus of Megrelian available online. The morphosyntactic annotation specially developed for this project will be considered as a base for the sketch grammar and online dictionary.

The annotation will be carried out by means of tools already available worldwide with regards to language documentation, especially, FLEx will be used for the generation of sketch grammar and dictionary and ELAN for the annotation of audio and video files in accordance with the Leipzig Glossing Rules and Eurotype specifications.

The results of the project can be used as a learning resource for different educational programs focusing on teaching unwritten Kartvelian languages at different levels of higher education, for preparation of different workbooks and grammar sketches. The results of the project are also important for the popularization of unwritten Kartvelian languages.

References

Bakker, Dik, König, Ekkehard, Dahl, Östen, Haspelmath, Martin, Koptjevskaja-Tamm, Maria, Lehmann, Christian, Siewierska, Anna. 1993. Eurotyp Guidelines. European Science Foundation in Language Typology

Bleichsteiner, Robert. 1919. *Kaukasische Forschungen: t. Georgische und mingrelische Texte*. Volume 1 of Osten und Orient Osten und Orient, 1. Reihe, 1. Bd. Wien: Forschungsinstitut für Osten und Orient

Comrie, Bernard, Haspelmath, Martin, Bickel, Balthasar. 2008. *The Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-morphene Glosses.* Max Planck Institute for Evolutionary Anthropology

Dillman, Don A. 1978. Mail and Telephone Surveys: The Total Design Method, 1978. John Wiley: New York

Dillman, Don A. 2000. Internet and Mail Surveys: The Tailored Design Method, 2000. John Wiley: New York

Gamkrelidze, Thomas, Machavariani, Givi. 1965. *The system of sonants and ablaut in the Kartvelian languages*. Tbilisi: Science

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press

Lobzhanidze, Irina, Gersamia, Rusudan. 2022. *Mingrelian Converter.* https://irinalobzhanidze.com/megrelian/converter/converter.html . Accessed 15 June, 2022

Standardization, ISO. 1996. *Information and documentation — Transliteration of Georgian characters into Latin characters, No 9984.* https://www.iso.org/standard/17892.html. Accessed 15 June, 2022

William Goodrich (1979) Dillman, Don. *Mail and Telephone Surveys-the Total Design Method*. New York: John Wiley & Sons, 1978, Journal of Advertising, 8:1, 52

პროგრამა „ლექსიკოგრაფი“

ქეთევან დათუკიშვილი, ნანა ლოლაძე, მერაბ ზაკალაშვილი
ლინგვისტური ტექნოლოგიების ჯგუფი, საქართველო
datukishvili510@gmail.com, nana.loladze@tsu.ge, merabza@gmail.com

ლექსიკოგრაფიაში სულ უფრო აქტიურად გამოიყენება ციფრული ტექნოლოგიები: იქმნება სხვადასხვა სახის პროგრამული ინსტრუმენტები, რომლებიც უზრუნველყოფს ლექსიკოგრაფიული სამუშაოების ეფექტურად წარმართვას.

წარმოგიდგენთ ერთ-ერთ ამგვარ ინსტრუმენტს – პროგრამა „ლექსიკოგრაფს“. მისი საშუალებით შესაძლებელია როგორც ბეჭდური, ისე ელექტრონული ლექსიკონების მომზადება და გამოცემა. ამ პროგრამის მეშვეობით შეიქმნა და გამოიცა „ქართული ლექსიკონი“ ჯერ ბეჭდური (2014 წ.) და შემდეგ – ელექტრონული სახით (2019 წ.) (https://www.ganmarteba.ge/). ეს უკანასკნელი დღეისათვის მოიცავს 41432 სალექსიკონო ერთეულს. გრძელდება მისი შევსება ახალი მასალით.

პროგრამის მონაცემთა ბაზის სტრუქტურას საფუძვლად დაედო კონცეპტუალური მოდელი, რომელიც მოიცავს შემდეგ კომპონენტებს: განმარტება, ილუსტრაცია, გრამატიკული ინფორმაცია, სხვადასხვა კლასიფიკატორი (ფუნქციონირების სფერო, დარგი და მისთ.) და ა. შ. კომპონენტთა გარკვეული ნაწილი ერთიანდება ცალკე სტრუქტურულ ერთეულებში, რომლებსაც სექციები ვუწოდეთ. გვაქვს შემდეგი სექციები: დასაწყისი, ძირითადი და ფრაზები. ელექტრონული გამოცემისთვის ცალკე სექციებია ბმულები, თითოეულ სექციას აქვს თავისი კომპონენტები, რომლებიც განთავსებულია ცალ-ცალკე ველებში.

მონაცემთა ბაზაში მონაცემების შეყვანა ხდება ბაზის რედაქტორის საშუალებით, სიტყვა-სტატიის შესაბამისი ველების მიხედვით. ამგვარად სტრუქტურირებული მასალა იძლევა ინფორმაციის ავტომატურად მოძიებისა და მართვის საშუალებას, რაც აადვილებს ლექსიკონზე მუშაობის პროცესს. გარდა ამისა, შესაძლებელია მონაცემთა სტატისტიკური კვლევა, რაც მნიშვნელოვანია სამეცნიერო სამუშაოებისთვის.

ფორმატირების რედაქტორის მეშვეობით ხორციელდება ლექსიკონის გამოსაცემად მომზადება, კერძოდ: ინფორმაციის დალაგება გარკვეული მიმდევრობით, სხვადასხვა სექციასა თუ ველში განთავსებული ინფორმაციისათვის შესაბამისი სტილის (შრიფტის ზომა, ფერი...) შერჩევა ან სიმბოლოების გამოყენება და ა. შ.

პროგრამა „ლექსიკოგრაფი“ გარკვეული მოდიფიკაციით შესაძლოა გამოვიყენოთ ნებისმიერი ტიპის (განმარტებითი, ორთოგრაფიული, თარგმნითი და სხვ.) ლექსიკონის შესაქმნელად.

The Program “Lexicographer”

Ketevan Datukishvili, Nana Loladze, Merab Zakalashvili
Linguistic Technologies Group, Georgia
datukishvili510@gmail.com, nana.loladze@tsu.ge, merabza@gmail.com

Digital technologies are widely used in lexicography: numerous program tools are created, ensuring efficient lexicographic activities.

One of such tools is the program “Lexicographer” which enables creation and publication of both printed and electronic dictionaries. This tool has yielded “The Georgian Dictionary”, which was first issued in the printed form (in 2014) and, later, in the electronic form (in 2019) (https://www.ganma-rteba.ge/). Currently, the Dictionary embraces 41432 units and is still enriched with new material.

The database structure was built upon a conceptual model, consisting of the following components: definition, illustration, grammatical information, various classifiers (usage, field of functioning etc). Some components are united under separate structural units called sections. There are the following sections: the starting section, the key section and phrases. For the electronic version, there are additional sections like links. Each section has its components, given in separate fields.

The data are added to the database by means of database editor according to corresponding fields of the dictionary units. The material structured in this way enables automatic search and management of information. All this facilitates both compilation and usage of the Dictionary. Besides, it is possible to carry out statistical analysis of the data, which is of crucial importance for scientific research.

Preparation of the Dictionary for publication is possible due to formatting editor, namely: the information is provided in a certain order; appropriate style (font size, colour…) is selected for each section and field, different symbols are used and so on.

With certain modification, the program “Lexicographer“ can be used for the creation of dictionaries of any type (explanatory, orthographic, bilingual and so on).

ტექნოლოგიაზე დაფუძნებული სწავლება ინტერკულტურულ გარემოში: თანამშრომლობა საზღვრების გარეშე კვლევისა და განათლების სფეროში შვედეთს, უკრაინასა და საქართველოს შორის (TELICORE)

პროექტის პრეზენტაცია

დმიტრი დობროვოლსკი, ტორა ჰედინი, ლუდმილა პეპელი, ნატალია რინგბლუმი
სტოკჰოლმის უნივერსიტეტი, შვედეთი
dm-dbrv@yandex.ru, tora.hedin@slav.su.se, ludmila.poppel@slav.su.se, natasha.ringblom@slav.su.se

ლადა კოლომიეცი
კიევის ტარას შევჩენკოს სახელობის ეროვნული უნივერსიტეტი, უკრაინა
ladakolomiyets@gmail.com

მიხაილ კოპოტევი
ჰელსინკის უნივერსიტეტი, ფინეთი
mihail.kopotev@helsinki.fi

თინათინ მარგალიტაძე
ილიას სახელმწიფო უნივერსიტეტი, საქართველო
tinatin.margalitadze@iliauni.edu.ge

აკადემიური მობილობა და თანამშრომლობა ყოველთვის მნიშვნელოვანი საკითხები იყო უმაღლესი განათლების დარგში. ბოლოდროინდელმა მოვლენებმა და შეზღუდვებმა არა მხოლოდ ზეგავლენა მოახდინა საერთაშორისო თანამშრომლობაზე, არამედ აგრეთვე აჩვენა, რომ არსებობს მოთხოვნა უნივერსიტეტებს შორის თანამშრომლობის ახალ სისტემებზე და ახალ ფორმებზე, განსაკუთრებით კი გაზრდილ თანამშრომლობაზე ქვეყნებს შორის განათლებისა და კვლევის სფეროში არსებული გამოცდილების გაზიარების მიზნით.

პროექტის საერთო მიზანი იყო კვლევისა და სწავლების სფეროში თანამშრომლობის განვითარება სტოკჰოლმის უნივერსიტეტს, ეკონომიკის უმაღლეს სკოლასა (სანქტ-პეტერბურგი, რუსეთი) და ილიას სახელმწიფო უნივერსიტეტს (თბილისი, საქართველო) შორის. უკრაინაში რუსეთის შეჭრის შემდეგ თანამშრომლობა ეკონომიკის უმაღლეს სკოლასთან შეჩერებულია. ამჟამად პროექტი მუშაობას განაგრძობს ორ ახალ პარტნიორთან – კიევის ტარას შევჩენკოს სახელობის ეროვნულ უნივერსიტეტთან (უკრაინა) და ჰელსინკის უნივერსიტეტთან (ფინეთი) თანამშრომლობით. პროექტის ფარგლებში ჩვენ ვაგროვებთ პარტნიორთა სასწავლო დაწესებულებებში არსებულ პედაგოგიურ და საგნობრივ კომპეტენციას თეორიისა და მეთოდოლოგიის სფეროში და ვიღვწით კვლევისა და სწავლების ერთმანეთთან უფრო მჭიდროდ დასაკავშირებლად.

პროექტს აქვს შემდეგი ქვემიზნები:

(1) სხვადასხვა უნივერსიტეტებში მომუშავე მკვლევართა შორის თანამშრომლობის გაღრმავება და კვლევის ეთიკის სფეროში არსებული გამოწვევებისთვის პასუხის გაცემა – როგორც კვლევების, ასევე ესეების წერის მხრივ, და, როგორც მოსალოდნელი შედეგი, ახალი საერთაშორისო კვლევითი პროექტების შემუშავება.

(2) სამაგისტრო და სადოქტორო დონეებზე გაცვლითი პროგრამების დანერგვა სტოკჰოლმის უნივერსიტეტის სტუდენტებსა და პარტნიორი ქვეყნების უნივერსიტეტთა სტუდენტებს შორის.

(3) ერთობლივი კურსების დაარსება კორპუსზე დაფუძნებული სწავლისა და კვლევის მეთოდოლოგიისა და ეთიკის დარგში.

პროექტი განსაკუთრებით გამოყოფს კორპუსზე დაფუძნებული სწავლის მეთოდოლოგიას, რაც წამყვანი თემაა თანამედროვე გამოყენებით ენათმეცნიერებაში. ერთ-ერთი მიზანია საერთაშორისო გამოცდილების დაგროვება კორპუსული ანალიზური კვლევისა და ლექსიკოლოგიის სფეროში და ენის შესწავლის მექანიზმების შემუშავება. ტექნოლოგიაზე დაფუძნებული სწავლის (TEL) კონცეფცია ინფორმაციულ და საკომუნიკაციო ტექნოლოგიათა (ICT) მეთოდოლოგიებთან და მექანიზმებთან ერთად სწავლის ხელმისაწვდომ და მიმზიდველ შესაძლებლობებს ქმნის. ამ კონცეფციას ეფუძნება აგრეთვე ამჟამად მიმდინარე ონლაინკურსი „კორპუსული ლინგვისტიკა ენის სწავლისა და კვლევისათვის“. იგი განკუთვნილია საბაკალავრო, სამაგისტრო და სადოქტორო საფეხურის სტუდენტებისთვის, რომლებიც სპეციალიზდებიან ენათმეცნიერებაში და ტექნოლოგიაზე დაფუძნებულ სწავლებაში (TEL) და მიზნად ისახავს სამაგისტრო და სადოქტორო დონეებზე გაცვლითი პროგრამების დანერგვას სტოკჰოლმის უნივერსიტეტის სტუდენტებსა და პარტნიორი ქვეყნების, უკრაინისა და საქართველოს, უნივერსიტეტთა სტუდენტებს შორის. კურსი დაიწყო ონლაინ-სწავლების, -სწავლის, -ინტერაქციისა და -თანამშრომლობისთვის განკუთვნილი სასარგებლო ვირტუალური მექანიზმების დანერგვით. იგი განსაკუთრებულ ყურადღებას ამახვილებს ისეთ საკითხებზე, როგორებიცაა კრიტიკული აზროვნება აკადემიურ წერაში, კორპუსზე დაფუძნებული კვლევის მეთოდები, ტექსტური კორპუსებისა და კორპუსზე დაფუძნებული ინსტრუმენტების გამოყენება ენათა სწავლებისას; კორპუსების გამოყენება ლინგვისტურ კვლევაში, თარგმანთმცოდნეობაში, პოლიტიკური დისკურსის ანალიზში *და ა.შ.*

TELICORE: ინტერნეტ-მისამართი https://www.su.se/english/research/research-projects/telicore-technology-enabled-learning-tel-in-intercultural-environment

Technology Enabled Learning in Intercultural Environment: Cross-Border Cooperation and Exchange Between Sweden, Ukraine and Georgia in Research and Education (TELICORE)

Presentation of the Project

Dmitrij Dobrovolskij, Tora Hedin, Ludmila Pöppel, Natalia Ringblom
Stockholm University, Sweden
dm-dbrv@yandex.ru, tora.hedin@slav.su.se, ludmila.poppel@slav.su.se, natasha.ringblom@slav.su.se
Lada Kolomiets
Taras Shevchenko National University of Kyiv, Ukraine
ladakolomiyets@gmail.com
Mihail Kopotev
University of Helsinki, Finland
mihail.kopotev@helsinki.fi
Tinatin Margalitadze
Ilia State University, Georgia
tinatin.margalitadze@iliauni.edu.ge

Academic mobility and education exchange have always been important issues in higher education. Recent events and restrictions have not only affected the international cooperation but also shown that there is a demand on new networks and new forms of cooperation between universities, especially increased cross-border cooperation in order to share experiences in education and research.

The overall purpose of the project was to develop a research and teaching collaboration between Stockholm University, Higher School of Economics (St. Petersburg, Russia) and Ilia State University (Tbilisi, Georgia). After Russia's invasion of Ukraine cooperation with Higher School of Economics was suspended. At present the project continues its work in cooperation with two new partners – Taras Shevchenko National University of Kyiv (Ukraine) and the University of Helsinki (Finland). Within the project, we gather the pedagogical and subject competence in theory and the methodology that exists at partners' institutions and aim to establish a closer connection between research and teaching.

The sub-goals of the project are as follows:

(1) To deepen the collaboration between researchers from different universities and address challenges in research ethics – both in research and in essay writing, and as an anticipated result, to develop new international research projects.

(2) To develop exchange at the undergraduate and advanced levels between students at

Stockholm University and students in the partner countries.

(3) To develop joint courses in corpus-based learning and research methodology and ethics.

The project highlights a corpus-based learning methodology, which is a mainstream in modern applied linguistics. One of the aims is to gather international expertise in corpus analytic research and lexicology and develop tools for language learning. A Technology Enabled Learning (TEL) concept with Information and Communication Technology (ICT) methodologies and tools provides accessible and attractive learning opportunities. This concept is also the basis of the ongoing online course "Corpus linguistics for language learning and research". It addresses students on undergraduate, graduate and postgraduate levels majoring in Linguistics and TEL and aims to develop exchange at the undergraduate and advanced levels between students at Stockholm University and students in the partner countries of Ukraine and Georgia. The course started with introducing useful virtual tools for online teaching, learning, interaction and cooperation. It has particular focus on such issues as critical thinking in academic writing, methods of corpus-based research, using text corpora and corpus-based tools for teaching languages; using corpora in linguistic research, translation studies, political discourse analysis, etc.

TELICORE: web address https://www.su.se/english/research/research-projects/telicore-technology-enabled-learning-tel-in-intercultural-environment

ანდიურის საველე ჩანაწერების კორპუსი (წინასწარი შედეგები)

სამირა ვერჰეესი
დამოუკიდებელი მკვლევარი, ნიდერლანდები
jh.verhees@gmail.com
აიგულ ზაკიროვა, გიორგი მოროზი, ელენა სოკური
ეკონომიკის უმაღლესი სკოლა, რუსეთის ფედერაცია
aigul.n.zakirova@gmail.com, agricolamz@gmail.com, elena.o.sokur@gmail.com

მოხსენებაში ვისაუბრებთ სხვადასხვა მეთოდზე, რომლებსაც დღეს ვიყენებთ მცირე რესურსების მქონე იმ ენის კორპუსის შესაქმნელად, რომელიც არ არის აღწერილი. აგრეთვე, წარმოვადგენთ, ანდიური ენის კორპუსის საპილოტე ვერსიას Tsakorpus-ის პლატფორმაზე (Arkhangelskiy 2019).

ანდიური [ანდიური 1255] არის უმწერლობო ენა, რომელიც მიეკუთვნება აღმოსავლეთ კავკასიურ ენათა ოჯახს. ანდიური ენის დიალექტები შეიძლება დავყოთ ორ ჯგუფად: ზემო ანდიური (საუბრობენ სოფლებში: ანდი, აშალი, რიქვანი, ღაღათლი, ზილო, გუნხი და ჩანყო) და

ქვემო ანდიური (საუბრობენ კვანხიდათლისა და მუნიში). ზოგიერთი ნაირსახეობისთვის არსებობს გრამატიკული აღწერები, მაგალითად, სოფელ ანდის ნაირსახეობის აღწერები ეკუთვნის დირს (1906), კიბრიკსა და კოძასოვს (1990), რიქვანისა – სულეიმანოვს (1957). მრავალი დიალექტი აღწერა ცერცვაძემ (1965), ხოლო ღაღათლის აღწერა ეკუთვნის სალიმოვს (2010 (1968)). სხვა, კერძოდ, ქვემო ანდიურის ნაირსახეობები, ფაქტობრივად, აღუწერელია. გამონაკლისია მხოლოდ ცერცვაძის მიერ მოპოვებული მწირი მონაცემები (1965). მიუხედავად იმისა, რომ ზოგი ლექსიკური მასალა ანდიურის შესახებ ალისულთანოვას დისერტაციაში გხვდება და სიტყვების სიები სულეიმანოვმა (1957), კიბრიკმა და კოძასოვმა (1990) და სალიმოვმა მოამზადეს (2010), ანდიურის ლექსიკონი ჯერ არ გამოცემულა. მიმდინარე პროექტი არის რამდენიმე მკვლევრის მცდელობა, შეკრიბოს ის აუდიომასალები, რომლებიც 2015 წლიდან დღემდე სოფელ რიქვანში, ზილოში, მუნისა და კვანხიდათლიში ლინგვისტიკური საველე მოგზაურობების დროს ჩაიწერა.

მდინარე ვოლგის რეგიონში სხვა უფრო სკრუპულოზურად აღწერილი ენისთვის – მდელოს მარიულისთვის (http://lingconlab.ru/spoken_meadow_mari/search) – გამოყენებული ალგორითმით კორპუსი შემდეგნაირად შეიქმნა: მოხდა ტექსტისა და ბგერის იმპორტირება ELAN-ში, დანაწევრება, შეთანადება და კორექტირება. შემდეგ კორპუსის ტექსტების დიდი ნაწილისთვის გამოყენებული იყო მორფოლოგიური ანალიზატორი (Arkhangelskiy 2021). შედეგები გაანალიზდა, ხოლო ანალიზატორი შეიცვალა და გაუმჯობესდა. ანალიზატორის ახალი ვერსია შეტანილ იქნა ELAN-ის ფაილებში. შედეგები ხელით შემოწმდა და არასწორი წაიშალა.

მიუხედავად ამისა, ანდიური ენის კორპუსის შექმნის საქმეს ხელს უშლის დიალექტების სტატუსი, რომ ისინი აღწერილი არ არის და არ არსებობს დიალექტთაშორისი სტანდარტი, სრულყოფილი ლექსიკონი და გრამატიკული ანალიზატორი (პირველი მცდელობა ბუნტიანოვას ეკუთვნის (2020). წავაწყდით შემდეგ პრობლემებს: 1) მონაცემები არაერთგვაროვანია იმის მიხედვით, თუ როგორ ინახება და რამდენ დიალექტურ სახესხვაობას ასახავს; 2) ანოტირება ამ ეტაპზე ხელით უნდა გაკეთდეს; 3) რადგან ჩვენი ცოდნა ანდიური ენის დიალექტების შესახებ შეზღუდულია, ხანდახან არ ვიცით მოცემული სიტყვის ფორმის ზუსტი ანალიზი.

ვინაიდან ანდიური ენის ჩანაწერები მკვლევრების სხვადასხვა ჯგუფის მიერ გაკეთდა, ხოლო ტრანსკრიბირება და ანოტირება სხვა ადამიანებს ეკუთვნის (ზოგ შემთხვევაში მკვლევრები, ენის მატარებლები), მასალა სხვადასხვა ფორმატში ინახება: .TextGrid (Praat), .eaf (ELAN), ასევე, .docx და .txt ფორმატებში. მასალა უნდა გადავიყვანოთ ერთ ფორმატში *phonfieldwork*-ის გამოყენებით (Moroz 2020). ანდიური ენის დიალექტური კორპუსისთვის პროცესი ამგვარია: წინასწარ ვამუშავებთ ანოტირების ფაილებს, გადაგვყავს .eaf-ფორმატში, ვათანადებთ და გლოსირებას ხელით ვაკეთებთ. შემდეგ ფაილი Tsakorpus-ის პლატფორმის გამოყენებით ქვეყნდება ონლაინ (Arkhangelskiy 2019), რაც მარტივი გზაა იმისთვის, რომ ELAN-ის ფაილები ონლაინკორპუსის ინტერფეისად გადავაქციოთ.

ვინაიდან საქმე გვაქვს სხვადასხვა დიალექტთან (სხვადასხვა მკვლევართან), შეიქმნა მორფოლოგიური ანოტირების უნიფიცირებული და შედარებით მოქნილი სისტემა, მაგალითად, ერთი და იგივე გლოსა მივანიჭეთ ფორმალურად განსხვავებულ მორფემებს, რომლებსაც დიალექტურ შესაბამისობებად განვიხილავთ. თუ მორფოლოგიურ ანალიზში დარწმუნებული არ

ვიქნებით, ცალკე ველში მის მონიშვნას ვგეგმავთ, რათა ძიებისას ანოტირების „საეჭვო“ სტატუსი ადვილად ხელმისაწვდომი იყოს.

ლიტერატურა

Alisultanova, Mesedu A. 2010. *Leksika andijskogo jazyka* [Lexicon of Andi]. DNC RAN, Institut jazyka, literatury i iskusstva im. Cadasy. Doctoral dissertation.

Arkhangelskiy, Timofey. 2019. Corpora of social media in minority Uralic languages.

Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages.

125–140. Tartu, Estonia. http://volgakama.webcorpora.net/ Social_media_ corpora_IWCLUL2019_final.pdf (02.07.2022)

Arkhangelskiy, Timofey. 2021. Meadow Mari morphological analyzer (GitHub repository).

https://github.com/timarkh/uniparser-grammar-meadow-mari

Buntyakova, Valeria A. 2022. Sozdanie morfologičeskogo parsera dla andijskogo jazyka v

sisteme lexd i twol [Morphological Parser of Andi in lexd and twol] Cercvadze, Ilia I. 1965. *Andiuri ena* [The Andi language]. Tbilisi: Metsniereba.

Dirr, Adolf M. 1906. Kratkij grammatičeskij očerk andijskago jazyka [A brief sketch of Andi].

Sbornik materialov dlja opisanija mestnostej i plemën Kavkaza 36: Otdel IV. Tiflis:

Kavkazskij Učebnyj Okrug.

Kibrik, Aleksandr E., Kodzasov Sandro V. 1990. *Sopostavitelnoe izučenie dagestanskix jazykov.*

Imja, fonetika [A comparative study of the Daghestanian languages: nouns, phonetics].

Moscow: Izdatel'stvo MGU.

Moroz, George. 2020. *Phonetic fieldwork and experiments with phonfieldwork package*. https://CRAN.R-project.org/package=phonfieldwork, 02.07.2022.

Salimov, Xangerej S. 2010 (1968). *Gagatlinskij govor andijskogo jazyka* [The Gagatli dialect of

Andi]. Makhachkala: Institut jazyka, literatury i iskusstva im. G. Cadasy.

Sulejmanov, Jakov G. 1957. *Grammatičeskij očerk andijskogo jazyka. Na materiale govora s.*

Rikvani [A sketch of the grammar of Andi. Based on material from the dialect of the village Rikvani]. Moscow: Institut jazykoznanija akademij nauk sojuza SSR. Phd thesis.

A Corpus of Andi Field Recordings (Preliminary Results)

Samira Verhees
Independent Researcher, Netherlands
jh.verhees@gmail.com
Aigul Zakirova, George Moroz, Elena Sokur
HSE University, Russian Federation
aigul.n.zakirova@gmail.com, agricolamz@gmail.com, elena.o.sokur@gmail.com

In this talk we will discuss some methods we currently use and are developing to create a corpus of a lower resourced and underdescribed language. We will also present a pilot version of the Andi corpus on the Tsakorpus platform (Arkhangelskiy 2019).

Andi [andi1255] is an unwritten language of the East Caucasian language family. Andi dialects can be divided into two groups: Upper Andi (spoken in the villages of Andi, Ashali, Rikvani, Gagatli, Zilo, Gunkha and Chanko) and Lower Andi (spoken in Kvankhidatli and Muni). For some of these varieties grammatical descriptions exist, e.g. Dirr (1906) and Kibrik & Kodzasov (1990) for Andi; Sulejmanov (1957) for Rikvani; Cercvadze (1965) for various dialects; Salimov (2010 (1968)) for Gagatli. Others, in particular the Lower varieties, remain virtually undescribed, except for some scarce data provided in Cercvadze (1965). Although some lexical materials on Andi can be found is a dissertation by Alisultanova (2009), and wordlists are provided by Sulejmanov (1957), Kibrik & Kodzasov (1990), Salimov (2010), no dictionary of Andi has been published yet. The current project is an attempt of several researchers to bring together the audio materials that have been recorded in the course of linguistic field trips from 2015 to the present to the villages of Rikvani, Zilo, Muni and Kvankhidatli.

For a different and more thoroughly described language spoken in the Volga region, Meadow Mari (http://lingconlab.ru/spoken_meadow_mari/search), the algorythm used to build the corpus was as follows: the text and the sound were imported to ELAN, segmented, aligned and proofread. Then a morphological parser (Arkhangelskiy 2021) was used on the whole bulk of corpus texts. The results were analyzed and the parser was modified and improved. The new version of the parser was applied to the ELAN files. The resulting analyses were checked manually and the incorrect ones were deleted.

However, for Andi the task of building a corpus is hindered by the underdescribed status of Andi dialects and by the absence of a cross-dialectal standard, a full-fledged dictionary and a grammatical parser (though see a first attempt in Buntyakova (2022)). We face the following problems: 1) the data is heterogeneous both in how it is stored and in how much dialectal variation it features; 2) the annotation needs to be done manually at this stage; 3) due to our limited knowledge of the Andi dialects, sometimes we do not know what the correct analysis of a given word form is.

Since the Andi recordings were made by different groups of researchers, transcribed and annotated by different people (sometimes researches, sometimes native speakers), the material is stored in different file

formats: .TextGrid (Praat), .eaf (ELAN), and even .docx and .txt. The material has to be converted to a singular format using *phonfieldwork* (Moroz 2020). For the Andi dialectal corpus the pipeline is as follows: we preprocess the annotation files, converting them to .eaf, align them with the sound, and then we gloss them manually for now. After that the file is published online using the Tsakorpus platform (Arkhangelskiy 2019), an easy way to transform ELAN files into an online corpus interface.

As we deal with several different dialects (and different researchers), a unified and relatively flexible system of morphological annotation has been developed. For example, we have assigned the same gloss to formally different morphemes that we analyze as dialectal correspondences. When we are not sure about the morphological analysis, we plan to mark it in a separate tier, so that the "dubious" status of some annotations is accessible through search.

References

Alisultanova, Mesedu A. 2010. *Leksika andijskogo jazyka* [Lexicon of Andi]. DNC RAN, Institut jazyka, literatury i iskusstva im. Cadasy. Doctoral dissertation.

Arkhangelskiy, Timofey. 2019. Corpora of social media in minority Uralic languages.

Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages.

125–140. Tartu, Estonia. http://volgakama.webcorpora.net/ Social_media_ corpora_IWCLUL2019_final.pdf (02.07.2022)

Arkhangelskiy, Timofey. 2021. Meadow Mari morphological analyzer (GitHub repository).

https://github.com/timarkh/uniparser-grammar-meadow-mari

Buntyakova, Valeria A. 2022. Sozdanie morfologičeskogo parsera dla andijskogo jazyka v

sisteme lexd i twol [Morphological Parser of Andi in lexd and twol] Cercvadze, Ilia I. 1965. *Andiuri ena* [The Andi language]. Tbilisi: Metsniereba.

Dirr, Adolf M. 1906. Kratkij grammatičeskij očerk andijskago jazyka [A brief sketch of Andi].

Sbornik materialov dlja opisanija mestnostej i plemën Kavkaza 36: Otdel IV. Tiflis:

Kavkazskij Učebnyj Okrug.

Kibrik, Aleksandr E., Kodzasov Sandro V. 1990. *Sopostavitelnoe izučenie dagestanskix jazykov.*

Imja, fonetika [A comparative study of the Daghestanian languages: nouns, phonetics].

Moscow: Izdatel'stvo MGU.

Moroz, George. 2020. *Phonetic fieldwork and experiments with phonfieldwork package*. https://CRAN.R-project.org/package=phonfieldwork, 02.07.2022.

Salimov, Xangerej S. 2010 (1968). *Gagatlinskij govor andijskogo jazyka* [The Gagatli dialect of

Andi]. Makhachkala: Institut jazyka, literatury i iskusstva im. G. Cadasy.

Sulejmanov, Jakov G. 1957. *Grammatičeskij očerk andijskogo jazyka. Na materiale govora s. Rikvani* [A sketch of the grammar of Andi. Based on material from the dialect of the village Rikvani]. Moscow: Institut jazykoznanija akademij nauk sojuza SSR. Phd thesis.

შეთანადების (ალინირების) კონცეპტუალიზაციისათვის „ვეფხისტყაოსნის“ თარგმანების მრავალენოვან პარალელურ კორპუსში

მანანა თანდაშვილი
ემპირიული ენათმეცნიერების ინსტიტუტი
ფრანკფურტის გოეთეს უნივერსიტეტი, გერმანია
tandaschwili@em.uni-frankfurt.de

ჰუმანიტარული დარგების განვითარება 21-ე საუკუნეში წარმოუდგენელია დიდი მონაცემთა ბაზებისა და კვლევის თანამედროვე მეთოდების განვითარების გარეშე. ქართველოლოგიამ, როგორც ეროვნულმა მეცნიერებამ, ღირსეულად შეაბიჯა მესამე ათასწლეულში: დღეისათვის არსებული დიგიტალური რესურსებისა და ქართული ენისათვის განვითარებული ტექნოლოგიების ბაზაზე საფუძველი ჩაეყარა ქართველოლოგიის თვისებრივად ახალ ეტაპს – **დიგიტალურ ქართველოლოგიას**. შოთა რუსთაველის „ვეფხისტყაოსნის“ ხელნაწერების კორპუსისა (http://corpora.iliauni.edu.ge/?q =ka/node/11) და პოემის თარგმანების დიგიტალური კორპუსის **Rustaveli goes digital** (https://titus.uni-frankfurt.de/texte/caucasica/rustveli/rgd.html) შექმნით კი ჩვენ თვალწინ მიმდინარეობს **დიგიტალური რუსთველოლოგიის** ჩამოყალიბება, რომელიც სცდება ეროვნული მეცნიერების ფარგლებს: შოთა რუსთაველის მიერ მე-12 საუკუნეში შექმნილი პოემა მსოფლიოს 56 ენაზე არის თარგმნილი და უნიკალურ რესურსს წარმოადგენს როგორც მრავალენოვანი დიგიტალური კორპუსის შესაქმნელად, ისე ინტერდისციპლინური კვლევების განსახორციელებლად.

„ვეფხისტყაოსნის“ თარგმანების პარალელური კორპუსი **Rustaveli goes digital** დღეისათვის მოიცავს პოემის სრული ტექსტის დაპარალელებულ ვერსიას მსოფლიოს 20 ენაზე და საშუალებას იძლევა ემპირიულ ბაზაზე დაყრდნობით თანამედროვე ტექნოლოგიების გამოყენებით ვიკვლიოთ თარგმანმცოდნეობის ისეთი პრობლემური საკითხი, როგორიცაა ეკვივალენტობის პრობლემა თარგმანში. აღნიშნული საკითხის კორპუსლინგვისტური დამუშავებისათვის კი აუცილებელია თარგმანის სტრატეგიების ავტომატური კვლევის თეორიული და ტექნოლოგიური ჩარჩოს შექმნა.

წინამდებარე მოხსენებაში სადისკუსიოდ წარმოდგენილია დიგიტალური რუსთველოლოგიის ერთ-ერთი მნიშვნელოვანი ამოცანის – **შეთანადების (ალინირების) კონცეპტუალური ჩარჩო-პროგრამის** კონცეფცია.

შეთანადება კორპუსლინგვისტური ოპერაციაა, პარალელური კორპუსების აგების პროცესში ერთ-ერთ უმნიშვნელოვანეს და აუცილებელ ოპერაციას წარმოადგენს და გულისხმობს ორი (ან მეტი) ერთგვაროვანი ტექსტის ურთიერთდაკავშირებას სინტაქსურ (წინადადების, როგორც სინტაქსური ოდენობის, დაპარალელება), მორფოსინტაქსურ (ფრაზებად სეგმენტირებული წინადადების, როგორც წინადადების შემადგენლების, დაპარალელება) ან ლექსიკურ (სიტყვების, როგორც ლექსიკური ოდენობების, დაპარალელება) დონეზე. შესაბამისად, პარალელურ კორპუსებში გამოყენებულია შეთანადების სამი სახე: წინადადებების შეთანადება, ფრაზების შეთანადება და სიტყვების შეთანადება. განვიხილოთ თითოეული ცალ-ცალკე „ვეფხისტყაოსნის“ თარგმანების პარალელური კორპუსის მაგალითზე.

შეთანადების პირველ ეტაპზე „ვეფხისტყაოსნის“ თარგმანების პარალელურ კორპუსში პოემის თარგმანები ორიგინალ ტექსტთან სტრუქტურულად, სტროფების დონეზე დაპარალელდა, თუმცა შეთანადების ეს ეტაპი არ იძლევა წინადადებების ავტომატური დაპარალელების საშუალებას: მართალია, სტროფების წინადადებებად დაშლა ორიგინალ ტექსტში ძირითადად ემთხვევა სტრიქონებად დაშლას (სტრიქონები, როგორც წესი, წინადადებას ან წინადადების სინტაქსურად სეგმენტირებად ერთეულს – მთავარ ან დამოკიდებულ წინადადებას შეესაბამება), მაგრამ თარგმანებში ეს პრინციპი, როგორც ეს პოემის გერმანული, ინგლისური, რუსული და ესპანური თარგმანების სტრუქტურულმა შედარებამ გვიჩვენა, ყოველთვის არ არის დაცული. წყარო-ტექსტში ერთ სტრიქონში მოცემული ერთი წინადადება ხშირად განვრცობილია და „გადანაწილებულია“ ორ ან რამდენიმე სტრიქონში. ამ თვალსაზრისით განსაკუთრებით გამოირჩევიან ე.წ. თავისუფალი თარგმანები, როგორიცაა მაგალითად, კონსტანტინე ბალმონტის რუსული თარგმანი. შდრ.: **თხუთმეტისა წლისა ვიყავ, მეფე მზრდიდა ვითა შვილსა** *vs.* ***Мне пятнадцать лет уж было. Сердце было полно пыла. / Воля царская взрастила как царевича меня.*** მეორე თავისებურება, რომლითაც ბალმონტის თარგმანი გამოირჩევა სხვა თარგმანებისაგან, არის სტრიქონებს შორის „გადანაწილებული“ წინადადებები. მაგ.: Исполняя приказанье, вот рабы идут. **Шуршанье / Слышно ног, звенит бряцанье их доспехов**. Витязь встал.

წინადადებების შეთანადებას ეკვივალენტობის თვალსაზრისით ის უპირატესობა აქვს, რომ იგი საშუალებას გვაძლევს სამიზნე ენად მარტივად „აღმოვაჩინოთ“ შინაარსობრივი განვრცობები, როგორც ამას, მაგალითად, ლინ კოფინის ინგლისურ თარგმანში ვხვდებით: ა) ერთხელ ესე თქვა „ვა მეო“, სხვად არას მოუბარია / He said aloud only, "Woe is me, {**that I should find myself here**}", ბ) რა ცნა, მეფე მოვიდაო, ჰკრა მათრახი მისსა ცხენსა / When he saw the king, he struck his horse; {**what came next was strange but true**}. მსგავსი შემთხვევები სხვა თარგმანებშიც დასტურდება, მაგ., მარტინეცის ესპანურ თარგმანში: ა) სადამდისცა დღენი მესხნენ, ვეღარამან გამახარნეს! *vs.* nada podrá alegrarme hasta el fin de mi vida, nadie podrá consolarme: **que el mundo llore por mí**" *(ვერაფერი გამაბედნიერებს სიცოცხლის ბოლომდე, ვერავინ დამამშვიდებს:* ***დაე, სამყარომ დამიტიროს****)*

ფრაზების შეთანადება, წინადადებების შეთანადებისაგან განსხვავებით, საშუალებას იძლევა ორიგინალი და თარგმნილი ტექსტი მორფოსინტაქსურ დონეზე „ჩავშალოთ“, ვინაიდან ფრაზებად დაშლილი სინტაქსური კონსტრუქცია არსებითად შეესაბამება სინტაქსურ შემადგენლებს წინადადებაში (სუბიექტი, ობიექტი, გარემოება და ა.შ.). ფრაზების სტრუქტურული ანალიზი კი ფრაზების სტრუქტურული მოდიფიცირების პოვნიერებას აადვილებს. შდრ.: **შავი ცხენი** *vs.* **schwarz** wie Nacht, sein edles **Roß**. ხშირ შემთხვევაში ასეთი განვრცობები **პერიფრაზირებული ატრიბუტების** მეშვეობით გადმოიცემა. შდრ.:

მოჰგვარა **მონა გრძნეული**, შავი მართ ვითა ყორანი;
Trajo al **esclavo hechicero**, negro como un cuervo (დე ლა ტორეს ესპანური თარგმანი)
Был доставлен **раб**, как ворон, **тайны ведает он рока**. (ნუცუბიძის რუსული თარგმანი)

ფრაზებად შეთანადების პროცესში ყურადღება მიიქცია თარგმანებში სახელური ფრაზების სტრუქტურულმა მსგავსებამ, რომელიც არსებითად განსხვავდება შესაბამისი ფრაზისაგან წყარო ტექსტში. შდრ.:

შესხდეს [**მეფე და ავთანდილ**] მის ყმისა მისაწეველად.
[Avtandil y el rey] *se lanzan tras el valiente con rapidez, (მარტინეცის ესპანური თარგმანში*
[Avtandil y el rey] *se lanzan como centellas en pos del Caballero, (ბარეას ესპანური თარგმანი)*
На конях вдогонку мчатся ***[Автандил и царь Ростан]****. (ნუცუბიძის რუსული თარგმანი)*
Da bestiegen ihre Rosse ***[Awthandil und Rostewan]****. (ჰუპერტის გერმანული თარგმანი)*

მსგავსი შემთხვევების კვანტიტატიური ანალიზი თარგმანებში განსაკუთრებით მნიშვნელოვანია წყარო ტექსტის დადგენის თვალსაზრისითაც, ვინაიდან ცალსახად მიუთითებს თარგმნის პროცესში გამოყენებულ შუალედურ წყაროტექსტზე.

შეთანადების მესამე სახე – **სიტყვების შეთანადება** – ეკვივალენტობის სახეების კვლევის თვალსაზრისით ყველაზე მრავალფეროვანია. **ვერნერ კოლერის** ეკვივალენტობის თეორიაზე დაყრდნობით, სიტყვების შეთანადების კვალიფიკაციისას ვიყენებდით ეკვივალენტების ხუთ სახეს: დენოტაციურ, კონოტაციურ, პრაგმატულ, ტექსტ-ნორმატიულ და ექსპრესიულ ეკვივალენტებს. თუმცა, პროტოტიპულ კორპუსში სიტყვების შეთანადებამ გვიჩვენა, რომ ის ლექსიკური ერთეულები, რომლებიც ტექსტში გამოყენებულია, როგორც მხატვრული საშუალება, განსხვავებულ მიდგომას მოითხოვენ. მაგ., ლექსემა „მზე“ პოემაში გამოიყენება სამი ფუნქციით: 1) როგორც ასტროლოგიური სხეული, 2) როგორც მხატვრული საშუალება (მეტაფორა, შედარება, ჰიპერბოლა) და 3) როგორც ფიცილის ფორმულის კომპონენტი. შდრ.: *ა) მიმავალი ცასა შესტირს, ეუბნების, ეტყვის* ***მზესა:*** *„აჰა,* ***მზეო,*** *გეაჯები შენ, უმძლესთა მძლეთა მძლესა,“ ბ) მას* ***მზესა*** *ტანსა ემოს-ნეს ნარინჯის-ფერნი ჯუბანი. გ)* ***შენმან მზემან,*** *თავი ჩემი არვის ჰმართებს უშენოსა!*

შესაბამისად, შეთანადების დროს, ლექსემების ტეგირებისას, შევიმუშავეთ საგანგებო მეთოდი იმ ტიპის ლექსიკურ ერთეულებთან მიმართებით, რომლებიც ტექსტში გამოიყენებიან როგორც ლექსიკური, ისე ფუნქციური რეფერენციის გადმოსაცემად.

თარგმანების სტროფული შეთანადების პროცესში გამოიკვეთა ის პრობლემები, რომლებიც განსაკუთრებით თავისუფალი თარგმანების შეთანადებას ახლავს თან და სრულიად გამორიცხავს არა მარტო სიტყვების ან ფრაზების შეთანადებას, არამედ ორიგინალი და თარგმნილი ტექსტის შეთანადებას წინადადებების (და არც თუ იშვიათად, სტროფულ) დონეზეც კი. შდრ. ბალმონტის რუსული თარგმანი:

იყო **არაბეთს როსტევან**, **მეფე ღმრთისაგან** სვიანი,
მაღალი, უხვი, მდაბალი, ლაშქარ-მრავალი, ყმიანი,
მოსამართლე და მოწყალე, **მორჭმული**, განგებიანი,
თვით მეომარი უებრო, კვლა მოუბარი წყლიანი.

Был в **Арабии** певучей **царь** от **бога**, царь **могучий**,
Рати сильного — как тучи, вознесенный **Ростеван**.
Многим витязям бессменный знак и образ несравненный,
Птицезоркий, в зыби пенной всё увидит сквозь туман.

ასეთ შემთხვევაში დამატებით ვიყენებთ **მერი სნელ-ჰორნბის** თეორიას ფრეიმებისა და სცენების სემანტიკის შესახებ და ვახორციელებთ კონცეპტების შინაარსობრივ დაპარალელებას - ვადგენთ კონცეპტების ეკვივალენტობას, რომლის დროსაც ეკვივალენტური შეტყობინებები განსხვავებული კოდებით გადმოიცემა. ამჟამად მიმდინარეობს ორიგინალი და თარგმნილი ტექსტების შედარება ფრეიმებისა და სცენების სემანტიკური ეკვივალენტობის პარამეტრების დადგენის მიზნით.

On the Conceptualization of Alignment in the Multilingual Parallel Corpus of the Translations of “The Knight in the Panther’s Skin”

Manana Tandaschwili
Institute for Empirical Linguistics
Goethe University Frankfurt, Germany
tandaschwili@em.uni-frankfurt.de

The development of the humanities in the 21st century is unthinkable without the development of large databases and digital research methods. Kartvelology, as a national scholarship, entered the third millennium with dignity: on the basis of existing digital resources and technologies developed for the Georgian language, the foundation was laid for a qualitatively new stage of Kartvelology that may be called

Digital Kartvelology. With the creation of the manuscript corpus of Shota Rustaveli's "The Knight in the Panther's Skin" (http://corpora.iliauni.edu.ge/?q =ka/node/11) and the digital corpus of the epic's translations, "Rustaveli goes digital" (https://titus.uni-frankfurt.de/texte/caucasica/rustveli/rgd.html), we are witnessing the establishment of Digital Rustvelology, which goes beyond the scope of a national scholarship: the epic created by Shota Rustaveli in the 12th century has been translated into 56 languages and thus represents a unique resource for creating a multilingual digital parallel corpus and conducting transdisciplinary research.

The parallel corpus of "The Knight in the Panther's Skin" translations, "Rustaveli goes digital", currently contains a parallel version of the full text of the poem in 20 different languages of the world and allows us to explore such problematic issues of translation studies as the problem of equivalence in translation, using modern technology on an empirical basis. For the corpus linguistic process of the mentioned issue, it is necessary to create a theoretical and technological framework for automated research with a wide range of possibilities for examining various levels of translation equivalence. Below we present **the conceptual framework of alignment for solving** this important task of Digital Rustvelology.

Alignment is one of the most important and necessary operations in the process of building parallel corpora. Alignment allows two (or more) identical translation techniques to be compared at the sentence, phrase, or word level. Accordingly, three types of alignment are distinguished: a) **alignment of sentences**, b) **alignment of phrases**, c) **alignment of words**. For the prototype corpus, a method of aligning phrases was selected, allowing the equivalence to be explored at the phrase level.

In the first stage of alignment, the translations of the poem in the parallel corpus of "The Knight in the Panther's Skin" were structurally aligned with the original text at the level of stanzas, although this stage of alignment does not allow automatic parallelization of sentences: it is true that the division of stanzas into sentences in the original text basically coincides with the division into lines (lines, as a rule, correspond with a sentence or a syntactically segmentable unit of a sentence – a main or a dependent clause), but in translations this principle, as shown by a structural comparison of the German, English, Russian and Spanish translations of the poem, is not always maintained. A single sentence in a single line in the source text is often expanded and "split" into two or more lines. From this point of view, the so-called *free translations*, such as the Russian translation by Konstantin Balmont, are particularly distinguished, cf. **თხუთმეტისა წლისა ვიყავ, მეფე მზრდიდა ვითა შვილსა vs. Мне пятнадцать лет уж было**. Сердце было полно пыла. / **Воля царская взрастила как царевича меня.** The second feature that distinguishes Balmont's translation from other translations is the "fragmentation" of sentences between the lines, e.g. Исполняя приказанье, вот рабы идут. **Шуршанье / Слышно ног, звенит бряцанье их доспехов.** Витязь встал.

The advantage of **aligning sentences** from the viewpoint of equivalence is that it allows us to easily "discover" extensions in terms of content in the target language, as we find, for example, in Lynn Coffin's English translation: a) ერთხელ ესე თქვა „ვა მეო", სხვად არას მოუბარია / He said aloud only, "Woe is me, {**that I should find myself here**}", b) რა ცნა, მეფე მოვიდაო, ჰკრა მათრახი მისსა ცხენსა / When he saw the king, he struck his horse; {**what came next was strange but true**}. Similar cases are also confirmed

in other translations, for example, in the Spanish translation of Martinez: a) სადამდისცა დღენი მესხნენ, ვეღარამან გამახარნეს! vs. nada podrá alegrarme hasta el fin de mi vida, nadie podrá consolarme: **que el mundo llore por mí**" (*ვერაფერი გამაბედნიერებს სიცოცხლის ბოლომდე, ვერავინ დამამშვიდებს: **დაე, სამყარომ დამიტიროს***).

Phrase alignment, unlike sentence alignment, allows the original and translated text to be "decomposed" at the morphosyntactic level, since the syntactic construction decomposed into phrases essentially corresponds to the syntactic constituents in the sentence (subject, object, setting, etc.). The structural analysis of phrases makes it easier to find structural modifiers of phrases, cf. **შავი ცხენი** vs. **schwarz** wie Nacht, sein edles **Roß**. In many cases, such expressions are conveyed through paraphrased attributes, cf.:

მოჰგვარა **მონა გრძნეული**, შავი მართ ვითა ყორანი;
Trajo al **esclavo hechicero**, negro como un cuervo (de la Torre's Spanish translation)
Был доставлен **раб**, как ворон, **тайны ведает он рока**. (Nutsubidze's Russian translation)

In the process of alignment into phrases, the structural similarity of the noun phrases in the translations need to be spotlighted, which are essentially different from the corresponding phrase in the source text, cf.:

*შესხდეს **[მეფე და ავთანდილ]** მის ყმისა მისაწეველად.*
***[Avtandil y el rey]** se lanzan tras el valiente con rapidez,* (Martinez's Spanish translation)
***[Avtandil y el rey]** se lanzan como centellas en pos del Caballero,* (Barea's Spanish translation)
*На конях вдогонку мчатся **[Автандил и царь Ростан]**.* (Nutsubidze's Russian translation)
*Da bestiegen ihre Rosse **[Awthandil und Rostewan]**.* (Huppert's German translation)

Quantitative analysis of similar cases in translations is especially for determining the source text and thus the language, since it unambiguously indicates the source text used in the translation process.

The third type of alignment– **alignment of words** – is the most diverse in terms of the research of equivalence types. Based on Werner Koller's theory of equivalence, we used five types of equivalents when qualifying word alignment: denotative, connotative, pragmatic, textual, and expressive equivalents. However, the alignment of words in the prototypical corpus showed us that the lexical units used in the text as stylistic devices require a different approach. For example, the lexeme "მზე" is used in the poem with three functions: 1) as an astrological term, 2) as a stylistic device (metaphor, simile, hyperbole), and 3) as a component of a fictitious formula. Therefore, during alignment, when lexemes are selected, we have developed a special method of tagging in relation to the types of lexical units that will be used in the text to convey both lexical and functional reference. In the process of strophic alignment of translations, the problems that accompany the alignment of free translations and completely exclude not only the alignment of words or phrases but also the alignment of the original and the translated text even at the level of sentences (and not rarely, strophes) were highlighted. cf. Balmont's Russian translation:

იყო **არაბეთს როსტევან, მეფე ღმრთისაგან** სვიანი,
მაღალი, უხვი, მდაბალი, ლაშქარ-მრავალი, ყმიანი,

მოსამართლე და მოწყალე, **მორჭმული**, განგებიანი,
თვით მეომარი უებრო, კვლა მოუბარი წყლიანი.

Был в **Арабии** певучей **царь** от **бога**, царь **могучий**,
Рати сильного — как тучи, вознесенный **Ростеван**.
Многим витязям бессменный знак и образ несравненный,
Птицезоркий, в зыби пенной всё увидит сквозь туман.

In such cases, we additionally use **Mary Snell-Hornby's** theory about the scenes-and-frames semantics and carry out the content parallelism of concepts – we establish the equivalence of concepts, during which equivalent messages are conveyed by different codes. Currently, the original and translated texts are being compared in order to determine the parameters of semantic equivalence of scenes-and-frames.

კორპუსზე დაფუძნებული ენის პედაგოგიკა (უცხოური გამოცდილება და საქართველო)

ნათია კენჭიაშვილი
ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი, საქართველო
natia.kentchiashvili@gmail.com

წინამდებარე კვლევა მიზნად ისახავს განვიხილოთ ენის სწავლისა და სწავლების კორპუსზე დაფუძნებული მეთოდი. კვლევამ მოიცვა კორპუსზე დაფუძნებული ენის პედაგოგიკის როგორც საერთაშორისო გამოცდილება, ისე საქართველოში მისი განვითარების პერსპექტივები. ჩვენი კვლევის ფარგლებში თეორიული განხილვის საგანია კორპუსი არა როგორც კვლევის ინსტრუმენტი, არამედ როგორც ენის სწავლა-სწავლებისთვის ძირითადი და/ან დამხმარე რესურსი. ამდენად, კორპუსზე დაფუძნებული კვლევის მეთოდის გარდა, განიხილება კორპუსზე დაფუძნებული სწავლების მეთოდიც.

იმის გათვალისწინებით, რომ ინტერნეტსივრცეში დაგროვდა სხვადასხვა სახის ქართულენოვანი კორპუსი, მნიშვნელოვანია ამ ტიპის ელექტრონული რესურსის განხილვა სწორედ ქართული ენის სწავლა-სწავლების კონტექსტში. მით უფრო, რომ სხვა ენებისთვის შექმნილი კორპუსები აქტიურად გამოიყენება საგანმანათლებლო მიზნებისათვის. ამიტომ, კვლევის ფარგლებში, განსაკუთრებული ყურადღება უნდა გამახვილდეს ქართულენოვანი კორპუსების მიმოხილვასა და ამ მიმართულებით მათი გამოყენების მნიშვნელობაზე. არაერთი კვლევა ადასტურებს იმას, რომ ენის სწავლისა და სწავლების კორპუსზე დაფუძნებული მეთოდი ეფექტური და პროდუქტი-

ულია; მოსწავლეებს/სტუდენტებს აქვთ უშუალო წვდომა ენობრივ მონაცემებზე, რითაც ავითარებენ თავიანთ ენობრივ უნარ-ჩვევებს. განხილულმა კვლევებმა აჩვენა ისიც, თუ როგორ შეიძლება კორპუსზე დაფუძნებულმა მეთოდი იყოს ინტეგრირებული ენის სწავლისა და სწავლების სფეროში, მოსწავლეთა კვალიფიკაციის დონის მიუხედავად, განსახილველი მეთოდი გამოსადეგია როგორც დამწყებთათვის, ისე მასწავლებლებისთვის.

კორპუსის მეშვეობით შესაძლებელია სხვადასხვა სახის ენობრივი მასალის სწრაფად და ეფექტურად მოძიება, ასევე მასწავლებლებს ეძლევათ საშუალება, მოამზადონ და შეავსონ სასწავლო მასალები საინტერესო და ავთენტიკური მაგალითებით სხვადასხვა დროის, ავტორის, დარგისა თუ ჟანრის ქართულენოვანი ტექსტებიდან. კორპუსულ რესურსებსა და ინსტრუმენტებს შეუძლიათ მოგვაწოდონ ინფორმაცია, თუ რომელი სიტყვა / ფრაზა როგორ და რამდენად ხშირად გამოიყენება, გვაჩვენოს მათი გამოყენების სიხშირე, სად დასტურდება მათი გამოყენების შემთხვევები, რომელი სიტყვების გარემოცვაში დასტურდება კონკრეტული სიტყვა, რა კონტექსტში გვხვდება და რა მიზნებისთვის; თუმცა კორპუსს არ შეუძლია ცალსახა პასუხების გაცემა კითხვებზე, არამედ, ამის ნაცვლად, საჭიროა კორპუსის მეშვეობით მოძიებული ინფორმაციის შეფასება, პედაგოგის დახმარებით სწორი ინტერპრეტაცია და სასწავლო სახელმძღვანელოებში შეზღუდული რაოდენობით მოყვანილი მაგალითების უამრავი, სხვა ახალი მაგალითით შევსება, ასევე მოსწავლეების აქტიური ჩართვა თვითონ მაგალითების მოძიების პროცესში და სწავლების მეთოდოლოგიის მრავალფეროვნების უზრუნველყოფა. თუმცა, კორპუსებს არ შეუძლიათ პირდაპირ იმის თქმა, როგორ განვასხვაოთ სალიტერატურო ენის ნორმა და სწორი-არასწორი ფორმები ენაში. კვლევისას ასევე ყურადღება გავამახვილეთ კორპუსის გამოყენების სწორედ ამ მხარეზეც, კერძოდ, კორპუსის დახმარებით ჩვენ შეგვიძლია მოვიძიოთ სალიტერატურო ენის ნორმების თვალსაზრისით როგორც სწორი, ისე არასწორი ფორმები, რაც ასევე დამატებითი წყარო შეიძლება გახდეს სასწავლო მასალის შევსების პროცესში. იმისათვის, რომ გამოვიყენოთ კორპუსები კითხვებზე პასუხების გასაცემად, საჭიროა მასწავლებლის აქტიური ჩართულობა, რადგან კორპუსების გამოყენება მთლიანად მათ ხელშია.

დასკვნის სახით უნდა აღინიშნოს, რომ არსებული საერთაშორისო გამოცდილება, რომელიც უკავშირდება ენის სწავლების პროცესში კორპუსების გამოყენებას, სულ უფრო და უფრო აქტუალური ხდება. ქართული ენის კორპუსების შექმნის პარალელურად, მათ ჩართვას არა მხოლოდ კორპუსზე დაფუძნებული კვლევებისთვის, არამედ ენის სწავლების პროცესში ნამდვილად აქვს პერსპექტივა.

Corpus-based Language Pedagogy (Foreign Experience and Georgia)

Natia Kentchiashvili
Ivane Javakhishvili Tbilisi State University, Georgia
natia.kentchiashvili@gmail.com

The present study aims to review the corpus-based method in the field of language teaching and learning. The research has covered both the foreign experience of corpus-based language pedagogy and the perspectives of its development in Georgia. Within the framework of our research, the subject of theoretical discussion is the corpus not as a research tool, but as a basic and/or auxiliary resource for language learning and teaching. Thus, in addition to the corpus-based research method, the corpus-based teaching method is also considered.

Considering that we have accumulated a lot of different corpora for the Georgian language in the Internet space, it is important to discuss this type of electronic resource in the context of teaching and learning the Georgian language. Moreover, the corpora created for other languages are actively used for educational purposes. Therefore, in the framework of this study, special attention should be paid for the review of Georgian-language corpus and the importance of their use in this regard. Numerous foreign studies prove that the language-based learning and teaching corpus-based method is effective and productive; pupils/students have direct access to language data, thus enhancing and developing their language skills. The reviewed studies also show how the corpus-based method can be integrated into the field of language learning and teaching, regardless of the level of qualifications of the students. The method under consideration is especially useful for both beginners and teachers.

Through the corpus it is possible to quickly and efficiently find different types of linguistic material, as well as teachers are given the opportunity to prepare and supplement learning materials with interesting and authentic examples from Georgian texts of different times, authors, fields or genres. Corpus resources and tools can provide us with information about which word / phrase is used, how and how often, show us the frequency of their use, where their use cases are confirmed, in which words is a particular word attested, in what context it occurs and for what purposes, etc. However, the corpus cannot provide unequivocal answers to the questions, but instead, it is necessary to evaluate the information found through the corpus, correct interpretation with the help of the teacher, and fill the limited number of examples in the textbooks with many other new examples, as well as the active involvement of students in the process of finding examples and ensuring the diversity of teaching methodology. However, corpora cannot directly tell how to distinguish between literary language norm and correct-incorrect forms in language. During the research, we also focused on this aspect of the use of the corpus, in particular, with the help of the corpus, we can find both correct and incorrect forms in terms of the norms of the literary language, which

can also become an additional source in the process of filling in the educational material. In order to use the corpus to answer questions, requires active teacher involvement and corpus use is entirely in their hands.

In conclusion, it should be noted that the existing international experience related to the use of corpora in the language teaching process is becoming more and more relevant. In parallel with the creation of Georgian language corpora, their involvement not only in corpus-based research, but also in the language teaching process really has prospects.

გრამატიკულ ლექსიკონში ქართული და ინგლისური ენის მორფოლოგიური მახასიათებლების შესაბამისობა

ლიანა ლორთქიფანიძე, ნინო ამირეზაშვილი, ანა ჩუტკერაშვილი, ნინო ჯავაშვილი, ლიანა სამსონაძე, გიორგი ჩიკოიძე
არჩილ ელიაშვილის სახელობის მართვის სისტემების ინსტიტუტი,
საქართველოს ტექნიკური უნივერსიტეტი, საქართველო
L_Lordkipanidze@yahoo.com, ninomaskh@yahoo.com,
annachutkerashvili@yahoo.com, ninojavashvili@yahoo.com,
liasams@yahoo.com, gogichikoidze@yahoo.com

მოხსენება შეეხება ქართულ-ინგლისური გრამატიკული ლექსიკონის კომპაილერს, რომელიც იქმნება შოთა რუსთაველის საქართველოს ეროვნული სამეცნიერო ფონდის საგრანტო პროექტის ფარგლებში.

სახელმწიფო ენის სიცოცხლისუნარიანობისა და მდგრადობის უზრუნველსაყოფად ენობრივ ტექნოლოგიებს მნიშვნელოვანი როლი ეკისრება. მსოფლიო სულ უფრო მეტად ურთიერთქმედებს მანქანებთან ბუნებრივი ენით, ხოლო ენობრივი ტექნოლოგია ერთ-ერთი აუცილებელი კომპონენტია ავტომატიზებულ IT სისტემებში.

ბუნებრივი ენის დამუშავების NLP (Natural Language Processing) მნიშვნელოვანი საკითხია მწირი რესურსის მქონე ენებისათვის ავტომატური სისტემების განვითარება. NLP სისტემების უმეტესობა ხორციელდება კონტროლირებადი დასწავლის მეთოდების გამოყენებით. რადგან ეს სისტემები მოითხოვს დიდი რაოდენობის ანოტირებულ მონაცემებს, იგი გამოიყენება იმ ენებისთვის, რომლებისთვისაც უკვე შექმნილია დიდი მოცულობის ტექსტური კორპუსები. დღეს არსებული მცირე რაოდენობის ქართული ანოტირებული კორპუსები არაა საკმარისი NLP სისტემების განსახორციელებლად.

სხვადასხვა ენის გრამატიკულ ლექსიკონებს დიდი წარმატებით იყენებენ ტექსტური კორპუსების ანოტირებისათვის. არსებობს რუსული ენის გრამატიკული ლექსიკონი და მასზე აგებული კომპიუტერული სისტემა, პოლონური ენის გრამატიკული ლექსიკონი, ინგლისური ენის და სხვ. ასევე, ცნობილია რუსული ენის გრამატიკული ლექსიკონის კომპაილერები. ქართულ სივრცეში კი ასეთი ლექსიკონი ჯერჯერობით არ არსებობს.

ზემოხსენებული პროექტის – ქართულ-ინგლისური გრამატიკული ლექსიკონის კომპაილერი (GEGDiCo - The Compiler of the Georgian-English Grammatical Dictionary) – მიზანია უზრუნველყოს ქართული ენის გრამატიკული ლექსიკონის ძირითადი კომპონენტების შესაბამისობა საერთაშორისო სტანდარტებთან და ქართულ-ინგლისური გრამატიკული ლექსიკონის კომპაილერის შემუშავება.

ელექტრონული გრამატიკული ლექსიკონის დანიშნულებაა მომხმარებლისთვის ინფორმაციის მიწოდება სალექსიკონო ერთეულის იმ მორფოლოგიური და სინტაქსური მახასიათებლების შესახებ, რომლებსაც არსებითი მნიშვნელობა აქვთ გრამატიკულად სწორი ფრაზების ასაგებად. გარდა ამისა, ტექსტების მანქანური დამუშავებისას ასეთი ტიპის ლექსიკონები გამოიყენება ავტომატური მორფოლოგიური ანალიზის ინსტრუმენტად.

ლინგვისტი-მომხმარებელი ვებგვერდზე წინასწარი დარეგისტრირებისა და საჭირო აპლიკაციების ჩამოტვირთვის შემდეგ შეძლებს გრამატიკული ლექსიკონის კომპაილერის საშუალებით ქართული ენის სხვადასხვა ქვესისტემებისა და დიალექტების გრამატიკული ლექსიკონის შედგენას არაანოტირებული ტექსტური კორპუსებიდან. ინტერნეტსივრცეში დაიდება ქართველური ენების გრამატიკული ლექსიკონის კომპილირების სისტემა, რომლის ინსტრუმენტების გამოყენებით ენათმეცნიერებს შეეძლებათ სასურველი ენის დიალექტის გრამატიკული ლექსიკონის დამოუკიდებლად შედგენა. გრძელვადიან პერსპექტივაში შესაძლებელი გახდება ნებისმიერი ქართველური ენის ნებისმიერი სიტყვის ქართულ-ინგლისური თარგმანი.

ლექსიკონის თავისებურება და მნიშვნელობა, პირველ რიგში, მდგომარეობს მწყობრი მორფოლოგიური სისტემის შექმნაში, რომელიც ეყრდნობა ლექსიკონის ერთეულების ამომწურავ სიმრავლეს. ლექსიკონი გამდიდრებული იქნება გარკვეული სინტაქსური მონაცემებით, კერძოდ, ზმნურ-აქტანტური და პრედიკატულ-როლებრივი მახასიათებლებით. ასევე მოხდება სახელური ჯგუფის და ზმნური მახასიათებლების სისტემატიზაცია.

ლექსიკონის მენეჯმენტის თვალსაზრისით, სასარგებლოა სიტყვების შენახვა ლემებისა და ფლექსიური პარადიგმების სახით, რაც საშუალებას მოგვცემს სიტყვა დავამატოთ უბრალოდ ლემის მითითებით და შემდეგ, უკვე წინასწარ განსაზღვრული ფლექსიური პარადიგმიდან, ავირჩიოთ ფორმა. თუ შესაფერისი პარადიგმა არ არსებობს, შესაძლებელია განისაზღვროს ახალი. თუ შეცდომა გამოვლინდა ერთ ფლექსიურ პარადიგმაში, ასეთ შემთხვევაში ჩასწორება მხოლოდ ერთ ადგილას იქნება საჭირო. ამ ფენომენის დამუშავებისას გამოვლენილი კანონზომიერებები გრამატიკული ლექსიკონების აგებისას შეიძლება დაჯგუფდეს, რათა თავიდან ავიცილოთ ყოველი სიტყვისთვის ყველა ფორმის მიწერა. ეს მოსახერხებელი იქნება როგორც ანალიზისთვის, ისე წარმოქმნისთვის.

ამოსავალი სალექსიკონო ერთეულის ლექსიკური ფორმები შედგენილი იქნება ლემისგან

და მორფოლოგიური ტეგების მწკრივისგან. პირველი ტეგი განიხილება როგორც მეტყველების ნაწილის მარკერი, ხოლო დანარჩენი – როგორც ლექსიკური ქვეკატეგორიების მარკერები.

პარადიგმები, რომლებიც გამოიყენება ლექსიკონების ასაგებად, შეიძლება მიღებული იყოს ავტომატურად, ხელით ან შერეული მეთოდით - ავტომატურად და ხელით.

სისტემა მოგვცემს სიტყვის როგორც ფორმაწარმოების, ისე სიტყვაწარმოების სრულად აღწერის საშუალებას, რაც ქართული ენისთვის სიახლეა. ასევე დაგეგმილია გრამატიკული ლექსიკონის კომპაილერის მისადაგება ქართველური ენების ქვესისტემების/დიალექტების კორპუსების ანოტირებისათვის.

სირთულეს წარმოადგენს ერთგვაროვანი მახასიათებლების მქონე სიების შედგენა, განსაკუთრებით ზმნისათვის. მართალია, არსებობს "ქართული ენის ზმნური ფუძეების ლექსიკონი", მაგრამ მისი დახმარებით ზმნების კლასიფიცირება რთულია, როგორც პირდაპირი სიტყვაწყობით, ისე ინვერსიული რიგით. შეუძლებელია ავტომატური დამუშავების გარეშე ზმნების კლასიფიცირება მათი ყველა ფორმის მიხედვით. ლექსიკონის კომპაილერი დაეყრდნობა ავტომატურ მორფოლოგიურ პროცესორს.

გარდა ამისა, ცნობილია, რომ გრამატიკული ლექსიკონების მორფოლოგიური მახასიათებლებით უზრუნველსაყოფად აუცილებელია მათთვის ძირითადი

საკლასიფიკაციო პრინციპების შემუშავება. ინდოევროპული ენებისაგან განსხვავებით, ქართული ზმნის აგებულების ორმაგი, აგლუტინაციურ-ფლექსიური ბუნების გამო, გარდა ტრადიციული გრამატიკული მახასიათებლებისა, აუცილებელია ყველა ანომალიური ფონეტიკური ცვლილებებისათვის ახალი ნიშნების დაფიქსირება.

მორფოლოგიური მახასიათებლების შერჩევისას ჩვენ ვეყრდნობით EAGLES2 (Expert Advisory Group on Language Engineering Standards) სტანდარტს. შემუშავებულია მორფო-სინტაქსური პარამეტრები და მათი მარკერები სალიტერატურო ქართული ენის კორპუსის ფარგლებში[1]3. GEGDiCo სისტემის პროგრამულ უზრუნველყოფაში არის როგორც მახასიათებლების, ისე მათი მარკერების დამატების საშუალება. ქართული ენის ნებისმიერი ქვესისტემის გრამატიკული ლექსიკონის შედგენისას ლინგვისტ მომხმარებელს შესაძლებლობა აქვს დაამატოს გრამატიკული მახასიათებელი შესაბამისი მარკერით. ამჟამად მიმდინარეობს GEGDiCo სისტემაში ინგლისური ენის მახასიათებლების დამატება და ქართულ ენასთან მათი შესაბამისობაში მოყვანა.

მოხსენებაში განხილული იქნება გრამატიკულ ლექსიკონში ქართული და ინგლისური არსებითი, ზედსართავი და რიცხვითი სახელების, ნაცვალსახელის, ზმნისა და უდეტრების ფლექსიური და დერივაციული ფორმებისთვის საკლასიფიკაციო მახასიათებლების შემუშავებისა და მათი სისტემატიზაციის პროცესი.

[1] შოთა რუსთაველის საქართველოს ეროვნული სამეცნიერო ფონდის საგრანტო პროექტი – ქართული ენის კორპუსის სრული (მორფოლოგიური, სინტაქსური, სემანტიკური) ანოტირების სისტემა FR/463/4-105/12 (2013-2016)

ლიტერატურა

Weiss, D., Alberti, C., Collins, M., and Petrov, S. (2015). Structured training for neural network transition-based parsing. CoRR, abs/1506.06158.

Straka, M., Hajicˇ, J., and Strakova´, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorozˇ, Slovenia. European Language Resources Association.

Ballesteros, M., Goldberg, Y., Dyer, C., and Smith, N. A. (2016). Training with exploration improves a greedy stacklstmparser. In Proceedings of the Conference, 2016 on Empirical Methods in Natural Language Processing, pages 2005–2010, Austin, Texas, November. Association for ComputationalLinguistics.

http://www.ilc.cnr.it/EAGLES/home.html

Зализняк А. Грамматический словарь русского языка. Изд. «Русский язык», Москва. ГСРЯ. 1977

.http://starling.rinet.ru/cgi-bin/morphque.cgi?encoding=win.

http://sgjp.pl/leksemy/#13589/a

https://public.oed.com/how-to-use-the-oed/glossary/

http://www.solarix.ru/for_developers/bootstrap/compile_dictionary-en.shtml,

https://seelrc-iis.trinity.duke.edu/russdict/index.aspx?doc=2;

https://pymorphy2.readthedocs.io/en/0.2/internals/dict.html#id6.

ავტომატური განმარტებით-კომბინატორული ლექსიკონი როგორც ქართული ენის მოდელირე-ბის საფუძველი (2009-2011) №A 36-09.

ქართული ენის კორპუსის სრული (მორფოლოგიური, სინტაქსური, სემანტიკური) ანოტირების სისტემა (2013-2016) FR/463/4-105/12.

Matching the morphological characteristics of Georgian and English in the grammatical dictionary

Liana Lortkipanidze, Nino Amirezashvili, Ana Chutkerashvili, Nino Javashvili, Liana Samsonadze

Archil Eliashvili Institute of Control Systems, Georgian Technical University, Georgia
L_Lordkipanidze@yahoo.com, ninomaskh@yahoo.com,
annachutkerashvili@yahoo.com, ninojavashvili@yahoo.com,
liasams@yahoo.com, gogichikoidze@yahoo.com

The report is about the compiler of the Georgian-English grammar dictionary, which is being created within the framework of the grant project of the Shota Rustaveli Georgian National Science Foundation.

Language technologies play an important role in ensuring the viability and sustainability of the state language. The world is increasingly interacting with machines using natural language and language technology is one of the essential components in automated IT systems.

The development of automatic systems of low-resource languages is a significant issue for Natural Language Processing (NLP). Most NLP systems are constructed by using supervised learning methods. These systems require a large number of annotated data and thus are used for languages that already have big volume text corpora. Today available small number of Georgian annotated corpora is not enough for NLP system implementation.

Grammatical dictionaries of various languages are successfully used for text annotation. There is a grammar dictionary of Russian language as well as a computer system based on it; There are grammar dictionaries for Polish language, English language etc.; also, there are the compilers of the Russian grammar dictionary. Yet, there is no such dictionary for Georgian language.

The aim of the above-mentioned project – The Compiler of the Georgian-English Grammatical Dictionary (GEGDiCo) is to ensure the compliance of the main components of the Georgian grammar dictionary with international standards and to develop a Georgian-English grammar dictionary compiler.

The purpose of the electronic grammatical dictionary is to provide information on the morphological and syntactic characteristics of the dictionary unit, which are essential for structuring correct grammatical phrases. These types of dictionaries are used as a tool of automated morphological analysis while processing texts.

After pre-registering on the website and downloading the necessary applications, the linguist-user will be able to create a grammar dictionary of Georgian language various subsystems and dialects from non-annotated text corpora using a grammar dictionary compiler. The grammar dictionary compiling system of Kartvelian languages will be placed on the Internet. Linguists will be able to compile a grammatical dictionary of any dialect of desired language on their own by using its tools. In the long-term

plan, it will be possible to translate any word of any Kartvelian languages into Georgian or English language.

The peculiarity and the value of the dictionary is, first of all, creation of an orderly morphological system, which is based on the exhaustive set of dictionary units. The dictionary will be enriched with certain syntactic data, in particular, verb-actant and predicate-role characteristics. The vocabulary will be systematized by the noun phrase and verb phrase characteristics.

In terms of dictionary management, it is useful to keep words in the form of lemmas and inflective paradigms. That will allow adding a word simply by indicating a lemma and then choosing the form from the inflective paradigm, which is defined in advance. If a suitable paradigm does not exist, it is possible to define a new one. If a mistake is found in one of the inflective paradigms, there will only be one place for editing. While structuring grammatical dictionaries the patterns identified in the processing of this phenomenon can be grouped to avoid ascribing all forms to each word. This will be convenient for both analysis and generation.

The lexical forms corresponding to the title form of grammatical dictionary entry will be drawn from the lemma and morphological tags. The first tag is considered as a marker of part of speech, while the rest as the markers of lexical subcategories.

The paradigms used to construct dictionaries may be implemented automatically, manually or by the mixed method - automatically and manually.

The system will allow us to fully describe both the word formation and the word production, which is a novelty for the Georgian language. It is also planned to adapt a grammar dictionary compiler to annotate Kartvelian language subsystems / corpus dialects.

It is difficult to compile lists with similar characteristics, especially for verbs. It is true that there is “The Dictionary of Georgian Verb Stems", but with its help it is difficult to classify verbs, both in direct and in inverse word order. It is impossible to classify verbs according to all their forms without automatic processing. The dictionary compiler will be based on the automated morphological processor of modern Georgian language.

In addition, in order to provide grammatical dictionaries with morphological characteristics, it is necessary to develop basic classification principles for them. Unlike Indo-European languages, due to the dual agglutinative-flectional nature of Georgian verbs, in addition to traditional grammatical characteristics, it is necessary to fix new signs for all anomalous phonetic changes.

We select the EAGLES (Expert Advisory Group on Language Engineering Standards) standard when selecting morphological characteristics. Morpho-syntactic parameters and their markers have been developed within the framework of the literary Georgian language corpus[1]. GEGDiCo system software includes the ability to add both characteristics and their markers. When compiling a grammar dictionary

[1] The grant project of the Shota Rustaveli Georgian National Science Foundation. The Full (Morphological, Syntactical, Sematic) Annotation System of Georgian Language Corpora FR/463/4-105/12 (2013-2016)

of any subsystem of the Georgian language, the linguist-user has the opportunity to add a grammatical characteristic with the appropriate marker. At present, English language characteristics are being added to the GEGDiCo system and brought in compliance with the Georgian language.

The report will discuss the development and systematization of classification characteristics for flectional and derivational forms of Georgian and English nouns, adjectives and numerals, pronouns, verbs and adverbs in the grammar dictionary.

References

Weiss, D., Alberti, C., Collins, M., and Petrov, S. (2015). Structured training for neural network transitionbased parsing. CoRR, abs/1506.06158.

Straka, M., Hajicˇ, J., and Strakova´, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorozˇ, Slovenia. European Language Resources Association.

Ballesteros, M., Goldberg, Y., Dyer, C., and Smith, N. A. (2016). Training with exploration improves a greedy stacklstmparser. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2005–2010, Austin, Texas, November. Association for Computational Linguistics.

Zaliznyak, A. Russian Grammatical Dictionary. Moskva, 1977, (in Russian);

http://starling.rinet.ru/cgi-bin/morphque.cgi?encoding=win.

http://sgjp.pl/leksemy/#13589/a

https://public.oed.com/how-to-use-the-oed/glossary/

http://www.solarix.ru/for_developers/bootstrap/compile_dictionary-en.shtml,

https://seelrc-iis.trinity.duke.edu/russdict/index.aspx?doc=2;

https://pymorphy2.readthedocs.io/en/0.2/internals/dict.html#id6.

Automatic explanatory-combinatorial dictionary as the basis of modelling of the Georgian language (2009-2011) №A 36-09.

The Full (Morphological, Syntactical, Sematic) Annotation System of Georgian Language Corpora, (2013-2016) FR/463/4-105/12

ლექსიკის „სწავლება“ მანქანური თარგმნის პროგრამისთვის

თინათინ მარგალიტაძე
ილიას სახელმწიფო უნივერსიტეტი, საქართველო
tinatin.margalitadze@iliauni.edu.ge

მანქანური თარგმნის პროგრამისთვის განსაკუთრებულ სირთულეს, ლექსიკის „შესწავლის“ თვალსაზრისით, ქმნის ოთხი რამ: შესიტყვებები, პოლისემია, ომონიმია და იდიომები. ევროპული ენების მანქანური თარგმნის პროგრამებზე დაკვირვება ცხადყოფს, რომ ხშირ შემთხვევაში, პროგრამები წარმატებით ართმევს თავს ზემოხსენებულ სირთულეებს. ეს განსაკუთრებით ითქმის DeepL-ის პლატფორმის გამოყენებით შექმნილ პროგრამებზე. ქვემოთ წარმოდგენილია რუსულ-ინგლისური / ინგლისურ-რუსული მანქანური თარგმნის პროგრამით შესრულებული წყვილები (https://www.deepl.com/translator):

(1) Это было гвоздём программы
It was the highlight of the programme[1].
(2) Этот парень целый день бьет баклуши
This guy's been chugging along all day.
(3) It is raining cats and dogs
Идет дождь из кошек и собак
(4) I heard the hissing of a snake from the bushes.
Из кустов послышалось шипение змеи.
(5) Он страдает тяжелой болезнью
He suffers from a serious illness.
(6) A flock of birds fly together
Стая птиц летит вместе.
(7) A herd of swine fell into the sea
Стадо свиней упало в море.
(8) Это был известный ищейка
It was a famous bloodhound
Он был известным ищейкой
He was a famous bloodhound.

ამ მაგალითების ანალიზი გვიჩვენებს, რომ მანქანური თარგმნის პროგრამაში, ისეთი ენებისთვისაც კი, რომელთაც საკმაოდ დიდი რესურსები აქვთ და მილიონობით პარალელური წინადადების დაგროვებაა შესაძლებელი, ლექსიკის „შესწავლის“ პრობლემა ყოველთვის არ არის

[1] https://www.deepl.com/translator#ru/en

გადაწყვეტილი. პირველ წინადადებაში რუსული იდიომი *гвоздь программы* სწორადაა ინგლისურად თარგმნილი, თუმცა იმავეს ვერ ვიტყვით მეორე წინადადებაზე, სადაც რუსული იდიომი *бить баклуши* (უსაქმურობა) არასწორი ინგლისური ეკვივალენტითაა გადატანილი. მესამე მაგალითში, ცნობილი და საკმაოდ გაცვეთილი ინგლისური იდიომი: it is raining cats and dogs (თავსხმაა) სიტყვასიტყვითაა თარგმნილი. მე-4, მე-5, მე-6 და მე-7 მაგალითებში შესიტყვებები სწორი ეკვივალენტებითაა გადატანილი როგორც რუსულიდან ინგლისურად, ისე პირუკუ. საინტერესოა, რომ მე-5 მაგალითში რუსული შესიტყვება *тяжелая болезнь* სწორი ინგლისური შესატყვისითაა თარგმნილი და არა სიტყვასიტყვით heavy illness.

მე-8 მაგალითი პოლისემიის მაგალითია. რუსული სიტყვა *ищейка* მეძებარ ძაღლსაც აღნიშნავს და დეტექტივსაც. როგორც ვხედავთ, თარგმანში პროგრამამ ორივე შემთხვევაში რუსული პოლისემიური სიტყვა *ищейка* გადათარგმნა როგორც მეძებარი ძაღლი, მიუხედავად იმისა, რომ მეორე წინადადებაში არის მინიშნება ადამიანზე - он (он был известным ищейкой).

ლექსიკის „შესწავლის“ ეს პრობლემები კიდევ უფრო მკაფიოდ იჩენს თავს მწირი რესურსების მქონე ენების შემთხვევაში. ინგლისურ-ქართული მანქანური თარგმნის პროგრამა Google translate ინგლისურ-ქართულ წინადადებათა 1.3 მილიონი წყვილის საფუძველზეა შექმნილი. ქართული ენის მსგავსი რესურსების მქონე ენისათვის ეს არ არის ცოტა. მიუხედავად ამისა, პროგრამას სერიოზული ხარვეზები აქვს შესიტყვებების, პოლისემიის, ომონიმიის, იდიომების თარგმნის თვალსაზრისით.

(9) rough day, rough road
უხეში დღე, უხეში გზა
(10) ჩიტების გუნდი ერთად მიფრინავდა
A team of birds flew together
(12) She has delicate features
მას აქვს დელიკატური თვისებები
(13) წერწეტა ტანი აქვს
She has a short body
(14) Dogs of war in Ukraine
ომის ძაღლები უკრაინაში
(15) It is raining cats and dogs
კატები და ძაღლები წვიმს
(16) ეზოსთვის ახალი ბარი ვიყიდე
I bought a new bar for the yard
(17) გახარებულებმა მთა და ბარი მოვირბინეთ
Exited, we rushed to the mountain and the bar

ამ მაგალითებში კარგად ჩანს პრობლემები შესიტყვებების, პოლისემიის, ომონიმიის, იდიომების თარგმნისას.

ჩვენი ლექსიკოგრაფიული გუნდი 2018 წლიდან მუშაობს ინგლისურ-ქართული/ქართულ-ინგლისური მანქანური თარგმნის პროგრამის კონცეფციაზე (მარგალიტაძე, ფურცხვანიძე 2019ა,

2019ბ). ჩვენთვის, როგორც ლექსიკოგრაფებისათვის, განსაკუთრებით საინტერესო იყო და არის ლექსიკოგრაფიული მასალის ეფექტიანობა მანქანური თარგმნის პროგრამაში. ამგვარ პროგრამას და მასზე დაფუძნებულ აპლიკაციებს ესაჭიროება ბუნებრივი ენის სემანტიკის აღწერა, ენის სემანტიკის ყველაზე სრულყოფილი აღწერა კი ლექსიკონებშია მოცემული. ჩვენი ჰიპოთეზის მიხედვით, ჩვენი ლექსიკოგრაფიული გუნდის მიერ 35 წლის განმავლობაში შექმნილი *დიდი ინგლისურ-ქართული ონლაინლექსიკონის მასალა* მნიშვნელოვანი წყარო უნდა ყოფილიყო ინგლისურ-ქართული მანქანური თარგმნის პროგრამისთვის. მიმდინარე წლის მარტში მოგვეცა ჩვენი ჰიპოთეზის შემოწმების საშუალება, როდესაც ჩვენ მიერ მომზადებული 370 000 წინადადების წყვილით პირველად გაიწვრთნა ქართულ-ინგლისური მანქანური თარგმნის პროგრამა OpenNMT[1] მოდელისთვის[2]. აღნიშნული 370 000 წინადადებიდან 100 000 წინადადება იყო *დიდი ინგლისურ-ქართული ონლაინლექსიკონიდან* ამოღებული პარალელური წინადადებები. ჩვენდა გასაოცრად, პროგრამამ საკმაოდ კარგად „შეისწავლა“ სპეციფიკური ლექსიკა, მათ შორის შესიტყვებები და ხშირ შემთხვევაში უკეთეს და უფრო ზუსტ თარგმანს გვთავაზობს ქართულიდან ინგლისურად, ვიდრე Google translate, რომელიც 1.3 მილიონი წინადადების წყვილს ეფუძნება. ქვემოთ მოცემულია რამდენიმე მაგალითი, რომელიც აჩვენებს განსხვავებას Google translate-ის მიერ ქართულიდან ინგლისურად თარგმნილ წინადადებებსა და ჩვენი პროგრამის მიერ თარგმნილ წინადადებებს შორის.

(18) ღორების კოლტი ზღვაში გადავარდა:
Google translate: The pig colt fell into the sea.
ჩვენი მთარგმნელი: A herd of swine fell into the sea.

(19) მგლების ხროვა მას ყოველი მხრიდან უტევდა:
Google translate: A herd of wolves attacked him from all sides.
ჩვენი მთარგმნელი: A pack of wolves was attacking him from all sides.

(20) არწივი ცაში ლივლივებდა:
Google translate: The eagle was flying in the sky.
ჩვენი მთარგმნელი: The eagle was soaring in the sky.

(21) მდინარე ტყეში მორაკრაკებდა:
Google translate: The river was flowing in the forest.
ჩვენი მთარგმნელი: The river bubbled in the forest.

(22) ფარდები ქარში ფრიალებდა:
Google translate: The curtains were flying in the wind.
ჩვენი მთარგმნელი: Curtains fluttered in the wind.

(23) ჩიტების გუნდი ერთად მიფრინავდა:
Google translate: A team of birds flew together.
ჩვენი მთარგმნელი: A flock of birds flew together.

[1] https://opennmt.net/

[2] მონაცემების გაწვრთნაზე იმუშავა მონაცემთა ანალიზის მეცნიერმა ვახტანგ ელერდაშვილმა.

ამჟამად უფრო დეტალურად ვსწავლობთ მანქანური თარგმნის პროგრამის მიერ ლექსიკონის მასალის ათვისების წესებს. ამ წესების დადგენის შემდეგ გადამუშავდება *დიდი ინგლისურ-ქართული ონლაინლექსიკონის* მასალა მანქანური თარგმნის მიზნებისათვის. ჩვენი ვარაუდით, ეს მასალა ნახევარ მილიონ ინგლისურ-ქართულ პარალელურ წინადადებას მიაღწევს. ამავდროულად გროვდება წინადადებები ინგლისურ-ქართულ პარალელურ კორპუსში (მარგალიტაძე, მელაძე, ფურცხვანიძე 2022), რის შემდეგაც გაგრძელდება მონაცემთა წვრთნა ინგლისურ-ქართული მანქანური თარგმნის პროგრამისთვის.

ლიტერატურა

დიდი ინგლისურ-ქართული ონლაინლექსიკონი (მთავარი რედაქტორი თ. მარგალიტაძე). 2010. თბილისი : ლექსიკოგრაფიული ცენტრი. www.dict.ge

მარგალიტაძე, თ., მელაძე, გ., ფურცხვანიძე, ზ. (2022). ინგლისურ-ქართული პარალელური კორპუსი და მისი გამოყენება ქართულ ლექსიკოგრაფიაში. სამეცნიერო ჟურნალი *Lexikos,* ტ. 32 (2). https://lexikos.journals.ac.za/pub/article/view/1701

მარგალიტაძე, თ, ფურცხვანიძე, ზ. (2019 ა) „ქართული ენა ხელოვნურ ინტელექტზე დაფუძნებულ თარგმანის მოდელებში: ლექსიკოგრაფებისა და ბუნებრივი ენის დამუშავების სპეციალისტების თანამშრომლობა“. საერთაშორისო კოლოკვიუმი *ლექსიკოგრაფია გზის გასაყართან.* კონფერენციის ორგანიზატორები: თსუ ლექსიკოგრაფიის ცენტრი და EMLEX - ლექსიკოგრაფიის ევროპული სამაგისტრო პროგრამის კონსორციუმი. ოქტომბერი. https://margaliti.com/emlexweb.pdf

მარგალიტაძე, თ., ფურცხვანიძე, ზ. (2019 ბ) „მაღალი ხარისხის ლექსიკოგრაფიულ მონაცემთა ბაზის ეფექტურობა ინგლისურ-ქართული / ქართულ-ინგლისური მანქანური თარგმანის პროგრამის შექმნისათვის“. საერთაშორისო კონფერენცია *ენა და თანამედროვე ტექნოლოგიები V – ისტორიული და ეტიმოლოგიური ლექსიკოგრაფიის საკითხები.* კონფერენციის ორგანიზატორები: თსუ არნოლდ ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი, სოხუმის სახელმწიფო უნივერსიტეტი, სახელმწიფო ენის დეპარტამენტი, ჩერქეზული (ადიღეური) კულტურის ცენტრი. დეკემბერი.

On "Teaching" Vocabulary to Machine Translation Program

Tinatin Margalitadze
Ilia State University, Georgia
tinatin.margalitadze@iliauni.edu.ge

From the point of view of "teaching" lexis to the machine translation (MT) program, major difficulties are posed by collocations, polysemy, homonymy and idioms. The study of MT programs of European languages has revealed that, in the majority of cases, above-mentioned problems are dealt with quite successfully. This is particularly true of MT programs based on the DeepL platform. Below are given parallel Russian-English sentences, translated by the Russian-English / English-Russian DeepL translator (https://www.deepl.com/translator):

(1) Это было гвоздём программы
It was the highlight of the programme[1].
(2) Этот парень целый день бьет баклуши
This guy's been chugging along all day.
(3) It is raining cats and dogs
Идет дождь из кошек и собак
(4) I heard the hissing of a snake from the bushes.
Из кустов послышалось шипение змеи.
(5) Он страдает тяжелой болезнью
He suffers from a serious illness.
(6) A flock of birds fly together
Стая птиц летит вместе.
(7) A herd of swine fell into the sea
Стадо свиней упало в море.
(8) Это был известный ищейка
It was a famous bloodhound
Он был известным ищейкой
He was a famous bloodhound.

The analysis of these examples reveals that not all problems of "learning" vocabulary are solved in MT programs, even for languages with great resources and millions of parallel sentences as translation memories. For example, in the first sentence, Russian idiom *гвоздь программы* is translated into English with a very good equivalent. This cannot be said about the second example, where translation of the

[1] https://www.deepl.com/translator#ru/en

Russian idiom *бить баклуши* (to idle, to twiddle, to sit around) is not accurate. In the third example, a well known English idiom *it is raining cats and dogs* (it's downpouring) is translated literally. In the 4th, 5th, 6th and 7th examples, collocations are rendered very well both from Russian into English and vice versa. It is worth noting that in the example 5, Russian collocation *тяжелая болезнь* is translated with a relevant equivalent and not literally as *heavy illness*.

The 8th example deals with polysemy. The Russian word *ищейка* has 2 senses: a hound and a sleuth, detective. As can be seen from the translations, in both cases *ищейка* is rendered as a hound, although the second sentence has an indication that a human being is implied - *он* ('он был известным ищейкой').

The above-mentioned problems of "learning" lexis are even more conspicuous in the case of lesser-resourced languages like Georgian. The English-Georgian MT program Google translate is based on 1.3 million English-Georgian sentence pairs. For a language like Georgian, this amount of data cannot be considered to be scarce. Despite this, the MT program for Georgian has serious drawbacks in dealing with collocations, polysemy, homonymy and idioms.

(9) rough day, rough road
უხეში დღე, უხეში გზა
(10) ჩიტების გუნდი ერთად მიფრინავდა
A team of birds flew together
(12) She has delicate features
მას აქვს დელიკატური თვისებები
(13) წერწეტა ტანი აქვს
She has a short body
(14) Dogs of war in Ukraine
ომის ძაღლები უკრაინაში
(15) It is raining cats and dogs
კატები და ძაღლები წვიმს
(16) ეზოსთვის ახალი ბარი ვიყიდე
I bought a new bar for the yard
(17) გახარებულებმა მთა და ბარი მოვირბინეთ
Exited, we rushed to the mountain and the bar

The above-cited examples are literal translations of Georgian or English phrases and sentences and reveal very well the problems, related to the rendition of collocations, polysemy, homonymy and idioms.

Our lexicographic team has worked on the conception of the English-Georgian/Georgian-English MT program since 2018 (Margalitadze, Pourtskhvanidze 2019a, 2019b). For us, lexicographers, it has been especially interesting to investigate effectiveness of the lexicographic material in a MT program. Such programs and applications, based on them, need description of semantics of a human language, and dictionaries contain the most detailed and accurate description of semantics. According to our hypothesis, the material of our *Comprehensive English-Georgian Online Dictionary,* created during 35 years could be

an important source for an English-Georgian MT program. In March of the current year, we had a possibility to check our hypothesis, when 370 000 English-Georgian sentence pairs, prepared by us, were trained in the OpenNMT[1] model[2]. Out of the 370 000 sentence pairs, used for training, 100 000 sentences were extracted from the *Comprehensive English-Georgian Online Dictionary.* The results obtained as a result of training of the data surprised even us and exceeded our expectations. The program has learnt even very specific vocabulary quite well, and deals particularly well with collocations. From this point of view, our machine translation program, in some cases, provides more accurate translations from Georgian into English, than Google translate, which is based on the 1.3 million English-Georgian sentence pairs, as mentioned above. Below are quoted some examples which illustrate the difference in the English translations of Georgian sentences by the Google translate and our translator:

(18) ღორების კოლტი ზღვაში გადავარდა:
Google translate: The pig colt fell into the sea.
Our translator: A herd of swine fell into the sea.

(19) მგლების ხროვა მას ყოველი მხრიდან უტევდა:
Google translate: A herd of wolves attacked him from all sides.
Our translator: A pack of wolves was attacking him from all sides.

(20) არწივი ცაში ლივლივებდა:
Google translate: The eagle was flying in the sky.
Our translator: The eagle was soaring in the sky.

(21) მდინარე ტყეში მორაკრაკებდა:
Google translate: The river was flowing in the forest.
Our translator: The river bubbled in the forest.

(22) ფარდები ქარში ფრიალებდა:
Google translate: The curtains were flying in the wind.
Our translator: Curtains fluttered in the wind.

(23) ჩიტების გუნდი ერთად მიფრინავდა:
Google translate: A team of birds flew together.
Our translator: A flock of birds flew together.

At the present stage we investigate the rules, according to which English-Georgian MT program "learns" dictionary material. After the establishment of these rules, the *Comprehensive English-Georgian Online Dictionary* material will be processed anew for the MT program purposes. According to our estimates, dictionary material will amount to half a million English-Georgian parallel sentences. Simultaneously, we continue accumulation of English-Georgian sentences in the English-Georgian parallel corpus (Margalitadze, Meladze, Pourtskhvanidze 2022), developed by our lexicographic team. After this phase, training of the data will continue for English-Georgian MT program.

[1] https://opennmt.net/
[2] The data was trained by the data scientist Vakhtang Elerdashvili.

References

Comprehensive English-Georgian Online Dictionary (editor-in-chief T. Margalitadze). 2010. Tbilisi: Lexicographic Centre. www.dict.ge

Margalitadze, T., Meladze, G., Pourtskhvanidze, Z. (2022). English-Georgian Parallel Corpus and Its Application in Georgian Lexicography. *Lexikos,* vol. 32 (2). https://lexikos.journals.ac.za/pub/article/view/1701

Margalitadze, T, Pourtskhvanidze, Z. (2019a) 'The Georgian Language in AI-based Translation Models: Cooperation of Lexicographers and NLP specialists'. International conference *Lexicography at a Crossroads*, organized by TSU Lexicographic Centre and Consortium of European Master in Lexicography (EMLEX). https://margaliti.com/emlexweb.pdf

Margalitadze, T., Pourtskhvanidze, Z. (2019b). 'Effectiveness of High Quality Lexicographic Data for the Development of English-Georgian / Georgian-English Machine Translation Program'. International Conference *Language and Modern Technologies V - Issues of Historical and Etymological Lexicography*, organized by TSU Arnold Chikobava Institute of Linguistics, Sokhumi State University, State Language Department of Georgia, Circassian (Adyghean) Culture Center.

ქართულ ხმის სინთეზატორი

რატი სხირტლაძე, ლევან ლაშაური, ლევან შუღლიაშვილი, საბა კობახიძე
მონაცემთა ანალიზის ლაბორატორია, საქართველო
Rati2008@gmail.com, llevani@gmail.com, levan@shugliashvili.com,
s.kobakhidze@freeuni.edu.ge

ხმის სინთეზატორს აქვს მრავალმხრივი პრაქტიკული გამოყენება. მაგალითად, მას იყენებენ შეზღუდული მხედველობის თუ მეტყველების მქონე ადამიანები, ხმოვანი ასისტენტები, ეკრანის წამკითხველები, ავტომატური სატელეფონო სისტემები. იგი ასევე გამოიყენება აუდიო წიგნების შესაქმენლად და ენის დასწავლის გამარტივებისთვის.

ჩვენ მიერ შექმნილი ხმის სინთეზატორი ეფუძნება 2016 წელს გუგლის მიერ გამოქვეყნებულ კვლევას, რომელიც პრობლემის გადასაწყვეტად ნეირონული ქსელის გამოყენებას ეხება.

აღნიშნული მიდგომა გულისხმობს ხმის გარდაქმნას 80 განზომილებიან აუდიო სპექტროგრამად, რომელიც ქმნის ხმის „კადრებს" ყოველ 12.5 მილიწამში, და ასახავს ადამიანის ხმისთვის დამახასიათებელ სხვადასხვა მახასიათებლებს, მაგალითად, ინტონაციას.

ამ მოდელს ეწოდება Tacotron 2 და იგი წარმოადგენს WaveNet და Tacotron-ის კომბინაციას.

მოდელის გასაწვრთნელად ქართულ ენაზე შესრულებული აუდიოჩანაწერები გარდავქმენით ე.წ. მელ-სპექტროგრამებად (mel spectogram; mel - მელოდიის შემოკლებული აღნიშვნა). აღნიშნულ სპეტქოგრამა გამოიყენება ადამიანის ფიზიოლოგიური შესაძლებლობების უკეთ ასასახად. ჩვენ ახლოს მდგომ სიხშირეებს ერთგვარად აღვიქვამთ, რადგან არ გვესაჭიროება აბსოლუტური სიზუსტე. რაც უფრო მაღალია სიხშირე, იზრდება დიაპაზონი, რომლის ფარგლებშიც გვიჭირს სიხშირეების გარჩევა. ამიტომ ვიყენებთ ჰერცების მელებში გადაყვანის ლოგარითმულ გარდაქმნას.

მოდელი გავწვრთენით კონვოლუციური და განმეორებადი ნეირონული ქსელების გამოყენებით.

სრულყოფილი ხმის სინთეზატორის შესაქმნელად ასევე დაგვჭირდა ენისთვის დამახასიათებელი ისეთი დამხმარე მოდულების შექმნა, როგორებიცაა, მაგალითად, რიცხვების ტექსტად გარდაქმნა მათი შემდგომი გახმოვანების მიზნით.

დასკვნის სახით შეიძლება ითქვას, რომ ჩვენმა ჯგუფმა შეძლო მაღალი ხარისხის ხმის სინთეზატორის შექმნა ქართული ენისთვის. პროდუქტის დანერგვისა და პრაქტიკული გამოყენების პროცესში, ბუნებრივია, გავაგრძელებთ მის სრულყოფას.

Georgian Voice Synthesizer

Rati Skhirtladze, Levan Lashauri, Levan Shugliashvili, Saba Kobakhidze
Data Analysis Laboratory LLC, Georgia
Rati2008@gmail.com, llevani@gmail.com, levan@shugliashvili.com,
s.kobakhidze@freeuni.edu.ge

The voice synthesizer has many practical uses. For example, it is used by people with limited vision or speech, voice assistants, screen readers, and automated telephone systems. It is also used to create audiobooks and facilitate language learning.

Our voice synthesizer is based on research published by Google in 2016 about using a neural network to solve a problem.

This approach consists in converting the sound into an 80-dimensional audio spectrogram, which creates "frames" of the sound every 12.5 milliseconds and reflects various characteristics of the human voice, such as intonation.

This model is called Tacotron 2 and it is a combination of WaveNet and Tacotron.

To train the model, we converted the audio recordings made in the Georgian language into the so-called mel spectrograms (mel spectogram; mel - abbreviated comes from melody). The mentioned spectogram is used to better reflect the physiological capabilities of a person. Humans tend to perceive close frequencies as one because we don't need absolute precision. The higher the frequency, the greater the range within which it is difficult for us to distinguish frequencies. That's why we use the logarithmic conversion of hertz to mels.

We trained the model using convolutional and recurrent neural networks.

In order to create a perfect voice synthesizer, we also needed to create language-specific auxiliary modules, such as converting numbers to text for further vocalization.

In conclusion, it can be said that our group was able to create a high-quality voice synthesizer for the Georgian language. In the process of its implementation and practical use, we will continue to improve it.

ვირტუალური რეალობა, თამაშების გამოყენება ინკლუზიისთვის და ენის სწავლის ინტერაქცია

ტერეზა ტომაშევიჩი
მალმოს უნივერსიტეტი, შვედეთი
teresa.tomasevic@mau.se

ენის ონლაინსწავლების დროს სტუდენტებისთვის თანაკურსელებთან და მასწავლებლებთან ინტერაქციის შესაძლებლობის ნაკლებობა გამოწვევაა, რაც კოვიდ-19-ის პანდემიის დროს განსაკუთრებით ნათლად გამოჩნდა. ამ ნაკლებობის კომპენსაციისთვის 2021 წელს დავიწყეთ თამაშების გამოყენება (გეიმიფიკაცია) და ვირტუალური რეალობის ენის შესწავლის პროექტი, რომლის მიზანი იყო, შეგვექმნა:

1) სტუდენტებისთვის ვირტუალური სასწავლო გარემოს პროტოტიპი, რომელშიც ისინი ავატარებივით იმოქმედებდნენ, ანუ შეგვესწავლა სტუდენტის ქცევაზე ვირტუალური რეალობის გარემოს ეფექტი პანდემიის პირობებში;
2) ონლაინთამაშების პროტოტიპები, რომლებიც დაგვეხმარებოდა პედაგოგიური ამოცანების გადაჭრაში ზეპირი ინტერაქციის დროს და წარმოთქმაზე მუშაობისას.

ორივე ამოცანის მიზანი იყო, აგვენაზღაურებინა უნივერსიტეტის ტერიტორიაზე პანდემიის მიერ გამოწვეული შეზღუდვების გამო თანაკურსელებთან და მასწავლებლებთან ინტერაქციის შესაძლებლობის დანაკარგი.

პროექტის ფარგლებში თამაშის განვითარების საბაკალავრო პროგრამის სტუდენტებმა საბოლოო ინდივიდუალური პროექტის ფარგლებში შექმნეს ენის სწავლის თამაშების პროტოტიპები სამი სხვადასხვა ენისთვის. ვირტუალური რეალობის გარემოს პროტოტიპი შეიმუშავეს ინტერაქციის დიზაინის სამაგისტრო პროგრამის სტუდენტებმა, როგორც საბოლოო სამაგისტრო დისერტაციის პროექტი. ყველა პროტოტიპი მალმოსა და ვიადრინას უნივერსიტეტების (ოდერის ფრანკფურტი, გერმანია) სამიზნე ენების მასწავლებლებთან თანამშრომლობით შექმნეს მალმოს უნივერსიტეტის (მალმო, შვედეთი) სტუდენტებმა.

პროექტი მთავრდება 2023 წლის ივნისში. ჩვენი მიზანია, დარჩენილი დროის განმავლობაში პროდუქტები პედაგოგებთან და მოსწავლეებთან ერთად შევაფასოთ.

პრეზენტაციაში წარმოგიდგენთ პროექტს, ასევე, თამაშების პროტოტიპებს ენის სწავლისთვის და ვირტუალური რეალობის გარემოს პროტოტიპს.

Virtual Reality, Gamification for Inclusion and Language Learning Interaction

Teresa Tomašević
Malmö University, Sweden
teresa.tomasevic@mau.se

In online language teaching, the lack of opportunities for students to interact with fellow students and teachers is a challenge, something that became particularly clear during the Covid-19 pandemic. To compensate for that lack, a gamification and VR language learning project was initiated in 2021, with the objectives to create

1) a prototype of a virtual learning environment for students to act in as avatars or otherwise, exploring the (Virtual Reality) VR environment's effect on students' VR interaction behaviour during the Covid-19 pandemic, and
2) prototypes for online games, addressing the pedagogical challenge in training oral interaction and pronunciation online.

Both objectives were intended to compensate for the loss of opportunities to interact with fellow students and teachers on campus due to pandemic restrictions, and to be used in any online language teaching and learning.

Within the project, students from a Game Development Bachelor Programme have developed the intended prototypes for language learning games for three different languages as their individual final projects. The prototype for the VR environment was developed by students from a master programme in Interaction Design as their final master thesis projects. All the prototypes were developed by students from

Malmö University (Malmö, Sweden) in collaboration with teachers in the target languages from Malmö University and Viadrina University (Frankfurt an der Oder, Germany).

The project ends in June 2023, and during the remaining project time, the aim is to start testing and evaluating the products together with teachers and learners.

In this presentation, the project will be introduced, as well as the prototypes of the language learning games and the VR environment prototype.

სიტყვამნერგები ქართულისათვის და მათი გამოყენების ავკარგიანობა

ზაქარია ფურცხვანიძე, რიკარდო იუნგი
ემპირიული ენათმეცნიერების ინსტიტუტი,
ფრანკფურტის გოეთეს უნივერსიტეტი, გერმანია
pourtskhvanidze@em.uni-frankfurt.de, ric.jun@gmx.de

სიტყვის ჩანერგვის მეთოდი (Word embedding) გამოიყენება კორპუსლინგვისტური გზით გარკვეული სოციალური დისკურსების გამოვლენისა და აღწერისათვის. ბოლო დეკადაში მონაცემთა დიდი მასივების ავტომატური ანალიზი მზარდი პროგრესიით მიემართება ცნებას Word Embedding (სიტყვის ჩანერგვა). კონკრეტული ლექსემის ჩანერგვა მრავალგანზომილებიან ციფრულ სივრცეში, რომელიც მილიარდობით სიტყვა-ფორმის ვექტორულ რეპრეზენტაციას შეიცავს, იძლევა ჩანერგილი სიტყვის გაცილებით ზუსტი და კონტექსტიდან გამომდინარე სემანტიკური მნიშვნელობის დადგენის საშუალებას, ვიდრე ეს ოდესმე იყო შესაძლებელი.

დღესდღეობით, ქართული ენისათვის უკვე მინიმუმ ორი სიტყვამნერგია ღიად წვდომადი. მათი შექმნის ქრონოლოგიის გათვალისწინებით, პირველი სიტყვამნერგი სლოვაკეთის მეცნიერებათა აკადემიის ენათმეცნიერების ინსტიტუტში შეიქმნა. საუბარია რამდენიმე ათეული ენის მონაცემების შემცველ კორპუსზე GigaWord Web Corpora (Aranea), რომელიც 254M ტოკენის მოცულობის ქართული ინტერნეტის ენის კორპუსსაც შეიცავს. ეს მონაცემი 700T ცალკეული დოკუმენტისაგან 2014 წელს ჩამოიტვირთა ქსელიდან. რიგით მეორე სიტყვამნერგი შეიქმნა ფრანკფურტის გოეთეს უნივერსიტეტის ემპირიული ენათმეცნიერების ინსტიტუტის ერთი კვლევითი პროექტის ფარგლებში, სადაც ანალიზი 1,5 მილიარდ ქართულ სიტყვა-ფორმას ეყრდნობა. მესამე სიტყვანმერგი Embedding Projector წარმოადგენს იმ შემთხვევას, სადაც ნებისმიერი ენის მკვლევარს შეუძლია თავად გარდაქმნას საკუთარი კორპუსი ვექტორულ სივრცედ და ჩანერგოს მასში ნებისმიერი სიტყვა-ფორმა.

ამ სამი ინსტრუმენტის ბაზაზე პრეზენტაციაში ნაჩვენებია სოციალური დისკურსების ამოცნობის, მათი მნიშვნელობის გაგებისა და აგრეთვე მათი განსურათების მაგალითები. გარკვეული სოციოკულტურულად განსაზღვრული სოციალური იერარქიების მანიფესტაციას ხშირად წარმოადგენენ ენობრივი ფორმულირებები, რომლებიც სოციალური ინტერაქციის სიხშირის პარალელურად მაღალი ენობრივი სიხშირითაც გამოირჩევიან. უფრო ღრმა დაკვირვებით აღმოჩნდება, რომ ამგვარი ენობრივი ფორმულირებები სოციალური დისკურსების მარკერების ფუნქციას ითავსებენ. ამ მარკერების გამოყენება და მათი ადეკვატური გაგება საშუალებას აძლევს კომუნიკაციის თანამონაწილეებს გახდნენ აქტუალური დისკურსის თანაზიარნი. დისკურსის ენობრივი მარკერები აღძრავენ და ინარჩუნებენ დისკურსს. ამგვარი მარკერი-ფორმულირებების დამატებით ფუნქციას წარმოადგენს ლინგვისტური თვისებების გამყარება, რაც საბოლოოდ მათ ენობრივი ყალიბის ფორმას აძლევს. მაგალითად გამოყენების მაღალი სიხშირის მქონე ფორმულირება „ქუდზე კაცი“ წარმოადგენს ტრადიციული ქართული საომარი მობილიზაციის მოწოდების შემოკლებულ (ელიფსურ) ფორმას. მის უკან დგას მთელი სოციოკულტურული დისკურსი, რომლის გასაგებად საკმარისი არ არის, ვიცოდეთ რას ნიშნავს „ქუდი“, „კაცი“ და „-ზე“. ამ გამოხატულების ჩანერგვით მრავალგანზომილებიან ვექტორულ სივრცეში მივიღებთ სემანტიკურად ახლო სიტყვაფორმების ღრუბელს, რომლიც ავლენს მთლიანი დისკურსის ლექსიკურ გარემოს (Chart1.).

Chart 1. Word embedding with idiomatic expression /kudze_k□aci/

Semantic similarity of words

Language: ka . Use diacritics (if applicable).

Similar to the word(s):

+ ქუდზე_კაცი

☒	word	count
0.000	ქუდზე_კაცი	71 G A W
0.600	საპროტესტოდ	64 G A W
0.631	საომრად	747 G A W
0.647	რუსი_ჯარი	459 G A W
0.662	ლაშქარიც	75 G A W
0.673	სალაშქროდ	219 G A W
0.675	მიტინგზე	1213 G A W
0.675	დასარბევად	137 G A W
0.684	სამშვიდობოს	424 G A W
0.689	ალყიდან	67 G A W
0.692	სამალავებიდან	29 G A W

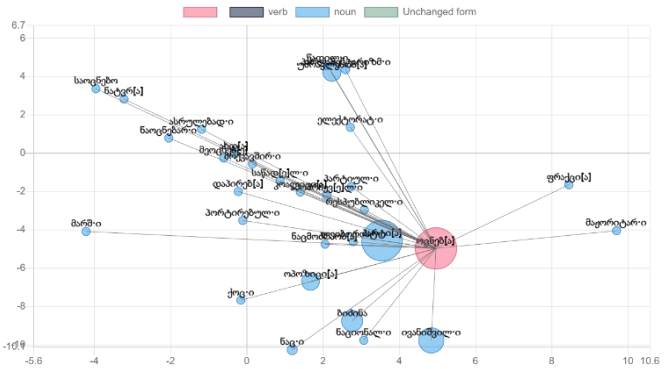
Chart 2.-ში მოცემულია სიტყვა-ფორმის „ოცნება“ ლექსიკოგრაფიული განსაზღვრება. მასთან შედარებით იმავე სიტყვა-ფორმის ჩანერგვა გვაძლევს მისი აქტუალური სოციოკულტურული მოხმარების ბაზაზე დაფუძნებულ მნიშვნელობას და შესაბამისად ავლენს ამ სიტყვაფორმით აღძრულ ქართული ენისათვის სპეციფიკურ დისკურსს (Chart 3.).

Chart 2. Definition of /ocneba/ „Dream“

ოცნება (ოცნებისა) 1. იმის წარმოდგენა, რისი განხორციელებაც სასურველია; სასიამოვნო, სანატრელ რამეზე ფიქრი. 2. გადატ. თვით ის, რაზედაც ან ვისზედაც ოცნებობენ, ოცნების საგანი.

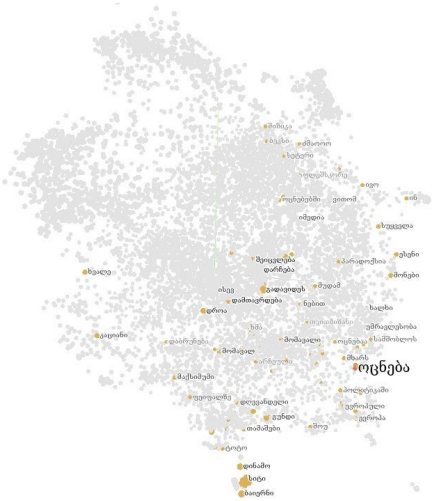
Chart 3. Embedding of /ocneba/ „Dream“

ნაცმოძრაობა	nacmoʒraob[a]
მეოცნებ[ე]	meocneb[e]
ოპოზიცი[ა]	opʼozici[a]
ნაციონალ·ი	nacional·i
პარტი[ა]	pʼart□i[a]
ქოც·ი	koc·i
ნაოცნებარ·ი	naocnebar·i
კოალიცი[ა]	k□□□oalici[a]
ნატვრ[ა]	nat□vr[a]
ახდ[ა]	axd[a]



Embedding Projector-ი წარმოადგენს მოქნილ სისტემას საკუთარი კორპუსული მონაცემების რეპრეზენტაციისათვის ვექტორულ სივრცეში და იქ ნებისმიერი სიტყვაფორმის ჩასანერგად. აქ შესასრულებელია წინარე სამუშაო იმ ფორმით, რომ მომხმარებელმა თავად უნდა განსაზღვროს ვექტორული პარამეტრები, გაწვრთნას ენობრივი მოდელი და მხოლოდ ამის შემდეგ მიმართოს ეს მოდელი ვექტორულ სივრცეზე. ამ საფეხურების გავლა შესაძლოა Gensim-ის პლატფორმაზე პროგრამირების მცირე ცოდნით ან შეგიძლიათ გამოიყენოთ ამ პრეზენტაციის ავტორების მიერ შექმნილი გარემო CoLabGoogle-ზე. Chart 4.-ში მოტანილია სიტყვა-ფორმა „ოცნება“-ს ვექტორული ღრუბელი Embedding Projector-ის ფარგლებში.

Chart 4. Embedding of /ocneba/ „Dream“ Projector



წარმოდგენილი სამი სიტყვამნერგი ქართული ენისათვის აჩვენებს ანალიზის სხვადასხვა შესაძლებლობას ერთი მეთოდის ფარგლებში. აქ გადამწყვეტი მნიშვნელობა ენიჭება გაწვრთნილი მასალის შინაარსს და ზომას.

სიტყვამნერგები იძლევა დიდი ტექსტური მასივების ავტომატური ანალიზის საშუალებას და ცალკეული სიტყვაფორმების, ფრაზების მიმართებას მათი გამოყენების აქტუალურ სოციო-კულტურულ დისკურსებთან.

ლიტერატურა

Bubenhofer, Noah. 2008. Diskurse berechnen? Wege zu einer korpuslinguistischen Diskursanalyse. In: Spitzmüller, Jürgen; Warnke, Ingo H. Methoden der Diskurslinguistik: sprachwissenschaftliche Zugänge zur transtextuellen Ebene. Berlin / New York: de Gruyter, S. 407-434.

Bubenhofer, Noah. 2009. Sprachgebrauchsmuster: Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. Berlin & New York: De Gruyter. S.30.

Bubenhofer, Noah 2013 Quantitativ informierte qualitative Diskursanalyse. Korpuslinguistische Zugänge zu Einzeltexten und Serien. In: Roth, Kersten Sven/Spiegel, Carmen (Hrsg.): Angewandte Diskurslinguistik. Felder, Probleme, Perspektiven. Berlin: Akademie-Verlag. S. 109-134. S.109.

Foucault, Michel 1974 Die Ordnung des Diskurses. Carl Hanser Verlag.

Goldberg, Yoav 2017 Neural Network Methods in Natural Language Processing (Synthesis Lectures on Human Language Technologies). Morgan & Claypool Publishers. P. 317.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean 2013 Distributed Representations of Words and Phrases and their Compositionality. In: Computation and Language. Pp. 1-9.

ინტერნეტწყაროები

http://unesco.uniba.sk/aranea/

https://www.cis.upenn.edu/~ungar/eigenwords/

https://wordembedding.spellchecker.ge/

https://research.facebook.com/fasttext/

https://radimrehurek.com/gensim/models/word2vec.html

https://nlp.stanford.edu/projects/glove/

https://projector.tensorflow.org/

https://ka.wikipedia.org/wiki/სიტყვამნერგი

Word Embedding Tools for Georgian - What are they good for?

Zakharia Pourtskhvanidze, Ricardo Jung
Institute for Empirical Linguistics, Goethe University Frankfurt, Germany
pourtskhvanidze@em.uni-frankfurt.de, ric.jun@gmx.de

Word embedding is used, inter alia, to identify and make visible certain social discourses corpus linguistically. In the last few years, the rapid development of automatic analysis of large amounts of data (Big Data) has centered around the term Word Embedding. Embedding a lexeme in a multidimensional digital space of billions of lexemes allows us to determine the semantic shading of the embedded meaning more precisely than ever before.

Meanwhile, there are two freely available Word Embedding tools for Georgian. According to chronological order, the first functional tool that emerges is the one that was constructed at the Institute of Linguistics at the Slovak Academy of Sciences. This is A Family of Comparable GigaWord Web Corpora (Aranea), which includes, among others, 254M tokens from the Georgian web. The data was crawled from the 700T documents in 2014.

Another word embedding tool was constructed as part of a research project on language and identity at the Institute of Empirical Linguistics in Frankfurt/M. Here, lemmas from Georgian are embedded in a database of 1.5 billion tokens and the results are output as a list or 2D graph.

In the presentation a third free tool Embedding Projector is shown, where the user himself can upload the pre-trained data and use it for the Word Embedding method. The data can be selected and compiled according to the user's own criteria.

Using the examples from all three tools, the scenarios of scientific use of the Word Embedding method will be shown and discussed. This includes the identification and visualization of key meanings of social discourses, determination of the precise contextual interpretations of neologisms as well as the synonym generating.

The maintenance of certain socio-culturally determined situational social hierarchies is based on the use of formulaic linguistic expressions, which according to the frequency of interactions, have high linguistic frequency. It can be observed that the socially current discourses lead such linguistic formulas as a kind of marking. Just the linguistic use of this marker makes clear for the participants of a communication the discourse that the current conversation is about. Thus, these discourse-marking linguistic expressions facilitate the initiation and conduct of a conversation. Another feature of such expressions is the establishment of a linguistic pattern, which is due to the formulaic nature of the expression. Thus, the high frequency idiomatic expressions cannot even be finished and fully realized for the corresponding discourse to be triggered. The expression /kudze kაci/ is a short form of the call to mobilization in the time of crisis or war. For understanding the discourse behind this expression, it is not enough to understand the separate meanings of /kudi/ "cap" and /kaci/ "man". Embedding this phrase in the multidimensional vectoral space generates the list of semantically similar terms that bring us closer to the meaning of the embedded phrase (Chart1.).

The generated list shows the dominance of discourse-triggering key concepts such as "to protest," "to war," or "Russian army," which reduce lexical meanings of "hat" or "man" to discourse-relevant significance.

Chart 1. Word embedding with idiomatic expression /kudze_kაci/

Semantic similarity of words

Language: ka . Use diacritics (if applicable).

Similar to the word(s):

+ ქუდზე_კაცი

☒	word	count	
0.000	ქუდზე_კაცი	71	G A W
0.600	საპროტესტოდ	64	G A W
0.631	საომრად	747	G A W
0.647	რუსი_ჯარი	459	G A W
0.662	ლაშქარიც	75	G A W
0.673	სალაშქროდ	219	G A W
0.675	მიტინგზე	1213	G A W
0.675	დასარბევად	137	G A W
0.684	სამშვიდობოს	424	G A W
0.689	ალყიდან	67	G A W
0.692	სამალავებიდან	29	G A W

In Chart 2. the definition of the lexeme "dream" is presented, as it is given in the dictionary. In contrast, the embedding of the same lexeme (Chart 3.) shows the meaning from the current usage not as an abstract cognitive activity, but completely politicized and integrated in the political discourse of Georgian.

Chart 2. Definition of /ocneba/ „Dream“

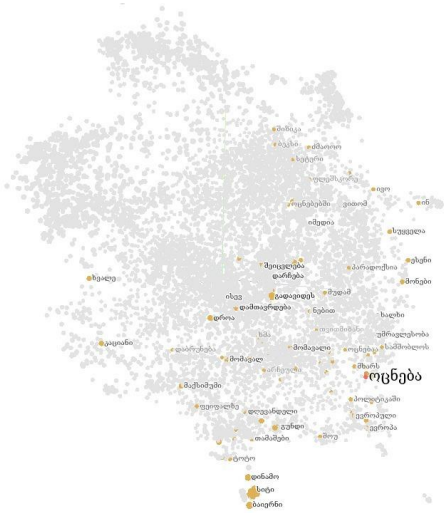
ოცნება (ოცნებისა) 1. იმის წარმოდგენა, რისი განხორციელებაც სასურველია; სასიამოვნო, სანატრელ რამეზე ფიქრი. 2. გადატ. თვით ის, რაზედაც ან ვისზედაც ოცნებობენ, ოცნების საგანი.

Chart 3. Embedding of /ocneba/ „Dream“

ნაცმოძრაობა	nacmoʒraob[a]
მეოცნებ[ე]	meocneb[e]
ოპოზიცი[ა]	opʼozici[a]
ნაციონალ·ი	nacional·i
პარტი[ა]	pʼart□i[a]
ქოც·ი	koc·i
ნაოცნებარ·ი	naocnebar·i
კოალიცი[ა]	k□□□oalici[a]
ნატვრ[ა]	nat□vr[a]
ახდ[ა]	axd[a]

A flexible option is the Embedding Projector tool, which is ready to upload your own trained data and analyze it from different points of view. The data must first be output as vector data so that the distances between the vectors can be calculated. The semantic similarity is based on this calculation. In Chart 4. shows the data from the Georgian-language "Tbilisi Forum", where the word /cximebi/ “Fats” is embedded. The clustered cloud of data points shows the composition of semantically similar concepts.

Chart 4. Embedding of /ocneba/ „Dream“Projector



Three-word embedding tools featured in the presentation describe the different perspectives of realizing word embedding for Georgian. They are based on differing data back and allow various degrees of complexity of the results.

The Word Embedding method gives us the possibility to machine process the large amounts of text and make their content structures accessible. Thus, an automatic classification of the content can be achieved more accurately. The presented three tools for Georgian process the different data of different sizes. However, they provide interpretable results and reproduce the size and content of the underlying data.

References

Bubenhofer, Noah. 2008. Diskurse berechnen? Wege zu einer korpuslinguistischen Diskursanalyse. In: Spitzmüller, Jürgen; Warnke, Ingo H. Methoden der Diskurslinguistik: sprachwissenschaftliche Zugänge zur transtextuellen Ebene. Berlin / New York: de Gruyter, S. 407-434.

Bubenhofer, Noah. 2009. Sprachgebrauchsmuster: Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. Berlin & New York: De Gruyter. S.30.

Bubenhofer, Noah 2013 Quantitativ informierte qualitative Diskursanalyse. Korpuslinguistische Zugänge zu Einzeltexten und Serien. In: Roth, Kersten Sven/Spiegel, Carmen (Hrsg.): Angewandte Diskurslinguistik. Felder, Probleme, Perspektiven. Berlin: Akademie-Verlag. S. 109-134. S.109.

Foucault, Michel 1974 Die Ordnung des Diskurses. Carl Hanser Verlag.

Goldberg, Yoav 2017 Neural Network Methods in Natural Language Processing (Synthesis Lectures on Human Language Technologies). Morgan & Claypool Publishers. P. 317.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean 2013 Distributed Representations of Words and Phrases and their Compositionality. In: Computation and Language. Pp. 1-9.

Links

http://unesco.uniba.sk/aranea/

https://www.cis.upenn.edu/~ungar/eigenwords/

https://wordembedding.spellchecker.ge/

https://research.facebook.com/fasttext/

https://radimrehurek.com/gensim/models/word2vec.html

https://nlp.stanford.edu/projects/glove/

https://projector.tensorflow.org/

განსაზღვრულობის დიაქრონიული ანალიზი ქართულში (ქართული ენის ეროვნული კორპუსზე დაყრდნობით)[1]

მარიამ ყამარაული
ხელნაწერთა კვლევის ცენტრი
ჰამბურგის უნივერსიტეტი, გერმანია
mariam.kamarauli@uni-hamburg.de

ქართულ დამწერლობას 1500 წელზე მეტი ხნის წერილობითი ტრადიცია გააჩნია: V საუკუნის უძველესი წარწერებიდან მოყოლებული ქართული ენა დღემდე უწყვეტად არის დოკუმენტირებული, რაც ემპირიულ მასალაზე დაყრდნობით ქართული ენის პერიოდიზაციისა და დიაქრონიული ანალიზის უნიკალურ შესაძლებლობას იძლევა. განსაზღვრულობის კატეგორია და მისი გამოხატვის ენობრივი საშუალებები, განსაკუთრებით არტიკლი, ერთ-ერთ ყველაზე საინტერესო

[1] ეს პუბლიკაცია არის სამეცნიერო პროექტის ნაწილი, რომელმაც მიიღო დაფინანსება ევროპის კვლევის საბჭოსგან (ERC) **ევროკავშირის ჰორიზონტი 2020** კვლევისა და ინოვაციების პროგრამის ფარგლებში (საგრანტო ხელშეკრულება No. 101019006).

პრობლემად არის მიჩნეული ქართველოლოგიაში. განსაზღვრული არტიკლები, რომლებიც ძირითადად ძველ ქართულში დასტურდება, ჩვენებითი ნაცვალსახელების სპეციფიკური გამოყენების გზით რეალიზდება მორფოსინტაქსური და ტიპოლოგიური ცვლილების ფონზე. მაგალითად, 1. ძველი ქართული ენა არტიკლის ფუნქციით ძირითადად მე-3 დეიქტური საფეხურის ჩვენებით ნაცვალსახელს ***იგი*** იყენებს, თუმცა იმავე ფუნქციით ასევე გამოიყენებიან სხვა ჩვენებითი ნაცვალსახელებიც (თუმცა გაცილებით ნაკლებად, ვიდრე *იგი*); 2. ჩვენებითი ნაცვალსახელის დისტრიბუცია სახელთან მიმართებით განსაზღვრავს მათ ფუნქციონირებას: სახელის მომდევნო (პოსტნომინალურ) პოზიციაში ისინი გამოიყენებიან განსაზღვრული არტიკლის ფუნქციით; სახელის წინა (პრენომინალურ) პოზიციაში კი ინარჩუნებენ ჩვენებითი ნაცვალსახელის ფუნქციას; 3. არტიკლი უმეტესად იკავებს მეორე (ე.წ. ვაკერნაგელის) პოზიციას სახელურ ფრაზაში (NP); არტიკლის აღნიშნული პოზიცია შეიძლება სახელური ფრაზის ბოლო პოზიციას წარმოადგენდეს, თუ სახელური ფრაზა მხოლოდ არსებითი სახელისა და არტიკლისაგან შედგება; 4. თუ NP-ის საზღვრული დგას სახელობით ბრუნვაში და მრავლობით რიცხვში, არტიკლი არ ეთანხმება მას რიცხვში და ჩვეულებრივ, მხოლობითი რიცხვის ფორმით რეალიზდება. თუმცა, მე-10 საუკუნიდან დასტურდება შეთანხმების განსხვავებული წესები, როდესაც მრავლობით რიცხვში მდგარ მსაზღვრელს არტიკლიც მრავლობით რიცხვში შეეწყობა, მაგალითად:

(1)	*და*	***მეფე-ნ-ი***	***იგი-ნ-ი***	***მათ-ნ-ი***	*რომელ-თა*
	And	king-PL-NOM	the-PL-NOM	their-PL-NOM	which-DAT.PL
	უპყრიეს		*ქალაქი*	*ჩუენ-ი*	
	conquer.S3PL.O3SG.PRES		city-NOM.SG	our-NOM.SG	

(*Timothy of Antioch*, 364, 11)

განსაზღვრულ არტიკლს შეუძლია, მაგალითად, NP-ის ბოლოსაც რეალიზდეს, მაგრამ ასეთ შემთხვევაში, ადგილი არა აქვს ძველი ქართული ბრუნების სისტემისათვის დამახასიათებელ განსაკუთრებულ ფენომენს *Suffixaufnahme*-ს (პოსტპოზიციური ატრიბუტული დეგენეტიური არსებითი სახელები, ზედსართავები ან ნაცვალსახელები, დამატებით ეთანხმებიან საზღვრულს ბრუნვასა და რიცხვში). მაგალითად:

(2)	*ხოლო*	*რაჟამ-ს*	*მოვიდეს*	***ო(ჳფალ)-ი***	***იგი***
	but	when-DAT.SG	come.S3SG.CONJ	lord-NOM.SG	the.NOM.SG
	ვენაჯ-ისა-ჲ		***მის***		
	vineyard-GEN.SG-NOM.SG		the.GEN.SG		

(Mt., 21, 40, Khanmeti Gospels)

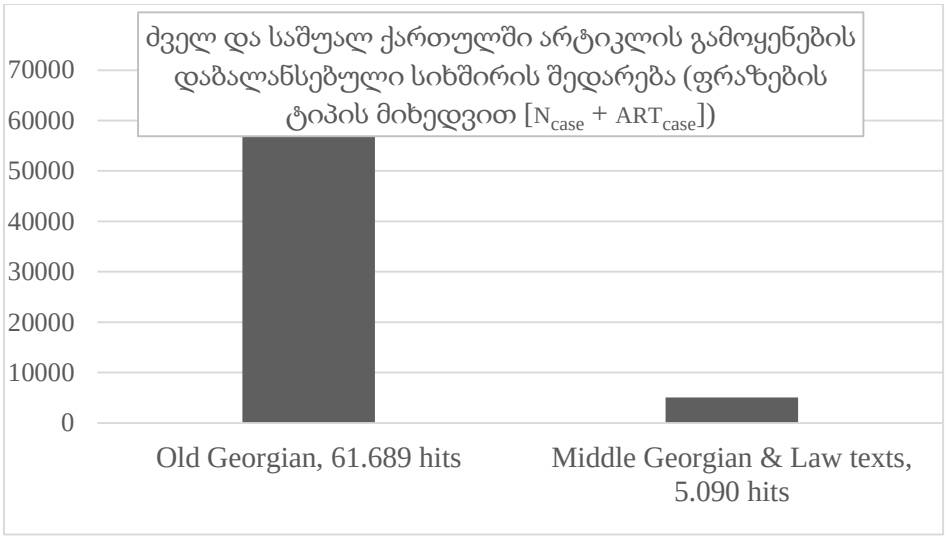
NP-ის *ო(ჳფალ)ი იგი ვენაჯისაჲ მის* ბოლო უშუალო შემადგენელი არის განსაზღვრული არტიკლი ***მის***, რომელიც ნათესაობით ბრუნვაში დგას. არადა, მოსალოდნელი იყო: *ო(ჳფალ)ი იგი*

ვენაჯისად მისი. მაგრამ, როდესაც არტიკლი სახელური ფრაზის ბოლო პოზიციას იკავებს, რაც ძალიან იშვიათად გვხვდება, ***Suffixaufnahme*** ხორციელება NP-ის ბოლო არაკლიტიკურ ელემენტზე, მაგ., არსებით სახელთან, ზედსართავთან ან კუთვნილებით ნაცვალსახელთან. ამგვარად, ძველ ქართულში დადასტურებული განსაზღვრული არტიკლის ბოლო მახასიათებელი ნიშანი ასე შეგვიძლია ჩამოვაყალიბოთ: იმ შემთხვევაში, როდესაც განსაზღვრული არტიკლი სახელური ფრაზის ბოლო პოზიციას იკავებს, ***Suffixaufnahme***-ს წესებს ემორჩილება ბოლო არაკლიტიკური ელემენტი, ხოლო მისი მომდევნო არტიკლი რეალიზდება ადნომინალური ბრუნვის ფორმით.

ჩვენებითი ნაცვალსახელები საშუალ ქართულში დისტრიბუციის იმავე წესებს გვიჩვენებენ ფუნქციური თვალსაზრისით, როგორც ძველ ქართულში: სახელურ ფრაზაში პრეპოზიციურად რეალიზებულები ფუნქციონირებენ როგორც ჩვენებითი ნაცვალსახელები, პოსტპოზიციურად რეალიზებულები კი როგორც განსაზღვრული არტიკლი. თუმცა, მკვეთრი ცვლილება შეინიშნება სიხშირული თვალსაზრისით, კერძოდ: ამგვარი სტრუქტურის სახელური ფრაზა **კვანტორი + არტიკლი + არსებითი სახელი** [QUANTIFIER + ARTICLE + NOUN; მაგ., *ყოველნი იგი სიტყუანი*, სადაც სახელი მრავლობით რიცხვში დგას, არტიკლი კი მხოლობითში] ძველ ქართულში გაცილებით ხშირად გვხვდება (1.269 კონტექსტი), მაშინ როდესაც, საშუალი ქართულის ქვეკორპუსში ამ ტიპის სტრუქტურა მხოლოდ 11-ჯერ არის დადასტურებული (საშუალი ქართული კორპუსი და იურიდიული ქვეკორპუსი).

სიხშირული მაჩვენებლის ამგვარი შემცირება იმით უნდა აიხსნას, რომ ჩვენებითი ნაცვალსახელის გამოყენება განსაზღვრული არტიკლის ფუნქციით ძველ ქართულში თავდაპირველად ბიბლიურ ტექსტებში (ზოგადად თარგმანებში) ჩნდება და კარგავს გრამატიკულ ფუნქციას საშუალ ქართულში, ვინაიდან იგი ქართული ენის შინაგანად განპირობებულ მახასიათებელს არ წარმოადგენდა. ეს კი იმაზე მიგვითითებს, რომ ჩვენებითი ნაცვალსახელის განსაზღვრებითი არტიკლის ფუნქციით გამოყენება ბიბლიის ტექსტების თარგმნით იყო გამოწვეული, რომელიც ტექსტის ზედმიწევნით ზუსტ, სიტყვასიტყვით თარგმანს მოითხოვდა. ძველ ქართულში ეს ფუნქცია ჩვენებითმა ნაცვალსახელმა აიღო თავის თავზე, ვინაიდან სხვა არც ერთი ელემენტი ისე ახლოს არ იყო სემანტიკურ-ფუნქციური თვალსაზრისით არტიკლთან, როგორც ჩვენებითი ნაცვალსახელი. ძველი ქართულისაგან განსხვავებით, რომელიც წერილობითი ენით შემოიფარგლებოდა, საშუალი ქართული სალაპარაკო ენას ასახავდა და არ იყენებდა არტიკლს, ვინაიდან არ გააჩნდა იგი. ჩვენი ვარაუდი, რასაკვირველია, ჰიპოთეტურია. ამიტომ კონკრეტული მტკიცებულებისთვის საგანგებო კვლევა ჩავატარეთ ქართული ენის ეროვნულ კორპუსში, სადაც ძველ და საშუალ ქართულში (იურიდიული ტექსტების ქვეკორპუსის ჩათვლით) ვეძებდით NP-ების ძალიან მარტივ ტიპს, რომელიც შედგება [N_{case} + ART_{case}]-ისგან. ძველი ქართულის ქვეკორპუსი გვიჩვენებს 61,689 შემთხვევას აღნიშნული NP სტრუქტურისთვის, მაშინ, როცა იგივე NP სტრუქტურა მცირდება 627-მდე საშუალ ქართულში და 1832-მდე სამართლის ტექსტების ქვეკორპუსში. რა თქმა უნდა, აქ გასათვალისწინებელია ქვეკორპუსის მოცულობა: მაშინ, როცა ძველი ქართული ენის ქვეკორპუსი მოიცავს 6 062 122 ტოკენს, საშუალი ქართულის ქვეკორპუსს ამ მოცულობის მეოთხედიც კი არ გააჩნია (1 432 262 ტოკენი). ნათქვამი ეხება სამართლის ტექსტების ქვეკორპუსსაც (1 495 985 ტო-

კენი). ამიტომ, კვლევაში გამოყენებული უნდა იყოს დამაბალანსებელი ფაქტორი 2.07 (ძველი ქართული ენის ქვეკორპუსის მოცულობა გაყოფილი საშუალი ქართულის ქვეკორპუსის მოცულობაზე პლუს სამართლის ტექსტების მოცულობა; 6,062,122 / 2,928,247 = 2,07). კვლევის შედეგად მიღებული მიმართებები ილუსტრირებულია ქვემოთ:



მაშინაც კი, თუ საშუალი ქართულისა და სამართლის ტექსტების ქვეკორპუსებიდან მიღებულ შედეგებს გადავამრავლებთ დამაბალანსებელ ფაქტორზე 2.07, პროტოტიპური ფრაზის [Ncase + ARTcase] კლება მაინც 91%-ზე მეტია – და, აქედან გამომდინარე, უნდა შეფასდეს, როგორც რადიკალური, იმისათვის, რომ არტიკლი მივიჩნიოთ ქართული ენის „ბუნებრივ“ და ორიგინალურ კომპონენტად. განსაზღვრული არტიკლებისგან განსხვავებით, განუსაზღვრელი არტიკლის (პოსტპოზიციური „ერთი“ (NOM.SG)) და განუსაზღვრელი ნაცვალსახელების (**რა(ი)მე** (NOM.SG), **ვინმე** (NOM.SG)) სიხშირე, პირიქით, გაიზარდა:

- რიცხვითი სახელის ***ერთი***, როგორც განუსაზღვრელი არტიკლის ფუნქციონირების სიხშირე მარტივ NP-ებში [Ncase + ერთიcase] ძველ ქართულ ქვეკორპუსში შეადგენს 2492 შემთხვევას (აბსოლუტური სიხშირე), ხოლო საშუალი ქართულისა და სამართლის ტექსტების ქვეკორპუსის შედეგების მიხედვით – 1647-ს (აბსოლუტური სიხშირე) → დამაბალანსებელი ფაქტორის გამოყენების შემდეგ განუსაზღვრელი არტიკლის სიხშირე გაიზარდა 36%-მდე (შეფარდებითი სიხშირე საშუალ ქართულში: 3409 შემთხვევა);
- მიუხედავად იმისა, რომ განუსაზღვრელი ნაცვალსახელები ***ვინმე*** (NOM.SG) და ***რაიმე*** (NOM.SG) ბრუნვაფორმების გათვალისწინებით 11315-ჯერ გვხვდება ძველი ქართული ენის ქვეკორპუსში, საშუალ ქართულსა და სამართლის ტექსტების ქვეკორპუსებში მისი სიხშირე დაეცა 8111-მდე (აბსოლუტური სიხშირე, კომბინირებული მონაცემი).

დამაბალანსებელი ფაქტორის გათვალისწინებით კი, შედეგები 8111-დან 16790-მდე იზრდება, რაც იმას ნიშნავს, რომ სიხშირე გაიზარდა 48%-ით.

ჩვენი ვარაუდით, განუსაზღვრელობა ყოველთვის იყო ქართული ენისათვის შინაგანად განპირობებული კატეგორია, განსაზღვრულობა კი არა. განუსაზღვრელობის კატეგორია ყოველთვის იყო წარმოდგენილი ნაცვალსახელებით ***რაიმე*** და ***ვინმე***, მაშინ როდესაც ჩვენებით ნაცვალსახელებს სჭირდებოდათ დამატებითი სინტაქსური მახასიათებელი: პრეპოზიციურად გამოყენებულნი, ისინი ფუნქციონირებენ, როგორც ჩვენებითი ნაცვალსახელები, პოსტპოზიციურად გამოყენებულნი კი - როგორც არტიკლი.

A Diachronic Analysis of Definiteness in Georgian Based on the GNC [1]

Mariam Kamarauli
Centre for the Study of Manuscript Cultures
University of Hamburg, Germany
mariam.kamarauli@uni-hamburg.de

The Georgian language can look back on an uninterrupted written tradition of more than 1500 years (the oldest written evidence are inscriptions from the 5th century CE), it is ideal for diachronic analyses. A distinction is usually made between an Old Georgian (approx. 5th - 11th century), a Middle Georgian (approx. 12th - 18th century) and a Modern Georgian period (since the 18th century). The topic of definiteness and its representatives, especially articles, is counted among the most interesting research areas. These definite articles, which were mostly present and functional in Old Georgian, are identical to demonstratives aside some morphosyntactic and typological changes, e.g. 1. as an article, Old Georgian mostly uses the demonstrative pronoun *igi* (3rd-level deictic) but the other demonstratives appear in this function, too (though much less frequently than igi); 2.when placed postnominal, they function as definite articles; in prenominal placement, they maintain their function as demonstratives; 3. articles mostly take the second position (also called Wackernagel position) in the NP (which can be simultaneously the last position if the NP consists only of a noun and an article); 4. within an NP in the nominative, if the head of the article is marked for plural, the article typically stays in the singular. Nonetheless, a few examples with the head and article agreeing in the plural can be found (from the 10th century):

[1] This publication is part of a project that has received funding from the European Research Council (ERC) under the **European Union's Horizon 2020** research and innovation programme (Grant agreement No. 101019006).

(1)	*da*	***mepe-n-i***	***igi-n-i***	***mat-n-i***	*romel-ta*
	and	king-PL-NOM	the-PL-NOM	their-PL-NOM	which-DAT.PL
	u̯p̣qries		*kalak-i*	*čuen-i*	
	conquer.S3PL.O3SG.PRES		city-NOM.SG	our-NOM.SG	

'And their kings (*lit.* **the kings of theirs**), who conquered our city.' (*Timothy of Antioch*, 364, 11)

The definite article can take, e.g., the last position within the NP but in that case, *suffixaufnahme* (the phenomenon of postpositioned attributive nouns (in the genitive), adjectives or pronouns adding (and thus duplicating) the case endings of preceding nouns) will not be executed:

(2)	*xolo*	*ražam-s*	*movides*	***owpal-i***	***igi***
	but	when-DAT.SG	come.S3SG.CONJ	lord-NOM.SG	the.NOM.SG
	venaq-isa-j		***mis***		
	vineyard-GEN.SG-NOM.SG		the.GEN.SG		

'But when **the owner of the vineyard** comes […]' (Mt. 21, 40, Khanmeti Gospels)

The last constituent of the NP *owpali igi venaqisaj mis* 'the owner (NOM.SG) (of) of the vineyard (GEN.SG + NOM.SG)' is the article *mis*, which is only marked for genitive singular (adnominal case marking). What we would expect is:

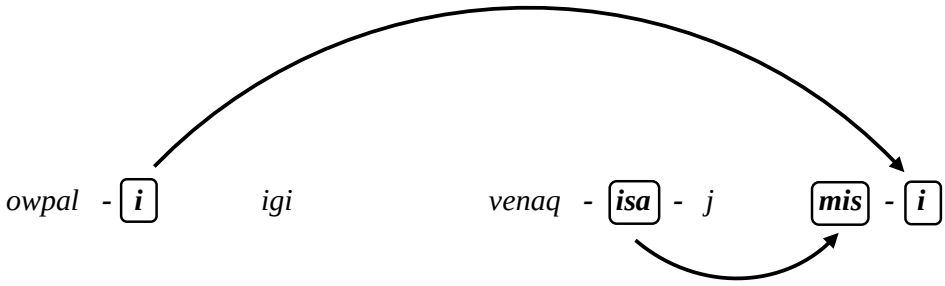


Figure 1

But obviously, Old Georgian does not tolerate *suffixaufnahme* on articles and finds a remedy in performing it on the last noun, adjective or possessive pronoun in the NP:

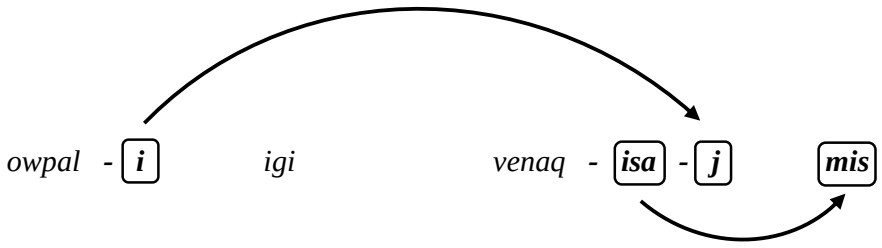


Figure 2

Thus, formulating the last feature of definite articles in Old Georgian: within an NP, where the article takes the last position, it is the last non-clitic element of the NP that receives the *suffixaufnahme*, and the article following it, receives only adnominal case marking.

Demonstratives in Middle Georgian differ in the same way as in Old Georgian when it comes to their placement: prenominally placed, they function as demonstratives, postnominally as articles. However, a drastic change is noticable in the frequency of definite articles: the NP structure [QUANTIFIER + ARTICLE + NOUN; e.g. *ǫovelni igi sit̆quani* 'all the words (NOM.PL)'], which is quite frequent in Old Georgian with 1,269 examples, decreases in Middle Georgian to 11 examples (Middle Georgian and Law texts subcorpora). This significant decrease could be explained with the loss of the definite article, which firstly appeared in biblical texts and other translations in Old Georgian but lost its grammatical function in Middle Georgian because it was not a "natural" component of the Georgian language. This presupposes that the article in Old Georgian emerged with the translations, for which it was needed: when the translators saw that the biblical texts of Ancient Greek had a definite article and intended their translations to be accurate (word-by-word), an element was needed that fulfilled the same function in Old Georgian, and no element was closer to that than the demonstrative pronoun. If the article in Old Georgian first appeared in translations, it may have remained limited to the written language, whereas Middle Georgian reflects the spoken language which had no article. It is true that this assumption remains hypothetical; for concrete proof separate research would have to be performed.

For a concrete analysis, I searched in the Old and Middle Georgian (including the subcorpus of Law texts) for the very simple type of NPs consisting of [N_{case} + ART_{case}]: the Old Georgian subcorpus shows 61,689 hits for the NP structure [N_{case} + ART_{case}] while the same NP structure is reduced to 627 hits in Middle Georgian and to 1,832 hits in the subcorpus for Law texts. Of course, the size of the subcorpora must be considered here: while the Old Georgian subcorpus comprises 6,062,122 tokens, the Middle Georgian subcorpus has not even a quarter of that amount (1,432,262 tokens), and the same is true for the subcorpus of Law texts (1,495,985 tokens). Thus, the **balancing factor 2.07** (the size of the Old Georgian subcorpus divided by the size of the Middle Georgian subcorpus plus that of the Law texts; 6,062,122 / 2,928,247 = 2.07) must be applied. The resulting relation is illustrated below:

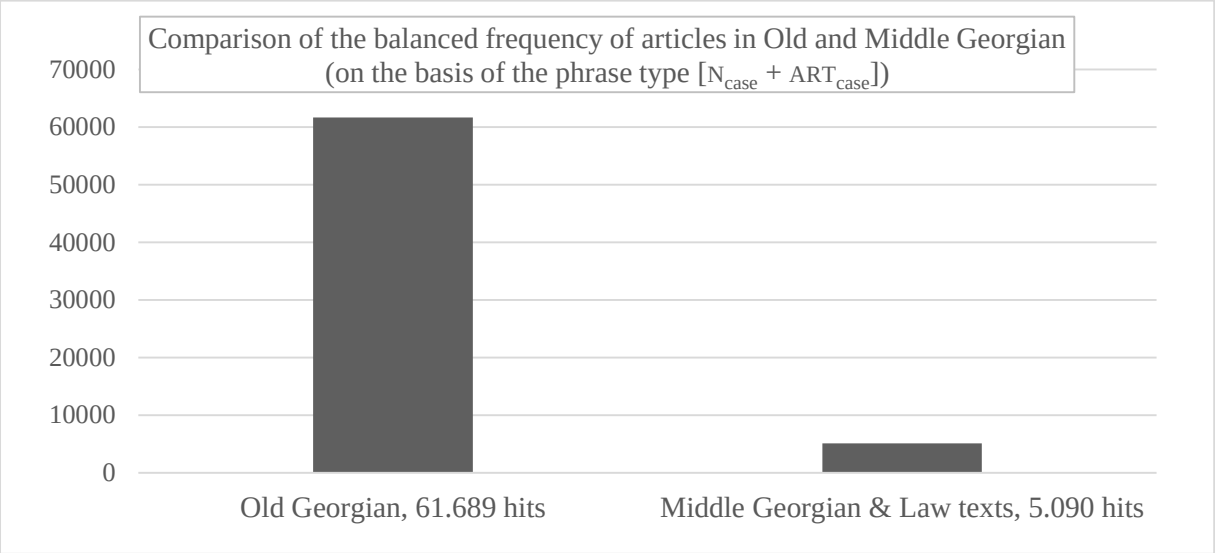


Figure 3 (Kamarauli 2022: 51)

Even if the results from Middle Georgian and Law texts subcorpora are multiplied with the balancing factor of 2.07, the decrease of the prototypical phrase [Ncase + ARTcase] is still at over 91% – and thus simply too radical if the article had been a "natural" and original component of the Georgian language.
In contrast to definite articles, the frequency of indefinite article (postpositioned numeral *erti* 'one (NOM.SG)') and indefinite pronouns (*ra(j)me* 'something (NOM.SG)', *vinme* 'someone (NOM.SG)') increased:

- the frequency of the numeral *erti* functioning as an indefinite article in simple NPs [Ncase + *erti*case] in the Old Georgian subcorpus amounts to 2,492 hits (absolute frequency), while the results from the Middle Georgian and Law texts subcorpora amount to 1,647 hits (absolute frequency) → after applying the balancing factor, the frequency of the indefinite article increased by over 36% (relative frequency in Middle Georgian: 3,409 hits);
- While the indefinite pronouns *vinme* 'someone (NOM.SG)' and *rame* 'something (NOM.SG)' with their declined forms appear 11,315 times in the Old Georgian subcorpus, the frequency dropped to 8,111 (absolute frequency) in the Middle Georgian and Law texts subcorpora (combined) – with the balancing factor, the results change from 8,111 hits to 16,790, which means that the frequency increased by 48% (the increase of the indefinite pronouns can be explained by the grammaticalized forms *vinme* and *rame* and their declension).

In my opinion, indefiniteness has always been an inherent category of the Georgian language while definiteness has not. The category of indefiniteness has always been represented by *ra(j)me* 'something (NOM.SG)', *vinme* 'someone (NOM.SG)', while the demonstrative pronouns *ese/ege/igi* needed a special syntactic feature (being placed postnominally) for their additional functions as definite article.

MLU-პროტოკოლის დადგენისათვის ქართულში

თინათინ ჭინჭარაული, ნინო ცინცაძე, ნინო დობორჯგინიძე, თეონა დამენია, თამარ კალხიტაშვილი

ილიას სახელმწიფო უნივერსიტეტი, საქართველო
tinatin.chincharauli@iliauni.edu.ge, nino.tsintsadze@iliauni.edu.ge,
nino_doborjginidze@iliauni.edu.ge, teona.damenia@iliauni.edu.ge,
kalkhitashvili.tamar@gmail.com

ბავშვის გრამატიკული განვითარების შესაფასებლად ძალზე მნიშვნელოვანია სისტემური კოდირების წესების არსებობა, რომლის მიხედვითაც ასახული იქნება კონკრეტული ენის თავისებურებები. ქართული ენის შემთხვევაში გასათვალისწინებელია, მისი აგლუტინაციური ბუნება და პოლიპერსონალური შეთანხმების არსებობა. ეს კიდევ ერთხელ უსვამს ხაზს კოდირების ერთიანი სისტემის შექმნის აუცილებლობას, რაც მნიშვნელოვანია ბავშვებში ქართული ენის ათვისების შესასწავლად და ენისა და მეტყველების თერაპიის თვალსაზრისითაც.

როჯერ ბრაუნის კლასიკური ნაშრომის „პირველი ენა: ადრეული ეტაპები“ (Brown 1973) მიხედვით, მორფემების რიცხვის გამოთვლა ერთი გამონათქვამის ფარგლებში (MLUm), ენის ათვისების საზომად ფართოდ გამოიყენებოდა ენის სპონტანური ნიმუშების შესასწავლად. ის ერთ-ერთ საიმედო მაჩვენებლად იყო მიჩნეული ინგლისური ენის ნიმუშების შესწავლისათვის (Brown, 1973; Miller & Chapman, 1981). მაგრამ ინგლისურისგან განსხვავებული მორფოლოგიის მქონე ენების შესწავლამ ცხადყო, რომ აღნიშნული საზომის უცვლელად გამოყენებისას ვლინდება გარკვეული სირთულეები. სწორედ ამიტომ დადგა ამგვარი კვლევის წარმოების აუცილებლობა. კვლევის მიზანს წარმოადგენს MLU-ს გამოთვლის მეთოდის შექმნა, რომლის გამოყენებაც შესაძლებელი იქნებოდა ქართულენოვან ბავშვებში ენის ათვისების ადრეული სტადიის შეფასების საზომად.

კვლევა ჩატარდა გრძივი სექციათაშორისი დიზაინის გამოყენებით. კვლევაში მონაწილეობა მიიღო ოთხმა ქართულენოვანმა, ტიპური ენობრივი განვითარების ბავშვმა, ორმა გოგონამ და ორმა ბიჭმა. ორი მათგანი კვლევის დასაწყისში იყო 12 თვის ასაკისა, ხოლო ორი – 36 თვის ასაკისა. კვლევის დასაწყისში ყველა მათგანი შეფასდა AEPS®-ით, რათა დადასტურებულიყო მონაწილეთა ტიპური განვითარება. აღსანიშნავია, რომ მიღებული მონაცემების სარწმუნოობის უზრუნველსაყოფად გამოყენებულია დამატებითი კრიტერიუმები: ენის ნიმუშების გადაღება ხდებოდა მშობლების ან აღმზრდელის მიერ ორი წლის განმავლობაში. მათ დეტალური ინსტრუქციები მიიღეს ვიდეო გადაღების შესახებ. ჩაწერა მიმდინარეობდა ყოველთვიურად 4-5 საათის განმავლობაში. ვიდეოგადაღება მიმდინარეობდა ბუნებრივ გარემოში, სპონტანურ სამეტყველო სიტუაციებში ოჯახის წევრებთან ან ახლო გარემოცვის პირებთან. ყველა მონაწილე მიჰყვებოდა ვიდეოს ჩაწერის ერთსა და იმავე სტრუქტურას: თვეში

ერთი კვირის განმავლობაში იღებდნენ ვიდეოს ჭამის, თამაშის, წიგნის კითხვის ან სხვა აქტივობების დროს. ჩაწერილი ფაილები შეგროვდა და მოხდა მათი ტრანსკრიბირება. მონაცემთა ბაზა მოიცავს თითქმის 480 საათის ჩანაწერებს, რომლებიც ტრანსკრიბირებულია CHAT-ის ფორმატში, CHILDES-ის სტრანდარტების შესაბამისად (MacWhinney & Snow, 2000). ჩანაწერების ხანგრძლივობა განსხვავებულია ბავშვებისა და სესიების მიხედვით.

გასაანალიზებლად აღებულია პირველი 100 გამონათქვამი, რომელიც სრულიად გასაგები და პროდუქტიულია, რაც ნიშნავს, რომ არის კონტექსტის შესაბამისი და აღქმადი.

კვლევის შედეგად დადგინდა MLU-ს გამოთვლის წესები და ეს წესები გამოყენებულია 18-დან 36 თვის ასაკის ბავშვების მეტყველებიდან აღებული გამონათქვამების შესაფასებლად.

მთელ რიგ მეთოდოლოგიურ გამოწვევებს შორის, ერთ-ერთი მთავარი გამოწვევა იყო სიტყვებისთვის ამოსავალი ფორმების განსაზღვრის საკითხი, რაც უმნიშვნელოვანესი საკითხია ბავშვებში MLU-ს გამოთვლისათვის. ქართული ენა მდიდარია ფლექსიური ფორმებით, ეს შესაძლოა იყოს სახელადი ან ზმნური ფორმები, ამიტომ მათთვის ამოსავალი ფორმის განსაზღვრა უნდა მოხდეს წინასწარ შეთანხმებული წესის მიხედვით. ნორმატიული გრამატიკისგან განსხვავებით, ბავშვის ენის შესწავლისას ამოსავალი ფორმა ყოველთვის არ ემთხვევა საბაზისო მორფოლოგიურ ფორმას, ლემას. არამედ ეს არის ადრეულ ასაკში ათვისებული პროდუქტიული ფორმა, რომელიც მაღალი სიხშირული მაჩვენებლით გამოირჩევა. მონაცემების მიხედვით, 12 თვიდან 18 თვემდე ასაკობრივ პერიოდში ბავშვების მეტყველებაში გამოყენებული სახელების 100% სახელობითი ბრუნვით არის გაფორმებული. ზმნების შემთხვევაში კი 13% არის აწმყოს ფორმა, ხოლო 87% კი ბრძანებითის ფორმები. ყოველივე ამის გათვალისწინებით, სახელებისთვის (არსებითი სახელი, ზედსართავი სახელი, რიცხვითი სახელი, ნაცვალსახელი) ამოსავალ ფორმად განისაზღვრა სახელობითი ბრუნვის ფორმები. ხოლო ზმნებისთვის ამოსავალ ფორმად განისაზღვრა არა მორფოლოგიურად საბაზისო ფორმა, არამედ სიხშირის მიხედვით გამორჩეული ფორმა. როგორც მოსალოდნელიც იყო, ადრეული ზმნური ფორმები ბავშვის მეტყველებაში წარმოდგენილია სუბიექტური წყობის აქტიური ფორმის, მეორე პირის დიალოგური ბრძანებითის/კითხვითი ფორმებით.

მეორე სადავო საკითხს წარმოადგენდა მიმღეობების საკითხი, რომლებსაც ზედსართავი სახელების მსგავსი ფუნქცია აქვთ კომუნიკაციაში, მაგრამ ზმნური წარმომავლობის არიან. სადავო იყო მაწარმოებელი აფიქსების დათვლის საკითხი. ქართულში მიმღეობის მაწარმოებლად გვევლინება აფიქსები: -ული, სა-ელი. მაგალითად: გაკეთებ-ული /gɑkˈɛtʰɛb-uli/, გა-სა-კეთებ-ელი /gɑ-sɑ-kˈɛtʰɛb-ɛli/. დისკუსიების შედეგად ჩამოყალიბდა მიდგომა, რომ რადგანაც ენის ათვისებისას არ ხდება ამ აფიქსების მექანიკური გამეორება, არამედ ბავშვი მათ იყენებს მხოლოდ გარკვეული გრამატიკულ ცოდნაზე დაყრდნობით, ამიტომ, ის შეიძლება ჩაითვალოს მორფოლოგიური განვითარების გარკვეული საფეხურის ინდიკატორად, შესაბამისად დამატებითი ქულით უნდა მოხდეს ამ აფიქსების აღნიშვნა.

მეთოდოლოგიურად პრობლემური აღმოჩნდა აგრეთვე კუმშვისა და კვეცის შემთხვევების დათვლის საკითხიც. ქართულში ამ პროცესებს მორფოლოგიური მნიშვნელობა აქვს

და გამოიხატება ხმოვანთა დაკარგვით რამდენიმე ბრუნვაში (ნათესაობითში, მოქმედებითსა და ვითარებით ბრუნვებში). დისკუსიის შედეგად გადაწყდა, ამ ფორმებს მიენიჭოს დამატებითი ქულა დათვლის პროცესში, რადგანაც ბავშვების მეტყველებაში მათი სწორად გამოყენება ასახავს მორფო-ფონოლოგიური განვითარების ეტაპს და სიტყვათა ბრუნებას უკავშირდება.

აღნიშნული წესების გამოყენებით მიღებულმა წინასწარმა შედეგებმა აჩვენა, რომ ქართულენოვანი და ინგლისურენოვანი ბავშვების შემთხვევაში MLU-ს მაჩვენებელი განსხვავებულია, რაც ქართული ენის მორფოლოგიური მარკერების მრავალფეროვნების გათვალისწინებით, მოსალოდნელიც იყო. ქართულენოვანი ბავშვის ქულა აღემატება ამავე ასაკის ინგლისურენოვანი ბავშვის ქულას. ბუნებრივია, ორი ბავშვის მონაცემები არ არის საკმარისი შედეგების განზოგადებისათვის. საჭიროა შემდგომი კვლევა, თუ გავითვალისწინებთ ებრაულ, ბასკურ და თურქულ შემთხვევებს, დიდი ალბათობით, ეს განსხვავება კვლავაც დადასტურდება, მაგრამ უნდა მოხდეს საკითხის უფრო სიღრმისეული შესწავლა და თვისებრიობის დაზუსტება.

შეჯამების სახით უნდა აღინიშნოს, რომ მიღებული პროტოკოლი და ქართულენოვანი ბავშვების მეტყველების კორპუსის შექმნა იძლევა შესაძლებლობას, სამომავლოდ განხორციელდეს ფართო სპექტრის კვლევები, რასაც ხელს უწყობს ტრანსკრიბირებული ჩანაწერების არსებობა, რომელიც ხელმისაწვდომი გახდება დაინტერესებული მეცნიერებისთვის.

Establishing the Mean Length of Utterance Protocol for the Georgian Language

Tinatin Chincharauli, Nino Tsintsadze, Nino Doborjginidze, Teona Damenia, Tamar Kalkhitashvili

Ilia State University, Georgia

tinatin.chincharauli@iliauni.edu.ge, nino.tsintsadze@iliauni.edu.ge,
nino_doborjginidze@iliauni.edu.ge, teona.damenia@iliauni.edu.ge,
kalkhitashvili.tamar@gmail.com

In order to evaluate language acquisition, it is crucial to have appropriate systematic coding rules which consider the peculiarities of a given language. In the case of the Georgian Language, a primarily agglutinative, ergative language with the poly personal agreement, creating a coding system is crucial for future studies of Georgian language acquisition, child language development, speech and language

disorders and so on.

Since Roger Brown's classical work “A first Language: Early Stages” (Brown, 1973), calculating the numbers of morphemes per utterance (MLUm) has been widely used to measure language acquisition in spontaneous language samples. Studies have shown MLU to be a reliable indicator of grammatical development in the early years (Brown, 1973; Miller & Chapman, 1981). However, the application of the MLU to languages other than English reveals some difficulties. One of the goals of this study was to create an MLU calculation method applicable to measuring early language acquisition in Georgian children.

The research was conducted using a longitudinal cross-sectional design based on the study's aim and objectives. Four monolingual Georgian-speaking children with typical language development took part in the study. At the outset of the study, two were twelve months old and two – thirty-six months old. There was a girl and a boy in each age group. At the beginning of the study, all of them were assessed with AEPS® to ensure the typical development of the participants. It should be noted that we used some additional inclusion criteria to ensure the validity of the obtained data. The language samples were regularly obtained from parents and primary caregivers for two years. Parents and primary caregivers received detailed instructions regarding video recording. Children were video-recorded each month for 4-5 hours in realistic, spontaneous speech situations with their mothers and other family members at their homes. All participants followed the same structure for video recording: they had one week each month to make videotapes during mealtimes, play, and book reading. The adults were given instructions to do at least 1-hour long recording every day of a chosen week. After completing the chosen week, recorded files were collected and transcribed by a research assistant. In summary, almost 480 hours of recording have been collected. Native speakers transcribed all data in CHAT format following the standards of the CHILDES (MacWhinney & Snow, 2000). The length of recordings varied across childrenand sessions.

The first hundred utterances have been analyzed by us from each fully intelligible and productive recording, which means these were contextually relevant memorized repetitions.

As a result of the study, we established the rules to calculate MLU and tried them on the utterances obtained from eighteen to thirty-six months.

Among a number of methodological challenges, one of the main challenges was to determine the base unit for verbs, as there is no unanimity among scholars regarding the base form of Georgian verbs. After consultations with linguists, psychologists and psycholinguists, the 2nd person singular of subjective conjugation was chosen as the base unit, since the form has a simple inflectional structure (in most cases, it does not take morphological markers) and a high rate of frequency.

Another debatable question concerned participles, which function like adjectives but derive from verbs. Specifically, the point at issue was the counting of Georgian participle-forming

affixes –uli, sɑ-ɛli, as in gɑk'ɛtʰɛb-uli, gɑ-sɑ-k'ɛtʰɛb-ɛli. After discussions, it was found that the affixes cannot be acquired merely through mechanical repetition and imitation and their appearance in children's speech can be seen as an indicator of their progress to a new level of language development.

The next methodological problem was to determine whether to include the cases of syncope and apocope in the morpheme count. In Georgian, these processes have a morphological value and are

manifested as the loss of a vowel in several case forms (the genitive, instrumental and adverbial). It was finally decided to assign an additional point to the correct use of syncope and apocope, as in Georgian these are not merely phonological processes but are specific to certain types of declension.

Our preliminary results pointed out a significant gap in MLU scores between Georgian and English speakers, which was to be expected. Given the diversity of Georgian morphological markers, even the minimum score at a certain age is higher than the corresponding score with English-speaking children of the same age. Naturally, two children's language data are not sufficient and further research is needed. If we consider the cases of Hebrew, Basque and Turkish, the gap is highly likely to persist, but its degree will be specified.

In sum, annotated and fully coded Georgian Child Language Corpus was created that will be available to all interested researchers. As the database includes transcribed material as well as video files, it can be applied across a wide range of studies.

EFL-ის სტუდენტებისთვის წერის კომპეტენციაზე კლავიატურის გავლენის შედარებითი კვლევა

ათენა ჯონსონი
ანგლოფონური კვლევების ცენტრი, პარიზ-ნანტერის უნივერსიტეტი, საფრანგეთი
ajohnson@parisnanterre.fr

საკლასო ოთახში, სადაც სტუდენტებს მოეთხოვებოდათ შენიშვნების ჩაწერა, კვლევის შედეგად აღმოჩნდა, რომ შენიშვნების ხელით დაწერა იმ ტვინის უბნების სტიმულირებას იწვევდა, რომლებიც დაკავშირებულია გრძელვადიან მეხსიერებასთან და სწავლასთან (van der Meer, 2017). კვლევამ ასევე დაადასტურა, რომ წერისა და რედაქტირების პროცესში კლავიატურაზე ბეჭდვა მოსწავლეებს აძლევდა საშუალებას, ნაკლებ სწორხაზოვნად ეწერათ და ამ გზით უადვილდებოდათ უფრო დახვეწილი ტექსტის შექმნა (Baker & Kinzer, 1998). წარმოდგენილ კვლევამდე კვლავ გაურკვეველი რჩებოდა წერის ორგვარ მეთოდს შორის ლინგვისტიკური სხვაობა EFL-ის იმ სტუდენტებისთვის, რომლებსაც უთხრეს, ორივე მეთოდით დაეწერათ.

წარმოდგენილი კვლევა მიზნად ისახავს ორი მეთოდით წერისას განსხვავებების შესწავლას ლინგვისტიკური, ლექსიკური და გრამატიკული მრავალფეროვნების გამოკვლევის გზით. ლინგვისტიკური განსხვავებების შესწავლისთვის მეთოდოლოგიურ ჩარჩოდ გამოყენებულია დუგლას ბიბერის 1988 და 1991 წელს გამოცემული ნაშრომები, რომლებშიც ავტორი განიხილავდა მეტყველებისა და წერის ნაირსახეობებს. ქვემოთ მოცემული მოვლენების გასაანალიზებლად წერის თითოეული მეთოდისთვის შეგროვდა და შედარდა საუნივერსიტეტო ასაკის იმ

58 მონაწილის მონაცემები, რომლებიც გადიოდნენ კურსს: „ინგლისური სპეციალური მიზნებისთვის“.

სტილისტიკური მოვლენა	ლექსიკური მოვლენა	გრამატიკული მოვლენა
ინფორმაციული ჩართულობა პიროვნული ჩართულობის საპირისპიროდ	ლექსიკური მრავალფეროვნება	ატრიბუტული ზედსართავი სახელი
ექსპლიციტური მინიშნება იმპლიციტურის საპირისპიროდ	ლექსიკური სირთულე	
დარწმუნების დაუფარავი გამოთქმა ფარული გამოთქმის საპირისპიროდ		
აბსტრაქტული ბუნება კონკრეტულის საპირისპიროდ		

ლიტერატურა

Baker, E., & Kinzer, C. K. (1998). “Effects of Technology on Process Writing: Are They All Good?” National Reading Conference Yearbook, 47, 428-440.

Biber, Douglas. “The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings”. Computers and the Humanities, vol. 26, no 5/6, 1992, p. 331-45.

Dahlström, Dahlström, & Boström Boström. “Pros and Cons: Handwriting Versus Digital Writing”. Nordic Journal of Digital Literacy, vol. 12, no 4, December 2017, p. 143-61.

Michael, R., & Tao, W. (2004). Effects of Handwriting and Computer-Print on Composition

Scores: A Follow-up to Powers, Fowles, Farnum, & Ramsey.

van der Meer, Audrey L. H., & F. R. (Ruud) van der Weel. “Only Three Fingers Write, but the Whole Brain Works†: A High-Density EEG Study Showing Advantages of Drawing Over Typing for Learning”. Frontiers in Psychology, vol. 8, May 2017, p. 706.

Warschauer, M., & Carla Meskill. “Technology and second language learning”. Handbook of Undergraduate Second Language Education, January 2000, p. 303-18.

A Comparative Study of the Impact of the Keyboard on Written Productions for EFL Students

Atheena Johnson
Centre de Recherches Anglophones - CREA - Université de Paris Nanterre, France
ajohnson@parisnanterre.fr

Within a classroom context where students were asked to take notes, research found that writing one's notes by hand stimulated areas in the brain that were tied to long-term memory and learning (van der Meer, 2017). Research also found that within the writing and editing process, keyboarding allowed leaners to write in a less linear manner and, in doing so, facilitated the production of a more revised text (Baker & Kinzer, 1998). What remained uncertain to the date of the current study was the linguistic indistinguishability of the two modes of production for EFL students when they were asked to produce writing in both modes.

This study was designed to explore the differences in the modes of production by examining stylistic, lexical and grammatical diversity. The methodological framework used to explore the linguistic differences were Douglas Biber's 1988 and 1991 works, which studied variations across speech and writing. One of each mode of production was collected from 58 university-aged, English for special purposes participants, and were compared to analyse the following phenomena.

Stylistic phenomena	**Lexical phenomena**	**Grammatical phenomena**
Informational versus personal involvement	Lexical diversity	Attributive adjectives
Explicit versus implicit referencing	Lexical sophistication	
Overt versus covert expression of persuasion		
Abstract versus concrete nature		

References

Baker, E., & Kinzer, C. K. (1998). "Effects of Technology on Process Writing: Are They All Good?" National Reading Conference Yearbook, 47, 428-440.

Biber, Douglas. "The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An

Overview of Methodology and Findings". Computers and the Humanities, vol. 26, no 5/6, 1992, p. 331-45.

Dahlström, Dahlström, & Boström Boström. "Pros and Cons: Handwriting Versus Digital Writing". Nordic Journal of Digital Literacy, vol. 12, no 4, December 2017, p. 143-61.

Michael, R., & Tao, W. (2004). Effects of Handwriting and Computer-Print on Composition Scores: A Follow-up to Powers, Fowles, Farnum, & Ramsey.

van der Meer, Audrey L. H., & F. R. (Ruud) van der Weel. "Only Three Fingers Write, but the Whole Brain Works†: A High-Density EEG Study Showing Advantages of Drawing Over Typing for Learning". Frontiers in Psychology, vol. 8, May 2017, p. 706.

Warschauer, M., & Carla Meskill. "Technology and second language learning". Handbook of Undergraduate Second Language Education, January 2000, p. 303-18.

სარჩევი – Contents

INTERNATIONAL CONFERENCE

LANGUAGE AND MODERN TECHNOLOGIES – 2022

საერთაშორისო კონფერენცია

ენა და თანამედროვე ტექნოლოგიები – 2022

INTERNATIONAL CONFERENCE

LANGUAGE AND MODERN TECHNOLOGIES – 2022

საერთაშორისო კონფერენცია

ენა და თანამედროვე ტექნოლოგიები – 2022

INTERNATIONAL CONFERENCE

LANGUAGE AND MODERN TECHNOLOGIES – 2022