ინგილოური ტექსტების კოლექცია მულტიმედიური კორპუსისთვის

მაია ბარიხაშვილი
თსუ არნ. ჩიქობავას სახ. ენათმეცნიერების ინსტიტუტი (საქართველო)
maiahereti@yahoo.com

ელექტრონული კორპუსის, როგორც ლინგვისტიკური კვლევისათვის მდიდარი რესურსის, გამოყენება სულ უფრო მზარდ მასშტაბებს იძენს.

ამჟამად მსოფლიოში აქტიური მუშაობა მიმდინარეობს მულტიმედიური კორპუსების შესაქმნელად. მულტიმედიური კორპუსი წარმოადგენს ელექტრონულ რესურსს, რომელიც ტექსტის გარდა მოიცავს ვიდეო და აუდიოჩანაწერებს. ასეთი სახის რესურსი საშუალებას გვაძლევს ვიკვლიოთ არა მხოლოდ საკუთრივ ენობრივი პროცესები, არამედ სხვადასხვა ვითარებაში გამოყენებული პარალინგვისტიკური საშუალებებიც (მაგ.: მიმიკა, ჟესტები, პოზა). მულტიმედიური კორპუსები პერსპექტიულია ვერბალური და არავერბალური კომუნიკაციის კვლევისათვის.

2006 წლიდან არნოლდ ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტში მუშაობა მიმდინარეობს ქართული დიალექტური კორპუსის (ქდკ) შესაქმნელად. ჩვენი კორპუსი აერთიანებს თითქმის ასი წლის მონაკვეთში დაფიქსირებულ ტექსტებს. ეს ტექსტები ჩაწერილია არა მხოლოდ სხვადასხვა მეთოდოლოგიითა და მიდგომით, მეტატექსტური და ლინგვისტიკური მახასიათებლების სხვადასხვაგვარი ფიქსაციით, არამედ სხვადასხვა ტექნიკური საშუალებითაც. ამ მასალაში განსაკუთრებული ადგილი უჭირავს მაგნიტურ მატარებელზე შენახულ ტექსტებს – აუდიო და ვიდეომასალას. ასეთი სახით დოკუმენტირებული რესურსი უნიკალურ საშუალებას წარმოადგენს ზეპირი მეტყველების შესასწავლად.

ქდკ-ის კონცეფციის შემუშავებისას ვხელმძღვანელობდით ტექსტური კოლექციის ერთგვაროვნების პრინციპით. ბუნებრივია, რომ მეოცე საუკუნის დასაწყისში ჩაწერილი ტექსტებისა და ოცდამეერთე საუკუნეში ციფრული ტექნიკით აღბეჭდილი მასალის ინფორმაციული ღირებულება თანაბარი ვერ იქნებოდა; ერთი მხრივ, ქრონოლოგიური დისტანცია უფრო ღირებულს და შეუცვლელს ხდიდა პირველს, ხოლო, მეორე მხრივ, ტექნოლოგიური საშუალებების წყალობით ფიქსირდებოდა ზეპირი დისკურსის ბევრი ისეთი ელემენტი, რომელიც უგულებელყოფილი და შეუმჩნეველი რჩებოდა ხელით ჩაწერის დროს.

აღნიშნული მიდგომის გამო ის ინგილოური ტექსტები, რომლებიც 2005, 2007, 2012 წლებშია მოპოვებული (მ. ბერიძე, მ. ბარიხაშვილი, ნ. სურმავა, ე. ნაპირელი) და ციფრულ ფორმატშია დოკუმენტირებული, კორპუსისთვის მომზადდა მანამდე გამოცემულ ტექსტებთან (გ. იმნაიშვილი, რ. ღამბაშიძე) მაქსიმალური რედაქციული მსგავსების დაცვით. უნიფიცირების გამო შეტანილი ცვლილებები ერთნაირად გავრცელდა ტექსტების როგორც ძველ, ისე ახალ კოლექციაზე.

ჩვენს ხელთ არსებული მასალები, რომლებიც მაგნიტურ და ციფრულ მატარებლებზეა ჩაწერილი, თავისუფლად შეიძლება გამოყენებულ იქნეს თანამედროვე მულტიმედიური კორპუსის შესაქმნელად, რომელშიც შენარჩუნებული იქნება თავდაპირველი ტექსტის მრავალი არსებითი ნიშანი. იმისათვის, რომ არ დაგვეკარგა რელევანტური ინფორმაცია ჩვენ მიერ მომზადე-

ბული მასალის პირველად დესკრიპტებში, აღვნიშნავდით ისეთ ენობრივ და არაენობრივ (პარალინგვისტიკურ) მახასიათებლებს, რომლებიც აუცილებელია ზეპირი კორპუსების პროსოდიული ნიშანდებისთვის (ხმოვანთა ტონური დაგრძელება, პაუზების გამოყოფა და სხვ.), ჰეზიტაციური აღწერისთვის (დაუსრულებელი სიტყვები, ჩახველება, ბგერითი პაუზა, ჩაცინება, სხვა სახის ემოციის გამოხატვა...), ფალსტარტების აღნიშვნისათვის, ექსპრესიის სხვადასხვა საშუალების გამოვლენისთვის (ბგერათა ექსპრესია, ჟესტიკულაცია, მიმიკა...).

ინგილოური დიალექტის ახალი ჩანაწერების ბაზაზე (ვ. აბაშვილი, მ. ბერიძე, მ. ბარიხაშვილის მიერ თავმოყრილი აუდიო-ვიდეო მასალა) შექმნილია ტექსტური კოლექციის პარალელური ვერსიები, რომელთაგან ერთი ქდკ-ის ბაზისთვისაა საერთო წესებით მომზადებული, მეორე კი წარმოადგენს ბაზას თანამედროვე ტიპის ზეპირი (მულტიმედიური) კორპუსის მონიშვნისთვის.

მოხსენებაში წარმოდგენილი იქნება ტექსტების ნიმუშები პირველად განაშიფრებში დაფიქსირებული იმ ენობრივი და არაენობრივი მახასიათებლების შენარჩუნებით, რომელთა ნიველირება მოხდა უნიფიკაციის დროს ქდკ-ის ბაზაში.

ტექსტების ამგვარი პირველადი დამუშავება ქმნის იმის საფუძველს, რომ სამომავლოდ ქდკ-ის წიაღში თანამედროვე ტიპის მულტიმედიური ქვეკორპუსი შეიქმნას.

A Collection of Ingilo Texts for a Multimedia Corpus

Maia Barikhashvili
Arn. Chikobava Institute of Linguistics, TSU (Georgia)
maiahereti@yahoo.com

Application of an electronic corpus, as a rich resource for linguistic research, has been considerably increasing. The process in point is accompanied with the development of corpora with respect to both their quantity and quality.

Currently, several multimedia corpora have been developed around the world. A multimedia corpus is an electronic resource which, alongside with texts, comprises video and audio recordings. Such a resource allows for the study of not only linguistic processes proper but also paralinguistic means (for instance, mimicry, gestures, posture) in various situations. Multimedia corpora have been perspective for the study of verbal and non-verbal communication.

Since 2006, the Georgian Dialect Corpus (GDC) has been developed at Arnold Chikobava Institute of Linguistics. Our corpus has amalgamated the texts, having been recorded within a span of a hundred years. The texts were recorded not only by means of different methodologies and approaches, different treatments of meta-textual and linguistic properties, but also by means of different technical devices. The

collection features taped texts – audio and video materials. The such documented resource is a unique opportunity for the study of oral speech.

While developing the concept of the GDC, we followed the principle of homogeneity of a textual collection. Naturally enough, informative values of the texts, recorded in the early 20th century, and of the materials, documented by means of digital devices in the 21st century, can not be equal; on the one hand, the chronological distance made the latter more valuable and unique, and, on the other, by means of technological devices, many such elements of oral discourse were documented that were neglected and unnoticed during manual recording.

Owing to the said approach, the Ingilo texts, collected and digitalized by us (M. Beridze, M. Barikhashvili, N. Surmava, E. Napireli) in 2005, 2007, 2012, were prepared for the corpus maintaining the maximum editorial similarity with the earlier published texts (G. Imnaishvili, R. Gambashidze). Due to the unification, the changes covered both the old and new collections.

The available materials, both taped and digitalized, can be used for the development of a contemporary multimedia corpus, preserving many essential features of an original text. In order not to lose relevant information in the original descripts of the material, prepared by us, we marked the linguistic and paralinguistic features, necessary for the prosodic tagging of oral corpora (pitch lengthening of vowels, identification of pauses, etc.), for the hesitational description (incomplete words, cough, sound pause, grin, expression of other kinds of emotion…), for the marking of false starts, for the identification of various means of expression (sound expressives, gesture, mimicry…).

Based on the new recordings of the Ingilo dialect (audio-video materials, collected by V. Kuzi-babashvili, M. Beridze, M. Barikhashvili), parallel versions of the textual collection were created, one of which is prepared with common rules for the base of the GDC, and another is a base for the tagging of a contemporary oral (multimedia) corpus.

The paper presents sample texts, preserving the linguistic and non-linguistic features in original transcripts, having been leveled during the unification in the Base of the GDC.

Such primary processing of texts makes a basis for the future development of a contemporary multimedia subcorpus within a framework of the GDC.

სამეცნიერო ტექსტების კორპუსების ლექსიკოგრაფიული პოტენციალი

ლარისა ბელიაევა
გერცენის სახ. სახელმწიფო პედაგოგიური უნივერსიტეტი (რუსეთის ფედერაცია)
Iauranbel@gmail.com

დღესდღეობით გამოყენებითი ლექსიკოგრაფია წარმოადგენს გამოყენებითი ლინგვისტიკისა და ენის ინჟინერიის (კომპიუტერული ლინგვისტიკის) სპეციალურ დარგს, რომელიც ემსახურება პრობლემების გადაჭრაზე ორიენტირებული ავტომატიზებული და ავტომატური ლექსიკონებისა და მონაცემთა ბაზების შექმნასა და მართვას. ლექსიკოგრაფიული სისტემების სისრულე და ადეკვატურობა გარკვეულწილად განსაზღვრავს სხვადასხვა შემადგენლობის, აგებულებისა და ფუნქციის ტექსტებიდან მოპოვებული ინფორმაციისა და ცოდნის დონესა და სანდოობას.

ლექსიკონის შედგენის თანამედროვე მიდგომა გულისხმობს წინასწარ სამუშაოს პარალელური ან შედარებითი კორპუსების შესაქმნელად, რომლებიც შეიძლება განვიხილოთ, როგორც მონაცემთა ბაზა კვლევითი და პრაქტიკული ლექსიკოგრაფიის პრობლემების გადასაწყვეტად. პარალელური ტექსტების კორპუსები სრულყოფილ ლექსიკოგრაფიულ მასალობრივ წყაროს წარმოადგენენ, რადგანაც ეს კორპუსები იგება დარგობრივი ტექსტების ვიწრო მონაცემებზე დაყრდნობით (სტატიები, მონოგრაფიები, კონფერენციის მასალები და მათი თარგმანები სხვა ენებზე). ამგვარი კორპუსები წინადადებების მიხედვით უნდა იქნეს შეთანადებული, რაც საშუალებას იძლევა, რომ გამოვავლინოთ და გავაანალიზოთ ტერმინები და მათი თარგმანები, შევაფასოთ მათი სტანდარტიზაციის დონე და თარგმანის შესაბამისობა ისევე, როგორც გარკვეულ სახესხვაობათა პრევალირება.

და მაინც, ამგვარი კორპუსის შექმნა ყოველთვის არ არის შესაძლებელი. ერთ-ერთ არჩევანს წარმოადგენს ის, რომ წყაროს სახით არსებული მასალა გადავაქციოთ ტექსტების კორპუსად, რომელშიც წარმოდგენილი იქნება თავდაპირველი ტექსტების პარალელური პრეზენტაციები, მათი მანქანური თარგმანები და რედაქტირების შემდგომი შედეგები. ეს რედაქტირებული მანქანური თარგმანები უნდა შეთანხმდეს შესაბამისი დარგების ექსპერტებთან. მნიშვნელოვანია, რომ ამგვარი კორპუსის ხარისხი და პოტენციალი ემყარება ექსპერტებთან თანამშრომლობას საწყისი მასალის შერჩევისა და მანქანური თარგმანების რედაქტირების პროცესში.

ლექსიკოგრაფიული ანალიზის შემთხვევაში წინადადებების მიხედვით შეთანადებული ტექსტი საშუალებას გვაძლევს შევადაროთ საწყისი წინადადება, მისი მანქანური თარგმანი და საბოლოო წინადადების თარგმანი; ამგვარად, ჩვენ შეგვიძლია გამოვავლინოთ და აღვწეროთ ტერმინოლოგიური გამონათქვამების (ძირითადად, მსაზღვრელ-საზღვრულის) წყება თარგმანის დონეზე. იმისათვის, რომ მივიღოთ მანქანური თარგმანის შედეგი, მიზანშეწონილია მანქანური მთარგმნელი სისტემის იმ ლექსიკონების გამოყენება, რომლებიც შეიცავენ საჭირო ან შესადარებელ სიტყვებს ან გამოთქმებს.

მოხსენებაში ეს საკითხი განხილულია „ბოლონიის პროცესის“ ლექსიკონის შექმნის მაგალითზე. წინადადებების მიხედვით შეთანადებული ტექსტების გამოყენება საშუალებას გვაძლევს დავაზუსტოთ ლექსიკურ ერთეულთა თარგმანები დიდ ტექსტურ კოლექციებში. გარდა ამისა, ტექსტების კორპუსი შეიძლება გამოვიყენოთ დარგის სტრუქტურისა და მისი ტერმინოლოგიური სისტემის გამოსავლენად. ეს ნაჩვენებია კომპონენტ „უმაღლესი განათლების“ შემცველი კოლოკაციების მაგალითზე, რომლებიც ყველაზე ხშირია ბოლონიის პროცესის ტექსტების კორპუსში. ეს კორპუსი აიგო როგორც მცირე სამეცნიერო კორპუსი (500 000 სიტყვაფორმა), რომლის მიზანი იყო სწორედ ამ სფეროსათვის შექმნილი რუსულ-ინგლისური გლოსარიუმების შემადგენლობის, სტრუქტურისა და თარგმანების შემოწმება.

ლექსიკონის შედგენის პრობლემის გადაჭრა მოითხოვს, რომ ერთმანეთისაგან გავარჩიოთ: *ლინგვისტური ავტომატები*, რომელთა საშუალებითაც ეს პრობლემა უნდა გადაიჭრას ავტომატურად ან ნახევრადავტომატურად და *ლინგვისტური ავტომატები*, რომელთა მუშაობაში პროფესიონალი ლექსიკოგრაფი შეძლებს ჩართვას როგორც კორპუსის ნიშანდების, შეთანადებისა და ტექსტის ანალიზის, ასევე ლექსიკოგრაფიული პრობლემის გადაწყვეტის დონეზე. პარალელური კორპუსების გამოყენებისას მეორე ტიპი უფრო მიზანშეწონილია.

Lexicographic Potential of Scientific Text Corpora

Larisa Beliaeva
Herzen State Pedagogical University (Russian Federation)
Iauranbel@gmail.com

Nowadays applied lexicography is a special domain of applied linguistics and language engineering for creation and management of problem–oriented automated and automatic dictionaries and databases. Completeness and adequacy of lexicographic systems to a considerable extent determine the level and reliability of information and knowledge mining from the texts of various composition, structure and function.

A modern approach to dictionary creation assumes preliminary work with parallel or comparable text corpora, which can be considered as a database for solving both research and practical lexicographic problems. Parallel text corpora are a perfect source of lexicographical materials as these corpora are to be constructed on the basis of narrow data domain texts (articles, monographs, conference materials and their translations into other languages). Such corpora are to be sentence–by–sentence aligned, allowing for revealing and analysing of terms and their translations, evaluating their level of standardization and translation conformity as well as prevalence of special variants. However, creation of such a corpus is not always possible. One of the options is to create a source lexicographical material as text corpora with parallel presentation of initial texts, their machine translations and post-editing results. These edited machine translations are to be agreed with experts in the proper knowledge domains. It is important,

that the quality and potential of such corpus depends on cooperation with experts when selecting the source material and editing the machine translations.

In case of a lexicographical analysis, sentence-by-sentence text alignment allows for comparing initial sentence, its machine translation and the final sentence translation, thus we are able to reveal and describe the set terminological expressions (mostly noun phrases) on the translation level. In order to receive machine translation results, it is expedient to use the dictionaries of a MT system, containing the needed or comparable words and expressions.

The paper considers this process on the example of creating a dictionary on the Bologna procession domain. Use of the sentence-by-sentence aligned texts gives us an opportunity to specify translations of the lexical units in large text collections. Besides, a text corpus can be used for revealing the domain structure and its terminological system. This idea is shown on the examples of analysis for the collocations with component *higher education*, being the most frequent in the Bologna procession text corpora. This corpus was constructed as a small research corpus (500 000 wordforms), the aim of this corpus was to verify the composition, structure and translations included in different Russian-English glossaries developed for this very domain.

Solving the problem of dictionary creation requires to distinguish between linguistic automata with which this problem could be solved automatically or semi-automatically and linguistic automata in work of which a specialist – lexicographer could participate both on the level of corpora tagging and alignment and text analysis, and on the level of lexicographic problem solving. In terms of the application of parallel corpora, the second type is more expedient.

ქართული დიალექტური კორპუსის ახალი ლექსიკოგრაფიული რედაქტორი

მარინა ბერიძე, ლიანა ლორთქიფანიძე, დავით ნადარაია
თსუ არნ. ჩიქობავას სახ. ენათმეცნიერების ინსტიტუტი (საქართველო)
marine.beridze@gmail.com, l.lordkipanidze@yahoo.com, dnad@itex.ge

ქართული დიალექტური კორპუსის შექმნის პირველი ეტაპი მოიცავდა ძირითადი ტექსტური მასივის, მეტატექსტური ანოტირების სისტემისა და მასზე დამყარებული საძიებო-საცნობარო სისტემის დამუშავებას. კორპუსის გრამატიკული მონიშვნა და ლექსიკოგრაფიული კომპონენტი მხოლოდ კონცეფციურ დონეზე განიხილებოდა.

პროექტის მიმდინარე ეტაპი გულისხმობს კორპუსის მორფოლოგიურ მონიშვნასა და ლექსიკოგრაფიული ბაზის ფორმირებას. შესაბამისად, გადამუშავდა კონცეფციის „ლექსიკოგრაფიული“ ნაწილიც.

ქართული დიალექტური კორპუსის ახალ ლექსიკოგრაფიულ რედაქტორში გათვალისწინებულია „ქაღალდის ლექსიკონების“ ყველა თავისებურება. ამასთან, ის წარმოადგენს ძველი

ლექსიკონების მნიშვნელოვანი დამატებითი ინფორმაციით აღჭურვისა და ახალი ლექსიკონის შექმნის მოქნილ და ეფექტურ საშუალებას.

ლექსიკონების ახალი რედაქტორით შესაძლებელია: 1. მთავარ სიტყვას მიეწეროს გრამატიკული კლასის მახასიათებელი; 2. აღიწეროს ფონეტიკური ვარიაცია – რაც კორპუსში ვარიანტების ავტომატური ნიშანდების საშუალებას იძლევა; 3. შეიქმნას ინფორმაცია ლექსემის გრამატიკულ (ფორმაწარმოებით) ვარიაციაზე; 4. შეიქმნას ფრაზეოლოგიზმებისა და მყარი სიტყვაშეხამებების ლექსიკონები, რომლებიც კორპუსში ასეთი მასალის ნიშანდებისა და, შესაბამისად, ძიების საშუალებას მოგვცემს. 5. შეიქმნას დიალექტურ ფორმაუცვლელ სიტყვათა და მათი მრავალრიცხოვანი ფონეტიკური ვარიანტების ლექსიკონები, რომლებიც ასევე გამოყენებული იქნება კორპუსში ავტომატური ანოტაციისათვის.

პრინციპულად მიგვაჩნია, რომ გრამატიკული სტანდარტი ქდკ-ში მაქსიმალურად ითვალისწინებდეს ტრადიციულ ლინგვისტიკურ ნააზრევს. შემუშავდა გრამატიკული მახასიათებლების სია და მათი ლათინური ასოებისგან შემდგარი აბრევიატურები, რომლებიც შეესაბამება ყველაზე ცნობილ სტანდარტებს (EAGLES, Leipzig Glossing Rules...), თუმცა მთლიანად მაინც ვერ ეყრდნობა მათ და ორიენტირებულია ქართული ენის თავისებურებისა და ქართული საენათმეცნიერო ტრადიციის სრულყოფილად ასახვაზე.

ახალი რედაქტორი სალექსიკონო სტატიას მონაცემთა ბაზის სახით ინახავს, რომელიც დაკავშირებულია ცალკე კონფიგურირებადი სიების სახით არსებულ სხვადასხვა მახასიათებელთა სიმრავლესთან. ასე მაგალითად:

მთავარი სიტყვის ველი უკავშირდება შემდეგ ადმინისტრირებად სიებს:

1. მეტყველების ნაწილთა ჯგუფების ჩამონათვალი
2. გრამატიკული მახასიათებლების ჩამონათვალი
3. სემანტიკური მახასიათებლების ჩამონათვალი
4. ლექსიკური ტიპის განმსაზღვრელი მახასიათებლების ჩამონათვალი (ტერმინები, ფრაზეოლოგიზმები / მყარი შესიტყვებები...)
5. ენებისა და დიალექტების ჩამონათვალი (სიტყვის დიალექტური წარმომავლობის აღსაწერად)
6. უცხო ენების ჩამონათვალი (სიტყვის უცხოური წარმომავლობის აღსაწერად)
7. ლექსიკონების გამოცემების ჩამონათვალი

სალექსიკონო სტატიის სტრუქტურა ახალ რედაქტორში ასეთია:

1. მეთაური სიტყვა (გრამატიკული და სემანტიკური ინფორმაცია)
2. ზუსტი სალიტერატურო შესატყვისი (ასეთის არსებობის შემთხვევაში)
3. სალიტერატურო სიტყვის განმარტება (რამდენიმეს დამატების ფუნქციით)
4. განმარტების კომენტარი / შენიშვნა (ენციკლოპედიური ინფორმაცია)
5. სიტყვის ფონეტიკური ვარიაციები
6. სიტყვის გრამატიკული ვარიაციები (პარადიგმატული რეალიზაციები)
7. ილუსტრაცია
8. ილუსტრაციის წყარო
9. ილუსტრაციის კომენტარი (თარგმანი, პერიფრაზი...)

მეთაური სიტყვა უკავშირდება ყველა არსებულ დიალექტურ ლექსიკონსა და აგრეთვე ტექსტურ ბაზას. ეს იმას ნიშნავს, რომ ყველა სალექსიკონო სტატია, რომელშიც დასტურდება ეს სიტყვა მთავარ ფორმად და ყველა კონტექსტი, რომელშიც ეს სიტყვა შედის ტექსტურ ბაზაში _ ერთიანი კონკორდანსითაა გამთლიანებული.

ისევე როგორც მეთაურ სიტყვას, გრამატიკული ვარიაციის ველსაც აქვს კავშირი გრამატიკულ მახასიათებელთა სიებთან. გარდა ქართულ ლექსიკოგრაფიაში „რელევანტურად“ მიჩნეული გრამატიკული ვარიაციებისა (ზმნისთვის – მასდარი ან აწმყოს მესამე პირის ფორმა; სახელისთვის – მხ. რ. ნათესაობითის ფორმა) დიალექტურ ლექსიკონებში წარმოდგენილია მიმღეობის, თურმეობითის, კაუზატივის და სხვ. ფორმებიც. სალექსიკონო სტატიაში მათი გრამატიკული მახასიათებლებით ნიშანდება კორპუსში ანალოგიური ფორმების ნიშანდების საშუალებას იძლევა, ამასთან, ავსებს კორპუსის სიტყვიერ მასივს პარადიგმატული რეალიზაციებით, რომლებიც შეიძლება ტექსტებში დადასტურებული არც იყოს.

ამრიგად, ქდკ-ს ახალ ლექსიკოგრაფიულ რედაქტორში შექმნილი სალექსიკონო სტატია მოიცავს საკმაო ლინგვისტიკურ ინფორმაციას იმისთვის, რომ ის კორპუსში გრამატიკული ნიშანდების ერთ-ერთ საშუალებად იქნეს გამოყენებული. ხოლო რედაქტორში ორ და მეტწევრა ერთეულების (სხვადასხვა ტიპის მყარი სიტყვათშეხამებების) დამატებისა და აღწერის შესაძლებლობა ტექსტურ მასივში ამგვარი ფრაგმენტების ლინგვისტური ნიშანდების პრობლემის ერთადერთი გზა იქნება. ლექსიკონის ეს ნაწილი დესკრიპტორი სიის სახით დაუკავშირდება ტექსტურ ბაზას.

ქდკ-ს ახალი ლექსიკოგრაფიული კონცეფცია კორპუსზე დაყრდნობილი და კორპუსით „მართული“ ლექსიკონების შექმნის ეფექტური საშუალებაც იქნება.

A New Lexicographic Editor of the Georgian Dialect Corpus

Marina Beridze, Liana Lortkipanidze, David Nadaraia
Arn. Chikobava Institute of Linguistics, TSU (Georgia)
marine.beridze@gmail.com, l.lordkipanidze@yahoo.com, dnad@itex.ge

The initial stage of the creation of the Georgian Dialect Corpus (GNC) comprised the development of the basic text body, the system of meta-textual annotation, and the reference system based on it. The grammatical markup and a lexicographic component were considered only on a conceptual level. The current stage of the project envisages the morphological annotation of the corpus and the formation of its lexicographic base. Therefore, the *lexicographic* part of the conception was revised.

The new lexicographic editor of the Georgian dialect corpus incorporates all peculiarities of „paper dictionaries“. Besides, it is a flexible and effective means for equipping older dictionaries with significant additional information and for the creation of a new one.

By means of the new editor of dictionaries it is possible: 1. to ascribe a grammatical class property to a headword; 2. to describe phonetic variation, this providing an opportunity for the automated tagging of variants in the corpus; 3. to develop information about the grammatical (inflectional) variation of a lexeme; 4. to create dictionaries of collocations and idioms, enabling to tag such entities in the corpus, and, thus, to search; 5. to create dictionaries of dialectal uninflected words and of their numerous phonetic variants, to be also used for the automated annotation in the corpus.

It has been principal for us to make the grammatical standard in the GNC maximally consider the traditional linguistic thought. A list of grammatical properties and their abbreviations were developed, being in accordance with the most acknowledged standards (EAGLES, Leipzig Glossing Rules…); however, this system is not totally based upon them and strives to the complete reflection of the peculiarities of the Georgian language and of the Georgian linguistic tradition.

The new editor stores an entry as a database, being associated with a set of various properties as individually configured lookup lists. For instance,

A headword field is associated with the following lookup lists:

1. List of parts of speech
2. List of grammatical properties
3. List of semantic properties
4. List of lexical properties (terms, collocations/idioms…)
5. List of languages and dialects (for the description of the dialectal origin of a word)
6. List of foreign languages (for the description of the foreign origin of a word)
7. List of dictionary publications

In the new editor, the structure of a dictionary entry is the following:

1. headword (grammatical and semantic information);
2. exact equivalent from the language standard (if available);
3. definition of a word belonging to the language standard (with a function of adding of several ones);
4. comment to a definition (encyclopedic information);
5. phonetic variations of a word;
6. grammatical variations (realizations) of a word;
7. illustration;
8. source of illustration;
9. comment to an illustration (translation, paraphrase…);

A headword is associated with all existing dialect dictionaries and text base as well. This implies that all entries, evidencing a word as a headform and all contexts, including the word in a text base, are integrated within a single concordance.

Similarly to a headword, a field of grammatical variation is also associated with lists of grammatical properties. Alongside with the grammatical variations, considered „relevant“ in Georgian lexicography (a masdar or 3rd person present form for a verb, genitive singular for a noun), dialect dictionaries present participle, resultative, causative and other forms. Their encoding with grammatical

properties within an entry provides an opportunity for the markup of similar forms in the corpus. Besides, it complements the lexical body with paradigmatic realizations, possibly not attested in texts.

Hence, an entry, developed within the new lexicographic editor of the GNC, comprises enough linguistic information to be applied as one of the means of the grammatical tagging in the corpus. The opportunity of the addition and description of and bi- or tri-componential items (various types of collocations) within the text base will be the only way for the solution of the problem of the linguistic tagging of such fragments. This part of the dictionary will be associated with the text base immediately as a descriptor list.

The new lexicographic conception will also be an effective means for the development of corpus-based and corpus-driven dictionaries.

პოლიტიკური ტექსტების ქართული მთარგმნელობითი კორპუსი[1]

ხათუნა ბერიძე, გრიგოლ კახიანი
ბათუმის შოთა რუსთაველის სახელმწიფო უნივერსიტეტი (საქართველო)
beridze@illinois.edu, gkakhiani@gmail.com

მთარგმნელობითი კორპუსი, რომლის ბაზებშიც რესურსები ჯერჯერობით მხოლოდ ქართულ-ინგლისურ ენებზეა განთავსებული, თავს უყრის თანამედროვე პოლიტიკურ ლიტერატურას: პოლიტიკური მეცნიერების, ანალიტიკურ, საგაზეთო, საკანონმდებლო და სხვა ჟანრის ტექსტებს. კორპუსის ამჟამინდელი ბაზა მოიცავს:

(1) პოლიტიკური ტერმინებისა და კონცეპტების მონოლინგვურ განმარტებით ლექსიკონს; სამომავლოდ გათვალისწინებული კვლევების ხელშეწყობის მიზნით, სალექსიკონო ერთეულების განმარტებები სტრუქტურირებულია რამდენიმე სხვადასხვა წყაროდან. ამჟამად მუშავდება განმარტებების პარალელური სტრუქტურები;

(2) კონტექსტური მონაცემების ბაზებს, რომლებიც შედგება კომპარატიული და პარალელური ტექსტების სექციებისა და ქვესექციებისაგან.

კორპუსის საძიებო სისტემის მუშაობის პრინციპი ითვალისწინებს ლექსიკონიდან ტერმინის დეფინიციისა და კონტექსტის კომპარატიული და პარალელური კორპუსების სექციებიდან თანადროულ ძიებას და მიწოდებას, რაც უზრუნველყოფს მკვლევრებისა და/ან პოლიტიკის ენის შემსწავლელთათვის ძიების მიერ გენერირებული შედეგების მაღალ კოეფიციენტს. მაგალითად, ტერმინ „რეჟიმის“ ძიებისას, ამ ეტაპზე სალექსიკონო და კონტექსტური ბაზებიდან 113 ერთეულ მონაცემს მიიღებთ, ხოლო ტერმინ „დემოკრატიის“ ძიებისას – 231 ერთეულს, მათ შორისაა თანა-

[1] კორპუსს „მთარგმნელობითი“ ვუწოდეთ, ვინაიდან მასში განთავსებული სექციები მოიცავს როგორც პარალელურ, ისე კომპარატიულ ტექსტებს.

მედროვე, ბოლო ორი ათწლეულის განმავლობაში წარმოქმნილი ის ზედსართავები, რომლებიც აღნიშნულ ტერმინს პოლიტოლოგებმა დაუკავშირეს.

ამჟამად კორპუსი შედგება შემდეგი სექციებისა და ქვესექციებისაგან:

1. პარალელური კორპუსი	**2. ქართული კომპარატიული კორპუსი**	**3. ინგლისური / მულტილინგვური კომპარატიული კორპუსი**
1. ინგლისური ლიტერატურა თარგმანში	1. პოლიტიკა, სიახლეები	1. პოლიტიკა, სიახლეები
2. ქართული ლიტერატურა თარგმანში	2. პოლიტიკა, ინტერვიუ	2. პოლიტიკა, ინტერვიუ
3. პოლიტიკური ანალიზი	3. პოლიტიკური მეცნიერება	3. პოლიტიკური მეცნიერება
4. პოლიტიკური სიტყვა	4. პოლიტიკური ანალიზი	4. პოლიტიკური ანალიზი
5. პოლიტიკა, ლიტერატურა, თარგმანი	5. პოლიტიკა, ლიტერატურა, თარგმანი	5. პოლიტიკა, ლიტერატურა, თარგმანი
6. პოლიტიკა, სიახლეები		
7. პოლიტიკა, ინტერვიუ		
8. პოლიტიკური მეცნიერება		

კორპუსის ბაზები მხოლოდ წერილობითი ტექსტებისაგან შედგება. მასში განთავსებულია ქართული და ინგლისური პირველწყაროები და ინგლისურიდან ქართულად შესრულებული თარგმანები. კორპუსზე მუშაობის **საწყის ეტაპზე** შეგროვდა ინგლისურ ენაზე პოლიტიკური მეცნიერების შესახებ არსებული ტექსტები და პოლიტოლოგებთან ერთად ყურადღებით იქნა შესწავლილი. ამის შემდეგ ტექსტები დაიყო კომპარატიული და პარალელური კორპუსების საწყის რესურსებად.

აღნიშნულ ტექსტებს სამი ძირითადი ნიშანი აერთიანებთ: ყველა მათგანი თანამედროვეა, მათში დიდი რაოდენობითაა წარმოდგენილი ბოლო ორი ათწლეულის პერიოდში წარმოშობილი ინგლისურენოვანი ტერმინები, შინაარსობრივად კი შეეხება ყოფილი საბჭოთა ქვეყნების, მათ შორის, საქართველოს პოლიტიკური ტრანზიციის პერიოდს.

მომდევნო ეტაპზე მიმდინარეობდა: (ა) პარალელური კორპუსისათვის შერჩეული ტექსტების თარგმნა ქართულ ენაზე, ენის ეკოლოგიის დაცვის მიზნით ვცადეთ ტრანსლიტერაცია-ტრანსკრიბირების ნაცვლად გვეთარგმნა ის კონცეპტები, რომლებიც თარგმნას დაექვემდებარებოდა; (ბ) კონსულტირება ტერმინებისა და განმარტებების სალექსიკონო ბაზებისათვის დამუშავებისას; (გ) პოლიტიკასთან ასოცირებული ტექსტების მასივების დამუშავება, ჟანრობრივი დახარისხება, დაბალანსება და, შესაბამისად, კორპუსის სექციებად და ქვესექციებად დაყოფა; (დ) ფაილებისათვის საიდენტიფიკაციო ნუმერაციის მინიჭება. ამჟამად მცირე გუნდი მუშაობს კორპუსისათვის ტექსტების მოძიებისთვის, არსებული ტექსტების ანოტირებისთვის, ექსტრალინ-

გვისტური მონაცემების შედგენა-თანდართვისთვის საიდენტიფიკაციო ნუმერაციის მიხედვით. მანქანური პროგრამა, რომელიც ბაზებთან მუშაობას უზრუნველყოფს, სპეციალურად შეიქმნა აღნიშნული კორპუსისათვის.

ტექსტების განთავსების პარამეტრებია:

მოდუსი I

ინგლისური L1 - ქართული L2

პარალელური ტექსტები სტრუქტურირებულია ორენოვან პწკარედებად, სადაც შემდეგი პრინციპია დაცული: **1** L1 - 1 L2 / 2 L2; ეს უკანასკნელი ისეთი შემთხვევისათვის, როდესაც გრამატიკული ტრანსფორმაციის გამო 1 L1 გადმოცემულია 2 L2-ით. ინტერვიუს ტექსტის პარალელური სტრუქტურირების შემთხვევაში ტექსტი კოჰერენტულ აბზაცებად დავყავით, რომელშიც, შესაძლოა, ერთი, ორი ან სამზე მეტი წინადადებაც იყოს. კორპუსი არასრულია, ის მუშავდება მიმდინარე რეჟიმში. ამ ეტაპზე იგი უნიდირექციულია, უახლოეს მომავალში, კორპუსის ტექნოლოგიური და მანუალური დამუშავება, ინტერდისციპლინარული კვლევების განხორციელების მიზნით, პარალელური და კომპარატიული კორპუსების კვლევითი დიაპაზონის გაფართოებას ითვალისწინებს. კორპუსის გამოყენება შესაძლებელი გახდება შეპირისპირებითი ანალიზისათვის:

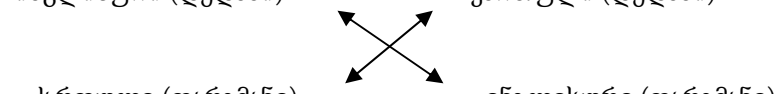
ა. კომპარატიული ინგლისური და ქართული პირველწყარო ტექსტებისა;

ბ. პირველწყარო ინგლისური ტექსტებისა თარგმანებთან;

ბ. პირველწყარო ქართული ტექსტებისა თარგმანებთან.

ბიდირექციული კორპუსისათვის ტექსტების მარკირება-ანოტირების შემდეგ ტექსტების განთავსების პარამეტრები იქნე

მოდუსი II



Georgian Translational Corpus of Politics

Khatuna Beridze & Grigol Kakhiani
Batumi State University (Georgia)
beridze@illinois.edu, gkakhiani@gmail.com

GTCP is a digital bilingual collection of contemporary political science and of texts related to politics. The current version of the corpus can retrieve: (1) multiple definitions for a political term from various sources in Georgian, and (2) their contextual usage. The corpus alignment allows maximum output for a searched term, combining all the contextual data to appear simultaneously with its definition. The context-fed

digital terminological dictionary is a custom-made learner-friendly technological resource, e.g. search of the term *regime* totals 113 hits, combining both terms and contextual usage samples. The search of the term *democracy* retrieves 231 hits, including contemporary adjectives introduced by the political scholars during the fourth wave of democratization: tutelary democracy, Consociational democracy, controlled democracy, restrictive democracy, Male democracy, Delegative democracy, Defective democracy etc. But it does not necessarily mean that all the terms appear with the similar richness of the output. The corpus-building process is in progress. Since the corpus contains both parallel and comparable texts, we keep to the term „translational corpus".

The GTCP consists of the balanced sections and subsections:

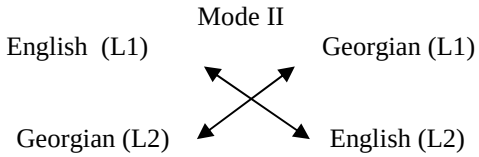
1. Parallel Corpora	**2. Georgian Comparable Corpora**	**3. English / multilingual Comparable Corpora**
1. English literature in translation (upcoming)	1. Political news	1. Political news
2. Georgian literature in translation (upcoming)	2. Politics, interview	2. Politics, interview
3. Political news analysis	3. Political science	3. Political science
4. Political speeches	4. Political news analysis	4. Political news analysis
5. Politics, literature, translation	5. Political speeches	5. Political speeches
6. Political news		
7. Politics, interview		
8. Political science		

The GTCP content consists of written texts, comprising translations from Georgian into English and vice versa. At the initial step, a large collection of scholarly texts on politics have been compiled, carefully examined for several months and filtered into comparable and parallel corpora. There are three major criteria unifying the texts: they are contemporary, contain large amount of newly-coined terms, and deal with the transition period of former Soviet states, Georgia among them. At the next step, I carried out custom-made translations, collecting the terms, consulting on their definitions, experimenting on the language ecology by means of resisting transcription or transliteration of the translatable concepts, specifically, avoiding new borrowings. In a team work, we worked out the design of the corpus and balanced the textual resources for each section. The software was specifically developed to the needs of the corpus. Currently, a small team is multitasking on (a) annotation of the texts, concurrently collecting new texts and selecting them for all sections of the corpus in a balanced manner, (b) documentation of the corpus texts, adding extratextual data. Customarily, the corpus uses a separate header file for each text; we plan to follow this tradition once we finalize the work on the meta-data. The recent documented files are accurately numbered, registered and attached to the specific section for the ease of their identification. The linguistics-oriented corpus study enables scholars of translation studies to compare and analyze similarities and differences of original and translated texts; hence, we aligned the sentences in the texts in Mode I:

English L1 to Georgian L2

All parallel lines of the texts are numbered successively; parallel texts are structured into coherent paragraphs, with possible ones, two or more than three sentences. This structuring method is customarily

employed for the alignment of the interviews. The structures of the rest of the parallel texts are arranged as 1 L1 to 1 L2 except where translation may apply grammatical transformation and break L1 grammatical structure into 2 in the L2. Currently, the corpus is unidirectional. Having marked up and annotated texts for the bidirectional corpus, the following text alignment mode will be introduced:



ეროვნული სამეცნიერო ბიბლიოთეკის ხელნაწერთა ციფრული არქივი

მანანა ბუკია, ირაკლი ღარიბაშვილი, ნინო ფანცხავა
თსუ ეროვნული სამეცნიერო ბიბლიოთეკა (საქართველო)
manbuki@rambler.ru; igar@hotmail.com, Nino.p99@gmail.com

თსუ ეროვნული სამეცნიერო ბიბლიოთეკის ფილმოთეკის განყოფილებაში დაცულია უნიკალური მასალა – საზღვარგარეთ არსებულ მანუსკრიპტორიუმებში შემონახული ქართული, ბერძნული, არაბული, სომხური ხელნაწერების მიკროფილმები.

ფონდი მოიცავს ათონის მთის, სინის მთის, იერუსალიმის მართლმადიდებლური საპატრიარქოს ბიბლიოთეკის ხელნაწერთა კოლექციებს, ასევე – ცალკეული ხელნაწერების მიკროფილმებს ოქსფორდის ბოდლის ბიბლიოთეკიდან, ნეაპოლის ბიბლიოთეკიდან, ბრიტანეთის მუზეუმიდან.

ხელნაწერთა მცირე ნაწილი აღწერილია და გამოცემულია, უმეტესი ნაწილი ჯერაც ელოდება მკვლევარს.

ამ უნიკალურ წყაროებზე მკვლევართათვის ხელმისაწვდომობის გაზრდისა და მათი დაცვის გაუმჯობესების მიზნით ბიბლიოთეკაში მიმდინარეობს ამ ხელნაწერთა გაციფრების პროექტი.

ხელნაწერთა მიკროფილმების დამუშავების ეს პროცესი ორი მიმართულებით მიმდინარეობს:

1. ხელნაწერთა ბიბლიოგრაფიული აღწერა-კატალოგიზაცია;
2. ციფრული ასლების არქივის შექმნა-გაციფრება და ბიბლიოთეკის ვებსერვერზე განთავსება.

ბიბლიოგრაფიული აღწერა, ერთი მხრივ, გულისხმობს სპეციალისტის მიერ ხელნაწერების შესწავლასა და სამეცნიერო დამუშავებას, მეორე მხრივ, დამუშავებული ინფორმაციის შეტანას ელექტრონულ მონაცემთა ბაზაში – ბიბლიოთეკის კატალოგში. კატალოგი შექმნილია ღია

პროგრამული უზრუნველყოფის – KOHA-ს საშუალებით. აღწერილობის მეთოდიკა დაფუძნებულია საერთაშორისო სტანდარტებზე (RDA, MARC21, DCRM).

ციფრული არქივი მუშავდება რამდენიმე ეტაპად: ხელნაწერთა გვერდების სკანირება (ელექტრონული გრაფიკული ასლების შექმნა), მეტამონაცემების დამატება და მონაცემთა ბაზაში ატვირთვა. შემდგომ ეტაპზე ვგეგმავთ გრაფიკული ასლების ტექსტურ ფორმატში გადაყვანას. აღნიშნული არქივი შექმნილია ღია პროგრამული უზრუნველყოფის Dspace-ის საშუალებით და დაფუძნებულია საერთაშორისო სტანდარტებზე (UNICODE, Dublin Core, OAI).

Digital Archive of Manuscripts of the National Scientific Library, TSU

Manana Bukia, Irakli Garibashvili, Nino Pantskhava
National Scientific Library, TSU (Georgia)
manbuki@rambler.ru, igar@hotmail.com, Nino.p99@gmail.com

The library film collection contains unique materials – microfilm copies of ancient Georgian, Greek, Arabic, Armenian manuscripts, preserved in foreign manuscriptoriums.

The collection includes collections of manuscripts from Mount Athos, Mount Sinai, Jerusalem Orthodox Patriarchate library and microfilms of manuscripts from the Bodleian library, Oxford, UK, Library of Naples, from the library of the British Museum.

A small portion of the manuscript has been studied and published (bibliography); most of them are still waiting for scholars.

In order to increase access to this unique source for scholars and improve preservation of manuscripts, the library works on this project of digitization.

This project of microfilmed manuscript processing implies two directions:

1. Bibliographic description – cataloging of manuscripts

2. Creation of digital copies of the archive and its web-publishing.

The bibliographic description involves, on the one hand, the specialist study of manuscripts and scientific bibliographic processing, and, on the other, input of processed information into the database -- the library catalog. The catalog is based on an open-source software - KOHA. The methodology of processing-description is based on international standards (RDA, MARC21, DCRM).

A digital Archive is being developed in several stages: scanning pages of manuscripts (graphical copies), adding metadata and upload to a database. Production of copies in plain text is planned for the next stage. The Archive is created by means of open-source software Dspace and entry is based on international standards (UNICODE, Dublin Core, OAI).

დონეებად დაყოფილი საკითხავი ქართული ტექსტების შედგენისა და რედაქტირების ფორმულისათვის

კახა გაბუნია, ნათია გორგაძე, რატი სხირტლაძე
ივანე ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)
სამოქალაქო ინტეგრაციისა და ეროვნებათშორისი ურთიერთობების ცენტრი (საქართველო)
ენის მოდელირების ასოციაცია (საქართველო)
kgabunia@cciir.ge, ngorgadze@cciir.ge, rati2008@gmail.com

2012 წლის ნოემბრიდან „სამოქალაქო ინტეგრაციისა და ეროვნებათშორისი ურთიერთობების ცენტრმა“ ამერიკის განვითარების სააგენტოს (USAID) დაფინანსებით დაიწყო მუშაობა პროექტზე: „მულტიწიგნიერების განვითარება არაქართულენოვან სკოლებში“. პროექტის ფარგლებში დაგეგმილია **დონეებად დაყოფილი საკითხავი ქართული ტექსტების** შექმნა არაქართულენოვანი მოზარდებისათვის, რომლებიც უმცირესობებით კომპაქტურად დასახლებულ რეგიონებში არაქართულენოვან სკოლებში სწავლობენ.

უპირველეს ყოვლისა, გასათვალისწინებელი იყო სპეციფიკა, რომელიც ქართულის, როგორც მეორე ენის, სწავლებას ახლავს თან; ეთნიკური უმცირესობების წარმომადგენელი ბავშვების უმრავლესობა სკოლაში შესვლისას საერთოდ ვერ ფლობს (ან ძალიან სუსტად ფლობს) სახელმწიფო ენას. ამ მხრივ, ცხადია, მათთვის კითხვის უნარების განვითარება სრულიად განსხვავებულ მიდგომას მოითხოვს და, უპირველესად, საკითხავი ტექსტების გააზრებულ და მეცნიერულად დასაბუთებულ შერჩევაზეა ორიენტირებული.

წიგნიერების ამოსავალი წერტილი კითხვის უნარის განვითარებაა, რაც მოიცავს გარკვეული კოგნიტური პროცესების ერთობლიობას – სიტყვების დეკოდირებითა და გაგებით დაწყებული, ტექსტის სიღრმისეული გააზრებით დამთავრებული. კითხვის უნარის განვითარება მრავალი კომპონენტისგან შედგება, რომლებიც ფონოლოგიურ, ორთოგრაფიულ, სემანტიკურ, სინტაქსურ და მორფოლოგიურ კომპონენტებს მოიცავს და აუცილებელ საფუძველს ქმნის გამართული და გააზრებული კითხვისათვის.

ბუნებრივია, ბავშვს უნდა მიეწოდოს გარკვეულ კანონზომიერებებზე დაფუძნებული საკითხავი მასალა, რათა კითხვის პროცესი მისთვის საინტერესო და, ამავე დროს, სასარგებლოც იყოს; სხვაგვარად რომ ვთქვათ, შერჩეული მასალა უნდა იყოს შემეცნებითი და, ამავდროულად, სასწავლო დანიშნულებაც ჰქონდეს. წიგნიერების განვითარებისათვის ევროპასა და ამერიკაში აპრობირებული და მიღებულია ე. წ. „დონეებად დაყოფილი წიგნები“.

ეს არის სირთულის მიხედვით დახარისხებული პატარ-პატარა ტექსტების ერთობლიობა, რომლებიც დაწყებითი საფეხურის მოსწავლეებისათვის არის განკუთვნილი და მათი კითხვითი გამოცდილების საფეხურებრივ ზრდას უზრუნველყოფს. ამ ტექსტების საშუალებით, „წარმატებული მკითხველის“ ჩამოყალიბებისათვის აუცილებელი უნარები თანდათანობით, ეტაპობრივად ვითარდება. დონეებად დაყოფილი მასალა უარყოფს „ერთ ყალიბში მოქცევის“ მიდგომას კითხვის უნარის განვითარებასთან მიმართებით და თითოეულ ბავშვს აძლევს საშუალებას, გაუმჯობესოს ეს უნარები ინდივიდუალური მახასიათებლების გათვალისწინებით.

საკითხავი მასალის დონეებად გადანაწილება სპეციალური **ფორმულების** საშუალებით ხდება. ის ითვალისწინებს სხვადასხვა ასპექტს, უპირველეს ყოვლისა კი – მკითხველის შესაძლებლობებს. ეს ასპექტები იმგვარად არის მოწესრიგებული, რომ მასალა მკითხველის კითხვის უნარს პატარ-პატარა, თანმიმდევრული ნაბიჯებით ავითარებს. ეს ფორმულა სხვადასხვა ავტორისა და გამომცემლობის მიერ შემუშავებული კრიტერიუმების ერთობლიობას წარმოადგენს.

თანამედროვე წიგნებში დონეებად დაყოფის კრიტერიუმები ბევრად მრავალფეროვანია და მკაცრად კონტროლირებადი ლექსიკითა და კონცეპტუალური მახასიათებლების მაღალი კონცენტრაციით ხასიათდება. ზოგადად, კრიტერიუმები, რომელთა მიხედვითაც წიგნების დაყოფა ხდება დონეებად, ტექსტური მახასიათებლების, კონტროლირებადი ლექსიკის, წინადადებების სიგრძისა და კომპლექსურობის, თემატიკისა და შინაარსის, შინაარსობრივი სტრუქტურისა და ენობრივი მახასიათებლების ერთობლიობას წარმოადგენს. ტექსტებს სამი განსხვავებული დონის სირთულის მიხედვით განიხილავენ. ესენია: სიტყვის, წინადადებისა და მონაკვეთის დონეები.

პროექტის ფარგლებში შემუშავდა დონეებად დაყოფის ქართული ენის სპეციფიკაზე მისადაგებული ორიგინალური ფორმულა, რომლის პროგრამულ პროდუქტად და ტექსტების რედაქტირების ინსტრუმენტად ქცევა ძალიან დაგვეხმარა ტექსტების ბაზის ფორმირებაში.

ფორმულაზე დაყრდნობით შეიქმნა პროგრამა, რომელმაც სპეციალურად შედგენილი ტექსტები შემდეგი პარამეტრების მიხედვით დაამუშავა:

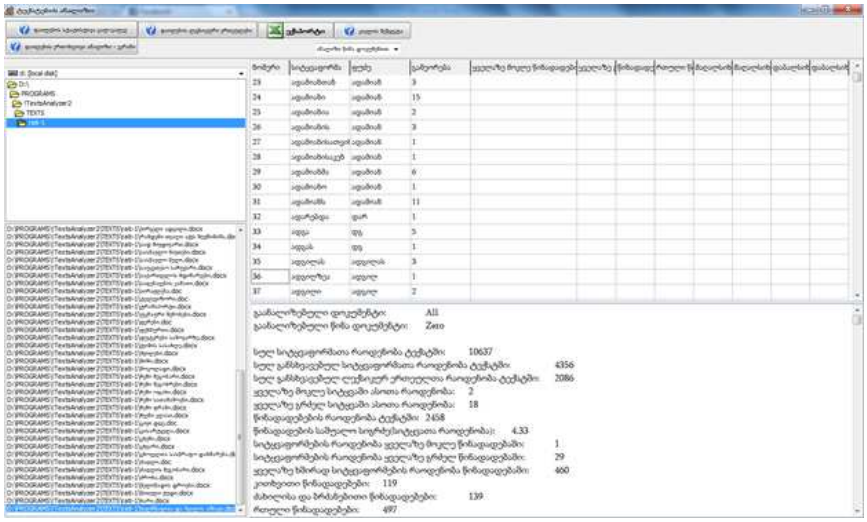
I. წინადადების დონე:

1. რამდენი წინადადებაა ტექსტში;
2. რამდენი კითხვითი წინადადებაა ტექსტში;
3. რამდენია ძახილისა და ბრძანებითი წინადადება;
4. რამდენი რთული წინადადებაა;
5. რამდენი სიტყვაფორმისგან შედგება ყველაზე მცირე წინადადება;
5. რამდენი სიტყვაფორმისგან შედგება ყველაზე გრძელი წინადადება;
6. წინადადების საშუალო სიგრძე;
7. შედარება წინა ტექსტთან (რამდენი წინადადებით მეტია ან ნაკლები).

II. სიტყვის დონე

1. რამდენი სიტყვაფორმაა სულ ტექსტში;
2. სიტყვაფორმების სიხშირე;
3. რამდენი ასოსგან შედგება ყველაზე გრძელი (ყველაზე მოკლე) სიტყვა;
4. შედარება წინა ტექსტთან (რამდენი სიტყვაფორმით მეტია ან ნაკლები);
5. რამდენი ლექსიკური ერთეულია ტექსტში;
6. რამდენი მაღალსიხშირული ერთეულია გამოყენებული წინასწარ მოცემული 1000 საბაზისო სიტყვიდან;
7. რამდენი დაბალსიხშირული ერთეულია გამოყენებული ტექსტში;
8. პროცენტული თანაფარდობა ტექსტში მაღალსიხშირული სიტყვებისა სიტყვათა სრულ რაოდენობასთან;
9. პროცენტული თანაფარდობა ტექსტში დაბალსიხშირული სიტყვებისა სიტყვათა სრულ რაოდენობასთან;

10. შედარება წინა ტექსტთან (რამდენი ახალი/განსხვავებული ლექსიკური ერთეულია წინა ტექსტთან შედარებით).



ნახ. 1

ტექსტების რედაქტირების თვალსაზრისით მოცემულმა პროგრამამ (იხ.) უაღრესად საჭირო ინსტრუმენტის სახე მიიღო, რომლის მეშვეობითაც ხორციელდება ტექსტების კორექტირება-რედაქტირება.

On the formula for developing and editing of Georgian reading texts divided into levels

Kakha Gabunia, Natia Gorgadze, Rati Skhirtladze
Ivane Javakhishvili Tbilisi State University (Georgia)
Center for Civil Integration and Inter-ethnic Relations (Georgia)
Language Modeling Association (Georgia)
kgabunia@cciir.ge, ngorgadze@cciir.ge, rati2008@gmail.com

Since November, 2012, „Center for Civil Integration and Inter-Ethnic Relations“, funded by USAID, started working on the project: „Development of Multi-literacy in Non-Georgian Schools“. As part of the project, it is planned to create Georgian reading texts divided into levels for non-Georgian-speaking adolescents, enrolled in non-Georgian schools in regions densely populated by ethnic minorities.

Initially, the specific character of teaching Georgian as a second language was to be taken into account; most of the children from ethnic Minorities cannot speak the national language (or speak it very poorly) when they start the school. With respect to that, the development of their reading skills naturally requires a distinct approach. This approach is mostly focused on choosing the texts in a sensible and scientifically proven way.

The starting point of Literacy is the development of reading skills, including a combination of several cognitive processes: from decoding of words to in-depth understanding of a text. The development of reading skills contains many components, including orthographic, semantic, syntactic and morphological ones. They make up a necessary basis for correct and meaningful reading.

Naturally enough, the reading material given to a child should be based on specific criteria to make the reading process interesting as well as educational; on other words, the chosen material should be educational and, at the same time, amusing. In the area of Literacy-development, Europe and USA have approved and accepted so called „books divided into levels."

These are unities of little texts, sorted according to difficulty, intended for children in beginning grades to guarantee the development of their reading skill step by step. Using these texts, the skills necessary for the development of a „successful reader" are acquired gradually. The divided material denies the „putting in same caliber" approach to the development of reading skills and allows each child to develop their abilities taking their individual characteristic into account.

The reading material is divided into levels using special formulas. It takes several aspects into account, first of all – the abilities of a reader. These aspects are ordered in such a way that the reading skill of a reader is developed in small, consistent steps. This formula is a combination of the criteria developed by different authors and publishers.

The criteria of dividing modern books are much more diverse, with a high concentration of strictly controlled vocabulary and conceptual characteristics. Generally, these criteria are a combination of textual characteristics, controllable vocabulary, length and complexity of sentences, themes and contents, the structure of the content and linguistic characteristics. The text difficulty is reviewed in three different aspects. These are the levels of words, sentences and sections.

As part of the project, an original formula for the division in levels, adapted to the Georgian language, was developed. Turning it into a software and text-editing tool helped us very much in forming the text database.

Based on the formula, a program was created, allowing for the development of special texts using these parameters:

I. Sentence level:

1. How many sentences are in a text;
2. How many interrogative sentences are in a text;
3. How many exclamatory sentences are in a text;
4. How many complex sentences are there;
5. How many word-forms does the smallest sentence contain;
6. Average length of a sentence;
7. Comparison to the last text (difference between sentence numbers);

II. Word level:

1. How many word-forms are in a text;
2. Frequency of word-forms;
3. How many letters does the smallest (largest) word have;
4. Comparison to the last text (difference between word-form numbers);
5. How many lexical units are in a text;
6. How many high-frequency units are used from 1000 given base words;
7. How many low-frequency words are used in a text;
8. The ration of high-frequency words and the total number of words;
9. The ratio of low-frequency words and the total number of words;
10. Comparison to the last text (how many different/new lexical units)

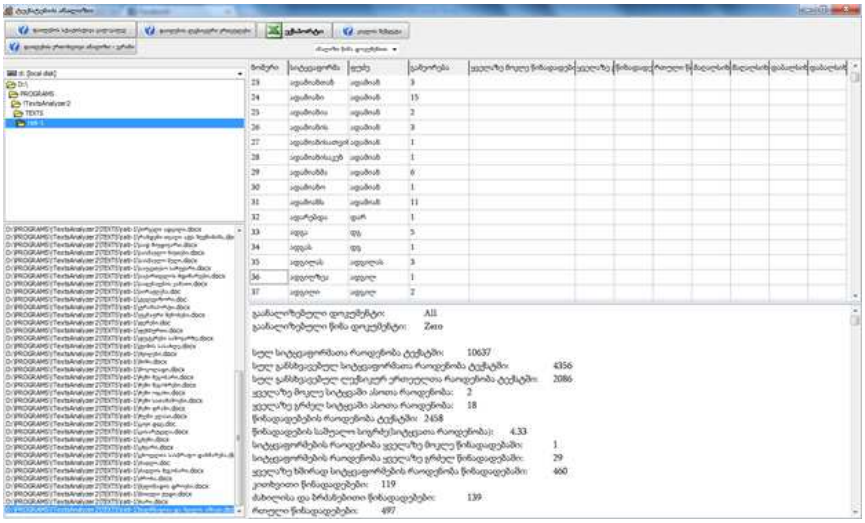


Fig. 1

As a text-editing software, this program (see picture 1) became a very useful tool for correcting and editing texts.

დიაქრონიული კორპუსები – ამოცანები, პრობლემები და მათი გადაწყვეტა

იოსტ გიპერტი
ფრანკფურტის გოეთეს უნივერსიტეტი (გერმანია)
gippert@em.uni-frankfurt.de

მოხსენება ეხება იმ სპეციფიკურ ამოცანებს, რომლებსაც წარმოგვიდგენს დიაქრონიული კორპუსები ანუ ისეთი კორპუსები, რომლებიც ისეა შედგენილი, რომ მოიცვას ენის ცვალებადობა ხანგრძლივი დროის განმავლობაში. ქართული ეროვნული კორპუსის პროექტზე დაყრდნობით გამოვლენილი და ჩამოთვლილია ის ძირითადი პრობლემები, რომლებიც დგება დიაქრონიული კორპუსის წინაშე სინქრონიულთან შედარებით (მაგალითად, სიტყვათა (ორთო)გრაფიული რეპრეზენტაციის ცვლილებებთან, სახელური და ზმნური ფორმების წარმოების ცვლილებებთან, სახელთა და ზმნათა მორფოსინტაქსში მომხდარ ცვლილებებთან დაკავშირებული პრობლემები და ა.შ.), ასევე წარმოდგენილია ამ პრობლემათა გადაჭრის გზები.

Diachronic corpora - tasks, problems and solutions

Jost Gippert
Goethe University Frankfurt (Germany)
gippert@em.uni-frankfurt.de

The paper addresses the peculiar tasks set by diachronic corpora, i.e., corpora that are designed to cover a language changing in the course of a longer period of time. On the basis of the Georgian National Corpus project, the main problems diachronic corpora have to face in comparison with synchronic corpora are outlined and exemplified (among others, problems of changes in the (ortho)graphic representation of words, changes in the formation of nominal and verbal forms, changes in the morphosyntax of nouns and verbs etc.), as well as solutions to overcome these problems.

ქართული ენის ელექტრონული სწავლების კურსი არაქართველთათვის

ქეთევან გოჩიტაშვილი, მარიამ მანჯგალაძე
ივანე ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)
თსუ არნ. ჩიქობავას სახ. ენათმეცნიერების ინსტიტუტი (საქართველო)
ketevan.gochitashvili@tsu.ge, mariam@ice.ge

ენის ელექტრონული სწავლების კურსების მზარდი რაოდენობა ადასტურებს, რომ საზოგადოების მხრიდან მათზე ინტერესი და მოთხოვნა ინფორმაციული და კომუნიკაციური ტექნოლოგიების შესაძლებლობების განვითარების შესაბამისად იზრდება.

ენის ტექნოლოგიებზე/კომპიუტერზე/მულტიმედიაზე დაფუძნებული სწავლების თვალსაჩინო უპირატესობაა ერთი კურსის ფარგლებში მრავალფეროვანი საშუალებებით: ტექსტის, აუდიო და ვიზუალური მასალის, გრაფიკის, ანიმაციის, ვიდეორგოლის გამოყენებით ცოდნის გადაცემა ტრადიციულ წერილობით ტექსტებსა და ზეპირმეტყველებასთან ერთად.

სასწავლო მასალის ნაკლებობა, ერთფეროვნება, არაავთენტურობა და ხელოვნური ხასიათი სწავლა/სწავლების პროცესს მნიშვნელოვნად აფერხებს. თანამედროვე ელექტრონული/დისტანციური საგანმანათლებლო ტექნოლოგიები კი იძლევა საშუალებას, სტუდენტს შევთავაზოთ საჭირო ინფორმაციის განუსაზღვრელი რაოდენობა.

ენის ელექტრონული სწავლების კურსს შეუძლია ენის შესწავლის რთული და ხანგრძლივი პროცესი შედარებით გააადვილოს. სტუდენტი არ იზღუდება დროსა და სივრცეში, აქვს შესაძლებლობა მიუბრუნდეს მასალას, გაიმეოროს, თავად განსაზღვროს საკუთარი დროის ლიმიტი.

პროექტი **„ქართული ენის ელექტრონული სწავლების კურსი“** ენის ფლობის დონეთა ევროპული სტანდარტის მიხედვითაა აგებული. A1 დონეზე 40 თემაა შემოთავაზებული და ყველა ინსტრუქცია და გრამატიკის ბლოკი წარმოდგენილია როგორც ქართულ, ასევე სომხურ და აზერბაიჯანულ ენებზე. კურსი შეიცავს 316 (რვა ტიპის) ინტერაქტიულ სავარჯიშოს, 142 საკითხავ ტექსტს, 702 ზეპირ და წერით სავარჯიშოს, 168 თვითშეფასების ტესტს, 125 აუდიოფაილს, 33 ანიმაციურ ფაილს, 12 ვიდეო და 161 გრაფიკულ ფაილს. იგი **პირველი უფასო ელექტრონული პროდუქტია,** რომელიც ხელს უწყობს აზერბაიჯანელი და სომეხი მოსახლეობის ინტეგრაციას სახელმწიფო ენის შესწავლის თვალსაზრისით.

გაკვეთილების/თემების ციკლი საშუალებას აძლევს კომპიუტერის ნებისმიერ მომხმარებელს, ასაკის შეუზღუდავად, ეტაპობრივად, საფუძვლიანად შეისწავლოს ქართული ენა – ლექსიკა, გრამატიკა, ფრაზეოლოგია; გამართოს მეტყველება, დახვეწოს წერის კულტურა.

მოსმენა/ლაპარაკის ბლოკი აუდიო ან ვიდეოფაილებისა და მასზე დაფუძნებული სხვადასხვა ტიპის დავალებებისგან შედგება. მოსასმენი ტექსტი კომუნიკაციურია (მეტწილად – დიალოგები). აუდიო ან ვიდეოტექსტის მოსმენის შემდეგ ენის შემსწავლელი ასრულებს დავალებებს (როგორც ინტერაქტიულს, ისე ინდივიდუალურს), რომლებიც მოსმენისა და ლაპარაკის უნარ-ჩვევების განვითარებაზეა ორიენტირებული.

კითხვა/წერის ბლოკი მოიცავს საკითხავ ტექსტებს და მათზე დაფუძნებულ სხვადასხვა ტიპის სავარჯიშოებს; ეს სავარჯიშოები კითხვისა და წერის უნარების განვითარებაზეა ორიენტირებული.

გრამატიკის ბლოკი პრაქტიკული გრამატიკის საკითხებს მოიცავს და ენის შემსწავლელს საშუალებას აძლევს გაეცნოს ამა თუ იმ გრამატიკული ცნებისა თუ კატეგორიის განმარტებას; სასწავლო კურსის ეს ნაწილი არასავალდებულო ხასიათს ატარებს და მხოლოდ ისეთი მომხმარებლისთვისაა განკუთვნილი, რომელიც ენის სისტემურ შესწავლაზეა ორიენტირებული. თეორიულ მასალას მოსდევს პრაქტიკული სავარჯიშოები, რომლებიც კურსის ავტორთა მიერ შექმნილ მასალაზეა დაფუძნებული და წარმოდგენილია ელექტრონულ კურსში.

ლექსიკის ბლოკი ორიენტირებულია პროგრამის მომხმარებლის მიერ სიტყვათა მარაგის შესწავლის გაიოლებაზე. ამ მხრივ განსაკუთრებით საინტერესოა სპეციალურად ენის შემსწავლელთათვის შექმნილი კროსვორდები. ელექტრონული კურსი მოიცავს სხვა ტიპის მრავალფეროვან დავალებებსაც.

ყოველ გაკვეთილს ახლავს **ახალი სიტყვებისა და ფრაზების ლექსიკონი**. კურსს ერთვის მთელი მასალის მიხედვით გაერთიანებული ხუთი თარგმნითი ლექსიკონი.

„ქართული ენის ელექტრონული სწავლების კურსში“ გამოყენებულია შემდეგი ***სასწავლო მეთოდები:*** *სპირალური განმეორების პრინციპი,* ევრისტიკული, აუდიალური და ვიზუალური მეთოდები; როლური და სხვა დიდაქტიკური თამაშები; ქმედებაზე ორიენტირებული სწავლება; ინდივიდუალური მუშაობა; ტექსტზე მუშაობის მეთოდი; ვერბალური, ანუ ზეპირ-სიტყვიერი მეთოდი, ახსნა-განმარტებითი მეთოდები. სტუდენტებს ვთავაზობთ მოდელირებულ სიტუაციებსაც.

გარდა საგნობრივი ცოდნისა, ელექტრონული კურსი შემსწავლელს პიროვნულ უნარებსაც უვითარებს. მაგალითად:

კომუნიკაციის უნარი:

- შეუძლია ზეპირი და წერილობითი კომუნიკაცია ქართულ ენაზე.
- იცავს კომუნიკაციის წესებს.

სწავლის უნარი:

- იძენს დამოუკიდებლად მუშაობის უნარ-ჩვევებს.
- შეუძლია საკუთარი ცოდნის მუდმივად და თანამიმდევრულად შემოწმება-შეფასება.

ღირებულებები:

- იცნობს ქართულ და ზოგადსაკაცობრიო კულტურულ ღირებულებებს.
- იძენს სოციოკულტურულ/ინტერკულტურულ კომპეტენციებს და ა. შ.

eLearning of the Georgian language for non-Georgians

Ketevan Gochitashvili & Mariam Manjgaladze
Ivane Javakhishvili Tbilisi State University (Georgia)
Arn. Chikobava Institute of Linguistics, TSU (Georgia)
ketevan.gochitashvili@tsu.ge, mariam@ice.ge

The increasing number of e-learning courses of various languages indicates that the public interest and demand is increasing together with the development of the capabilities of information and communicational technologies.

The obvious advantage of technology/computer/multimedia-based language learning is to give lessons using diverse means such as texts, audio and visual materials, graphics, animations, video clips and at the same time the traditional written texts and speech.

The lack of educational materials, monotony, no authenticity and the artificial nature make the teaching/learning process hampered. Present-day electronic/distance educational technologies allow us to provide students with the unlimited amount of the necessary information.

eLearning courses can facilitate a difficult and long process of the language learning. Students are not limited in time and space, have the ability to return to the materials, to repeat them or to determine their own time limit.

The project „eLearning Course of Georgian" was created according to the European standards. A1 level has 40 topics, and all the instructions and the grammar block are presented both in Georgian and either in Azerbaijani or Armenian. At A2 and B1 levels have 30-30 topics. The course contains 316 (eight types) interactive exercises, 142 texts for reading, 702 oral and written exercises, 168 self-evaluation tests, 125 audio-files, 33 animated files, 12 video and 161 graphic files. It is the first free electronic product which contributes to the integration of Azerbaijani and Armenian population in terms of language learning.

The set of lessons enables any user of a computer to learn Georgian essentially, on a step-by-step basis: grammar, vocabulary, and phraseology; improve conversational skills; learn and improve writing skills.

Listening/speaking block contains audio and video files and various tasks, based on them. As a rule, a text for listening is communicative (mostly, dialogues). After listening to an audio- or video-text, a learner is given tasks (both interactive and individual), designed for the development of listening and conversational skills.

Reading/writing block contains texts for reading and various exercises, based on them; the exercises are designed for the sake of the development of reading and writing skills. At the initial stage (first 6 lessons), a special emphasis is laid upon on the development and enhancement of motor skills.

Grammar block contains issues of practical grammar and gives an opportunity to learners to get acquainted with definitions of various grammatical notions and categories; this part of the course is optional and is meant for learners aiming at the systemic acquisition of the language. The theoretical

data are followed by practice exercises, based on the materials created by the course authors and thus included in it.

Vocabulary block is designed to make the acquisition of the lexical stock easier for learners. There are crosswords, specially created for language learners. The eLearning course contains a variety of other tasks.

Every lesson is accompanied by **a dictionary of new words and phrases**, occurring in an individual lesson. Totally, five bilingual dictionaries are included in the project.

In the „eLearning Course of Georgian" we have used the following methods: the spiral repetition principle, heuristic, audio and visual methods, role-playing and other didactic games, action-oriented teaching, individual work, text-oriented methods, verbal or oral methods and explanation methods. We also offer students some simulated situations.

Besides the subject knowledge, eLearning courses develop personal skills. For example:

Communication skills:

He/She

- is able to communicate in oral and written Georgian.
- follows the communication rules.

Learning skills:

He/She

- acquires an ability to work independently.
- is able to check his knowledge constantly and consistently.

Values:

He/She

- is familiar with the Georgian and worldwide cultural values.
- acquires socio-cultural / intercultural competences etc.

ლექსიკონების შედგენის პროგრამული ინსტრუმენტი

ქეთევან დათუკიშვილი, ნანა ლოლაძე, მერაბ ზაკალაშვილი
ლინგვისტური ტექნოლოგიების ჯგუფი (საქართველო)
k_datukishvili@yahoo.com, nanaloladze@yahoo.com, GILC@Wanex.ge

თანამედროვე ტექნოლოგიები საშუალებას იძლევა შეიქმნას სპეციალური ინსტრუმენტები, რომლებიც უზრუნველყოფენ ლექსიკოგრაფიული სამუშაოების ეფექტურად წარმართვას.

ლინგვისტური ტექნოლოგიების ჯგუფში შეიქმნა პროგრამა „ლექსიკოგრაფი", რომლის დახმარებითაც მომზადდა სასწავლო ტიპის განმარტებითი ლექსიკონი - „ქართული ლექსიკონი". წარმოვადგენთ ამ პროგრამის სტრუქტურასა და მუშაობის პრინციპს.

პროგრამა „ლექსიკოგრაფი“ საშუალებას იძლევა, ერთი მხრივ, ეფექტურად დამუშავდეს სალექსიკონო მასალა, მეორე მხრივ, განხორციელდეს მასალის ფორმატირება როგორც ბეჭდური, ისე ელექტრონული გამოცემებისათვის.

პროგრამის საშუალებით სალექსიკონო სტატიები იქმნება მონაცემთა ბაზის სახით. ბაზა შედგენილია სალექსიკონო სტატიის სტრუქტურის შესაბამისად. კერძოდ, სიტყვა-სტატია დაშლილია სტრუქტურულ ელემენტებად. პროგრამაში თითოეული ამ ელემენტისათვის გამოყოფილია ცალკე ველი და ინფორმაცია შეგვყავს ველების მიხედვით. წარმოდგენილია შემდეგი **ველები:**

- განმარტება
- ილუსტრაცია
- გრამატიკული ფორმა (სახელებისათვის – მხოლობითი რიცხვის ნათესაობითი ბრუნვის ფორმა, ზმნებისთვის – მხოლობითი რიცხვის მესამე პირის ფორმა)
- მეტყველების ნაწილი
- ფუნქციონირების სფერო (საუბრ., ძვ. და სხვ.) და ა.შ.

კონკრეტულ სიტყვასთან შეიძლება მითითებული იყოს ინფორმაცია ერთი, ორი ან მეტი ველის მიხედვით. ველებში განთავსებული ინფორმაციის შეერთებით ვიღებთ სალექსიკონო სიტყვა-სტატიას.

ზოგიერთი სიტყვა პოლისემიურია, ზოგი სიტყვის ბუდეში განიმარტება ფრაზეოლოგიური გამოთქმები ან ამ სიტყვის ფუძიდან ნაწარმოები ახალი ლექსიკური ერთეულები. თითოეული ეს ნაწილი წარმოადგენს სტატიის შემადგენელ ავტონომიურ ბლოკს. მათ ქვესტატიებს ვუწოდებთ. პროგრამაში წარმოდგენილია შემდეგი **ქვესტატიები:**

- პოლისემია
- ნაწარმოები სიტყვები
- ფრაზეოლოგიური გამოთქმები

ბაზაში ინფორმაციის შეყვანისას პირველ რიგში ვქმნით სტატიას, შემდეგ ვირჩევთ ქვესტატიას, თითოეულ ქვესტატიას აქვს ველების ჩამონათვალი და ვავსებთ ამ ველებს.

სალექსიკონო მასალის ამგვარად მომზადება იძლევა ინფორმაციის ავტომატურად მოძიებისა და მართვის საშუალებას, რაც აადვილებს როგორც განმარტებების შექმნის, ისე რედაქტირების პროცესს. შესაძლებელია სალექსიკონო სტატიების დახარისხება ყველა იმ პარამეტრის მიხედვით, რომლებიც ენიჭება ამა თუ იმ ქვესტატიასა თუ ველს. მაგალითად, შეიძლება მოვიძიოთ სტატიები მეტყველების ნაწილის მიხედვით, ფუნქციონირების სფეროს მიხედვით და ა. შ.

მონაცემთა ბაზაში ამგვარი სახით წარმოდგენილი მასალა კარგ საფუძველს წარმოადგენს ელექტრონული ლექსიკონებისთვის. ბეჭდური გამოცემებისთვის კი პროგრამა „ლექსიკოგრაფს“ აქვს სპეციალური რედაქტორი ლექსიკონის (წიგნის) დიზაინის შესაქმნელად.

დიზაინის რედაქტორით შესაძლებელია სხვადასხვა სახის სალექსიკონო ფორმატის მიღება. მოცემული ველებისათვის (განმარტება, ილუსტრაცია და სხვ.) შეგვიძლია შევქმნათ ნებისმიერი ფორმატი (შრიფტის ზომა, სახეობა და ა. შ.). გარდა ამისა, თითოეულ ველს შეგვიძლია მივანიჭოთ სპეციალური აღნიშვნა. მაგალითად: სინონ., ანტონ. და სხვ. დიზაინის რედაქტორი უზრუნველყოფს შერჩეული ფორმატის მიხედვით მთელ ლექსიკონში შესაბამისი ცვლილებების შეტანას ავტომატურად. ყოველივე ეს აადვილებს ლექსიკონზე მუშაობის პროცესს და გვიცავს მექანიკური შეცდომებისაგან.

პროგრამით შესაძლებელია ლექსიკონზე მუშაობის პროცესში სტატისტიკური მონაცემების გათვალისწინება: რამდენი სტატიაა შექმნილი, რამდენი ფრაზეოლოგიური გამოთქმაა, რამდენ სიტყვასთან დასტურდება გადატანითი მნიშვნელობა და ა.შ.

პროგრამა იძლევა სამუშაოს ადმინისტრირების საშუალებას. მუშაობა შესაძლებელია ინტერნეტის მეშვეობით. ამდენად, მონაცემთა ბაზასთან ერთდროულად მუშაობს რამდენიმე ლექსიკოგრაფი. ყველა მონაწილეს აქვს მინიჭებული გარკვეული უფლებები. ბაზაში განთავსებული ინფორმაციის ნახვა შეუძლია ყველას, ხოლო ინფორმაციის შეცვლა (დამატება ან რედაქტირება) თითოეულს შეუძლია მხოლოდ მისთვის განკუთვნილ მონაკვეთში. ამგვარი მიდგომა ხელს უწყობს სამუშაო პროცესის ორგანიზებულად მართვას.

პროგრამა არის უნივერსალური ინსტრუმენტი, რომლითაც შესაძლებელია ნებისმიერი ტიპის (განმარტებითი, ორენოვანი და ა.შ.) ლექსიკონის შექმნა. პროგრამის ზოგადი სტრუქტურა მოიცავს ყველა ტიპის ლექსიკონის სავარაუდო სქემას. ლექსიკონის სპეციფიკის მიხედვით შესაძლებელია ქვესტატიების ანდა ველების ჩამონათვალის კორექტირება. ამ ფორმატით შეყვანილი ლექსიკონების საერთო ბაზა მოგვცემს ლექსიკონთა მონაცემების შედარებისა და ინფორმაციის ადვილად მართვის შესაძლებლობას.

A Software Tool for Dictionary Compilation

Ketevan Datukishvili, Nana Loladze, Merab Zakalashvili
Linguistic Technologies Group (Georgia)
k_datukishvili@yahoo.com, nanaloladze@yahoo.com, GILCE@Wanex.ge

Modern technologies enable us to create special tools making it possible to conduct the lexicographic work in an automated way.

The computer software „Lexicographer“ was created by the group of linguistic technologies. The software allowed compiling of a learner's explanatory dictionary, called „Georgian Dictionary“. Here we present the structure and working principles of the software.

The software „Lexicographer“ allows to process the dictionary material in a convenient editing interface, on the one hand, and to fulfil data formating on the other, both in print and electronic publications, on the other.

This software helps to create dictionary entries in the form of a database. The base is founded on the structure of a dictionary entry. Structural elements of an entry are outlined. A separate field is marked out for each element and the information enters according to those fields. The following **fields** are presented:

- definition
- illustration
- grammatical forms (for nouns – genitive case, sing.)
- part of speech

- functioning area (coll., old and others), etc.

A certain word may fill one, two or more fields. Assembling the information, given in the fields, provides for an entry of the dictionary.

Sometimes a word is polysemantic; collocations or nouns, derived from stems, are defined in the nest of same words. Each of such parts represents an autonomous block – component of an entry. We refer to them as sub-entries. There are the following **sub-entries** in the software:

- polisemnatic meanings
- derived words
- collocations

In order to input the information into the base, we initially create an entry, then choose a sub-entry. Each sub-entry is provided with a list of fields and we fill in those fields.

Preparing the material for the dictionary by means of this method, allows automatic searching and managing of information, facilitating both processes – defining and editing. It becomes also possible to sort dictionary entries according to all of the parameters, assigned to a sub-entry of a field. For instance, we can obtain entries according to a part of speech belonging, according to usage, etc.

The data, thus presented in the database, make a good foundation for electronic dictionaries. As for providing for print publications, the software „lexicographer" has the special design editor.

Using the design editor allows receiving different kinds of dictionary formats. We can create any format (font size, type, etc.) for given fields (definition, illustration, and others). Besides, each field can be assigned a special marking, such as syn. ant., etc. In accordance with the sample represented in the format, the design editor ensures automatic preparation of the overall dictionary material. In addition, numbering of homonyms and polysemantic words is automatically performed. All the above mentioned facilitate the process of working on the dictionary and are a good protection against possible mistakes.

The software helps observe statistics during the working process on the dictionary: how many entries have been created so far, how many phrasal expressions are given, how many words are evidenced with figurative meaning, etc.

The software allows for administration of the working process. It is possible to work on the software using any computer, through the internet. Thus, several lexicographers are able to work simultaneously, in the zone assigned to each of them. All participants have their rights. Anybody can view the information placed in the database, while each one can make changes in the information (adding or editing) only in the zone assigned to them. Such an approach ensures organized administration of activities.

This software is a universal tool, by means of which it is possible to compile any kind of a dictionary (explanatory, bilingual, etc.). The general structure of the software includes the schemes (drafts) for any types of dictionaries. It is available to correct sub-entries or the list of fields according to peculiar properties of a dictionary. The overall base of the dictionaries transferred to the digital devices in such format will enable us to compare the data of different dictionaries and to easily manage the information.

ელექტრონულ კორპუსებზე დაფუძნებული ახალი ქართულ-ინგლისური სასწავლო ლექსიკონი

სოფიკო დარასელია
ივანე ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)
sopod@yahoo.com

XX საუკუნის 80-იანი წლებიდან მოყოლებული ელექტრონული რესურსებისა და ენათა ეროვნული კორპუსების შექმნამ მთლიანად შეცვალა დამოკიდებულება როგორც ენების კვლევის, ისე ლექსიკონების შედგენისადმი.

მოხსენებაში განხილული იქნება ქართულ სინამდვილეში არსებული კორპუსების გამოყენების შესაძლებლობები „ქართულ-ინგლისური სასწავლო ლექსიკონის“ შედგენისას. კერძოდ, ვაჩვენებთ, როგორ ვიყენებთ ლექსიკონზე მუშაობისას ქართულ ანოტირებულ კორპუსს (GEKKO), ადამ კილგარიფისა და დევიდ თაგუელის „სიტყვათშეხამებითი მოდელების პროგრამას“ (Sketch Engine) და თინათინ მარგალიტაძის საერთო რედაქტორობით გამოცემულ „დიდ ინგლისურ-ქართულ ონლაინლექსიკონს“.

XX საუკუნის 20-იანი წლებიდან საფუძველი ჩაეყარა ლექსიკოგრაფიის ახალ ჟანრს, რომელსაც სასწავლო ლექსიკოგრაფია ეწოდა. ამ ჟანრს მისთვის დამახასიათებელი ლექსიკოგრაფიული პრინციპები ჩამოუყალიბდა, რომელთა შორის ერთ-ერთი უმნიშვნელოვანესია ლექსიკური ერთეულების სიხშირული პრინციპით შერჩევა. ქართულ-ინგლისური და, ზოგადად, ქართულ-ევროპული ტიპის სასწავლო ლექსიკონებისათვის, მნიშვნელოვანი ეტაპია ლექსიკონის ქართული ნაწილის დამუშავება, რაც სასწავლო ლექსიკოგრაფიის ჟანრული თავისებურებებიდან გამომდინარე, ელექტრონული კორპუსების გამოყენების გარეშე შეუძლებელია.

მოხსენებაში დეტალურად იქნება ნაჩვენები ქართულ-ინგლისურ სასწავლო ლექსიკონზე მუშაობის ეტაპები და მეთოდები ზემოხსენებული ელექტრონული კორპუსების გამოყენებით. განსაკუთრებული ყურადღება ენიჭება ქართულ შესიტყვებებს, რომელთა შერჩევაც ასევე სიხშირული პრინციპებით ხდება კორპუსებიდან.

მაგალითისათვის განვიხილოთ სიტყვა „სათუთი“.

არნოლდ ჩიქობავას საერთო რედაქტორობით გამოცემულ ქართული ენის განმარტებით ლექსიკონში სიტყვა **სათუთის** სამი სინონიმია გამოყოფილი: **ნაზი, ნებიერი** და **აზიზი.** საილუსტრაციო მასალაში დამოწმებულია მისი გამოყენების 6 შემთხვევა (აკაკი წერეთლის, ლეო ქიაჩელის, იოსებ გრიშაშვილის, ლავრენტი არდაზიანისა და გალაკტიონ ტაბიძის ნაწარმოებებიდან), რაც, რასაკვირველია, სრულყოფილად ვერ ასახავს აღნიშნული ქართული სიტყვის დღევანდელ მნიშვნელობებს.

ქართული ენის განმარტებითი ლექსიკონის მიხედვით გვხდება **სათუთის** შემდეგი შესიტყვებები:

- სათუთი ბავშვი
- სათუთი ქალი
- სათუთი ქალიშვილი

- სათუთი გაზაფხული
- სათუთი რითმა
- სათუთი გრძნობა

პაულ მოირერის კორპუსში სიტყვა **სათუთის** გამოყენების **126** შემთხვევა გვხვდება, მათ შორის იმ მნიშვნელობებით, რომლებიც არ არის ასახული ქართული ენის განმარტებით ლექსიკონში და, ბუნებრივია, ვერც იქნებოდა, რადგან განმარტებით ლექსიკონზე მუშაობა 60-იან წლებში დასრულდა და მასში ვერ აისახა ქართული ენის განვითარების შემდეგი ეტაპები. მაგალითად:

- სათუთი პროცესი
- სათუთი სიცოცხლე
- სათუთი დღე
- სათუთი პრობლემა და ა.შ.

ადამ კილგარიფისა და დევიდ თაგუელის სიტყვათშეხამებითი მოდელების (Sketch Engine) პროგრამაში კი ამ სიტყვის **134** დადასტურება გვაქვს :

- სათუთი ნაკეთობა
- სათუთი საქმე
- სათუთი სული
- სათუთი თემა და ა.შ.

ყოველივე ზემოაღნიშნულიდან **სათუთის დამატებით მნიშვნელობებად გამოიყოფა: წრფელი** (*სათუთი სული*), **დელიკატური, საფრთხილო, საჩოთირო** (*სათუთი პრობლემა, სათუთი თემა*), **გაფრთხილება რომ სჭირდება, ადვილად მტვრევადი/მსხვრევადი/მყიფე** (*სათუთი ნაკეთობა*) და ა.შ.

ამ მნიშვნელობების გაუთვალისწინებლად ახალი ტიპის სასწავლო ლექსიკონის შექმნა წარმოუდგენელია და ვფიქრობთ, რომ ქართულ-ევროპული ლექსიკონებისათვის ქართული ნაწილის მომზადება სწორედ კორპუსებიდან ამოღებული მაგალითების გაანალიზებით უნდა დაიწყოს და მას დაეფუძნოს.

„ქართულ-ინგლისური სასწავლო ლექსიკონისათვის“ შექმნილი ქართული ნაწილი მომავალში წარმატებით შეიძლება იქნეს გამოყენებული სხვა ქართულ-ევროპული ტიპის სასწავლო ლექსიკონების შედგენისათვის.

Corpus-Based New Georgian-English Learner's Dictionary

Sophiko Daraselia
Ivane Javakhishvili Tbilisi State University (Georgia)
sopod@yahoo.com

Since the late 1980s, the advent of electronic text databases and national corpora totally changed the approaches both to language and to dictionary compilation.

In the present paper, I will discuss the application of existing Georgian corpora during the compilation of a Georgian-English Learner's Dictionary, viz. *Georgian Annotated Corpus (GEKKO)* by Paul Meurer, *Sketch Engine* by Adam Kilgariff and David Tugwell and *Comprehensive English-Georgian Online Dictionary* under the general editorship of Tinatin Margalitadze.

Since the 1920s, a new genre of lexicography – dictionaries for learners started to take shape. The genre gradually developed its characteristic lexicographic principles, one of the key features being the frequency – a focus on the more frequent words and the more frequent meanings of words. One of the important stages in the compilation of Georgian-English Learner's Dictionary is creation of the Georgian content. Given the specificities of the genre, this task cannot be undertaken without application of electronic corpora.

The paper will demonstrate the stages and methods of defining the entries of Georgian-English Learner's Dictionary using the above mentioned electronic corpora. Special attention will be paid to Georgian collocations that will be selected and extracted from corpora according to their frequency. Let us discuss Georgian word სათუთი [satuti] as an instance.

In *Explanatory Dictionary of the Georgian Language* (in 8 volumes under the general editorship of Arnold Chikobava), the entry სათუთი [satuti] has only three synonyms: ნაზი [nazi], ნებიერი [nebieri] და აზიზი [azizi], and only 6 hits (from Akaki Tsereteli, Leo Kiacheli, Ioseb Grishashvili, Lavrenti Ardaziani, and Galaktion Tabidze); it is obvious that above mentioned cannot fully reflect present-day Georgian usage.

According to *Explanatory Dictionary of the Georgian Language*, there are the following collocations including the Georgian word **სათუთი**:

- სათუთი ბავშვი [satuti bavšvi] → spoilt child
- სათუთი ქალი [satuti kali] → gentle woman
- სათუთი ქალიშვილი [satuti kališvili] → gentle young woman
- სათუთი გაზაფხული [satuti gazapxuli] → gentle spring
- სათუთი რითმა [satuti ritma] → soft rhyme
- სათუთი გრძნობა [satuti grĵnoba] → gentle feeling

In Paul Meurer's *Georgian Corpus GEKKO*, there are 126 hits of the Georgian word სათუთი, including those evidenced in *Explanatory Dictionary of the Georgian Language*, and it is only natural,

since the dictionary was compiled back in the 1960s and the subsequent stages of the development of Georgian are not reflected therein, such as:

- სათუთი პროცესი [satuti p'roc'esi] → delicate process
- სათუთი სიცოცხლე [satuti sic'oc'xle] → gentle life
- სათუთი დღე [satuti dğe] → gentle day
- სათუთი პრობლემა და ა.შ. [satuti p'roblema] → delicate problem, etc.

In *sketch Engine* by Adam Kilgarriff and David Tugwell, there are 134 hits of the entry:

- სათუთი ნაკეთობა [satuti nak'etoba] → fragile structure
- სათუთი საქმე [satuti sakme] → delicate business
- სათუთი სული [satuti suli] → gentle /sincere soul
- სათუთი თემა და ა.შ. [satuti tema] → delicate/tricky topic, etc.

Based on the aforementioned, the additional meanings of the word are singled out: **წრფელი** (*სათუთი სული*) [c'rpeli] - gentle /sincere soul; **დელიკატური, საფრთხილო, საჩოთირო** (*სათუთი პრობლემა, სათუთი თემა*), [delik'at'uri, saprtxilo, sačotiro] delicate problem, tricky topic; **გაფრთხილება რომ სჭირდება, ადვილად მტვრევადი / მსხვრევადი / მყიფე** (*სათუთი ნაკეთობა*), [gaprtxileba rom sč'irdeba, mt'vrevadi, msxvrevadi, mq'ipe] to be treated with caution, fragile, brittle → fragile structure, etc.

The compilation of this new type learner's dictionary is impossible without the above mentioned meanings, and we believe that the preparation process of the Georgian word-list for Georgian-European dictionaries should be started with the analyses of the corpus driven data and based thereupon.

In future, the Georgian content prepared for „Georgian-English Learner's Dictionary" can be successfully applied to the compilation of other Georgian-European learner's Dictionaries.

„ვეფხისტყაოსნის" პარალელური (ქართულ-ინგლისური) კორპუსი

ნინო დობორჯგინიძე
ილიას სახელმწიფო უნივერსიტეტი (საქართველო)
nino_doborjginidze@iliauni.edu.ge

საპრეზენტაციოდ წარმოდგენილი პორტალი – **„ვეფხისტყაოსნის" პარალელური (ქართულ-ინგლისური) კორპუსი** 2012 წელს შეიქმნა ილიას სახელმწიფო უნივერსიტეტის ლინგვისტურ კვლევათა ცენტრში. იგი მიეძღვნა პოემის პირველი ნაბეჭდი გამოცემის საიუბილეო თარიღს.

„ვეფხისტყაოსნის" კორპუსი კვლევის სრულიად ახალი ეტაპის დასაწყისია. პროექტზე მუშაობისას გამოვლინდა დღემდე უცნობი თორმეტი ხელნაწერი (მათ შორის ერთი ილუსტრირებული). ისინი არათუ გამოყენებული არ იყო პოემის კრიტიკული ტექსტის დადგენისას, აქა-

მდე აღწერილიც კი არ ყოფილა. მონაცემთა ელექტრონული დოკუმენტირებისას თავიდან გადაისინჯა აქამდე შესწავლილი ყველა რესურსი, როგორც ბეჭდური გამოცემები და ხელნაწერები, ასევე პოემის ირგვლივ შექმნილი ლიტერატურა. გასწორდა/აღინუსხა რიგი უზუსტობანი, რომლებიც ცალკეულ გამოცემებსა და აღწერილობებში იყო, ხელახლა მომზადდა ხელნაწერების აღწერილობები.

პროექტის განხორციელებით შესაძლებელი გახდა პოემის ტექსტური (ხელნაწერები, გამოცემები) და ვიზუალური (ილუსტრაციები, კალიგრაფია, ყდა) მემკვიდრეობის დოკუმენტირება თანამედროვე ლინგვისტური და ტექნოლოგიური სტანდარტების მიხედვით; ამასთანავე, პოემის კორპუსულ გამოცემაში განთავსდა „ვეფხისტყაოსანთან“ დაკავშირებული სამეცნიერო-კვლევითი და ბიბლიოგრაფიული ლიტერატურის ელექტრონული ვერსიები, პოემასთან დაკავშირებული დღემდე შეუსწავლელი მეტამონაცემები (ხელნაწერთა მინაწერები).

პორტალში წარმოდგენილია შემდეგი ძირითადი ბლოკები:

1. **ტექსტების ბლოკი.** იგი ორი ნაწილისგან შედგება: ა. „ვეფხისტყაოსნის“ გამოცემები, ბ. „ვეფხისტყაოსნის“ ხელნაწერები. ამ ბლოკში შესულია პოემის ყველა ქართულენოვანი ბეჭდური გამოცემისა და საქართველოში დაცული ხელნაწერების ელექტრონული ვერსიები.

2. **„ვეფხისტყაოსნის“ პარალელური კორპუსის ბლოკში** განთავსებულია პოემის ყველა ქართულენოვანი გამოცემის, ხელნაწერებისა და ინგლისურენოვანი თარგმანების გათანაბრებული ტექსტები. ამ უკანასკნელში შეტანილია მარჯორი უორდროპის, სტივენსონისა და ნათელა ურუშაძის თარგმანები, სულ ხუთი გამოცემა. სპეციალური შემზღუდველების გამოყენებით ბლოკში ძიება შესაძლებელია როგორც ცალკეული გამოცემების, ხელნაწერებისა და ინგლისური თარმანების, ასევე ყველა არსებული ტექსტური რესურსის მონაცემთა მიხედვით.

რადგან „ვეფხისტყაოსნის“ პარალელურ კორპუსში ბეჭდური გამოცემების გარდა განთავსდა პოემის შემცველი ხელნაწერებიც (მათ შორის სხვადასხვა ეპოქაში „ვეფხისტყაოსნის“ სიუჟეტზე შექმნილი პოეტური და პროზაული ნაწარმოებები), კორპუსში შეუძლებელი აღმოჩნდა საეტალონო გამოცემის განსაზღვრა, რადგანაც გამოცემები და ხელნაწერები ერთმანეთისაგან განსხვავდებიან როგორც სტროფების რაოდენობის მხრივ, ასევე შინაარსობრივადაც. შესაბამისად, პროექტის მსვლელობისას შეიქმნა დამატებითი პროგრამული რესურსი Corpuser-ი, რომლის მეშვეობით მოხერხდა პოემის სხვადასხვა გამოცემისა და ხელნაწერის მიხედვით სტროფების ნახევრადავტომატურად გათანაბრება და ტექსტების გარეანოტირება. რაც შეეხება „ვეფხისტყაოსნის“ კორპუსის მორფოლოგიურ ანოტირებას, იგი განხორციელდება მას შემდეგ, რაც დასრულდება თანამედროვე ქართული ენის მორფოლოგიური ანალიზატორის ტესტირების პროცესი (ეს ამოცანა ამჟამად ხორციელდება სხვა პროექტის ფარგლებში).

3. **ვიზუალური მემკვიდრეობის ბლოკი** მომზადდა ხელნაწერთა ეროვნულ ცენტრთან თანამშრომლობით. „ვეფხისტყაოსნის“ პარალელურ კორპუსში განთავსდა ხელნაწერთა ეროვნულ ცენტრში დაცული პოემის ხელნაწერების ნიმუშები: ილუსტრაციების, კალიგრაფიის, ყდების ფოტოასლები. კორპუსის მომხმარებელს შეუძლია დაათვალიეროს, შეადაროს ერთმანეთს ან შემდგომი კვლევისთვის გამოიყენოს თითოეული ხელნაწერის პირველადი მონაცემები.

4. **სამეცნიერო ლიტერატურის** ბლოკში განთავსებულია პოემის შესახებ არსებული ძირითადი სამეცნიერო ლიტერატურის ელექტრონული ვერსიები.

5. **ბიბლიოგრაფიის ბლოკში** წარმოდგენილია „ვეფხისტყაოსნის“ ბიბლიოგრაფიის ბეჭდური გამოცემების ელექტრონული ვერსიები.

6. **„ვეფხისტყაოსნის“ ხელნაწერთა მინაწერების** ბლოკში განთავსებულია უნიკალური მასალა, პოემის ძირითად ტექსტზე სხვადასხვა პერიოდში დართული მეტატექსტები, რომელთა უდიდესი ნაწილი პირველად ქვეყნდება.

7. **საგანმანათლებლო რესურსები და ლექსიკონები.** პროექტის ფარგლებში პირველად შეიქმნა „ვეფხისტყაოსნის“ ჩაშენებული ლექსიკონი. იგი ერთვის პოემის აკადემიური გამოცემის ტექსტს. გრძელდება მუშაობა სპეციალური პროგრამების მეშვეობით სხვადასხვა ტიპის ლექსიკონების შექმნაზე.

Parallel (Georgian-English) Corpus of the poem *Vepkhistkaosani* („The Knight in the Panther's Skin“)

Nino Doborjginidze
Ilia State University (Georgia)
nino_doborjginidze@iliauni.edu.ge

The *Vepkhistkaosani* parallel (Georgian-English) corpus was created in 2012 at the Centre for Linguistic Research, Ilia State University. It was dedicate to the anniversary of the first publication of the poem.

The corpus marks a new stage in Rustaveli studies. During the work on the project, 12 previously unknown manuscripts were discovered, one of them illustrated, which were not only unused in establishing the critical text of the poem, but had not even been described. While creating an electronic database, all hitherto studied resources, whether printed or hand-written, as well as all relevant academic literature were re-examined. A number of errors found in some versions and descriptions were put right/registered and new descriptions were prepared.

The project enabled the documenting of the textual (manuscripts, editions) and visual (illuminations, handwriting, cover) legacy of the poem to modern linguistic and technological standards. The corpus also contains electronic academic literature, bibliography and metadata (adscripts) related to the poem.

There are the following main blocks in the portal:

1. **The text block,** consisting of two parts: a. *Vepkhistkaosani* editions; b. *Vepkhistkaosani* manuscripts. The block contains electronic versions of all published editions of the poem in the Georgian language and of manuscripts preserved in Georgia.

2. **The** p**arallel corpus block** stores balanced texts of all available Georgian editions and manuscripts of the poem and of its English translations by Marjory Wardrop, R. H. Stevenson and Natela Urushadze (5 English editions in all). Special filters enable search within individual editions, manuscripts and English translations as well as across the whole corpus.

As no methodologically updated critical edition of the poem is so far available (not all manuscripts were considered in earlier academic editions and neither was the stemma established), an additional software, *Corpuser*, was created to enable semi-automatic balancing of verses from various text versions and external annotation of texts. As concerns the morphological annotation of the corpus, it will be developed once the

testing of modern Georgian morphological analyzer is completed (the testing is being implemented under a different project).

3. **The visual legacy block** was developed in cooperation with the National Manuscript Centre. Photocopies of manuscripts kept in the Centre – illustrations, handwriting patterns, covers – were uploaded in the parallel corpus. The corpus user can look through and compare or apply for further research the primary data for each manuscript.
4. **The academic literature block** contains electronic versions of main research works on *Vepkhistkaosani*.
5. **The Bibliography block** presents electronic versions of the printed bibliography of the poem.
6. ***Vepkhistkaosani* adscript block** offers unique materials – metatexts added to the main body of the poem in various periods. Most of them are published for the first time
7. **Educational resources and dictionaries.** This project is the first to offer a built-in glossary to the poem. Its trial version is attached to the poem's academic edition. Various types of dictionaries are being developed with the help of special software solutions.

ზმნის სემანტიკური მოდელირება: პრობლემები და გადაწყვეტის გზები

ლალი ეზუგბაია, თედო უთურგაიძე, მარიამ მანჯგალაძე, რატი სხირტლაძე
თსუ არნ. ჩიქობავას სახ. ენათმეცნიერების ინსტიტუტი (საქართველო)
ენის მოდელირების ასოციაცია (საქართველო)
ezugbaia@ice.ge, mariam@ice.ge, rati2008@gmail.com

ქართული ენის მოდელირების პროცესში განსაკუთრებით თვალსაჩინო გახდა რამდენიმე პრობლემა, რომელთა ახლებურად გააზრებისა და პრაქტიკულად გადაწყვეტის გარეშე შეუძლებელი იქნება სინტაქსური და სემანტიკური ანალიზატორების შექმნა. კერძოდ, ქართული ზმნის მოდელირებას ართულებს არა მხოლოდ გრამატიკული კატეგორიების სიმრავლე, არამედ სემანტიკური მრავალფეროვნება, ხშირ შემთხვევაში პოლისემიურობა, რასაც განაპირობებს ზმნისწინი (ფორმალური სემანტიკა) და ზმნური შესიტყვების კონტექსტი (კონტექსტური სემანტიკა).

ზმნის უღლების ყველა კატეგორიის მიხედვით მოდელირება და სათანადო პროგრამის შექმნა ჯერ კიდევ არ არის საკმარისი პირობა ენაში არსებულ ზმნურ ფორმათა სრულად ასახვისათვის. კერძოდ, **აქტუალურია ყველა იმ ზმნური სემანტიკური ქვესისტემის აღწერა, რომელიც ზმნისწინის დართვით წარმოიქმნება.** ეს შეუძლებელიც იყო ტრადიციული მეთოდებით კვლევისას. მხოლოდ მანქანური დამუშავების შემდეგ გახდა თვალსაჩინო ზმნისწინის თავისებურებები, რადგან რეალურად არც ერთი ზმნური ფუძე არ დაირთავს იმავე ზმნისწინებს ისეთივე დანიშნულებით, როგორითაც ნებისმიერი დანარჩენი ფუძე. შეიძლება ითქვას, რომ ქართულ ზმნათა უმრავლესობა თავისებურია ზმნისწინთა სისტემასთან მიმართებით.

ქართული ზმნის სემანტიკური მოდელირება უკავშირდება სხვა პრობლემასაც. აუცილებელია **ზმნათა სემანტიკური ჯგუფების** გამოყოფა მათი ლექსიკური მნიშვნელობის მიხედვით. ასეთი ჯგუფებია: მოძრაობა-გადაადგილების, გარდაქცევითობის, საუბრის, მსჯელობის, ფიქ-

რის, აზროვნების, ზრუნვის, ენერგიის ჩადების, ხედვის, გაცვლა-გამოცვლის, გრძნობა-ემოციის, მოკითხვა-ჩაკითხვის, შებრუნება-შემობრუნების, შეხვედრა-დახვედრის, გაგზავნა-გამოგზავნის, სადაობა-მყოფობის და სხვა სემანტიკური ველები, რომლებიც ფრაგმენტულადაა დამუშავებული ქართულ სამეცნიერო ლიტერატურაში. იმისათვის, რომ ქართული ზმნის სემანტიკური კლასიფიკაცია მიესადაგოს მანქანური მოდელირების ამოცანებს, რაც შეიძლება ზუსტად უნდა დადგინდეს არსებული ლექსიკური ჯგუფები და მათი შემადგენელი ერთეულები.

ზმნის სემანტიკური მოდელირება ემყარება ჩვენ მიერ შექმნილ ქართული ენის ფორმალიზებული მოდელის პროგრამულ პაკეტს, რომელიც მოიცავს ქართული ენის სპელჩეკერსა და ანალიზატორს. ანალიზატორის საშუალებით მომხმარებელს შეუძლია მორფემებად დაშალოს სიტყვაფორმა, დაადგინოს მისი გრამატიკული კატეგორია. ავაგეთ 592 ზმნური და 72 სახელური მოდელი, რომლებიც, შესაბამისად, 28,982,157 და 108,761,163, სულ 137,743,320 განსხვავებულ სიტყვაფორმას აწარმოებენ. თითოეული მოდელი წარმოდგენილია ფუძეებისა და პრეფიქს-სუფიქსთა წყვილების ბაზით. ხაზგასმით აღვნიშნავთ, რომ პრეფიქსებისა და სუფიქსების ბაზისაგან განსხვავებით ზმნისწინთა განაწილება ხდება სათითაოდ ზმნური ფუძეების, გვარისა და ვერსიის გრამატიკული კატეგორიების გათვალისწინებით.

ზმნისწინის მიხედვით სემანტიკური მოდელირება გულისხმობს:

ა) ისეთი კომპიუტერული სისტემის შექმნას, რომელიც ზმნათა ფუძეების კონკრეტული სემანტიკური და გრამატიკული მნიშვნელობებისთვის მოგვაწვდის საჭირო ზმნისწინებს;

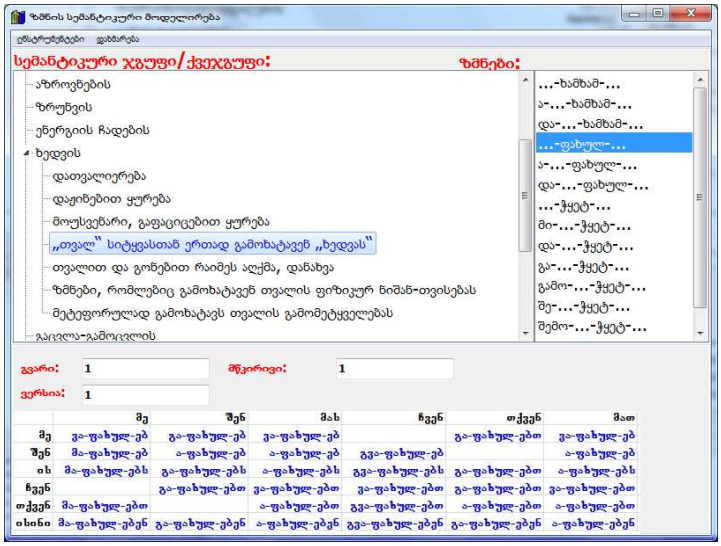
ბ) ზმნათა ბაზების დახასიათებას ფლექსიური ზმნისწინების მიხედვით შესაბამისი უღლების პარადიგმების გათვალისწინებით;

გ) ზმნათა ბაზების დახასიათებას დერივაციული ზმნისწინების მიხედვით შესაბამისი უღლების პარადიგმებითა და მნიშვნელობებით;

დ) ზმნურ ფუძე-ძირთა სემანტიკური ჯგუფების გამოყოფას.

ამ ამოცანების შესრულებას ახლავს გარკვეული სირთულეები, კერძოდ, ზმნისწინთა ფუნქციებისა და ზმნების სემანტიკური ჯგუფების სრული აღწერა ქართული ენის ეროვნული კორპუსის არარსებობის პირობებში ძნელია. თავისთავად, ენის გამოყენების არეალი ფართოა. ახალი დარგების განვითარების კვალდაკვალ ჩნდება ახალი ტერმინები, სიტყვები იძენენ ახალ მნიშვნელობას კონტექსტის მიხედვით. ამ პროცესებისთვის თვალყურის მიდევნება ძნელია და არცაა ჩვენი ამოცანა. აქედან გამომდინარე, ამ ეტაპზე ჩვენ შემოვიფარგლებით იმ ლექსიკური და სემანტიკური მონაცემებით, რომლებიც დადასტურებულია „ქართული ენის განმარტებით ლექსიკონში“ (მათ შორის ახალი რედაქციის 1-ლ და მე-2 ტომებში), ასევე ორთოგრაფიულ ლექსიკონში. გამოვიყენებთ სხვა ნორმატიულ ლექსიკონებსაც, მაგრამ განსაკუთრებით აღვნიშნავთ ბ. ფოჩხუას „თანამედროვე ქართული ენის იდეოგრაფიულ ლექსიკონს, I“, რომელიც დღემდე ამ ტიპის ერთადერთი ნაშრომია. ასევე ვითვალისწინებთ ტიპოლოგიური კვლევის შედეგებსაც, რასაც განსაკუთრებული მნიშვნელობა ენიჭება სწორედაც მანქანური თარგმანის პრობლემების გადასაჭრელად.

პროგრამას, რომლის საშუალებითაც განხორციელდება ზმნების კლისიფიკაცია სემანტიკური ჯგუფებისა და ქვეჯგუფების მიხედვით, დაახლოებით ასეთი სახე ექნება:



სემანტიკური ჯგუფები და ქვეჯგუფები ე.წ. „ხის“ სტრუქტურით იქნება აღწერილი. აღნიშნული სტრუქტურა საშუალებას იძლევა თითოეული განშტოება დაიშალოს ქვეგანშტოებებად, ისინი, თავის მხრივ, კიდევ დაიშალონ ქვეგანშტოებებად და ა.შ.

პროექტის ფარგლებში მიზნად ვისახავთ ორდონიანი ხის სტრუქტურის აგებას, თუმცა, საჭიროების მიხედვით, ცალკეულ შემთხვევაში არ გამოვრიცხავთ უფრო მეტი დონის გამოყენებასაც.

Semantic Modeling of the Verb: Problems and Solutions

Lali Ezugbaia, Tedo Uturgaidze, Mariam Manjgaladze, Rati Skhirtladze
Arn. Chikobava Institute of Linguistics, TSU (Georgia)
Language Modeling Association (Georgia)
ezugbaia@ice.ge, mariam@ice.ge, rati2008@gmail.com

The process of modeling Georgian has made it clear that the creation of syntactic and semantic spell-checkers cannot be carried out without considering some problems in a new way and practically resolving them. In particular, the modeling of the Georgian verb is complicated not only due to the multitude of grammatical categories, but also to semantic diversity and mainly polysemanticity, caused by a preverb (formal semantics) and verbal phrase context (context semantics).

Modeling of verb conjugation according to all categories and creation of the corresponding software is not sufficient to fully reflect the verb forms, available in the language. Specifically, **it is significant to describe all the semantic subsystems of a verb, formed by adding of a preverb.** This was not

possible when old methods were used in the course of research. It was only after the data were machine processed, when the peculiarities of the preverb became evident, as, actually, no other verb can take the same preverb with the same function, as any other stem. It may be said that each Georgian verb is peculiar in relation to the preverb system.

The semantic modeling of Georgian verb is associated with another problem. Distribution of verbs into semantic groups on the basis of their lexical meaning is necessary. These groups are: movement-relocation, transformation, conversation-discussion, thinking, consideration, care, putting energy into smth., viewing, exchanging, feeling-emotion, meeting, sending, being-presence and other semantic groups, having only partially been considered in the Georgian scholarly literature. In order to adjust the semantic classification of verbs to the tasks of machine modeling, the existing lexical groups and their constituent items should be established as precisely as possible.

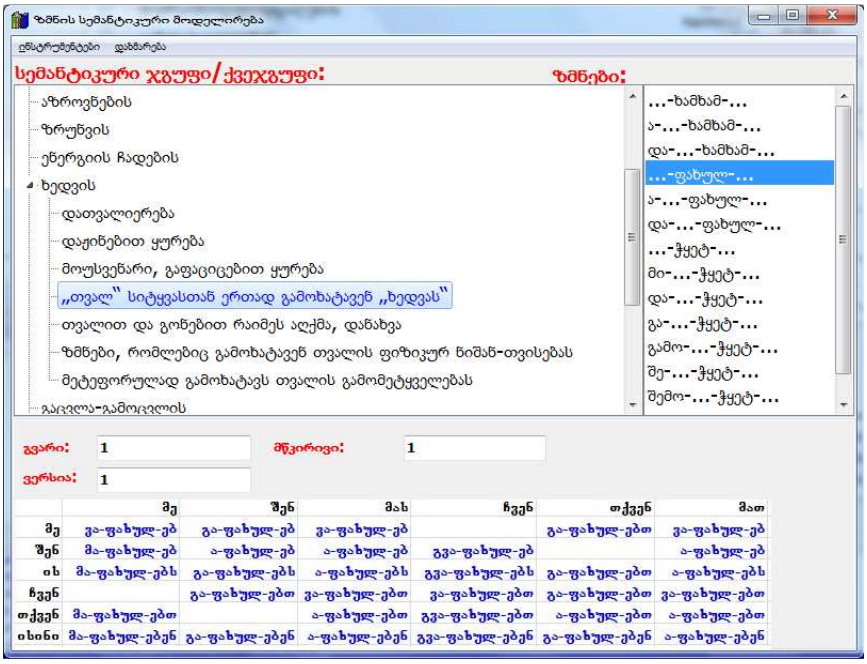
The semantic modeling of the verb is based on the software package of the Georgian orthographic spell-checker and the so called parser. Albeit the parser is able to break a word-form into morphemes, define its grammatical category (which part of speech it is; its case, if it is a substantive; its conjugation form, if it is a verb, etc). We have built 592 verbal and 72 nominal models forming 28.982.157 verbal and 108.761.163 nominal word-forms. In total, they create 137.743.320 different word-forms. Each model is represented with the base of a stem and prefix-suffix pairs. We would like to emphasize that, unlike prefix and suffix base, preverbs have been distributed one by one, taking into consideration verbal stems and grammatical categories of voice and version.

Semantic modeling according to preverb involves:

a) creation of the computer system which will provide necessary preverbs for specific semantic and grammatical meanings of verbal stems;
b) definition of the verbal bases according to preverbs considering appropriate conjugation paradigms;
c) characterization of verbal stems according to derivational preverbs with appropriate conjugation paradigms and meanings;
d) division of verbal stems into semantic groups.

There are a number of difficulties associated with the task; specifically, it is difficult to describe thoroughly the functions of preverbs and verbal semantic groups as the national corpus of Georgian does not exist. The area of language usage is broad. New terms emerge as new fields develop. Words acquire new meanings in various contexts. It is really difficult and it is not our task to observe these processes. Therefore, at this stage, we concentrate on lexical and semantic data, elicited from *The Explanatory Dictionary of the Georgian Language* (including the first and the second volumes of the new edition) and *Spelling Dictionary*. We will use other standard dictionaries. We will refer to *Ideographic Dictionary of Modern Georgian* as the only one of this type of work to date. We also take into consideration the results of the typological study as it is of a particular importance to solve the problems of machine translation.

The software, by means of which the classification of verbs will be made according to semantic groups and subgroups, will have the following structure:



Semantic groups and subgroups will be described as a *tree* structure. This structure enables each branch to be separated into sub-branches, etc.

Within the framework of the project, we aim to build a two-level *tree* structure, however, in some cases, we do not rule out to use more levels.

ზოგადიდან კონკრეტულამდე: ასპექტის ვარირება სლავურ ენებში

რუპრეხტ ფონ ვალდენფელსი
ბერნის უნივერსიტეტი (შვეიცარია)
waldenfels@issl.unibe.ch

პარალელურ კორპუსზე (ParaSol) დამყარებულ გამოკვლევებში (Waldenfels 2012, 2012a) ეკვივალენტური კორპუსული დამოწმებების მანუალური შეფასების მიხედვით შესაძლებელი გახდა სტივენ დიკისეული (2000) კლასიფიკაციის დადასტურება, კერძოდ, მან დაყო სლავური ასპექტი აღმოსავლურ და დასავლურ ჯგუფებად მისი გამოყენების მიხედვით ბრძანებითში (იხ. Benacchio 2010) და ასევე ნამყოს უარყოფით ფორმებში (Dickey & Kresin 2009).

წინამდებარე მოხსენებაში აღწერილი იქნება გაუმჯობესებული სისტემა, რომელშიც გამოყენებულია ავტომატური მორფოსინტაქსური ანოტირება და ლემატიზაცია (იხ. Waldenfels 2011)

სიტყვათა შეთანადებასთან ერთად (Tiedemann 2003), რათა მოხდეს კორპუსიდან ნიშან–თვისებათა ვარიაციების წარმოქმნის პროცედურის ავტომატიზაცია. ამ ავტომატური პროცედურის შედეგები შეიძლება პირდაპირ შევადაროთ ოქროს სტანდარტს, სახელდობრ, კორპუსის თავდაპირველ, ხელით ანოტირებას.

რაც შეეხება ასპექტის ფუნქციონირებას, ამ სისტემის საშუალებით შესაძლებელია ასპექტის ვარირების მოდელების შესწავლა უფრო ფართო პერსპექტივით, რადგანაც ასპექტის ფუნქციონირების სხვადასხვა გრამატიკული კონტექსტი (მაგ., ბრძანებითი და ნამყოს უარყოფითი) შეიძლება უფრო ადვილად განვმარტოთ და მათი პროფილები ერთმანეთს შევადაროთ.

ასპექტის ფუნქციონირების შედარება სხვადასხვა გრამატიკულ კონტექსტსა და ენასთან მიმართებით საშუალებას გვაძლევს, ემპირიული კუთხით მივუდგეთ გრამატიკული ანალიზის ძირითად საკითხს: თუკი მრავალ ენაში ასპექტის ფუნქციონირების მიხედვით არსებული *განსხვავებები* მჭიდროდ კორელირებს სხვადასხვა გრამატიკულ კონტექსტთან, ეს იქნება არგუმენტი საიმისოდ, რომ ამ კატეგორიის ერთი, უნიფიცირებული ფუნქცია მივანიჭოთ თითოეულ ენას, რადგანაც ამ კატეგორიის ფუნქცია მკაფიოდ სახესხვაობს ენათა მიხედვით, ვიდრე ფუნქციური კონტექსტების მიხედვით. სლავური ენების ასპექტის შემთხვევაში ჩემ მიერ მიღებული შედეგები ამ ჰიპოთეზას ამართლებს.

Seeing the tree for the woods: aspect variation in Slavic

Ruprecht von Waldenfels
University of Bern (Switzerland)
waldenfels@issl.unibe.ch

In a series of studies based on the parallel corpus ParaSol (Waldenfels 2012, 2012a), Dickey's (2000) broad division of Slavic aspect use into an Eastern and Western Group in respect to aspect use in the imperative (see Benacchio 2010) as well as in negated past events (Dickey and Kresin 2009) could be confirmed relying on the manual evaluation of equivalent corpus attestations.

In this paper, I describe an improved system where automatic morphosyntactic annotation and lemmatization using a variety of systems (see Waldenfels 2011) is used in combination with word alignment (Tiedemann 2003) to automatize the procedure of deriving feature variation from the corpus. The results of this automatic procedure can be directly compared against a gold standard, viz. the earlier, manual annotation of the corpus.

In respect to aspect use, this system allows to investigate the patterns of aspect variation from a broader perspective, since difference environments for aspect use (e.g., imperatives and negated past above) can be more easily defined and their profile compared.

This comparison of aspect use across environments and languages makes it possible to approach a basic issue of grammatical analysis from an empirical perspective: if the *differences* of aspect use across

many languages correlate strongly across different environments, this is an argument for positing a single, uniform function of this category for each language, since then the function of the category varies more strongly from language to language than from functional context to functional context. For Slavic aspect, my results suggest that this hypothesis holds.

References:

Dickey, S.M. (2000): *Parameters of Slavic Aspect: A Cognitive Approach.* Chicago.
Dickey, S.M. and S.C. Kresin (2009) Verbal aspect and negation in Russian and Czech. *Russian Linguistics* 33:121-176.
Tiedemann, Jörg, 2003. Combining Clues for Word Alignment. In *Proceedings of the 10th Conference of the European Chapter of the ACL (EACL03)* Budapest, Hungary, April 12-17, 2003
von Waldenfels, Ruprecht. 2012. Aspect in the imperative across Slavic - a corpus driven pilot study. In A. Grønn & A. Pazelskaya (eds.): The Russian Verb. *Oslo Studies in Language* 4, 141–154.
von Waldenfels, Ruprecht. Subm. Explorations into variation across Slavic: taking a bottom-up approach. In: B. Szmrecsanyi, B. Wälchli (ed.): *Aggregating dialectology and typology: linguistic variation in text and speech, within and across languages.* Mouton de Gruyter

LUNDIC –ლინგვისტურ და კულტურულ მონაცემთა ბაზის ახალი პროექტი

კარინა ვამლინგი
მალმეს უნივერსიტეტი (შვედეთი)
karina.vamling@mah.se

LUNDIC (ენისა და კულტურის ლუნდისეული ციფრული ატლასი) წარმოადგენს მონაცემთა ბაზის ახალ პროექტს, რომელიც ხორციელდება ლუნდის უნივერსიტეტში (შვედეთი). მოხსენებაში გამოკვეთილია პროექტის პრინციპები და საკვლევი სფერო, რომელიც სხვა ჯგუფებთან ერთად კავკასიურ ენებსაც შეიცავს.

მონაცემთა ბაზაში შედის სამი ძირითადი კომპონენტი: ტიპოლოგიური მასალა (სიტყვათა რიგი, ძირითადი/პერიფერიული პირების მარკირება, ზმნის ტიპოლოგია), ლექსიკური მასალა (სვოდეშის სიები და კულტურის ლექსიკური მასალა) და კულტურული მასალა (არქეოლოგიური, ეთნოგრაფიული და ექსპერიმენტული მასალა). ამგვარად, პროექტის ერთ-ერთ მიზანს წარმოადგენს პლატფორმის შექმნა საენათმეცნიერო და არასაენათმეცნიერო მასალების ინტეგრირებისათვის. გარდა კავკასიისა, მასალების შეგროვება ემყარება საველე სამუშაოებს ამაზონიაში, ევროპაში, ცენტრალურ აზიაში, ტაივანში, სამოაში და იაპონიაში. თავმოყრილ მასალას მინი-

ჭებული აქვს გეოინფორმაციული სისტემის (GIS) კოდები, რათა შესაძლებელი გახდეს მათი კარტოგრაფირება და ანალიზი.

პროექტს აფინანსებენ მარკუს და ამალია ვალენბერგების ფონდი და კრაფორდის ფონდი. პროექტს ხელმძღვანელობს გერდ კარლინგი. პროექტის ვებგვერდია: http://project2.sol.lu.se/lundic/ collaborators.html.

LUNDIC – a new linguistic and cultural database project

Karina Vamling
Malmö University (Sweden)
karina.vamling@mah.se

Lund Digital Atlas of Language and Culture – LUNDIC – is a new database project based at Lund University, Sweden. The paper outlines the principles and scope of the project, which among other groups, includes Caucasian languages.

The database holds three core components: Typological data (including word order, marking of core/peripheral arguments, verbal typology), Lexical data (Swadesh lists and culture vocabulary data) and Cultural data (archaeological, ethnographic and experimental data). One of the aims of the project is thus to create a platform for the integration of linguistic and non-linguistic data. Apart from the Caucasus, the data collections based on fieldwork in the areas of Amazonia, Europe, Central Asia, Taiwan, Samoa, and Japan. The collected data is GIS-coded in order to allow for geographical mapping and analysis.

The project is funded by the Marcus and Amalia Wallenberg Foundation and the Crafoord Foundation. Project leader is Gerd Carling. For web page and collaborators, see: http://project2.sol.lu.se/lundic/ collaborators.html.

ლიტვური ენა ციფრულ ხანაში

იოლანტა ზაბარსკაიტე
ლიტვური ენის ინსტიტუტი (ლიტვა)
jolanta.zabarskaite@lki.lt

ლიტვური ენა ერთ–ერთი ნაკლებად გავრცელებული ევროპული ენაა. ამ ენაზე მხოლოდ ოთხ მილიონამდე ადამიანი საუბრობს და მათი უმრავლესობა ლიტვის რესპუბლიკაში ცხოვრობს. ლიტვური, როგორც სახელმწიფო ენა, ლიტვის რესპუბლიკის ყველა მოქალაქის წერითი და ზეპირმეტყველების ენაა. 2010 წლის აღწერის მიხედვით, ლიტვის მოსახლეობა 3,3 მილიონს

აღწევს, რომელთა შორის არიან: 2,7 მილიონი ლიტველი (84%), ასევე პოლონელები (6.1%), რუსები (4.9%), ბელორუსები (1.1%), უკრაინელები (0.6%), ებრაელები (0.1%), გერმანელები (0.1%), ლატვიელები (0.1%), თათრები (0.1%), კარაიმები და ბოშები.

დაახლოებით 500000 ლიტვურად მოლაპარაკე ცხოვრობს ლიტვის ფარგლებს გარეთ. მასზე საუბრობენ ლიტვური ეთნიკური უმცირესობის წარმომადგენლები ბელორუსიაში, პოლონეთში, ლატვიაში, ასევე მრავალრიცხოვანი ემიგრანტული თემები აშშ–ში, კანადაში, დიდ ბრიტანეთში, ირლანდიაში, ესპანეთში, სამხრეთ ამერიკაში და ა.შ. მოლაპარაკეთა რაოდენობის მიხედვით, ლიტვური ენა მსოფლიოში 144–ე ადგილზეა.

ლიტვის ევროკავშირში შესვლის შემდეგ ლიტვური ენა თავისი განვითარების ახალ ფაზაში შევიდა: ევროკავშირის ოფიციალური ენის სტატუსის მინიჭებამ უზრუნველყო მისი გამოყენება და გავრცელება გაერთიანების მრავალენოვან სივრცეში.

ლიტვურ ენაზე არსებული საჯარო სერვისების რაოდენობა იზრდება. ეს კი გაზრდის ლიტვურის ხვედრით წილს. უფრო მეტიც, მიმდინარეობს მუშაობა საიმისოდ, რომ შემცირდეს ციფრული იზოლირება და უფრო ხელმისაწვდომი გახდეს ტექნოლოგიები, მათ შორის, შეზღუდული შესაძლებლობების ადამიანებისათვის.

იზრდება ახალი ამბების პორტალების, ლიტვის პერიოდული გამოცემებისა თუ სამეცნიერო ჟურნალების ვებსაიტების პოპულარობა. მათ შორის ყველაზე გამორჩეულებია www.epaveldas.lt, www.emokykla.lt, www.emokymas.lt და სხვები. იგეგმება პორტალის შექმნა, რომელზეც განთავსდება ადვილად ხელმისაწვდომი ენობრივი რესურსები და ტექნოლოგიები, რომლებიც ამ უკანასკნელ ხანს ინიცირებული პროგრამის „ლიტვური ენა საინფორმაციო საზოგადოებაში“ ფარგლებში შეიქმნა. პროგრამას კოორდინაციას უწევს ლიტვური ენის სახელმწიფო კომისია. მისი ამოცანებია ლოკალიზაცია, რესურსებისა და ინსტრუმენტების შექმნა, დოკუმენტირება და ა.შ.

ტრანსპორტისა და კომუნიკაციის სამინისტროსთან არსებული საინფორმაციო საზოგადოების განვითარების კომიტეტი პასუხისმგებელია პროგრამის „ლიტვური ენა საინფორმაციო საზოგადოებაში“ მეორე ფაზის (2010–2015) განხორციელებაზე. პროგრამა უზრუნველყოფს ისეთი საინტერნეტო პორტალის შექმნას, რომელზეც ხელმისაწვდომი იქნება არსებული ენობრივი რესურსები და ტექნოლოგიები, რომლებსაც დაემატება აქამდე არსებული და ახლადშექმნილი ენობრივი რესურსები; გააუმჯობესებს ASR და TTS ტექნოლოგიებს, მანქანური თარგმანის ახალ ინსტრუმენტებს, გააუმჯობესებს და განავითარებს სემანტიკურ და სინტაქსურ ანალიზს და საძიებო ინსტრუმენტებს.

კომპიუტერულ ლინგვისტიკასთან და ენის ტექნოლოგიებთან დაკავშირებული კურსები იკითხება სხვადასხვა პროგრამის ფარგლებში ვილნიუსის უნივერსიტეტსა და ვიტაუტას მაგნუსის უნივერსიტეტში. ვილნიუსის უნივერსიტეტის კაუნას ჰუმანიტარულ მეცნიერებათა ფაკულტეტის მაგისტრატურაში ისწავლება აუდიოვიზუალური თარგმანი, ხოლო ვიტაუტას მაგნუსის უნივერსიტეტში იგეგმება კომპიუტერული ლინგვისტიკის სამაგისტრო კურსების შექმნა.

სალიტერატურო ლიტვური ენისათვის არსებობს არაერთი ტექნოლოგია თუ რესურსი. ლიტვური ერთ-ერთია ევროპის ე.წ. არაკომერციულ ენებს შორის, რის გამოც მის წინაშე ისეთი საინფორმაციო გამოწვევები და სიძნელეები დგას, რომლებიც დამახასიათებელია ნაკლებად გავრცელებული ენებისათვის. **ლიტვური ენის ტექნოლოგიების განვითარება დიდადაა დამოკი-**

დებული სხვა ქვეყნების გამოცდილებასა და დახმარებაზე და საერთაშორისო თანამშრომლობაზე (პროექტები, როგორებიცაა META-NET, META-NORD და სხვები). მეორე მხრივ, განვითარებადი ენობრივი ტექნოლოგიები წარმოადგენენ უმნიშვნელოვანეს ელემენტს ლიტვური ენის ფუნქციონირების, აღიარებისა და შესწავლის განმტკიცების პროცესში ისევე, როგორც ლიტვური კულტურის გავრცელებისათვის მრავალენოვანი ევროპის მასშტაბით.

The Lithuanian Language in the Digital Age

Jolanta Zabarskaite
Institute of Lithuanian Language (Lithuania)
jolanta.zabarskaite@lki.lt

The Lithuanian language is one of the least commonly used European languages. Only about four million people speak it, and most of them live in the Republic of Lithuania. Lithuanian, as the state language, is a written and spoken language for all citizens of the Republic of Lithuania. Based on the 2010 census, Lithuania's population totals about 3.3 million, including roughly 2.7 million ethnic Lithuanians (84 %), besides, Poles (6.1 %), Russians (4.9 %), Belarusians (1.1 %), Ukrainians (0.6 %), Jews (0.1), Germans (0.1), Latvians (0.1), Tatars (0.1), Karaits and Roma, among others.

There are about 500,000 Lithuanian-speakers outside Lithuania. The language is spoken by Lithuanian ethnic minorities in Belarus, Poland, Latvia, as well as vast emigrant communities in US, Canada, UK, Ireland, Spain, South America, etc. According to the number of speakers, the Lithuanian language places 144th in the world.

After Lithuania's accession to EU, the language entered a new phase of development: its newly acquired status of an official EU language ensured its usage and dissemination across the multilingual space of EU.

Number of Lithuanian public services is growing. More public services will be transferred to the web, expanding the volume of Lithuanian content on the Internet. Moreover, efforts are made to reduce digital isolation and to make technologies user-friendly and easily accessible for people with disabilities.

Popularity of news portals, websites of Lithuania's main periodicals, of some journals has been increasing. The most noteworthy are www.epaveldas.lt, www.emokykla.lt, www.emokymas.lt, etc. There are plans to develop a portal, hosting easily accessible linguistic resources and technologies designed under the program *Lithuanian Language in the Information Society*, launched recently. The program is coordinated by the State Commission of the Lithuanian Language and deals with localisation, resource and tool creation, documentation, etc.

The Information Society Development Committee under the Ministry of Transport and Communications is responsible for the second phase of the program *Lithuanian Language in the*

***Information Society*, 2010-2015.** The program provides for the creation of an Internet portal with free access to all the available language resources and technologies, augmentation of the existing and newly created linguistic resources, improvement of the ASR and TTS technologies, new MT tools, improvement and development of semantic and syntactic analysis and search tools.

Some CL- and LT-related courses are taught as part of other studies at Vilnius University and Vytautas Magnus University. Kaunas Humanities Faculty (UV) offers master courses of audiovisual translation, and VMU plans to launch master courses of computational linguistics.

A number of technologies and resources are available for Standard Lithuanian. It is among the so-called non-commercial European languages, therefore facing the IT challenges and difficulties, typical of the development of a less widely used language. **The development of the Lithuanian LT relies heavily on the experience of and assistance from other countries and international cooperation (META-NET, META-NORD projects, etc).** On the other hand, developing language technologies is the most important element in the process of strengthening of functioning, recognition and learning of the Lithuanian language as well as the dissemination of the Lithuanian culture across the multilingual Europe.

KWIC-ი დან KRIC -ამდე
კომპლექსური ლინგვისტური ძიების კორპუსული ადაპტაცია

მანანა თანდაშვილი, ზაქარია ფურცხვანიძე
ფრანკფურტის გოეთეს უნივერსიტეტი (გერმანია)
tandaschwili@em.uni-frankfurt.de , pourtskhvanidze@em.uni-frankfurt.de

1. პრობლემის ზოგადი დახასიათება

ქართველოლოგიის ფუნქციონირება 21-ე საუკუნეში და საქართველოს მეცნიერული ინტეგრაცია გლობალიზებულ მსოფლიო სამეცნიერო სივრცეში თვისობრივად ახალი, კორპუსულ კვლევებზე დამყარებული ენათმეცნიერების განვითარებაზე არის დამოკიდებული. მხოლოდ თანამედროვე ტექნოლოგიური სამომხმარებლო ინტერფეისით აღჭურვილი დიდი მოცულობის ენობრივი რესურსები იძლევა საშუალებას თეორიული პოსტულატები გადავამოწმოთ და მათ ადეკვატური დასკვნების კვალიფიკაცია მივანიჭოთ.

„ქართული ენის ეროვნული კორპუსი“, როგორც საერთაშორისო სამეცნიერო პროექტი, აერთიანებს ქართული ენისათვის შექმნილ ელექტრონულ კორპუსულ რესურსებს - პროფ. ი. გიპერტისა და მ. თანდაშვილის მიერ ინიცირებულ TITUS-ისა და ARMAZI-ის ქართულ რესურსებს (ფრანკფურტის უნივერსიტეტი), პაულ მოირერის მიერ შექმნილ GEKKO-ს რესურსებს (ბერგენის უნივერსიტეტი) და მარინე ბერიძის მიერ შექმნილ „ქართული დიალექტური კორპუსის“ რესურსებს. კორპუსების პრაქტიკული გამოყენების ეფექტურობას ზოგადად სამომხმარებლო ინტერფეისის მოქნილობა და ლინგვისტურ პრობლემატიკაზე ორიენტირებული ძიების სტრატეგიის არსებობა განაპირობებს. როგორც TITUS-ისა და ARMAZI-ის, ისე GEKKO-ს საძიებო სის-

ტემები ლინეარული ტიპის საძიებო სისტემები გახლავთ. ამგვარი ძიების შემთხვევაში ადგილი აქვს საძიებო ფორმების შემცველი კონტექსტების გენერირებას, სადაც საძიებო ელემენტი აისახება, როგორც KWIC. ძიების ამგვარ ტიპს ჩვენ ვუწოდებთ **ლინეარულ ძიებას** და მას განვიხილავთ როგორც „ძიების უხეშ ფორმას“.

წინამდებარე თეზისში აღწერილია ძიების ალტერნატიული ტიპი, რომელიც არა ცალკეული ენობრივი ფორმების, არამედ კომპლექსური ლინგვისტური რელაციების შემცველი კონტექსტების გენერირების საშუალებას მოგვცემს.

2. ლინეარული ძიება, როგორც კვლევის პირველი საფეხურის ოპერაცია

კორპუსში კვლევის პროცესში განხორციელებული ძიების ლინეარული სახე საძიებო ელემენტების (როგორც ლექსიკური, ისე ფუნქციური) პირველად სელექციას გვთავაზობს. საძიებო სისტემა ახდენს ლექსიკური ან ფუნქციური ელემენტის შემცველი ყველა კონტექსტის გენერირებას. შემდეგი ეტაპი მოითხოვს სელექცირებული კონტექსტების შემდგომ სისტემატიზაციას - მათ ფუნქციურ დიფერენციაციას.

მაგალითისათვის განვიხილავთ **„ზედ“** და **„შიგ“** ნაწილაკების ფუნქციური ანალიზის კორპუსული კვლევის შედეგებს.

კვლევის პირველ ეტაპზე ლინეარული ძიების გზით ხდება შესაბამისი ემპირიული ბაზის შექმნა - ***ზედ***: 4916 კონტექსტი, ***შიგ***: 3046 კონტექსტი. პირველადი დაკვირვების შედეგად მიღებული დასკვნების მიხედვით **„ზედ“** და **„შიგ“** ნაწილაკების პრაგმატული ფუნქცია შემდეგნაირად გამოიკვეთა:

ა) **ზედ** რეალიზდება, როგორც ფსევდოანაფორული ელემენტი, რომელიც დეიქსისის ფუნქციასაც ითავსებს:

ბ) **ზედ** რეალიზდება, როგორც კატაფორული ელემენტი და მოცემულ მაგალითში დირექციონალური სემანტიკის გარდა მოდალური ელემენტის ფუნქციასაც ითავსებს:

მოდალურობის გარდა ასევე დასტურდება სხვა ფუნქციებიც, როგორიცაა აფირმატივი, პრედიკატით გამოხატული ქმედების ინტენსიფიკაცია, ინფორმაციის სტრუქტურირება ფოკუსირების გზით.

ამგვარი დებულებები ასახავენ ჰიპოთეტურ ვარაუდებს, რომელთა ვერიფიცირება სიგნიფიკანტური რაოდენობის ემპირიული მონაცემების ბაზაზე მეთოდური აუცილებლობაა.

3. ინტერლინეარული ძიება, როგორც კვლევის მეორე საფეხურის ოპერაცია

კორპუსში საანალიზო ნაწილაკების **„ზედ“** და **„შიგ“** ძიება, როგორც ცალკეული ელემენტებისა ფორმულით **[„ზედ|შიგ“],** ახდენს შესაბამისი კონტექსტების გენერირებას - **ზედ**: 4916 + **შიგ**: 3046. კონტექსტების მეთოდური სელექციის მეთოდად შეირჩა საძიებო ცნებების რეგულარულ გამოხატულებებად ფორმულირება ისე, რომ ისინი ასახავდნენ (1)-სა და (2)-ში ნავარაუდებ სტრუქტურებს.

ნაწილაკების „ზედ“ და „შიგ“ **ანაფორული** გამოყენება მიემართება მათი შემცველი კონსტრუქციების ფორმალურ სინტაქსურ, resp. ტოპოლოგიურ მახასიათებელს:

ა) მათი დისტრიბუცია უშუალოდ ზმნის (ფინიტური პრედიკატის) წინ ან

ბ) მათი დისტრიბუცია უშუალოდ არათანდებულიანი ნომინალური ფრაზის წინ.

ამდენად, საძიებო ველში ფორმულირდება არა ცალკეული ფორმა, არამედ ფორმ(ებ)ისა და რელაციების კომბინაცია:

„ზედ|შიგ“ []? [features=„V.*“]

ან

„ზედ|შიგ“ []? [features=„N.*“&„!=x“], მაშინ როცა x=შესაბამის თანდებულს.

ამ რეგულარულ გამოხატულებაზე საძიებო ცნების სახით კორპუსი რეაგირებს 2461 კონტექსტით, რომლის ცენტრალურ ელემენტს წარმოადგენს არა მარტო Key Word (KWIC), არამედ Key Relation („KRIC“).

პარტიკლების „ზედ“ და „შიგ“ **კატაფორული** გამოყენება მიემართება მათი შემცველი კონსტრუქციების ფორმალურ ლექსიკურ მახასიათებელს: ისინი კორესპონდირებენ კონტექსტის სხვა ელემენტებთან, რომლებიც ბოლოვდებიან საძიებო ნაწილაკების სუფიქსუ“რი ინვარიანტებით: „-ზე“ და „-ში“. ამგვარად, მარკირებული ელემენტები მოსდევენ „ზედ“ და „შიგ“ ნაწილაკებს უშუალოდ ან გაშუალებულობის სხვადასხვა ხარისხით[1] და ამ გზით ახორციელებენ კატაფორულ რეფერირებას. კატაფორულობის ამგვარი კონკრეტული შემთხვევა თავსდება შემდეგ რეგულარულ გამოხატულებებში:

„ზედ“ []{0,3} „.*ზე“ (**274** კონტექსტი)

„შიგ“ []{0,3} „.. ში“ (874 კონტექსტი)

ძიების ამგვარ სახეს ჩვენ ვუწოდებთ **ინტერლინეარული ძიების სახეს,** ვინაიდან მისი განხორციელება მოითხოვს ინტერლინეარული ანალიზის სპეციალურ დონეს, რომელიც შეიცავს ტოკენის კატეგორიზაციას - AOS (Art of speech) და მორფოლოგიურ სეგმენტაციას.

From KWIC to KRIC:
Adjustment of Corpus Queries to Complex Linguistic Queryions

Manana Tandaschwili & Zakharia Pourtskhvanidze
Goethe University Frankfurt (Germany)
tandaschwili@em.uni-frankfurt.de, pourtskhvanidze@em.uni-frankfurt.de

1. First observations

The functioning of Kartvelology in the 21th century and its integration into the global scientific space depends on the development of modern research methods and tools, which are currently identified with Corpus Linguistics. Use of powerful technologies has permitted the empirical verification of theoretical assumptions and adequacy of linguistic description.

[1] ჩვენი დაკვირვებით, არ უმეტეს 3 ტოკენით.

The GNC arose from international cooperation between different researcher groups and projects. The Georgian Language Corpus GEKKO compiled by Paul Meurer currently contains approximately 130 million tokens. GEKKO consists of the historical corpus TITUS developed by Jost Gippert and examples of the modern Georgian language from newspapers, works of fiction, etc. One part of GNC is also envisaged „The Georgian Dialect Corpus" curated by Marine Beridze. Mentioned electronic resources are used as a basis for the Georgian National Corpus, which is currently a work in progress project funded by Volkswagen Foundation.

The effectiveness of corpus use depends on flexible user interface design. Any such design requires the existence of a general strategy thar is focused on linguistic needs. The aforementioned resources present, to use our term, „linear query systems".

This paper describes an alternative form of query based on the use of regular expressions, which allows the generation not only of contexts with single forms, or chains of single forms, but contexts with complex grammatical relations.

2. „Linear query" as first step operation

„Linear query" performs the first selection of search keywords (lexical and functional elements), including corresponding contexts. The next step of selection must concern further systematization due to functional differentiation.

A thorough analysis of the particles ***zed*** 'on' and ***šig*** 'in/into' leads to five hypotheses concerning their linguistic functions in Georgian.

„Linear query" of those elements generates 4 916 contexts for „ზედ" and 3046 for „შიგ". The particles ***zed*** 'on' and ***šig*** 'in/into' emphasize follow referential functions

a. Quasi-anaphoric usage with secondary expression of deixis
b. Cataphoric usage

Three more functions that can be determined are the following: Affirmative, Modal usage and usage as focus particle.

The analyses in (2.1) and (2.2) reflect resumptions, which must be verified from an empirical point of view.

3. **"Interlinear query" as second step operation**

An empirical examination of this hypothesis requires a differentiated corpus query, given that a search for ზედ|შიგ yields 5008 hits in the annotated GEKKO corpus. We use the possibilities of the annotated parts of GEKKO (about 20 million tokens) and combine different properties and items into one search term.

The quasi-anaphoric usage of the particles ***zed*** 'on' and ***šig*** 'in/into' is characterized by formal syntactic markers:

a. the particles ***zed*** 'on' and ***šig*** 'in/into' occur in the syntactic position immediately before the Verb;
b. the particles ***zed*** 'on' and ***šig*** 'in/into' occur in the syntactic position immediately before the Noun or NP, which has no corresponding suffix.

4. Conclusion

We do not consider the development of a language corpus an end in itself. The implementations of many recent corpus projects show a picture in which the programming aspect of their development plays a key role. This leads to problems with usability of these corpora.

The formulation of complex regular expressions requires specific competencies in programming. Not all linguists have the knowledge necessary to transform grammatical rules into regular expressions. Therefore, we provide a strategic proposal in order to develop these language corpus projects further. We believe that the view of the corpus user should stand at the centre of planning and implementation. We consider it essential to concentrate the whole corpus development process on the user and the criterion of wide usability.

For the further development of the Georgian National Corpus, two solutions may be considered: (1) The development of a library of complex expressions, and/or (2) the creation of a set of simple buttons on the user interface to allow the user to combine expressions in the way indicated above.

References:

GEKKO - *www.iness.uib.no/gekko*

Pourtskhvanidze Z., Fokuspartikeln und Wortstellung im Georgischen. PhD Script. 2011 Frankfurt (M).

Tandaschwili M., Lokation und Deixis im Georgischen –I. Funktional-semantische Analyse von ***zed*** und ***šig***, ProGeorgia, 2013 (im Druck)

TITUS - *http://titus.uni-frankfurt.de*

ლიტვური სიტყვების აქცენტუაციის ნორმების პრინციპების შესწორება და მათი გადატანა ლექსიკონებში

ვიდას კავალიაუსკასი

ლიტვის პედაგოგიური უნივერსიტეტი (ლიტვა)

vk1119@gmail.com

ლიტვურ ენაში მახვილი თავისუფალია და არაფიქსირებული. მახვილი სიტყვას შეიძლება მოუვიდეს ნებისმიერ მარცვალზე. როდესაც სიტყვები ფორმას იცვლიან, იგი შეიძლება ერთი მარცვლიდან მეორეზე გადავიდეს გარკვეული წესების მიხედვით. მახვილი ასევე ერთმანეთისაგან მიჯნავს სიტყვათა ლექსიკურ და გრამატიკულ მნიშვნელობებს. ამიტომაც ყველა ლიტვური სიტყვა მახვილის დართვითაა წარმოდგენილი ლექსიკონებში. თითოეულ სახელს ახლავს ინფორმაცია იმის შესახებ, რომელ აქცენტუაციურ პარადიგმას მიეკუთვნება იგი. მიუხედავად ამისა, ლექსიკონებში დადგენილ აქცენტუაციურ ნორმებს ყოველთვის მხარს არ უჭერს ცოცხალი მეტყველება – ზოგიერთ შემთხვევაში სასაუბრო ენასა და დიალექტებში განსხვავებული აქცენტუაციური მოდელები გვხვდება. ცოცხალ მეტყველებასა და დადგენილ ნორმებს შორის არსებული შეუსაბამობა პერიოდულად ხვდება ლიტვური ენის სახელმწიფო კომისიის დღის წესრიგ-

ში. 2003–2005 წლებში მისმა სტრუქტურულმა დანაყოფმა – „წარმოთქმისა და აქცენტუაციის ქვეკომისიამ“ – მოამზადა 17 რეკომენდაცია, რომელთა მიხედვითაც სხვადასხვა მეტყველების ნაწილში შემავალი სიტყვების აქცენტუაციის შესწორებული ნორმები იყო წარმოდგენილი. ყველა ახლადდამტკიცებული აქცენტუაციის ნორმა შეტანილ იქნა „ახალი ლიტვური ენის ლექსიკონის“ გადამუშავებულ და განახლებულ გამოცემაში, რომელიც გამოიცა 2012 წელს (აქცენტუაციის ნორმები დაამუშავა და განაახლა წინამდებარე ნაშრომის ავტორმა). ეფექტური, სწრაფი და წარმატებული კოდიფიცირების უზრუნველსაყოფად საჭიროა ცოცხალი მეტყველების წარმომადგენლობითი მასალა ლიტვური ენის გავრცელების სრული არეალიდან. სამწუხაროდ, სასაუბრო ენის აქცენტუაციური ტენდენციები დღემდე საკმაოდ ეპიზოდურად იყო შესწავლილი და კვლევები ინფორმანტთა ძალზე მცირე რაოდენობას ეყრდნობოდა; ისინი გარკვეულწილად მოძველებულიც იყო. ამიტომაც, 2011 წელს დაიწყო ახალი პროექტი „ახალგაზრდების აქცენტუაციური ტენდენციები,“ რომელიც დააფინანსა ლიტვური ენის სახელმწიფო კომისიამ. მისი დასრულება დაგეგმილია 2013 წელს. წინამდებარე პროექტი წარმოადგენს პირველ ცდას ლიტვური ენათმეცნიერების ისტორიაში, შეისწავლოს ახალგაზრდების აქცენტუაციური ტენდენციები კომპლექსური თვალსაზრისით. ამგვარი კვლევების მნიშვნელობა ეჭვგარეშეა: ისინი გამოავლენენ კოდიფიკაციის მიმართულებას და გარკვეულ ტენდენციებს აქცენტუაციის განვითარებაში. მაშასადამე, მოსალოდნელია, რომ პროექტი არა მარტო წარმოადგენს მითითებებს კოდიფიკატორებისათვის, არამედ აქცენტუაციის ვარიანტების სისტემური კვლევის დასაწყისის მაუწყებელიც იქნება. პროექტის შედეგები დიდაქტიკური თვალსაზირისითაც იქნება მნიშვნელოვანი, რადგანაც დაგეგმილია სხვადასხვა რეგიონში მცხოვრები ინფორმანტების გამოკითხვა. მოსალოდნელია, რომ გამოკითხვის მასალები საშუალებას მოგვცემს გავაანალიზოთ, თუ როგორ ვასწავლოთ სპეციფიკურ დიალექტებზე მოსაუბრე მოსწავლეებს. ამასთან ერთად პროექტი დიდძალ მასალას მიაწვდის დარგობრივი ენების მასწავლებლებს.

პროექტი მიზნად ისახავს პოტენციური ვარიანტების (სიტყვებისა და მათი ფორმების) აქცენტუაციური ტენდენციების გამოვლენას ახალგაზრდების სასაუბრო მეტყველებაში. აქცენტუაციური ვარიანტების შესახებ დიდძალი მასალა მოგროვდება ლიტვის სხვადასხვა კუთხეში; ამგვარად, წარმოდგენილი პროექტი უშუალო კავშირშია პროგრამასთან „სალიტერატურო ლიტვური ენის, დიალექტებისა და სხვა ენობრივი სახესხვაობების ფუნქციონირებისა და ცვლილების კვლევა,“ რომლის მიზანია წარმომადგენლობითი მასალის თავმოყრა ენის ფუნქციონირებისა და ცვალებადობის შესახებ. საბოლოო შედეგი იქნება მოპოვებული მასალის ანალიზი, შედარება და აქცენტუაციის პრობლემური შემთხვევების გამოვლენა. პროექტი დასრულდება ვრცელი რეკომენდაციების შემუშავებით, რომლებიც უნდა გადაეცეს ლიტვური ენის სახელმწიფო კომისიას. წარმოთქმისა და აქცენტუაციის ქვეკომისია განიხილავს ამ მასალებს და მიიღებს ოფიციალურ რეზოლუციებს გარკვეული სიტყვებისა ან მათი ფორმების აქცენტუაციის ნორმათა შესწორების შეთავაზების თაობაზე. ამიტომაც, ამ პროექტს აშკარად პრაქტიკული ღირებულებაც აქვს. ლიტვური ენის ინსტიტუტის ლექსიკოლოგები ამჟამად მუშაობენ „სალიტერატურო ლიტვური ენის ლექსიკონის“ აბსოლუტურად ახალი ვერსიის შედგენაზე. ვიმედოვნებთ, რომ პროექტის მასალებზე დაყრდნობით შემუშავებული გადამუშავებული აქცენტუაციური ნორმები ახალ ლექსიკონშიც იქნება დაფიქსირებული.

On the Principles of Correction of Accentuation Norms of Certain Lithuanian Words and their Transfer to Dictionaries

Vidas Kavaliauskas
Lithuanian University of Educational Sciences (Lithuania)
vk1119@gmail.com

A stress in Lithuanian is free and not fixed. A stress can occur on any syllable of a word. When words are inflected, it may shift from one syllable to another according to certain rules. A stress also delimits lexical meanings or grammatical meanings of words. Therefore, all Lithuanian words are stressed in dictionaries. Each substantive is accompanied by the information about which accentuation paradigm it belongs to. However, the accentuation norms, established in dictionaries, are not always supported by actual usage – in some cases, different accentuation patterns occur in spoken language and dialects. The problem of inconsistency between actual usage and established norms is periodically on the agenda of the State Commission of the Lithuanian Language. In 2003–2005, its structural unit – Subcommission of Pronunciation and Accentuation – prepared 17 recommendations revising the accentuation norms of words belonging to different parts of speech. All newly approved accentuation variants were transferred into the seventh revised and updated edition of *The Dictionary of Modern Lithuanian Language,* published in 2012 (the accentuation norms were revised and updated by the author of this paper). To ensure the effective, speedy and successful codification, representative data of actual usage accumulated in a targeted manner from the total area of the Lithuanian language are required. Unfortunately, accentuation tendencies of spoken language have to date been analyzed rather episodically and cover a rather low number of informants; they are also slightly outdated. Hence, a new project *Youth Accentuation Tendencies,* funded by the State Commission of the Lithuanian Language, was launched in 2011. It is expected to be completed in 2013. The present project is a first attempt in the history of Lithuanian linguistics to investigate youth accentuation tendencies in a complex manner. The importance of such surveys is unquestionable: they reveal the trend of codification and certain tendencies in the development of accentuation. Therefore, it is likely that the present project will not only serve as the guidelines for accentuation codifiers but will also mark the beginning of the systematic research of accentuation variants. The results of the project should also be important in a didactic aspect. Whereas it is planned to interview the informants from different regions, it is expected that the survey data will provide possibilities for the analysis of how the pupils, speaking a specific dialect, should be taught. In addition, the project will provide plenty of data for teachers of the language for specific purposes. The aim of the project is to identify the accentuation tendencies of potential variants (words and their forms) in the spoken language of the youth. Plenty of data on accentuation variants will be collected from various locations in Lithuania; thus, the present project is immediately associated with the program *Research of Functioning and Change of Standard Lithuanian, Dialects and Other Language Varieties 2011–2020*, aiming at accumulating of representative data about the functioning and change of the language. The ultimate result will be the analysis and comparison of the data obtained and the

identification of problematic accentuation cases. The project will end with the drafting of extensive recommendations, to be submitted to the State Commission of the Lithuanian Language. The Subcommission of Pronunciation and Accentuation will consider the data and adopt official resolutions on the proposal of the correction of accentuation norms of certain words or their forms. Therefore, the project renders an obvious practical value. The lexicologists of the Institute of the Lithuanian Language are currently in progress of compiling an absolutely new *Dictionary of Standard Lithuanian*. It is expected that the would-be changes of the revised accentuation norms of words, adopted on the basis of the data obtained in the project, will contribute to the new dictionary.

ზმნური ფუძეების კვლევისათვის დიალექტებში დიალექტური კორპუსის მიხედვით

ციცინო კვანტალიანი, რუსუდან ლანდია
თსუ არნ. ჩიქობავას სახ. ენათმეცნიერების ინსტიტუტი (საქართველო)
tsitsino.nino.kvantaliani@gmail.com, landiarus@gmail.com

ქართული ენის ზმნური ფუძეებისა[1] და ქართული ენის სახელზმნური ფუძეების[2] ლექსიკონების შედგენის პროცესში ივარაუდებოდა მუშაობის გაგრძელება დიალექტურ მასალაზე და ასევე დიალექტური ლექსიკონების შედგენა. რა თქმა უნდა, ეს ძალიან შრომატევადი და ხანგრძლივი სამუშაო იქნებოდა. ქართულ დიალექტურ კორპუსზე (ქდკ) დაყრდნობით დღეს უკვე შესაძლებელია ამ ამოცანის რეალიზება. დიალექტური კორპუსი კვლევის გაფართოების საშუალებას იძლევა, რითაც შესაძლებელი გახდება დიალექტებში ზმნური და სახელზმნური ფუძეების შესწავლა არა მარტო ფორმობრივ-სტრუქტურული, არამედ სემანტიკური თვალსაზრისითაც.

წინამდებარე მოხსენებაში განვიხილავთ მხოლოდ ზმნურ ფუძეებს და წარმოვაჩენთ მათი შემდგომი კვლევის მოსამზადებელ სამუშაოებს.

ქართული ენის ზმნური ფუძეების ლექსიკონის ხუთგრაფიანი სქემიდან (I. სალექსიკონო ერთეული, ზმნური ფუძე; II. აწმყო-მყოფადის ფუძე /სალექსიკონო ერთეული + თემის ნიშანი/; III. ზმნისწინები; IV. ვნებითის ტიპი /პრეფიქსიანი, სუფიქსიანი თუ უნიშნო/; V. ხმოვანპრეფიქსები), რომელიც ზმნის ფორმაწარმოების შესაძლებლობებს უკეთ წარმოაჩენს, ამჯერად ჩვენ მხოლოდ სამ გრაფაზე შევჩერდებით: 1) ზმნური ფუძე, 2) ზმნის აწმყო-მყოფადის ფუძე (სალექსიკონო ერთეული + თემის ნიშანი), 3) ზმნურ ფუძესთან დაკავშირებული ზმნისწინები.

[1] გ. გოგოლაშვილი, ც. კვანტალიანი, დ. შენგელია. ქართული ენის ზმნური ფუძეების ლექსიკონი, თბილისი, 1989 წ.

[2] გ. გოგოლაშვილი, ც. კვანტალიანი, დ. შენგელია. ქართული ენის სახელზმნური ფუძეების ლექსიკონი, თბილისი, 1991 წ.

დიალექტური ზმნური ფორმები ძირითადად მიჰყვება და იმეორებს სალიტერატურო ქართული ენის ზმნურ ფორმაწარმოებას, თუმცა ფორმაწარმოების საშუალებების შერჩევისა და ურთიერთშეხამების ვარიაციულობა ქმნის მნიშვნელოვან დიალექტურ სხვაობას.

ზმნური ფორმაწარმოების მიხედვით დიალექტური ვარიაციულობის რამდენიმე შემთხვევა გამოიყოფა:

1) დიალექტებში სალიტერატურო ენისაგან არ განსხვავდება არც ფუძე და არც ზმნის ყალიბი. განსხვავებას ქმნის თემის ნიშანთა ფონეტიკური ვარიანტები.

ყალიბი: V-ზმნური ფუძე - თემის ნიშანი

ა-გინ-ებ	აგინებს, შიაგინა, შამააგინებს
ა-გინ-ეფ	აგინეფს, მაგინეფს
ა-გინ-ავ	დააგინავს, გააგინავს, გვაგინავდე (=შეაგინებს)
ა-გინ-ამ	აგინამს

2) დიალექტებში სალიტერატურო ქართული ენისაგან არ განსხვავდება არც ზმნური ფუძე, არც ზმნის ყალიბი, მაგრამ ზმნისწინი განსხვავებულია და აწარმოებს ახალ ზმნურ ფორმას, რომელიც სალიტერატურო ქართულ ენაში არ დასტურდება. იცვლება ზმნის მნიშვნელობაც.

ყალიბი: ზმნური ფუძე - თემის ნიშანი

ფარ-ავ ფარავს, დაფარავს

მაგრამ

ამოფარავს = დაფარავს

სალიტერატურო ქართულში **ამოფარავს** ფორმა არ იწარმოება.

„პურისათვის ჯერ მიწას დავხნავთ, მერმედ დავფარცხავთ, მერმედ პურს **ამოვფარავთ**“ (ქვ.იმერ.).

3) დიალექტებში არ იცვლება ზმნური ფორმა, არც ზმნის მნიშვნელობა, მაგრამ იცვლება ყალიბი:

ყალიბი: V -ზმნური ფუძე - თემის ნიშანი (ვარიანტებით)

ა-ფას-ებ	დააფასებს
ა-ფას-ეფ	დააფასეფს
ა-ფას-ენ	დააფასენდა (სამცხ.ჯავახ.)

ყალიბი: ზმნური ფუძე - თემის ნიშანი

ჰ-ფას-ავ დაჰფასავდნენ = დააფასებდნენ (მთ.-გუდამაყრ.)

4) დიალექტებში შეიცვალა ზმნური ფუძე (ან მორფემის ფონეტიკური / მორფოლოგიური ვარიანტის დართვით, ან კუმშვით), ზმნის ყალიბიც და მნიშვნელობაც უცვლელია.

ყალიბი: V -ზმნური ფუძე - თემის ნიშანი (ვარიანტები)

1)	ა-ფარ-ევ	იფარევენ, აფარევენ
	ა-ფარ-ეფ	გადავაფარეფთ, დააფარეფ
	ა-ფარ-ენ	დაგვაფარენდა, შეგვაფარენდა
	ა-ფარ-ემ	დავაფარემდით, დავაფარემ
	ა-ფარ-ან	თავ აფარანდნენ (სამცხ.-ჯავახ.).

2). V-შეკუმშული ფუძე - თემის ნიშანი

ა-ფრ-ი-ს = აფარებს მიწას აფრის (სამცხ.ჯავახ.).

ამგვარად, დიალექტური კორპუსის მიხედვით ზმნურ ფუძეთა წარმოების ძალიან საინტერესო სურათი იკვეთება. არანაკლებ საინტერესოა დიალექტური კორპუსის მიხედვით ზმნისწინის ფონეტიკური ნაირსახეობების შესწავლა.

მოხსენებაში წარმოდგენილი იქნება დიალექტური ზმნური ფუძეების წარმოებასთან დაკავშირებული საკითხების კორპუსული კვლევის კონკრეტული შედეგები.

On the Study of Verbal Stems in Dialects Based on the Georgian Dialect Dictionary

Tsitsino Kvantaliani & Rusudan Landia
Arn. Chikobava Institute of Linguistics, TSU (Georgia)
tsitsino.nino.kvantaliani@gmail.com, landiarus@gmail.com

During the process of compiling of the dictionaries of verbal stems of Georgian[1] and of masdar Georgian[2], we were planning to continue our work on dialects and to compile dialect dictionaries. It certainly was going to be a very labor-consuming and long-term activity. Thanks to the GDC, presently this objective can be accomplished. The dialect corpus allows for broadening of research in order to examine verbal and masdar stems in dialects both in terms of their formal and structural and semantic properties.

The present paper addresses only verbal stems, shedding light upon the preparatory activities for their future research.

Of the 5-item scheme of *A Dictionary of Verbal Stems of Georgian* (I. entry, verbal stem; II. Present-Future stem (entry + thematic marker); III. Preverbs; IV. type of passive (prefixed, suffixed, or unmarked); V. vowel prefixes), representing well the inflectional opportunities of the verb, currently we will concentrate only on three items: 1) verbal stem; 2) Present-Future stem of a verb (entry + thematic marker); 3) preverbs attached to a verbal stem.

Dialect verb forms mostly follow and repeat the verb-formation pattern of Standard Georgian; however, variability of the selection and co-occurrence of word-formation means yields in dialect differences.

According to the verb-formation, several cases of dialect variability have been identified:

1) As different from the language standard, dialects do not evidence the difference in terms of either a stem or a verb pattern. Phonetic variations of thematic markers make the difference. Pattern: V - verbal stem - thematic marker

a-gin-eb	agineb, šiagina, šamaaginebs
a-gin-ep	agineps, magineps

[1] Gogolashvili, G., Kvantaliani, Ts., Shengelia, D. *A Dictionary of Verbal Stems of Georgian*. Tbilisi, 1989.
[2] Gogolashvili, G., Kvantaliani, Ts., Shengelia, D. *A Dictionary of Masdar Stems of Georgian*. Tbilisi, 1991.

a-gin-av daaginavs, gaaginavs, gvaginavde (=šeaginebs)
a-gin-am aginams

2) Dialects do not exhibit differences from the standard with respect to either a verbal stem or a verb pattern; however, a preverb is distinct and derives a new verb form, not attested in Standard Georgian. A verb me aning is also changed.

Pattern: verbal stem - thematic marker

par-av paravs, daparavs

but

amoparavs = daparavs

Standard Georgian does not derive the form **amoparavs**.

3) in dialects, neither a verb form nor a verb meaning undergo any change, but the meaning is modified:

Pattern: V - verbal stem - thematic marker (with variants)

a-pas-eb daapasebs
a-pas-ep daapaseps
a-pas-en daapasenda (Meskhetian-Javakhetian)

Pattern: verbal stem - thematic marker

h-pas-av dahpasavdnen – daapasebdnen (Mtiulian-Gudamakrian)

4) In dialects, a verbal stem was modifies (either by means of taking on a phonetical/morphological variant of a morpheme or by means of syncopation); a verb pattern and meaning are not modified.

Pattern: V - verbal stem - thematic marker (with variants)

1) a-par-ev ipareven, apareven
a-par-ep gadavaparept, daaparep
a-par-en dagvaparenda, šegvaparenda
a-par-em davaparemdit, davaparem
a-par-an tav aparandnen (Meskhetian-Javakhetian)

2) V – syncopated stem – thematic marker

a-pr-i-s = aparebs mic'as apris (Meskhetian-Javakhetian)

Thus, a very interested picture outlines in terms of the formation of verbal stems according to the dialect corpus. A corpus-based study of phonetic varieties of preverbs seems to be nonetheless interesting.

The paper will present specific results of the corpus-based study of the issues associated with the formation of verbal stems in dialects.

კორპუსული მონაცემების წვლილი და მნიშვნელობა ენისა და გენდერის კვლევაში

ზაალ კიკვიძე
თსუ არნ. ჩიქობავას სახ. ენათმეცნიერების ინსტიტუტი (საქართველო)
zaalk@yahoo.com

ენისა და გენდერის კვლევის სფეროში შედის საკითხები, რომლებიც ერთიანდება ორი ფართო მიმართულების ფარგლებში: 1. როგორ მეტყველებენ ქალები და მამაკაცები; 2. რა საშუალებებით არიან ისინი წარმოდგენილი ენაში. ამ ორი მოცულობითი თემით ნაგულისხმევი საკითხების კვლევისას ემპირიულ მასალად სხვადასხვა სახის რესურსი გამოიყენება გენდერლინგვისტიკაში; და მაინც, ბუნებრივია, რომ ამ საქმეში კორპუსს მნელად თუ შეედრება რომელიმე სხვა რესურსი.

როდესაც ენისა და გენდერის საკითხების კვლევაში კორპუსის ჩართვაზე ვსაუბრობთ, ხაზი უნდა გაესვას იმ ფაქტს, რომ მისი გამოყენება არა მხოლოდ იმიტომ დგება დღის წესრიგში, რომ იგი ხელმისაწვდომია, არამედ იმიტომაც, რომ მისი საშუალებით ობიექტური და მრავალმხრივი ინფორმაციის მიღებაა შესაძლებელი.

კორპუსის რესურსების ჩართვა გენდერლინგვისტიკურ კვლევა-ძიებაში ჯერჯერობით შედარებით ახალი ხილია და ძალიან ცოტა ნაშრომი არსებობს, რომლებშიც ზემოხსენებული მიმართულებებიდან ძირითადად მხოლოდ პირველია (როგორ მეტყველებენ ქალები და მამაკაცები) განხილული.

რაც შეეხება მეორეს (რა საშუალებებით არიან წარმოდგენილი ქალები და მამაკაცები ენაში), აქ კორპუსის გამოყენების საკითხი მხოლოდ ახლა დგება დღის წესრიგში. ამ მხრივ არსებული ტრადიცია ძირითადად ინგლისური ენის მასალის ანალიზს ეყრდნობა, რის შედეგადაც შემუშავებულია შემდეგი დებულება: ენა ანთროპოცენტრული ფენომენია, რადგან მას ადამიანები გამოიყენებენ და, შესაბამისად, მასში დაუნჯებულია ადამიანთა მსოფლხედვა; ამასთანავე, ენა ანდროცენტრულიცაა, რადგან საუკუნეების განმავლობაში მამაკაცები წარმოადგენდნენ გაბატონებულ სოციალურ ჯგუფს, რის შედეგადაც ენაში მამაკაცური მსოფლხედვაა დაუნჯებული. ამგვარი დასკვნა განმტკიცებულია შესაბამისი საილუსტრაციო მასალით. არსებული გამოცდილება კარგ ამოსავალს წარმოადგენს მსგავსი ენობრივი მასალის შესამოწმებლად სხვა ენებში გენდერული რეპრეზენტაციის თვალსაზრისით. რომანულსა და სხვა გერმანიკულ ენებში არ არის გამოვლენილი პრინციპული განსხვავებები ინგლისურთან შედარებით. ქართული, როგორც გენეტიკურად და ტიპოლოგიურად განსხვავებული ენა, გონივრულ არჩევანს წარმოადგენს ამგვარი შეპირისპირებისათვის.

სხვადასხვა სახის ლექსიკოგრაფიული წყაროებისა და კორპუსული რესურსების („ქართული დიალექტური კორპუსი“, თსუ არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი) და „ქართული კორპუსი“ (ბერგენის უნივერსიტეტი) გამოყენებით წინამდებარე მოხსენებაში განხილულია ენის ანდროცენტრულობასთან დაკავშირებული ერთეულები შეპირისპირებით ასპექტ-

ში. კორპუსული ანალიზის შედეგად მნიშვნელოვანი რაოდენობრივი პარამეტრებიცაა გამოვლენილი, რაც გამორიცხავს შემთხვევითობას.

მიუხედავად საფუძველთდამდები განსხვავებებისა, ინგლისურ და ქართულ მასალაში უამრავი საერთო ნიშანი ჩნდება, თუმცა ქართულში არსებული ვითარება აქარწყლებს დებულებას ანდროცენტრული მიკერძოების უნივერსალურობის შესახებ ენაში, რაც იმას ნიშნავს, რომ კორპუსზე დამყარებული კვლევის შედეგად ენობრივი ფარდობითობის დამადასტურებელი ფაქტების გამოვლენაც გახდა შესაძლებელი.

Contribution and Significance of Corpus Data to/for Language and Gender Studies

Zaal Kikvidze
Arn. Chikobava Institute of Linguistics, TSU (Georgia)
zaalk@yahoo.com

Language and Gender Studies comprise issues within the framework of two broad directions: 1. how women and men speak, and 2. how they are represented in language. Investigations of these two very comprehensive areas have applied various kinds of resources as empirical data; and yet, a corpus, naturally enough, can hardly be rivaled in such an endeavor.

Whenever the involvement of a corpus in language and gender research is concerned, it should be emphasized that its application is addressed not only because it is available but also because it is possible to extract fair and diverse information by means of it.

The involvement of corpus resources in language and gender research is a relatively new fruit to date, and very few studies exist, which, of the aforementioned two directions, mostly the former (1. how women and men speak) has been dealt with.

As for the latter one (2. how they are represented in language), the application of corpus data has been a very recent issue on the agenda. The existing tradition has been predominantly based on the data from English, having led to the advancement of the following provision: language is an anthropocentric phenomenon as far as it is used by humans, and, thus, it is a carrier of the human worldview; besides, language is androcentric as well as far as, for centuries, men have been a dominant social group, having resulting in the dominance of the male worldview in language. This provision has been supported by corresponding illustrative data. The said experience can serve as a good outset for examining of similar linguistic data in other languages in terms of gender representation. Romance and other Germanic languages do not exhibit any principal differences with English. Georgian, as a genetically and typologically distinct language, is a reasonably choice for such a contrast.

Based on various lexicographic sources and corpus resources ("Georgian Dialect Corpus" (Arnold Chikobava Institute of Linguistics, TSU) and "Georgian Corpus" (Bergen University)), the present paper

verifies linguistic units, associated with androcentrism, in a contrastive aspect. The corpus analyses help indentify significant quantitative dimensions to exclude random outcomes.

Notwithstanding the fundamental distinctions, the English and Georgian data demonstrate a number of common features; however, the Georgian data refuted the provision about the universal character of androcentric bias in language, this implying that the corpus-bases analysis succeeded in the detection of the facts of linguistic relativity.

სასკოლო ლექსიკონები საქართველოს პარლამენტის ეროვნული ბიბლიოთეკის ლექსიკონთა ერთიან ციფრულ ბაზაში

გიორგი კილაძე, ნათია შენგელაია
საქართველოს პარლამენტის ეროვნული ბიბლიოთეკა (საქართველო)
giorgi_x2000@yahoo.com

თანამედროვე ბიბლიოთეკები, განსაკუთრებით ეროვნული ბიბლიოთეკები, დღეს აღარ განიხილება, როგორც წიგნადი ფონდის შენახვის ადგილი და მათი ფუნქცია მხოლოდ მკითხველთა მომსახურებით არ ამოიწურება.

წიგნადი ფონდის შენახვისა ან მკითხველთა მომსახურებაზე საუბრისას მხედველობაში გვაქვს როგორც ნაბეჭდი, ასევე გაციფრებული წიგნები.

საბიბლიოთეკო სისტემის განვითარების თანამედროვე ეტაპზე დიდმა ბიბლიოთეკებმა, და განსაკუთრებით ეროვნულმა ბიბლიოთეკებმა, სხვა სახის ფუნქცია და დანიშნულება შეიძინეს – განათლების, კულტურის, სოციალურ სფეროში უფრო ფართოდ ჩართვის აუცილებლობა.

ეროვნული ბიბლიოთეკის საგანმანათლებლო სივრცეში უფრო ფართოდ ჩართვაში ძირითადად ორი მიმართულება მოიაზრება:

I. ეროვნულ ბიბლიოთეკაში უკვე არსებული რესურსების სწრაფად და მარტივად მიწოდება საგანმანათლებლო დაწესებულებებისათვის (სკოლები, უნივერსიტეტები);

II. ახალი საგანმანათლებლო და შემეცნებითი რესურსების შექმნა და მიწოდება.

კონფერენციაზე წარმოვადგენთ ორ ახალ ელექტრონულ რესურს:

1. „ვეფხისტყაოსნის“ ლექსიკონი - მომხმარებელს მიეწოდება ტექსტში გამოყენებული უცნობი ან დასაზუსტებელი სიტყვებისა და ფრაზების განმარტებები.

2. „სასკოლო ლექსიკონი“ – სასკოლო სახელმძღვანელოებში გამოყენებული ძნელად გასაგები სიტყვებისა და ფრაზების განმარტებები.

დღეისათვის დასახელებული ლექსიკონები მკაცრად სტრუქტურიზებული არ არის და შევსების პროცესშია.

School Glossaries in entire digital base of National Parliamentary Library of Georgia

George Kiladze & Natia Shengelaia
National Parliamentary Library of Georgia (Georgia)
giorgi_x2000@yahoo.com

Modern libraries, such as national libraries, are no longer considered a place for storing of books and their function is not limited to those of serving to readers.

When discussing book repositories and/or services for readers, we mean both print books and digitized ones.

At the present stage of development of the library system, the Great Libraries, as well as the National Libraries, acquired other functions – the necessity of participation in education, culture, social services.

For the participation of the National Library in the education domain, mainly two aspects are considered:

I. The National Library has the resources to deliver quickly and easily to educational institutions (schools and universities);

II. Creation and delivery of new educational information - bibliographic and cognitive resources.

We will present two new electronic resources: 1. Dictionary of *The Knight in the Panther's Skin* - customers will be provided with definitions of unknown words and phrases, occurring in the texts or with words that need to be clarified. 2. *School Glossary* - definitions of hard-to-understand words and phrases, occurring in school textbooks.

At present, the aforementioned dictionaries are not strictly structured and are in the process of completion.

დისტანციური სწავლება: მიღწევები და გამოწვევები

მანანა კობაიძე
მალმეს უნივერსიტეტი (შვედეთი)
kobaidze@comhem.se

ცნობილია, რომ ონლაინკურსებმა და პროგრამებმა დიდი ხანია საბოლოოდ დაიმკვიდრა ადგილი განათლების სისტემაში. სტუდენტების უფრო და უფრო დიდი ნაწილი ირჩევს ამ ფორმას. დისტანციურ განათლებას რამდენიმე უპირატესობა აქვს. ის ეკონომიურია – სტუდენტს არ სჭირდება სხვა ქალაქში საცხოვრებლად გადასვლა, ადვილი ხდება სწავლისა და სამსახურის,

ასევე, სწავლისა და ოჯახური ცხოვრების შეთავსება. შეზღუდული უნარების მქონე პირებისთვისაც ეს ფორმა უაღრესად მოსახერხებელია. სასწავლო მასალა უმეტესად ინტერნეტით მიეწოდება სტუდენტს, იქნება ეს ლექციების ვიდეო თუ აუდიო ჩანაწერები, ტექსტები, ინტერაქტიული სავარჯიშოები თუ სამეცნიერო ლიტერატურა, რომლებიც ციფრულ ბიბლიოთეკებშია განთავსებული.

დისტანციურ განათლებას ხშირად ირჩევენ ის პირები, რომლებსაც უკვე აქვთ პროფესია, მაგრამ ესაჭიროებათ დამატებითი განათლების მიღება და ამა თუ იმ მიზეზის გამო არ შეუძლიათ, რომ დაესწრონ ე.წ. კამპუს კურსებს.

გამოკვლევები ადასტურებენ, რომ ონლაინკურსების შემდეგ სტუდენტები ხშირად უკეთეს შედეგს აღწევენ, ვიდრე საუნივერსიტეტო კურსების შემდეგ. მიზეზი ალბათ ის არის, რომ მხოლოდ მაღალი მოტივაციის სტუდენტები ამთავრებენ ონლაინკურსებს.

დღეს ჩვეულებრივი ონლაინკურსების გვერდით გაჩნდა ახალი, ე.წ. მასობრივი ონლაინკურსები (MOOC), რომლებზეც შესაძლოა ერთდროულად ათიათასობით სტუდენტი სწავლობდეს. სტანფორდის უნივერსიტეტი, ჰარვარდის უნივერსიტეტი და სხვა უაღრესად პრესტიჟული სასწავლებლები სთავაზობენ მსურველებს ასეთ კურსებს.

ზოგი მკვლევარი ამ ცვლილებას რევოლუციურს უწოდებს და ადარებს ინდუსტრიულ რევოლუციას. ყველაზე ცნობილი პროვაიდერები არიან Udacity, Coursera, Khan Academy და edX-ი (http://chronicle.com/article/Major-Players-in-the-MOOC/138817/).

თითქოს პარადოქსია, მაგრამ მასობრივ ონლაინკურსებზე მცირდება სტუდენტებს შორის ურთიერთობის შესაძლებლობა. გამოკვლევები ადასტურებენ, რომ სტუდენტთა დიდი ნაწილისთვის მნიშვნელოვანია სოციალური ურთიერთობა თანაკურსელებთან. ასეთი ურთიერთობისთვის ჩვეულებრივი ონლაინკურსების სტუდენტებისთვის გამოიყენება ფეისბუკის ჯგუფები და სხვა საშუალებანი, მაგრამ მასობრივ ონლაინკურსებზე ამა თუ იმ საკითხის განხილვის დროს კომენტარების უსაზღვრო რაოდენობა იყრის თავს, რაც ნამდვილი დიალოგისთვის დაბრკოლებას ქმნის. ენობრივი ბარიერებიც იჩენს თავს. სულ ახლახან შეიქმნა კიდევ ერთი პლატფორმა მასობრივი კურსებისათვის, ნოვოედი. ამ ახალი პლატფორმის ძირითადი განმასხვავებელი ის არის, რომ ის სტუდენტებს ერთმანეთთან ინტენსიური თანამშრომლობის საშუალებას აძლევს.

რა თქმა უნდა, ასეთ კურსებს თავიანთი ნაკლიც აქვთ. მასობრივი ონლაინკურსების მასწავლებლები ერთდროულად ყველასთვის მისაწვდომნი და ყველასთვის მიუწვდომელნი არიან. სტუდენტისთვის შესწორებების მიწოდებისა და შეფასების პრობლემა რთულია. ლექციების მომზადებაც განსაკუთრებით დიდ დროს მოითხოვს. საჭირო შედეგის მისაღწევად ხშირად საჭიროა რამდენიმე სპეციალისტის თანამშრომლობა (http://harvardmagazine.com/2013/05/ harvardx-and-edx-online-learning-update).

ციფრული რესურსების მატებასთან ერთად მატულობს პლაგიატის შესაძლებლობაც. სტუდენტების მხრივ პლაგიატის თავიდან ასაცილებლად შექმნილია მონაცემთა ბაზა, სადაც სტუდენტები აგზავნიან თავიანთ ნამუშევრებს და ავტომატური კონტროლის შემდეგ მასწავლებელი იღებს ინფორმაციას შესაძლო პლაგიატის შესახებ.

მოხსენებაში წარმოდგენილი იქნება ბოლო დროის გამოკვლევების მიმოხილვა ონლაინსწავლების დადებითი და უარყოფითი მხარეებისა და განვითარების პერსპექტივების შესახებ.

Distance learning and teaching: achievements and challenges

Manana Kobaidze
Malmö University (Sweden)
kobaidze@comhem.se

Online courses and programs have already an established place within education system. More and more students are choosing this form. Distance education has several advantages:

- It is economical; students don't have to move to another city. Educational materials are provided to the students via the Internet, be it video or audio recordings of lectures, texts or interactive exercises. Reference literature, recommended for students, is preferably accessible in digital libraries.
- It gives easy access to learning and administrative service.
- It makes it easy to combine learning with family life or work.
- This form is very convenient for persons with disabilities.

Distance Education is often the best form for those students who already have a profession, but require additional education, and for some reasons are not able to attend so-called Campus courses. The latest research confirms that students in online courses often achieve better results than those in campus courses. The reason probably is that only highly motivated students complete their online courses.

Along with ordinary online courses, a new type of online courses, so-called Massive Open Online Courses (MOOCs) have appeared. Tens of thousands of students study on such courses. Stanford University, Harvard University and other highly prestigious educational institutions offer such courses.

Several researchers consider this change revolutionary and compare it with the industrial revolution. The best-known MOOC providers are Udacity, Coursera, Khan Academy and edX (http://chronicle.com/ article/Major-Players-in-the-MOOC/138817/).

It seems a paradox, but students on MOOC have less opportunity for interaction to each other than students on ordinary online courses. Research confirms that interaction to course mates is important for large part of the students. For such interaction, Facebook or other means for social interaction are used for online students, but on MOOCs a tremendous number of comments appear during discussions, which create obstacles for genuine dialogue. Differences in language knowledge level among students have also been reported among obstacles. Recently a new platform for massive online courses, NovoEd has appeared. The main distinctive feature of this new platform is that it allows students to intensive cooperation with each other.

Of course, online education has negative sides too. Tutors on MOOCs are accessible for thousands of students but, at the same time, not to individual students. It's problematic to give feedback to students and to evaluate their results. It has been reported that planning and making of lectures and other course materials for MOOCs is especially time consuming and requires team work http://harvardmagazine.com/ 2013/05/harvardx-and-edx-online-learning-update

Usage of digital resources increases the possibility of plagiarism among students. Special databases have been designed for detecting and deterring plagiarism. Students are asked to send their works to such databases where the automatic control is being conducted. Afterwards, the teacher receives information about possible plagiarism.

The presentation will include an overview of recent research concerning advantages and disadvantages of online teaching and learning, as well as its development potential.

მორფოლოგიური და ლექსიკური პარამეტრების ავტომატური გამოვლენა N-გრამებში[1]

მიხაილ კოპოტევი
ჰელსინკის უნივერსიტეტი (ფინეთი)
mihail.kopotev@helsinki.fi

უცხოური ენების შემსწავლელების წინაშე სისტემატურად დგას სიტყვათა თანაპოვნიერების პრობლემა: რომელი ბრუნვა და სიტყვაფორმები გამოვიყენოთ მოცემული წინდებულის მომდევნოდ? რომელი მათგანი უნდა დავიზეპიროთ და რომელ მათგანშია დაცული წესი? ამგვარი შეკითხვების სიხშირე მიუთითებს იმაზე, რომ მნიშვნელოვანი გამოწვევის წინაშე ვდგავართ: ჩვენ უნდა ავაგოთ ისეთი სისტემა, რომლის საშუალებითაც კომპლექსური პასუხის მიღება ავტომატურად იქნება შესაძლებელი.

მოხსენებაში წარმოდგენილია ალგორითმი, რომელიც საშუალებას აძლევს მომხმარებელს შემოიტანოს საძიებო ყალიბი, თავს უყრის მრავალსიტყვიან გამონათქვამებს, რომლებიც ესადაგება ამ ყალიბს და შემდეგ ახდენს მათს რანჟირებას ერთი ნიშნის მიხედვით. აღნიშნულის მიღწევა შესაძლებელი ხდება მრავალსიტყვიან გამონათქვამებში სიტყვაფორმებსა და მათს მახასიათებლებს შორის არსებული ყველა შესაძლო მიმართების ძალის ურთიერთგადამრავლების გზით. ალგორითმი თავს უყრის მოცემული ყალიბის მორფოლოგიური კატეგორიების სიხშირეს ერთიან შკალაზე იმისათვის, რომ შეირჩეს სტაბილური კატეგორიები და მათი დამახასიათებელი ნიშნები. თითოეული მეტყველების ნაწილისათვის და თითოეული მათგანის კატეგორიებისათვის ჩვენ გამოვითვლით კულბაკ-ლაიბლერის ნორმალიზებულ დივერგენციას მოდელში კატეგორიის დისტრიბუციასა და მთლიანად კორპუსში მის დისტრიბუციას შორის. ყველაზე დიდი დივერგენციის მქონე კატეგორიები ყველაზე მნიშვნელოვნადაა მიჩნეული. ამ კატეგორიების გამორჩეული ღირებულებები დალაგებულია სიხშირის კოეფიციენტის მიხედვით. ამის შედეგად ჩვენ ვიღებთ მოცემული ყალიბის მორფოსინტაქსურ პროფილებს, რომელშიც შედის ამ ყალიბის ყველაზე სტაბილური კატეგორიები და მათი დამახასიათებელი ნიშნები.

[1] თანამშრომლობისათვის დიდ მადლობას მოვახსენებ ლ. პივოვაროვას, ნ. კოჩეტკოვას, რ. იანგარბერს.

What's Next? Automatic Morphological and Lexical Prediction in N-Grams[1]

Mikhail Kopotev
University of Helsinki (Finland)
mihail.kopotev@helsinki.fi

Instructors teaching a foreign language are regularly asked how words co-occur: What cases and word forms appear after a given preposition? Which ones should I learn by rote and which ones follow rules? The persistence of such questions indicates that this is an important challenge to be addressed—we should aim to build a system that can automatically generate an integrated answer.

This paper presents an algorithm that allows the user to issue a query pattern, collects multi-word expressions (MWEs) that match the pattern, and then ranks them in a uniform fashion. This is achieved by quantifying the strength of all possible relations between the tokens and their features in the MWEs. The algorithm collects the frequency of morphological categories of the given pattern on a unified scale in order to choose the stable categories and their values. For every part of speech, and for all of its categories, we calculate a normalized Kullback-Leibler divergence between the category's distribution in the pattern and its distribution in the corpus overall. Categories with the largest divergence are considered to be the most significant. The particular values of the categories are sorted according to a frequency ratio. As a result, we obtain morphosyntactic profiles of a given pattern, which includes the most stable category of the pattern, and their values.

მონაცემთა ბაზის სტრუქტურირების საკითხი „მეგრული ტექსტების ელექტრონული კორპუსის“ მიხედვით

მაია ლომია, რუსუდან გერსამია

ივანე ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)
საქართველოს უნივერსიტეტი (საქართველო)
maia-lomia@mail.ru, rgersamia@iliauni.edu.ge

დღეს ძალზე აქტუალურია კომპიუტერული ტექნოლოგიების ჩართვა ბუნებრივი ენების ლინგვისტური დამუშავების პროცესში. სამეცნიერო პროექტის – „მეგრული ტექსტების ელექ-ტრონული კორპუსი“ – ფარგლებში შეიქმნა ორი ნაწილისაგან შემდგარი მონაცემთა ბაზა: 1. **ექ-სპედიციის მონაცემთა ბაზა**, რომელიც მოიცავს 2007–2011 წლებში სამეგრელოს ყველა არეალში

[1] I deeply thank my collaborators L. Pivovarova, N. Kochetkova, R. Yangarber.

მოპოვებულ ჟანრულად და თემატურად მრავალფეროვან მასალას. 147 ინფორმატორის ჩანაწერების ქრონომეტრაჟი 103 საათია. ეს არის თანამედროვე მეგრული მეტყველების ამსახველი დოკუმენტირებული ბაზა, რომელიც განთავსდება ინტერნეტმისამართზე: *http://www.klr.ge* 2. **გლოსირებული მონაცემთა ბაზა** აერთიანებს გამოცემულ მეგრულ ტექსტებს კილოური სისრულის, თემატური და ჟანრობრივი მრავალფეროვნების შენარჩუნებით (მთლიანადაა ალ. ცაგარლისა და ი. ყიფშიძის ტექსტები, შერჩევითაა: მ. ხუბუას ტექსტები, ტ. გუდავას მეგრული პოეზიის ნიმუშები, ქართული ხალხური სიტყვიერების (მეგრული ტექტების) მე–2 ტომიდან (კ. დანელია, ა. ცანავა) – გამოცანები, ანდაზები, შელოცვები.

წინასწარ შერჩეულ მეგრულ ტექსტებს ცალ-ცალკე ენიჭებათ პირობითი კოდი (მაგალითად, ალ. ცაგარლისა – A, ი. ყიფშიძისა – B და ა. შ., რითაც შეიძლება იდენტიფიცირება ტექსტის გამოცემასთან). აბრევიატურა შეიცავს შემდეგ მონაცემებს: ტექსტის გამოცემა, ტექსტისთვის ბაზაში მინიჭებული რიგითი ნომერი, ტექსტის გამომცემლის გვარის ინიციალი და ტექსტის ნომერი გამოცემაში. მაგალითისთვის, ერთ-ერთი აბრევიატურის ჩანაწერი ასეთია: B.5.Q.V. საკუთრივ აბრევიატურის ან აბრევიატურის ქვევით „განმარტების“ სახელწოდებით მოცემული ბმულის გააქტიურებისას გამოდის დამხმარე ფანჯარა და შეიცავს გლოსირებული ტექსტების არაენობრივ მახასიათებლებს; ესენია: *ექსტრალინგვისტური, გეოლინგვისტური, ტიპოლოგიურ–თემატური*. არაენობრივ მახასიათებელთა ნუსხას დაემატა ცნობები იმ გამოცემის/გამოცემების შესახებ, რომელშიც/რომლებშიც ეს ტექსტი თავიდანვე იყო მოცემული ან გადაბეჭდილი.

ტექსტი დაყოფილია წინადადებებად, რომელთაც ენიჭებათ სპეციალური ინდექსი: 1, 2, 3... თუკი წინადადება მოცულობით დიდია, მაშინ იყოფა ნაწილებად (ძირითადად ორად ან სამად, ცალკეულ შემთხვევებში შეიძლება მეტადაც). ნაწილებად დაყოფისას წინადადების ინდექსი არ იცვლება, ნაწილები ზუსტდება სპეციალური სიმბოლოებით: a,b,c… და ინდექსის საბოლოო სახე ასეთია: 1a, 1b, 1c. პროგრამის მენიუში წარმოდგენილია კომპონენტები. თითოეული მათგანი წარმოადგენს ბმულს, რომლის გააქტიურებისას გამოდის ფანჯარა შესაბამისი ინფორმაციით:

- მეგრული ტექსტების სარჩევი
- გლოსირების საერთაშორისო სტანდარტი
- შემოკლებათა განმარტებები (მორფემათა შესაბამისი გლოსები)
- სიმბოლოთა განმარტებები

საკუთრივ მეგრული ტექსტი მუშავდება საანოტაციო ერთეულებად შერჩეული შემდეგი პარამეტრების მიხედვით:

- წინადადების ჩანაწერი
- მორფემებად დაშლილი მიმდევრობები
- მორფემის შესაბამისი გლოსები
- მეტყველების ნაწილების მიხედვით ანოტირებული ჩანაწერი
- ქართული თარგმანი (ანბანური ჩანაწერი)
- ინგლისური თარგმანი

მორფემებად დაშლილი მიმდევრობები ჩაწერილია საერთაშორისო სტანდარტით მიღებული სიმბოლოების მიხედვით. ჩვენს მონაცემთა ბაზაში გამოყენებულია რამდენიმე:

- > < (კონფიქსი)
- < > (ინფიქსი)
- = (კლიტიკა)
- – (მორფემათა საზღვარი)
- . (გრამატიკული კატეგორიის საზღვარი)
- : (ფორმისა და ფუნქციის შეთანადება/შეხამება)
- + (მორფემის/მორფემების მიერთება ფუძესთან)

ზემოხსენებული პროექტის ფარგლებში მონაცემთა ბაზის სტრუქტურა და საანოტაციო პარამეტრები განსაზღვრულია იმგვარად, რომ წარმოდგენილი მასალა ხელმისაწვდომია არა მხოლოდ ქართველი ან ქართულის მცოდნე მკვლევრებისთვის, არამედ ნებისმიერი დარგის მეცნიერისთვის.

The Issue of Database Structuring According to *Electronic Corpus of Megrelian Texts*

Maia Lomia & Rusudan Gersamia
Ivane Javakhishvili Tbilisi State University (Georgia)
University of Georgia (Georgia)
maia-lomia@mail.ru, rgersamia@iliauni.edu.ge

Currently, it is very essential to involve computational technologies in natural language processing. Within the framework of the research project *Electronic Corpus of Megrelian Texts*, a bipartite database was created: 1. **Expedition database**, comprising the genre and thematically diverse data, collected in all the parts of Megrelia in 2007-2011. The length of the recordings of 147 informants is 103 hours. This is a documented base, reflecting the present-day Megrelian speech, to be placed on the following website: *http://www.klr.ge* 2. **Base of glossed data**, comprising published Megrelian texts, maintaining dialectal complexity, thematic and genre diversity (all the texts by A. Tsagareli and I. Qipshidze; selected: texts by M. Khubua, pieces of Megrelian poetry by T. Gudava, excerpts from *Georgian Folklore (Megrelian Texts)*, Vol. 2 (K. Danelia, A. Tsanava) – riddles, proverbs, incantations).

Previously sampled Megrelian texts are given individual codes (for instance, by A. Tsagareli – A, by I. Qipshidze – B, etc., allowing to identify a text with its publication). An abbreviation consists of the following data: text publication, number in the text base, an initial letter of the publisher and text number in publication. For instance, one of the abbreviations is the following: B.5.Q.V. By clicking on either an abbreviation or the title „Definition" under it, an additional window opens, containing non-linguistic properties of glossed texts; they are: *extralinguistic, geolinguistic, typological and thematic*

ones. The list of properties was appended by the information about the publication(s) in which the text(s) initially appeared.

A text is divided into sentences which are given special indices: 1, 2, 3... If a sentence is large, it is sub-divided into parts (mostly, into two or three parts; in individual cases, in more parts). Their indices do not change whenever they are sub-divided; parts are specialized with special symbols: a, b, c...; hence, an index is eventually shaped as: 1a, 1b, 1c. The software menu presents various components. Each of them is a link, after clicking on which a window, containing the appropriate information, opens:

- Contents of Megrelian texts
- International glossing standards
- Clarifications of shortening (glosses for morphemes)
- Clarifications of symbols

A Megrelian text proper is processed in accordance with the following dimensions selected as tags:

- sentence record
- sequences, broken up into morphemes
- glosses for morphemes
- part-of-speech tagging record
- Georgian translation (alphabetic record)
- English translation

The sequences, broken up into morphemes, are recorded in accordance with the available international standards. Our database makes use of the following symbols:

- > < (confix)
- < > (infix)
- = (clitic)
- – (morpheme boundary)
- . (grammatical category boundary)
- : (form and function alignment)
- + (stem taking on morpheme(s))

Within the framework of the aforementioned project, the structure and annotation dimensions of the database have been determined so as to make the presented materials accessible not only for Georgian and Georgian-speaking researchers but also for any scholar interested in it.

თანამედროვე ქართული ენის მორფოლოგიური ანალიზატორი და გენერატორი

ირინა ლობჟანიძე
ილიას სახელმწიფო უნივერსიტეტი, ლინგვისტურ კვლევათა ცენტრი (საქართველო)
irina_lobzhanidze@iliauni.edu.ge

თანამედროვე ქართული ენის მორფოლოგიური ანალიზატორი შეიქმნა სასრული პოზიციის ავტომატების გამოყენებით. სასრული პოზიციის ტექნიკური საშუალებები გამოიყენება სხვადასხვა ენის ფონოლოგიისა და მორფოლოგიის კომპიუტერული აღწერის დროს.

ანალიზატორი შემუშავდა შოთა რუსთაველის ეროვნული სამეცნიერო ფონდის მიერ დაფინანსებული პროექტის (AR/320/4-105/11) ფარგლებში. სისტემა მოიცავს თანამედროვე ქართული ენის მეტყველების ნაწილების მორფოლოგიურ თვისებებს. მორფოტაქტიკა კოდირებულია ლექსიკონების, ხოლო ცვლილებები - რეგულარული გამოსახულებების სახით. სხვადასხვა ტექსტზე გადამოწმებული რესურსი გამოიყენება ტოკენიზაციის, ლემატიზაციისა და ტაგირებისათვის.

თანამედროვე ქართული ენის ანალიზატორი იქმნება სასრული პოზიციის ტექნიკური საშუალებების გათვალისწინებით (Beesley K.R., Kartunnen L. 2003, Koskenniemi, K. 1983 და ა.შ.), xfst-სა და lexc-ის გამოყენებით.

სტატიაში წარმოდგენილია ანალიზატორის სტრუქტურა, სამუშაო მოდული და განხილულია მისი შემდგომი განვითარების საფეხურები.

Morphological Analyzer and Generator of Modern Georgian Language

Irina Lobzhanidze
Ilia State University (Georgia)
irina_lobzhanidze@iliauni.edu.ge

The Morphological Analyzer of Modern Georgian Language was developed using finite-state automata. Finite state techniques have been applied successfully in computational phonology and morphology of the world's major and minor languages. The Analyzer was developed within the framework of the project (AR/320/4-105/11) financed by the Shota Rustaveli National Science Foundation. The system encodes the morphology of all inflected parts-of-speech of Modern Georgian. The morphotactics is encoded in the lexicons and word mutations are encoded in regular expressions. This resource evaluated against different texts is used for tokenizing, lemmatising and tagging.

Following approaches of finite state techniques (Beesley K.R., Kartunnen L. 2003, Koskenniemi, K. 1983, etc.), a morphological analyzer of Modern Georgian has been created using xfst and lexc tools.

We present the structure and working module of the analyzer and describe the further stages of its development.

პროგრამული პლატფორმა ელექტრონული ქართული ორენოვანი ლექსიკონებისათვის

თინათინ მარგალიტაძე, გიორგი ქერეჭაშვილი
ივანე ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)
tinatin@margaliti.ge; hello@dictionary.ge

თსუ ლექსიკოგრაფიულ ცენტრში შემუშავდა პროგრამული პლატფორმა დიდი ინგლისურ-ქართული ონლაინლექსიკონისათვის და უფრო მცირე მოცულობის ინგლისურ-ქართული დარგობრივი ლექსიკონებისათვის.

2012 წელს კი ლექსიკოგრაფიულმა ცენტრმა დაიწყო მუშაობა ე.წ. სამომხმარებლო აპლიკაციაზე და ამ ეტაპისათვის დასრულებულია მუშაობა პროგრამის პირველ ვერსიაზე.

დღეისათვის ელექტრონული და ონლაინლექსიკონები აკმაყოფილებენ შემდეგ მოთხოვნებს:

- ლექსიკონებში ძიება ხორციელდება როგორც საკვანძო სიტყვის, ასევე შესიტყვებების მიხედვით;
- ლექსიკონებს აქვს ორი საძიებო სისტემა: ძიება სიტყვა-სტატიის მიხედვით და ძიება ლექსიკონის მთელ ტექსტში;
- მომხმარებელს საშუალება აქვს გაფილტროს მონაცემები დარგების მიხედვით;
- ლექსიკონს აქვს შიდა ბმულები;
- ლექსიკონებს აქვს ამოსაბეჭდი ვერსიის ტექსტის ფორმატირების ფუნქცია;
- სიტყვის ძიება არ არის დამოკიდებული მის გრამატიკულ ფორმასა და საძიებო ველში განლაგებაზე. სიტყვის ძიება ხდება საძიებო ველში მისი არასრული / ნაწილობრივი შეყვანის შემთხვევაშიც, ხოლო სიტყვის არასწორად აკრეფის შემთხვევაში პროგრამა მკითხველს სთავაზობს სავარაუდო სწორ ვარიანტებს;
- ლექსიკონების საშუალებით შესაძლებელია ქართული ტერმინების მოძიებაც, რაც მათ, ნაწილობრივ, ქართულ-ინგლისური ლექსიკონის ფუნქციასაც ანიჭებს;
- ელექტრონული ლექსიკონი ინსტალაციის შემდეგ მუშაობს კომპაქტდისკისა და ინტერნეტის გარეშე;
- ელექტრონულ ლექსიკონს მალე ექნება ინტერნეტიდან განახლებული ვერსიის გადმოტვირთვის ფუნქცია.

ონლაინლექსიკონს სამომხმარებლო მხარესთან ერთად აქვს მართვის პანელი მრავალი დანიშნულებით. ის მოიცავს:

- ლექსიკონის სიტყვანის დათვალიერების, რედაქტირების, ახალი სიტყვების დამატების ფუნქციებს;
- მომხმარებლების მიერ განხორციელებული ძიების ლოგებს;
- რედაქტორისთვის საჭირო ინსტრუმენტებს – გენერატორებსა და კონვერტორებს.

Software Platform for Electronic Georgian Bilingual Dictionaries

Tinatin Margalitadze & Giorgi Keretchashvili
Ivane Javakhishvili Tbilisi State University (Georgia)
tinatin@margaliti.ge, hello@dictionary.ge

Lexicographic Centre at TSU has developed the software platform for Comprehensive English-Georgian Online Dictionary and for smaller English-Georgian specialized dictionaries.

In 2012, Lexicographic Centre started working on the so-called desktop application of its online dictionaries and the first version of the program is being implemented and tested.

At present, Electronic and Online Dictionaries meet the following requirements:

- Search throughout the Dictionaries is possible both by key words and by word-combinations;
- The Dictionaries have two search systems: search by entries (exact match), and full text search;
- Users are able to filter out data by specific subjects;
- The Dictionaries have inner links /cross references;
- The Dictionaries have the function of formatting the text of printable version;
- Word search does not depend on a grammatical form of the word or on its position inside the search field. A word is searched also when typed incompletely / partially. In case of incorrect typing, the software suggests probable correct versions;
- The Dictionaries enable the search for Georgian words as well, thus imparting to them certain features of a Georgian-English dictionary;
- After installation, the Electronic Dictionary functions without the support of CD-ROM and the Internet; and
- Soon the Electronic Dictionary will have the function of downloading updated versions of dictionaries from the Internet.

The Control panel of the Dictionary has the following functionalities:

- Dictionary vocabulary management, including the functions of viewing and editing the dictionary vocabulary, as well as the functions of adding new entries;
- Generation and conversion tools necessary for editors;
- Logs of searches made by users.

სამეცნიერო ტექსტების ინგლისურ-ქართული პარალელური კორპუსი (არიანე ჭანტურიას თარგმანების ბაზაზე)

თინათინ მარგალიტაძე, გიორგი ქერეჭაშვილი
ივანე ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)
tinatin@margaliti.ge; hello@dictionary.ge

მოხსენებაში წარმოდგენილი იქნება თსუ ლექსიკოგრაფიის ცენტრის ახალი პროექტი, რომელზეც ერთ წელზე მეტია მიმდინარეობს მუშაობა. პარალელური კორპუსების შექმნის ერთერთი უმნიშვნელოვანესი მიზანი ლექსიკოგრაფიაა, მათი გამოყენებაა ორენოვანი ლექსიკონების შესაქმნელად. თსუ ლექსიკოგრაფიის ცენტრის მთავარი პროექტი – დიდი ინგლისურ-ქართული ლექსიკონი ასეთი პარალელური კორპუსის გარეშე შეიქმნა და, ძირითადად, ინგლისური ენის განმარტებით ლექსიკონებსა და ინგლისურ-რუსულ ლექსიკონებს იყენებდა წყაროებად. ეს გამოწვეული იყო არა მხოლოდ ადამიანური რესურსის ნაკლებობით, არამედ თვით თარგმნილი ლიტერატურის ხარისხითაც. ყველასათვის კარგადაა ცნობილი, რომ საბჭოთა პერიოდში მხატვრული ლიტერატურა იშვიათად ითარგმნებოდა დედნიდან, ქართული თარგმანები ინგლისური ლიტერატურის რუსული თარგმანებიდან სრულდებოდა. ამასთანავე, ქართველი მთარგმნელების უმრავლესობა დედნის ტექსტს ხშირად საკმაოდ თავისუფლად უდგებოდა და ასევე თარგმნიდა. ორიგინალიდან შესრულებული ადეკვატური თარგმანების რაოდენობა კი არ იყო საკმარისი ვრცელი კორპუსების შესაქმნელად და ლექსიკონის შედგენაში მათ გამოსაყენებლად.

მას შემდეგ, რაც დასრულდა მუშაობა ინგლისურ-ქართულ ლექსიკონზე, გადაწყდა პარალელური კორპუსის შექმნა, რომელსაც ლექსიკოგრაფიის ცენტრი გამოიყენებს ერთი მხრივ, დიდი ინგლისურ-ქართული ლექსიკონის მეორე რედაქციისათვის, მეორე მხრივ, კი ცენტრის ორენოვანი დარგობრივი პროექტებისათვის. პარალელური კორპუსის პროგრამა და მეთოდოლოგია ასევე გამოყენებული იქნება ცენტრის მიერ ინიცირებულ სხვა ევროპულ-ქართულ ორენოვან პროექტებში.

ქართული მთარგმნელობითი საქმიანობის ფონზე სრულიად გამოირჩევა ცნობილი ქართველი მთარგმნელის, რედაქტორისა და ლექსიკოგრაფის, ენციკლოპედიური განათლების მქონე მეცნიერის არიანე ჭანტურიას მთარგმნელობითი საქმიანობა, რომელიც უკვე ათწლეულებს ითვლის. მისი თარგმანები მოიცავს მეცნიერების პრაქტიკულად ყველა დარგს და ქართული ენიდან ინგლისურ ენაზე თარგმნილი ლიტერატურის ყველაზე ავთენტურ, სარწმუნო პარალელურ კორპუსს ქმნის.

სამეცნიერო ტექსტების ინგლისურ-ქართული პარალელური კორპუსის, „არიანე ჭანტურიას კორპუსის“, მშენებლობა დავიწყეთ სამეცნიერო სტატიების რეზიუმეების პარალელური ქართულ-ინგლისური ტექსტებით. აღნიშნული ტექსტების ერთი ნაწილი თავმოყრილია საქართველოს მეცნიერებათა აკადემიის „მოამბის“ ნომრებში. არიანე ჭანტურია 1967 წლიდან 16 წელი თანამშრომლობდა საქართველოს მეცნიერებათა აკადემიასთან და ინგლისურად თარგმნიდა ან რედაქტირებას უკეთებდა სტატიებზე დართულ ინგლისურ რეზიუმეებს. იმ პერიოდში „მოამბე“ თვეში ერთხელ გამოდიოდა. შესაბამისად, ჩვენს ხელთ არის სამეცნიერო ჟურნალის 192 ტომი

(1967 - 1983), 8832 სტატია, რომლებიც მეცნიერების თითქმის ყველა დარგს მოიცავს. საწყის ეტაპზე, როდესაც ინგლისურ-ქართული პარალელური კორპუსის მშენებლობისათვის დიდი გამოცდილება არ არსებობს, რეზიუმეები საინტერესოა მრავალი თვალსაზრისით: ტექსტი მოკლეა, რაც მაქსიმალურად გამორიცხავს შეცდომას, ტექსტის სტილი მკაცრია, რაც კორპუსში გადვილებს პარალელური წინადადებების შეთანადებას, რეზიუმეს ტექსტი გაჯერებულია სამეცნიერო სტატიაში გამოყენებული ტერმინოლოგიით.

მომდევნო ეტაპზე, როდესაც დაგროვდება სათანადო გამოცდილება, კორპუსი განივრცობა არიანე ჭანტურიას მიერ თარგმნილი წიგნების პარალელური ინგლისურ-ქართული ტექსტებით.

პარალელური კორპუსი თანამედროვე სტანდარტებისა და მოთხოვნილებების შესაბამისად შემუშავებული ვებაპლიკაციაა, რომლის მოხმარებისთვის სრულიად საკმარისია ინტერნეტი და სტანდარტული ვებბრაუზერი. კორპუსის ძრავი დაწერილია PHP პროგრამულ ენაზე, ხოლო კორპუსის ტექსტური და სისტემური ბაზები განთავსებულია MySQL მონაცემთა ბაზაში. ინტერფეისებში აგრეთვე გამოყენებულია მცირეოდენი JavaScript.

ვებაპლიკაცია მოიცავს მომხმარებლისა და ადმინისტრირების ფუნქციებსა და ინტერფეისებს, რაც კორპუსის მოხმარების, მომსახურებისა და ადმინისტრირების უნიკალურ ინტეგრირებულ და დინამიკურ რესურსს ქმნის.

დღეისათვის დასრულებულია მუშაობა კორპუსის მართვის პანელზე, რომელიც მოიცავს შემდეგ ძირითად ფუნქციებს:

- კორპუსების ჯგუფების შექმნასა და რედაქტირებას;
- კრებულების შექმნასა და რედაქტირებას;
- ტექსტური წყვილების დამატებასა და რედაქტირებას;
- ტექსტის ავტომატურ დანაწევრებას, გათანაბრებასა და ტექსტური წყვილების გენერირებას შემდგომი რედაქტირების შესაძლებლობით.

მართვის პანელში გათვალისწინებულია ისეთი დამხმარე ბაზების დამატება/რედაქტირების შესაძლებლობა, როგორებიცაა: ნაწარმოების ტიპი, ჟანრი, ავტორი, მთარგმნელი და ა.შ. ასევე ხელმისაწვდომია ავტომატური გენერატორები და სხვადასხვა ინსტრუმენტი, რაც ამცირებს მასალის დამატებისა და დამუშავების სისწრაფესა და შემდგომ რედაქტირებას.

დამუშავების პროცესშია მომხმარებლის ინტერფეისი, რომლის დანიშნულებაა კორპუსის მონაცემთა ბაზაში სასურველი ქართული ან ინგლისური სიტყვის ან შესიტყვების ძიება და შესაბამისი შედეგების შემცველი პარალელური ტექსტური წყვილების გამოტანა. ძიება განხორციელდება ორმხრივად - როგორც ორიგინალში, ასევე ნათარგმნ ტექსტში. შესაძლებელი იქნება ძიება როგორც მთლიან მასალაში, ასევე სხვადასხვა პარამეტრის მიხედვით, მათ შორის ნაწარმოების ჯგუფების, კრებულების, ტიპების, ჟანრების, ავტორების, წლებისა და ა.შ. ფილტრაციით. ძიების შედეგებში ასევე ხელმისაწვდომი იქნება ინფორმაცია ნაწარმოების შესახებ, მათ შორის ავტორის, თარგმანის ავტორის, გამოცემის წელისა და ა.შ. შესახებ.

English-Georgian Parallel Corpus of Scientific Texts
(Based on the Translations by Arrian Tchanturia)

Tinatin Margalitadze & Giorgi Keretchashvili
Ivane Javakhishvili Tbilisi State University (Georgia)
tinatin@margaliti.ge, hello@dictionary.ge

The paper presents a new project of Lexicographic Centre at TSU, the work on which has been underway for more than one year. One of the aims of drawing up parallel corpora is their application to bilingual lexicography. The major project of Lexicographic Centre – Comprehensive English-Georgian Dictionary was compiled without such a parallel corpus and was largely based on monolingual dictionaries of the English language, as well as English-Russian dictionaries. This was conditioned not only by the lack of human resources but also by the quality of translations. It is a well-known fact that, during the Soviet period, European authors were rarely translated into Georgian from the original; instead, Russian translations from English and other European languages were used as a source. At the same time, the majority of Georgian translators treated original texts rather freely. The number of adequate translations, executed from the original, was not large enough for creating a parallel English-Georgian corpus and for its application in dictionary-making.

After the completion of the Comprehensive English-Georgian Dictionary, the decision was made to commence the work on an English-Georgian parallel corpus. The corpus will be used by Lexicographic Centre in the second edition of the Comprehensive English-Georgian Dictionary, as well as in bilingual specialized dictionary projects. The software and methodology, developed in the course of the work on the corpus, will be applied in other European-Georgian bilingual projects, initiated by the Centre.

Against the backdrop of Georgian translational activities, a famous Georgian translator, lexicographer and scholar, Dr. Arrian Tchanturia clearly stands out as a highly qualified translator. His translational work of many decades covers practically all fields of knowledge and constitutes the most adequate, authentic and reliable parallel corpus of texts translated from English into Georgian.

The construction of the English-Georgian parallel corpus of scientific texts, „Arrian Tchanturia's Corpus" has started by parallel, English-Georgian theses of scientific articles. Part of these texts are accumulated in the issues of *Bulletin of the Academy of Sciences of Georgia*, where Dr. Arrian Tchanturia was an editor of English abstracts from 1967 till 1983. 8 832 articles have been selected from 192 volumes of *Bulletin* (1967 – 1983), covering practically all fields of science. At the initial stage, when there is not much experience of the work on a parallel corpus, abstracts are interesting from many points of view: texts are brief, maximally excluding the probability of an error, the style of an abstract is laconic, which will ease the task of aligning parallel sentences in a corpus, the text of abstracts is full of scientific terminology, used in articles.

At the later stage of the work on the Corpus, it will include larger texts and books translated by Dr. Tchanturia.

The Parallel Corpus is a modern web based application, easily accessible through standard web browsers. The corpus engine is built on PHP/MySQL technologies and uses some JavaScript for front-end interfaces.

This web application combines user and administrative interfaces and functionalities, thus making it an integrated and dynamic resource for usage and managing operations.

Control Panel of the Parallel Corpus includes the following functions:

- Management functionalities of text groups
- Management functionalities of text sets
- Management functionalities of text couples
- Raw text submission and processing functionalities, that include automatic break down of texts by sentences, sentence alignment, generation of couples and further manual alignment options.

Parallel Corpus Control Panel includes options for managing complementary data essential to Corpus, such as: text types, genres, authors, translators etc. Various automating tools are also available, that contribute to saving time while formatting, submitting and data management.

Parallel Corpus User Interface, which is currently under development, will hold Parallel Corpus database search functions against Georgian and/or English words/word combinations and display matching text couples. Search will be bidirectional and conducted on both original and translated material data. It will be possible to filter data by groups, sets, types, genres, authors, years etc. Along with matched text couples, search results will hold additional information about displayed material, such as title, author, translation author, publishing date etc.

ქართული ჟესტური ენის ელექტრონული ლექსიკონი

თამარ მახარობლიძე
თსუ არნ. ჩიქობავას სახ. ენათმეცნიერების ინსტიტუტი (საქართველო)
ateni777@yahoo.com

ყრუები საქართველოში წარმოადგენენ ლინგვისტურ უმცირესობას, რომელთა უფლებების დაცვას სახელმწიფო ცდილობს შესაბამისი კანონებით. იმისთვის, რომ სწავლება ქართულ ჟესტურ ენაზე რეალურად შესაძლებელი გახდეს, საჭიროა ამ ენის გამოკვლევა და შესაბამისი სამეცნიერო სამუშაოების ჩატარება, კერძოდ, ლექსიკონების შექმნა, გრამატიკული კატეგორიების დეტალური ანალიზი, ენის ნორმების დადგენა, მეთოდიკის შემუშავება, ენის დონეების განსაზღვრა, და ა.შ. მსოფლიოს წამყვან უნივერსიტეტებში წარმატებით მიმდინარეობს ჟესტური ენების კვლევები. საქართველოში კი ამ მიმართულებით ახლა იდ-

გმება პირველი ნაბიჯები. საბჭოთა პერიოდში ქართული ჟესტური ენა განიცდიდა რუსულის გავლენას, არ არსებობდა არანაირი სახელმძღვანელო ან სხვა ტიპის წიგნი საკუთრივ ქართული ჟესტური ენის შესახებ. რუსული ენობრივი დომინანტის კვალი ნათლად ჩანს როგორც ძველ ანბანში, ასევე ლექსიკაშიც. ამჟამად მიმდინარეობს ჟესტურ ენათა ნაციონალიზაცია და დაწყებულია რეინტეგრაციის პროცესი. ლექსიკისაგან განსახვავებით, ჟესტური ენის გრამატიკული დონეები შედარებით თავისუფალია რუსულის ზეგავლენისა და ნასესხობებისაგან. თუმცა მიმდინარე ეტაპზე ლექსიკაში მომხდარი ცვლილებები ყველაზე უკეთ ასახავს ნაციონალიზაციის პროცესს.

ყრუთა თემების წევრები კრეატიული ბუნების ხალხია და ხშირად ისინი თავად ქმნიან ჟესტებს. ჟესტური ენის ცოდნა ყრუთათვის აუცილებელია მათივე საზოგადოებაში სრულფასოვანი წევრობისთვის. ყრუთა თემის წევრებისათვის დამახასიათებელია ბილინგვიზმი – ჟესტური და სამეტყველო ენების ცოდნა. ქართული ჟესტური ენის ელექტრონული ლექსიკონის შექმნა არის ამ თემისთვის სასიცოცხლო მნიშვნელობის საკითხი, რადგანაც ყრუთა და სმენადაქვეითებულ ადამიანთა ყველა პრობლემა კომუნიკაციას უკავშირდება.

ქართული ჟესტური ენის ლექსიკონში სათანადო ლინგვისტური ანალიზით წარმოდგენილი იქნება 4000 ერთეული გრაფიკული სახითა და ვიდეოფორმატით. ონლაინვიდეოფაილებით და ვიდეორგოლებით უკეთ აღიწერება კინეტიკური ენობრივი ერთეულები, რაც გაუადვილებს დაინტერესებულ პირებს ამა თუ იმ ჟესტის სწორად გაგებას. ამ ტიპის ლექსიკონისათვის და ჟესტური ენის დოკუმენტირებისათვის ელექტრონულ ფორმატს უკეთესი ალტერნატივა არ გააჩნია.

გამოსაკვლევია ქართული ჟესტური ენის ლექსიკური დონის დიდი ნაწილი. სამუშაო პროცესი მოიცავს პარალელურ რეჟიმში მიმდინარე ორ ფაზას: ენის წყაროებთან მუშაობასა და გამოკვლეული ლექსიკური ერთეულების ვიზუალიზაციას – ფოტოსურათისა და გრაფიკული მხატვრობის შექმნას, ვიდეოფორმატირებასა და მასალის კომპიუტერიზაციას. ელექტრონული ლექსიკონის პროგრამული ნაწილი უზრუნველყოფს მასალის სისტემურ წვდომას. ენობრივი ანალიზი გულისხმობს ჟესტის დესკრიფციულ ანალიზსა და ტიპოლოგიზაციას, ანუ სხვა ჟესტურ ენებთან მიმართებით მასალის გადააზრებას.

ამ ტიპის ელექტრონული პროდუქტი ყრუ ადამიანებს დაეხმარება საკომუნიკაციო პრობლემების გადაჭრაში, მათი სოციალური და ეკონომიკური მდგომარეობის გაუმჯობესებაში და დიდ როლს შეასრულებს სამოქალაქო საზოგადოებაში მათი ინტეგრაციისათვის.

Electronic Dictionary of the Georgian Sign Language

Tamar Makharoblidze
Arn. Chikobava Institute of Linguistics, TSU (Georgia)
ateni777@yahoo.com

Deaf and hard of hearing people are a linguistic minority in Georgia, whose biggest problem is the lack of communication. The state tries to solve their problems by means of proper laws. In order to overcome the problem of communication for deaf and hard of hearing people, it is necessary to provide the scientific investigations of the Georgian sign language, studying its grammar, compiling dictionaries, establishing the language standards, defining the language levels, working on the methodology of teaching of the Georgian sign language, etc. Many leading Universities worldwide are successfully working on sign languages whereas, in this country, we are just making the first steps. In the Soviet period, sign languages were under the influence of Russian. This influence is easily detectable in the earlier Georgian dactyl alphabet, having been totally based on the Russian one. The process of nationalization began everywhere throughout the post-Soviet space, and sign languages are reintegrating. Although the grammar of the Georgian sign language was free from the Russian elements, unlike the lexical level, the process of reintegration is better reflected on the lexical level of the language.

Usually, deaf people are very creative and they frequently invent new signs, and their family members use those individual signs not knowing the existing sign language in their country. Such people cannot communicate with other members of the deaf community, and they stay isolated because of the communication problem. That is why it is very important for deaf and hard of hearing people to know the sign language and to be full-fledged members of their own community. Usually these people are bilingual using both languages -- spoken and sign. Thus, taking into consideration the aforementioned, an electronic dictionary of the Georgian sign language is of a vital interest for these people as far as all of their problems are associated only with communication.

In the electronic dictionary of the Georgian sign language, about 4000 lexical items will be presented in the alphabetical order, based on spoken Georgian. The website with graphics, photos and video-files for the each sign will be easily accessible for interested individuals. For this type of a product and for the documentation of sign languages, the electronic format does not have a better alternative.

The most of the lexical base of the Georgian sign language should be investigated. The working process has two parallel phases: working with various language sources and visualization of the lexical items – creating photos, graphics, video-formatting the material and digitalization of the results. The proper computer software will provide a systemic access to the dictionary. The linguistic method is combined. On the one hand, we will apply descriptive methods and, on the other, we will take into account the existing typological parameters for the description of the lexical items.

The electronic dictionary of the Georgian sign language will help deaf and hard of hearing people to resolve the communication problems, thus improving their social and economical conditions and integrating them in local civil society.

-ყე ელემენტი ქართული ენის დიალექტებში *ქდკ*-ს მონაცემების მიხედვით

ელენე ნაპირელი
თსუ არნ. ჩიქობავას სახ. ენათმეცნიერების ინსტიტუტი (საქართველო)
elene.napireli@gmail.com

ქართული დიალექტური კორპუსის მორფოლოგიური მონიშვნის პროცესში მნიშვნელოვანია ფორმაუცვლელი სიტყვების, მორფემებისა და მოდალური ელემენტების მონიშვნისა და ომონიმური ფორმების გარჩევის საკითხი.

ამ თვალსაზრისით რამდენიმე ძირითადი პრობლემა იკვეთება: ფონეტიკური ვარიანტების იდენტიფიცირება; ომონიმური ფორმების გარჩევა; მრავალფუნქციური ელემენტების იდენტიფიკაცია.

ჩვენ შევისწავლეთ **-ყე** ნაწილაკისა და თანდებულის და მათი ფონეტიკური ვარიანტების (**-ყენ**, **-კენ**, **-ყ**... **-კე**, **-წყენ**, **-ყენა**, **-ყი**...) რეპრეზენტაცია ქართული დიალექტების კორპუსში.

როგორც ცნობილია, **-ყე** მრავალფუნქციური ნაწილაკია და საკმაოდ პროდუქტიულია ქართული დიალექტების ერთ ნაწილში.

-ყე სუფიქს-მორფემოიდი თავს იჩენს ქართული ენის წერილობით ძეგლებში XI საუკუნიდან: „ნუ გეშინინ-ყე, ნუცა სძრწით“, „უკუეთუ გიყუარს-ყე უფალი, ღმერთი თქუენი“, „ესმა-ყე, ვითარმედ მახლობლობით არიან იგინი“... თჳთ ადრითგან მამაობისა ეწერა-ყე საებისკოპოსოდ“ (დავით აღმაშენებლის ანდერძი შიო მღვიმის მონასტრისადმი), „და ჰქონდა-ყე ხუაშიადი მოაბადისი დიდებულთა და იტყოდეს...“ („ვისრამიანი“)... თუმცა ახალ სალიტერატურო ქართულში ასეთი ფორმები არ დამკვიდრებულა და მხოლოდ დიალექტებში შემოგვრჩა.

საენათმეცნიერო ლიტერატურაში (არნ. ჩიქობავა; ი. გიგინეიშვილი, ვ. თოფურია, ი. ქავთარაძე; ბ. ჯორბენაძე, მ. კობაიძე, მ. ბერიძე; კ. ტუიტი; კ. კუბლაშვილი; ნ. სურმავა...) გამოთქმული შეხედულებების თანახმად, **-ყე** ელემენტისთვის გამოყოფილია შემდეგი ფუნქციები:

1. თანდებულის, რომელიც დაერთვის მიცემით და ნათესაობით ბრუნვის ფორმებს და გამოხატავს გამოსვლითობას (დაწყებითობას); მიმართულებას ვინმესკენ ან რაიმესკენ; აღნიშნულია, რომ ფერეიდნულში მას კიდევ უფრო აქვს გაფართოებული ფუნქცია და ზოგ შემთხვევაში (ნაცვალსახელებთან, ზმნიზედებთან, სახელებთანაც) შეიძლება შეგვხვდეს **-ზე**, **-თვის**, **-გან**, **შესახებ**... თანდებულების პოზიციაში. მათ გვერდით ჩვეულებრივია მაგ. -ზე თანდებულიანი სახელები (დ. ჩხუბიანიშვილი).

2. ნაწილაკის, რომელიც დაერთვის ზმნას და გამოხატავს: მიცემითში დასმული (ირიბი) ობიექტის მრავლობითობას; მიცემითში დასმული რეალური სუბიექტის მრავლობითობას (ინვერსიულ ფორმებში); სახელობითში დასმული ობიექტის მრავლობითობას; სახელობითსა და მოთხრობითში დასმული სუბიექტის მრავლობითობას;

3. სუფიქს-მორფემოიდის, რომელიც დაერთვის ზმნას და გამოხატავს მოქმედების მრავალგზისობას (ცვლის „ხოლმე“ ნაწილაკ-მორფემოიდს).

აღსანიშნავია, რომ ქართული დიალექტური კორპუსის მიხედვით, აღნიშნული ელემენტის გეოგრაფიული გავრცელების არე კიდევ უფრო ფართოვდება, ვიდრე ეს სამეცნიერო ლიტერატურაშია მითითებული.

კორპუსის მორფოლოგიური მონიშვნა, ბუნებრივია, აუცილებელად გულისხმობს -**ყე** ელემენტისა და მისი ფონეტიკური ვარიანტებისა და ფუნქციური დიფერენციაციის იდენტიფიცირებას.

გამოქვეყნებულ დიალექტურ ტექსტებში -ყე ხან წინამავალ სიტყვასთან ერთად იწერება, ხან დეფისით და ხან ცალკე. ქდკ-ში ტექსტური მასივის უნიფიკაციის ეტაპზე -**ყე** ყოველთვის სიტყვასთან ერთადაა გადმოცემული, რათა კონკორდანსში მისი ყველა რეალიზაცია ერთად იყოს დაფიქსირებული.

მოხსენებაში წარმოდგენილია ქართული ენის დიალექტებში -**ყე** ელემენტის ფუნქციისა და დისტრიბუციის შესახებ სამეცნიერო ლიტერატურაში დამოწმებული ინფორმაციის ვერიფიკაციის ცდა კორპუსის მასალის მიხედვით. ყურადღება გამახვილდება ისეთ საკითხებზე, როგორებიცაა:

1. -**ყე** ნაწილაკისა და -**ყე** თანდებულის ფონეტიკური ვარიანტები
2. -**ყე** ნაწილაკისა და -**ყე** თანდებულის გავრცელების გეოგრაფია
3. -**ყე** ნაწილაკის ფუნქციური იდენტიფიკაცია
4. -**ყე** თანდებულის ფუნქციური იდენტიფიკაცია

წარმოდგენილი შედეგები გამოყენებული იქნება ქართული დიალექტური კორპუსის მორფოლოგიური ანალიზის სრულყოფისათვის.

The element *-q'e* in Georgian dialects based on the Georgian Dialect Corpus

Elene Napireli
Arn. Chikobava Institute of Linguistics, TSU (Georgia)
elene.napireli@gmail.com

In the process of the markup of the Georgian Dialect Corpus (GDC), the issues of tagging and distinguishing of homonymous forms of uninflected words, morphemes and modals have been significant.

In terms of the above said, several principal problems have been outlined: identification of phonetic variations; distinction of homonyms; identification of a multifunctional element.

We investigated the representation of the particle and postposition ***-q'e*** particle and postposition and their phonetic variations (-**q'en**, -**k'en**, - **q'**…-**k'e**, -**wq'en**, -**q'ena,** -**q'i**…) in the Georgian Dialect Corpus.

As is known, the ***-q'e*** is a multifunctional particle and is quite productive in some Georgian dialects.

The suffix-morphemoid ***–q'e*** occurs in Old Georgian written sources beginning from the 11th century: „nu gešinin-q'e, nuca sjrc'it", „uk'uetu giq'uars-q'e upali, ğmerti tkueni", „esma-q'e, vitarmed maxloblobit arian igini", „tuit adritgan mamaobisa ec'era-q'e saebisk'oposod" (*The Will of King David the Builder to Shio Mgvime Monastery*), „hkonda-q'e xuašiadi moabadisi didebulta da it'q'odes" (*Visramiani*). However, such forms did not gain currency in Modern Standard Georgian and occur only in dialects.

Based on the linguistic scholarly literature (Arn. Chikobava, I. Gigineishvili, V. Topuria, I. Kavtaradze, B. Jorbenadze, M. Kobaidze, M. Beridze, K. Kublashvili, K. Tuite, N. Surmava…), the following functions have been identified for the particle ***-q'e***:

1. It is taken on by dative and genitive case forms and expresses direction towards someone or something;
2. It is taken on by a verb and expresses: plural of a dative (indirect) object; plural of a dative real subject (inversive forms); plural of a nominative object; plural of nominative and ergative subjects;
3. The suffix-morphemoid, taken on by a verb and expressing repeatedness of an action (substituting for the suffix-morphemoid **xolme**);
4. In Fereidanian, it has extended its function and, in some cases (with pronouns, adverbs, and substantives as well) may occur in the positions of the postpositions **-ze, -t'vis, -gan, šesaxeb**... For instance, **-ze**-suffixated substantives are normal with them (D. Chkhubianishvili).

It should be noted that, according to the Georgian Dialect Corpus, the geographical distribution of the element is much wider.

The morphological annotation of the corpus necessarily implies identification of the element -***q'e*** and its phonetic variations and of their functional differentiation.

In published dialect texts, the ***-q'e*** is sometimes spelt solidly with a preceding word, sometimes with a hyphens, and sometimes separately. In GDC, at the stage of the unification of the text body, the -***q'e*** is always spelt solidly with a word in order to make all of its occurrenced be documented together.

The paper is an attempt of a corpus-based verification of the information about the function and distribution of the element ***-q'e***, provided in scholarly literature. The following issues will be focused:

1. Phonetical variations of the particle ***-q'e*** and the postposition ***-q'e***
2. Geographic distribution of the particle ***-q'e*** and the postposition ***-q'e***
3. Functional identification of the particle ***-q'e***
4. Functional identification of the postposition ***-q'e***

The results will be used for the improvement of the morphological analyses of the Georgian Dialectal Corpus.

ციფრული ბაზებისა და კორპუსული ლინგვისტიკის მეთოდიკის გამოყენება ტექსტოლოგიასა და ლიტმცოდნეობაში

მაია ნინიძე
თსუ შოთა რუსთაველის ქართული ლიტერატურის ინსტიტუტი (საქართველო)
maianinidze@yahoo.com

კომპიუტერული ტექნიკისა და ტექნოლოგიების სწრაფმა განვითარებამ ჰუმანიტარული კვლევები ახალი გამოწვევების წინაშე დააყენა. შესაძლებლობების გაფართოებამ მოითხოვა კვლევის მასშტაბებისა და სიღრმის ზრდაც. პრაქტიკული ტექსტოლოგიის უმნიშვნელოვანესი პროდუქტი - თხზულებათა აკადემიური გამოცემა, რომელიც თავს უყრის კლასიკოსთა ნაწერებს და აღჭურავს მას მრავალმხრივი სამეცნიერო აპარატით, კიდევ უფრო ინფორმაციული და ამომწურავი თუ არ გახდა, ვერ დააკმაყოფილებს თანამედროვე მკითხველის გაზრდილ მოთხოვნებს. ამის მიღწევა კი შეუძლებელია უახლესი ტექნოლოგიების მაქსიმალური გამოყენების გარეშე.

2007 წელს ილია ჭავჭავაძის თხზულებათა აკადემიური გამოცემის (ოცტომეული) მე-15 ტომის მზადებისას ანონიმური სტატიების ატრიბუციის მიზნით ტექსტოლოგიის ცენტრში დავწერეთ ციფრული ტექსტური კორპუსების შედარებითი ანალიზი. რამდენადაც წერილები დაბეჭდილი იყო გაზეთ „დროების“ სარედაქციო გვერდზე ავტორის მითითების გარეშე, იმ დროს კი ამ პერიოდულ გამოცემას ორი რედაქტორი ჰყავდა - სერგეი მესხი და ილია ჭავჭავაძე, შევქმენით ამ ორი ავტორის პუბლიცისტური წერილების ციფრული ბანკი და ჩავატარეთ შედარებითი კვლევა, რის შედეგადაც გამოიკვეთა, რომ პრეფიქს-სუფიქსების, თანდებულების, ზმნიზედებისა და სხვა დამხმარე სიტყვების სპეციფიკური გამოყენებით, ცალკეულ ლექსიკურ ერთეულთა სიხშირითა და ენობრივ-სტილური მახასიათებლებით ანონიმური ტექსტები ბევრად უფრო ახლოს იყო ილია ჭავჭავაძის სტატიებთან, ვიდრე სერგეი მესხისასთან.

ტექსტის ატრიბუცია მრავალმხრივ, კომპლექსურ კვლევასა და არგუმენტირებას მოითხოვს და ლინგვისტური კორპუსების ამგვარი შედარებითი ანალიზი საბოლოო დასკვნის გამოსატანად საკმარისი არ არის, მაგრამ, უნდა ითქვას, რომ იგი კარგი დამხმარე საშუალებაა. ანონიმური წერილების რაოდენობა როგორც XIX საუკუნის, ისე შემდგომი პერიოდის ქართულ პრესაში საკმაოდ დიდია და სხვადასხვა მწერლის თხზულებების ციფრული კორპუსების შექმნა, ვფიქრობ, ხელს შეუწყობს ამ ტექსტების ავტორთა დადგენასაც.

ატრიბუციული კვლევებისათვის განსაკუთრებული მნიშვნელობა აქვს ყველა ცნობილი ავტორის ხელნაწერი ნიმუშების სკანირებული ციფრული ბაზების შექმნას. ეს დახმარებას გაგვიწევდა ხელნაწერთა ავტორების იდენტიფიცირების, არქივების მოწესრიგებისა და აკადემიურ გამოცემათა სრულყოფის საქმეში.

XIX საუკუნის ქართველ მწერალთა ეპისტოლური მემკვიდრეობის გამოცემაში წარმოდგენილ ავტორთა უთარიღო წერილების დათარიღებისას ძალზე დაგვეხმარა ეპისტოლეთა ჩვენ მიერვე შედგენილი ციფრული კორპუსები. საძიებო სისტემები დიდი მოცულობის ტექსტებშიაც კი მნიშვნელოვნად აიოლებს ამა თუ იმ სიტყვის, სახელის თუ ციფრის პოვნას. ეს საშუალებას

გვაძლევდა, გვეწარმოებინა წერილების შედარებითი ანალიზი და დაგვედგინა უთარიღო ტექსტების შექმნის დრო. ამ ტექნოლოგიის მნიშვნელოვანი დამსახურებაა ის, რომ უკვე მოხერხდა მხოლოდ გრიგოლ ორბელიანის 83 უთარიღო წერილის დათარიღება.

ციფრული ტექსტური კორპუსების მეშვეობით გამარტივებულია პოეტურ ნაწარმოებებში ზმასიტყვაობის კვლევაც. ეს ხერხი პირველად გამოვიყენეთ „ვეფხისტყაოსანთან“ მიმართებით. გამოთქმული იყო მოსაზრება, რომ „ვეფხისტყაოსნის“ ტექსტში სიტყვათა გასაყარზე არაერთგან არის შეფარული იმ ორი პიროვნების სახელი - „თამარ“ და „დავით“, ვისაც ეძღვნება პოემა. ამ მოსაზრების ოპონენტები ამტკიცებდნენ, რომ სიტყვათა გასაყარზე ეს სახელები ამავე სიხშირით ავტორის ინტენციის გარეშეც შეიძლებოდა გამოკვეთილიყო. იმის შესამოწმებლად, თუ რამდენად რეალურია ასეთი შემთხვევითობა, შევადგინეთ პოეტური ტექსტების ციფრული კორპუსი (შევიტანეთ რუსთველის ეპოქასთან მიახლოებული რაც შეიძლება მეტი ტექსტი) და ჩავატარეთ კვლევა Microsoft Word-ის საძიებო სისტემა Find-ის მეშვეობით. ზემოხსენებულ სახელებს ვეძებდით პირველ და მეორე, მეორე და მესამე, მესამე და მეოთხე, მეოთხე და მეხუთე ასოებს შორის „ჰარის“ ჩართვის საშუალებით. გამოიკვეთა, რომ მომიჯნავე სიტყვათა გასაყარზე თამარისა და დავითის სახელები შეიძლებოდა შემთხვევითაც გამოსახულიყო, მაგრამ, რამდენადაც ეს მოვლენა რუსთველის პოემაში 3-ჯერ და უფრო მეტი სიხშირითაც გვხვდება, ვიდრე ჩვენ მიერ შედგენილ ბევრად უფრო დიდ კორპუსში, საფიქრებელია, რომ პოემის ტექსტთან მიმართებით უმრავლეს შემთხვევაში იგი ზმაა და არა ამ ორი მეფის სახელთა შემადგენელი ასოების შემთხვევითი განლაგება.

ციფრული საძიებო სისტემების ეფექტურად გამოყენება შეიძლება ინტერპოლაციების დასადგენად, პოეტურ ტექსტებში ალიტერაციებისა და რითმების საკვლევად და სხვ. შესაძლებლობები ამ მხრივ ძალზე დიდია. შესაბამისად, ახალი ტექნოლოგიებისადმი კრეატიული მიდგომა ჰუმანიტარულ მეცნიერებათა წარმატებული განვითარების მნიშვნელოვანი წინაპირობაა.

Use of Digital Bases and Corpus Linguistic Methods in Textual Criticism and Literary Studies

Maia Ninidze
Shota Rustaveli Institute of Georgian Literature, TSU (Georgia)
maianinidze@yahoo.com

Rapid development of computer techniques and technologies created new challenges for the humanity studies. Broadening of the possibilities requires increase of the research scales and profoundness. The most important product of the practical textual criticism – academic edition of works, accumulating full collections of the classic authors' texts and providing them with apparatus criticus

should become even more informative and comprehensive in order to comply with the growing requirements of present-day readers. This can be achieved only by means of up-to-date technologies.

In 2007, while working on the 15-volume collection of the academic edition of Ilia Chavchavadze's works, the scientific center of textual criticsm at the Institute of Georgian Literature implemented the technology of using a digital corpus for the purposes of anonymous text attribution. As far as the anonymous texts were published on the editorial page of the newspaper *Droeba* in the time when the periodical had two editors – Ilia Chavchavadze and Sergei Meskhi, we compiled well-known publications of these two authors in two digital corpora and carried out comparative analysis. The studies detected that from the point of the specific use of affixes, prepositions, adverbs and other auxiliary words, as well as frequency of some lexical units and other linguo-stylstic features, the anonymous texts were much closer to Ilia Chavchavadze's than to Sergei Meskhi's articles.

The attribution of a text requires all-round, complex study and argumentation. Hence, a comparative analysis of such a corpus is not sufficient for coming to the final conclusion; however, it is a good additional means. The number of the anonymous articles in the 19th century periodicals is quite large, and the compilation of the digital corpus of different writers will be very helpful for the identification of authors.

It is also very important for attribution investigations to create digital bases of the scanned calligraphy samples of all famous Georgian writers. It will help in the identification of authors, systematization of archives and improvement of academic editions.

Digital bases of the epistolary heritage of the 19th century Georgian authors helped us greatly in the temporary identification of undated letters. Digital search systems make it easy to find necessary words, names and numbers even in large texts. This enabled us to carry out comparative research, to identify the sequence of letters and, thus, to date them. It is partially thanks to this technology that we managed to date 83 private letters only by Grigol Orbelini.

Digitalization makes also much easier finding puns in large poetic texts. The technology was worked out and used by us in Rustaveli Studies. There existed a research, claiming that the names of the two addressees of the poem – Tamar and David are constructed at the junction of two adjacent words in different passages of the poem *The Knight in the Panther's Skin*. The study did not get much feedback as far as it was supposed that these names might be constructed by chance, without the author's special intention. In order to check whether these are puns or just accidental sequences of letters, we undertook a statistical study. We compiled a much larger digital corpus of poetic texts most of which were written in the period close to Rustaveli, and, by means of „Microsoft Word" „Find" system, counted the number of cases when the sequence of the letters – „tamar" and „david" were constructed at the junction of adjacent words. We did it by means of inserting space at different places – between the first and the second, the second and the third, the third and the fourth, the fourth and the fifth letters. It was revealed that such sequence of letters might appear by chance only three times rarely than it is in Rustaveli's poem. Hence, we came to the conclusion that a greater number of cases are real puns and not accidental sequences of letters.

Digital bases and search systems may be effectively used for studies of interpolations, alliterations, rhymes, etc. The possibilities are vast and inexhaustible. Accordingly, a creative attitude towards new technologies is one of the best guarantees for the development of present-day humanities.

ქართული ენის სინონიმთა ლექსიკონის შედგენის საკითხისათვის

სერგეი პოტიომკინი
მოსკოვის სახელმწიფო უნივერსიტეტი (რუსეთის ფედერაცია)
prolexprim@gmail.com

სინონიმები წარმოადგენენ ერთი და იმავე ენის ორ ან მეტ სიტყვას ან სიტყვათშეხამებას, რომელთაც აქვთ ამა თუ იმ საგნობრივი სფეროს იდენტური ან თითქმის იდენტური მნიშვნელობა. მიუხედავად იმისა, რომ მათ დიდი მნიშვნელობა ენიჭებათ სხვადასხვა თვალსაზრისით (მათ შორის, უცხოური ენებისა და დედაენის სწავლება, თარგმანი, ბუნებრივ ენათა ავტომატური დამუშავება), სინონიმების გამოვლენის ამოცანა კვლავაც რთულია და ლექსიკოგრაფიაში ჯერ კიდევ არ შემუშავებულა საყოველთაოდ აღიარებული მიდგომა. როგორც ცნობილია, თეზაურუსები წარმოადგენენ სინონიმთა უაღრესად გავრცელებულ წყაროს. მიუხედავად იმისა, რომ ასეთი რესურსები, რომლებიც კვალიფიციური ლინგვისტების მიერაა მომზადებული (იხ. ინგლისური ენის ყველაზე განვითარებული თეზაურუსი WordNet http://wordnet.princeton.edu), ჩვეულებრივ უზრუნველყოფენ სინონიმიის გამოვლენის მაღალ ხარისხს, თეზაურუსის აგებისათვის საჭირო ხელით შესასრულებელი სამუშაოს მოცულობა, ლექსიკის არასრული დაფარვა და შეზღუდული მისაწვდომობა ამნელებენ მის გამოყენებას და მიუთითებენ სინონიმთა გამოვლენის პროცესის ავტომატიზაციის აუცილებლობაზე. მოხსენებაში შემოთავაზებულია სინონიმთა იდენტიფიკაციის ახალი მეთოდი მანქანურად წაკითხვად ორენოვან ლექსიკონებზე (ქართულ–ინგლისური და ქართულ–რუსული) დაყრდნობით. ჩვენს განკარგულებაშია „დიდი ქართულ–ინგლისური ლექსიკონი," რომელშიც ეკვივალენტთა 70000 მეტი წყვილი შედის (*A Comprehensive Georgian-English Dictionary*. Garnett Press, - 1675 pp.) და შედარებით მომცრო ქართულ–რუსული სიხშირული ლექსიკონი (http://mydisk.ge/download.php?id=hynuremaz) – ეკვივალენტთა დაახლოებით 4000 წყვილი. უაღრესად სასურველია ლექსიკოგრაფიულ ბაზაში ისეთი ქართულ–რუსული და რუსულ–ქართული ლექსიკონების ჩართვა, როგორებიცაა დ. ჩუბინაშვილის ქართულ–რუსული ლექსიკონი (თბილისი, 1984, – 914 გვ., შეიცავს 40000 სიტყვას), მ. კანკავას ქართულ–რუსული ლექსიკონი (თბილისი, 2001, – 435 გვ., შეიცავს თანამედროვე ქართული ენის 20000–ზე მეტ სიტყვას), ა. ტოროტაძის მოკლე რუსულ–ქართული ლექსიკონი (თბილისი, 1969, – 835 გვ., შეიცავს 32000 სიტყვას), რომელთა ელექტრონული ვერსიები ჩვენ ვერ მოვიპოვეთ.

მომავალში ორენოვანი ლექსიკონების რაოდენობა შეიძლება გაიზარდოს, კერძოდ, მათ მიემატოს გერმანულ–ქართული ლექსიკონი (Fähnrich, H. *Kartwelisches Etymologisches*

Wörterbuch. Leiden/Boston: Brill, 2007, - 874 s.).

წარმოდგენილი მეთოდი ემყარება იმ დაკვირვებებს, რომელთა მიხედვითაც, ერთი და იმავე სიტყვის ლექსიკონში დაფიქსირებული უცხოენოვანი ეკვივალენტები, როგორც წესი, სინონიმები არიან, თუმცა მხოლოდ ორენოვან ლექსიკონებზე დამყარებულ მეთოდს დაბალი რელევანტურობა და სიზუსტე აქვს. ამ პრობლემის ნაწილობრივი გადაჭრისათვის ლექსიკოგრაფიულ ბაზაში უნდა მოხვდეს სინონიმთა ერთენოვანი ლექსიკონები (ინგლისური და რუსული). ეს საშუალებას მოგვცემს, ჩავრთოთ მეტი ლექსიკური ერთეული, მაგრამ აქ საჭირო გახდება გამოყოფილი კვაზისინონიმების სიზუსტის შემოწმება. გამოვლენილი სინონიმები შეფასდება სხვა ორენოვან ლექსიკონთან შეჯერების გზით. შეფასების მიხედვით ჩანს, რომ ჩვენი მიდგომით დამაკმაყოფილებელ შედეგებს ვიღებთ. მომავალში დაგეგმილი გვაქვს რესურსის _ „ქართული სინონიმები“ _ ინტერნეტში განთავსება და ონლაინ კენჭისყრის ჩატარება, თუ რამდენად მისაღებია სინონიმური წყვილი. ვიკიპედიის პრინციპზე აგებულ კენჭისყრაში მონაწილეობას ვთავაზობთ ქართულენოვან სპეციალისტ ლექსიკოგრაფებს, ასევე ინტერნეტის მომხმარებელთა ფართო წრეებს. მათი მონაწილეობის გასააქტიურებლად გათვალისწინებულია ენობრივი თამაშის შექმნა.

On the Development of the Georgian Synonyms Dictionary

Sergey Potemkin
Moscow State University (Russian Federation)
prolexprim@gmail.com

Synonyms are two or more words or phrases of the same language, having the same or the similar meaning within a certain subject area. Despite of the importance of synonyms for a variety of applications, including native and foreign languages learning, translation, automatic processing of natural language (NLP), the task of identifying synonyms remains a complex one and has no general solution in lexicography. Thesauri are probably the most common source of synonyms. While these resources are prepared by qualified linguists, see the most developed thesaurus of English WordNet (http://wordnet.princeton.edu/), typically provide high quality detection of synonymy, the amount of manual work needed to build a thesaurus, incomplete coverage of vocabulary, and restrictions on access, make their usage difficult and exhibits the need for the automation of the synonyms extraction process. This paper proposes a new method for the identification of synonyms based on bilingual (English-Georgian and Georgian-Russian) machine-readable dictionaries. We have in our disposal a large English-Georgian dictionary - more than 70,000 pairs of equivalents (*A Comprehensive Georgian-English Dictionary*. Garnett Press, - 1675 pp.) and a relatively small Georgian-Russian dictionary (http://mydisk.ge/download.php?id=hynuremaz) - about 4,000 pairs of equivalents. The inclusion of the

Georgian-Russian and Russian-Georgian dictionaries in the database is highly desirable: (D.I. Chubinashvili (Chubinov), *Georgian-Russian Dictionary*. Tbilisi: Sabchota Sakartvelo, 1984. - 914 pp., contains 40 thousand words; Kankava M.V. *Georgian-Russian Dictionary*. Tbilisi: Sakartvelos Matsne, 2001. - 435 pp. ISBN: 99928-28-22-6, contains more than 20,000 words of the modern Georgian language; Torotadze A.G. *A Brief Russian-Georgian Dictionary*. Tbilisi: Sabchota Sakartvelo, 1969. – 835 pp., contains 32 thousand words), for which we could not find the machine-readable versions.

In the future, the set of bilingual dictionaries will be replenished, in particular by the German-Georgian dictionary (Fähnrich H. Kartwelisches Etymologisches Wörterbuch. – Leiden/Boston: Brill, 2007. - 874 s.).

The proposed method is based on the observation that the foreign-language equivalents of a word tend to be synonymous, but the method is only based on bilingual dictionaries, has low recall and precision. A partial solution to this problem is to use the monolingual dictionaries of synonyms (English and Russian). This can increase the coverage of vocabulary, but requires additional precision checks of the selected quasi-synonyms. Extracted synonyms are estimated by involving another bilingual dictionary. Evaluation results show that our approach achieves satisfactory results. It is planned to allocate resource "*Georgian Synonyms*" on the Internet and to conduct the on-line voting on the admissibility of synonymous pairs. The native Georgians, as the professionals - lexicographers and the wide circles of the Internet users should be attracted to voting process on the principles of Wikipedia. To increase their participation some sort of the language game will be developed.

სახელდებითი ერთეულების ფუნქციური ტიპოლოგია ენის სემანტიკურ მოდელირებასთან კავშირში

ნინო სანაია
სოხუმის სახელმწიფო უნივერსიტეტი (საქართველო)
nsanaia@yahoo.com

სახელდებითი ერთეულების ფუნქციური ტიპების კავშირი ამ ერთეულების სემანტიკურ კომპიუტერულ ანალიზსა და, ზოგადად, ენის მიკროსემანტიკურ მოდელირებასთან კავშირში წარმოადგენს ნ. არუთიუნოვას (არუთიუნოვა, 1978) „მეტაფორის ფუნქციური ტიპების“ თეორიის განვითარებას. სახელდებით ერთეულებში ვგულისხმობთ ნომინაციურ (დენოტაციურ) ერთეულებს, ხოლო მიკროსემანტიკურ მოდელირებაში – ამ ერთეულების სემანტიკური მიკროველის (დუხაჩეკი, 1960) ანალიზს, მაკროველის (სემანტიკური ველის ფართო გაგებით) საპირისპიროდ.

არუთიუნოვას თეორია წარმოადგენდა მეტაფორის ფუნქციური ტიპების დიფერენციაციას, გამონათქვამში ნომინაციური ერთეულების კომუნიკაციური ფუნქციების (ნომინაციურისა და პრედიკაციულის) საფუძველზე. თუმცა ამ ნაშრომს წინ უძღოდა ვ. გაკის ოდნავ უფრო მცი-

რეკომპონენტიანი კლასიფიკაცია, რომელიც გულისხმობს: მაიდენტიფიცირებელ (ნომინაციურ), ექსპრესიულ და დამხმარე-სტრუქტურულ (служебно-строевая функция) ფუნქციებს (გაკი 1977: 248).

ნ. არუთიუნოვამ გამოყო ორი ძირითადი კომუნიკაციური ფუნქცია: ნომინაციური და პრედიკაციული და შემოგვთავაზა ამ ფუნქციათა უფრო ფართო სპექტრი მეტაფორის მაგალითზე. ესენია: 1) მაიდენტიფიცირებელი, ანუ ნომინაციური (არუთიუნოვა 1978: 333); 2) ხატოვანი (არუთიუნოვა 1978: 334); 3) კოგნიტური (იქვე); 4) განმაზოგადებელი (იქვე);

მოგვიანებით ამ ტიპოლოგიას ვ. თელიამ და ე. ოპარინამ დაუმატეს კიდევ ერთი სახეობა – კონცეპტუალური მეტაფორა, რომელიც ენაში არა მხოლოდ ახალ საგნებს და ფენომენებს აღნიშნავს, არამედ ერთი საგნის სხვადასხვა კუთხით დანახულ მრავალ ცნება-კონცეპტს ქმნის (თელია 1988: ოპარინა 1988).

აზროვნების აღმნიშვნელი მეტაფორული შესიტყვებების კვლევისას ჩვენ შევაჯერეთ გაკისა და არუთიუნოვას კლასიფიკაციები, დავუმატეთ ჩვენს საკვლევ მასალაში გამოკვეთილი ახალი ტიპები და მივიღეთ შემდეგი სურათი: 1) მაიდენტიფიცირებელი; 2) დამხმარე-სტრუქტურული ; 3) განმაზოგადებელი ; 4) კოგნიტური; 5) ექსპრესიული; 6) ხატოვანი; 7) შეფასებითი; ემოციური; 8) სტილისტური; 9) სოციოკულტურული (ეროვნული) (სანაია, 2009).

თითოეული ტიპის მიკროსემანტიკურმა მოდელირებამ, რომლის ამსახველი ჩვენ მიერ შედგენილი ცხრილი **Excel** ან **Access** პროგრამებშია წარმოდგენილი (სანაია 2012), დაგვანახა, რომ თითოეულ ტიპს სემების გარკვეული რაოდენობის ნაკრები შეესაბამება, ანუ მეტაფორის სემური შემადგენლობა და მისი ფუნქციური ტიპი ურთიერთდამოკიდებულ კავშირში აღმოჩნდა.

ვინაიდან არუთიუნოვას თეორია მეტაფორის ფუნქციური ტიპების შესახებ ზოგადად სახელდებითი ერთეულების კომუნიკაციური ტიპებიდან გამომდინარეობს, ჩვენ გადავწყვიტეთ ეს ტიპოლოგია ისევ ზოგადნომინაციურ ჩარჩოებს დავუბრუნოთ და ვფიქრობთ, ექსპერიმენტმა გაამართლა. მივიღეთ შემდეგი სურათი: 1) მაიდენტიფიცირებელი ტიპი შეიცავს მხოლოდ სემანტიკურ ბირთვს – მაიგივებელ (არქისემა) და კატეგორიალურ ნიშანს, თუ ეს ჰიპერონომია და დიფერენციალურსაც, თუ იგი სემანტიკური ველის ჩვეულებრივი წევრია; 2) დამხმარე-სტრუქტურული – ასევე შეიცავს მხოლოდ სემანტიკურ ბირთვს; 3) განმაზოგადებელი – მხოლოდ ბირთვს; 4) კოგნიტური – მხოლოდ ბირთვს; 5) ექსპრესიული – ბირთვს, დიფერენციალურს და შიდაფორმას (ზოგჯერ დიფერენციალური სემა და შიდაფორმა ემთხვევა ერთმანეთს), 6) ხატოვანი – ასევე პრესუპოზიციულ ხატს, თუკი იგი მკვეთრად არის გამოხატული; 7) შეფასებითი – შეიცავს კიდევ ერთი სემით მეტს და ეს არის შეფასებითი კონოტატური სემა; 8) ემოციური ტიპი შეიცავს ყველა წინაჩამოთვლილ სემას და კიდევ ერთს – ემოციურს; 9) სტილისტურს ემატება კონოტატური ინფორმაცია სტილური მარკირების შესახებ; 10) სოციო-კულტურული (ეროვნული) ტიპი ყველა დანარჩენის გარდა შეიცავს კიდევ ერთ სემას, რომელიც წარმოგვიდგენს კონოტაციურ სოციოკულტურულ ინფორმაციას.

ამგვარად, ჩვენი კვლევა წარმოდგენს ცდას იმის დასადასტურებლად, რომ სახელდებითი ერთეულების ტიპური კლასიფიკაცია მჭიდროდ არის დაკავშირებული მათ სემურ შემადგენლობასთან და ეს კავშირი შესაძლებელია აისახოს ენის მაკროსემანტიკური მოდელის არქიტექტონიკაში, ე. ი. ერთეულები დალაგდეს სემური რაოდენობრივი შემადგენლობის მიხედვით,

ეს კავშირი შესაძლებელია აისახოს მიკროსემანტიკური მოდელის ამსახველ ცხრილში კიდევ ერთი აღწერის ახალი ზონის დამატებით, რომელსაც ფუნქციური ტიპი შეიძლება ეწოდოს.

წარმოდგენილი მოდელი დამუშავებისა და დახვეწის პროცესშია.

Functional Typology of Denotation Units in Connection with Semantic Modeling

Nino Sanaia
Sokhumi State University (Georgia)
nsanaia@yahoo.com

The paper addresses functional types of denotation units with semantic computational analysis of these units and, in general, micro-semantic modeling of language. The research presents the development of the theory of „Functional Types of Metaphor“ by N. Arutyunova (Arutyunova, 1978). By denotation units we mean nomination (denotation) units and by micro semantic modeling we mean the analysis of the semantic micro-field (Dukhacheki, 1960) of these units as opposed to macro-field (semantic field in the broadest sense).

Arutyunova's theory presented the differentiation of functional types of metaphor on the basis of the communicative functions (denotation and predication) of denotation units in the expression. However, this paper was preceded by V. Gak's classification with a slightly little component which is as follows: Identifier (denotation), expressive and auxiliary-structural functions (Gak, 1977: 248).

N. Arutyunova distinguished between two basic communicative functions: denotation and predication. He offered a wide spectrum of these functions exemplified by metaphor. These are: 1. Identifier (denotation) (Arutyunova, 1978: 333); 2. Figurative (Arutyunova 1978: 334); 3. Cognitive; 4. Generalize. Later, V. Telia and E. Oparina added one more type - conceptual metaphor to this typology which marks not only new subjects and phenomena in the language but creates lots of concepts viewed in different sides by one subject (Telia, 1988; Oparina, 1988).

We studied the metaphorical correspondents of thought, we summarized the classifications of Gak and Arutyunova, added new types mentioned in our research material, arriving at the following result: 1. Identifier; 2. Auxiliary – structural; 3. Generalized; 4. Cognitive; 5. Expressive; 6. Figurative. 7. Evaluative; Emotional; 8. Stylistic; 9. Socio-Cultural (National) (Sanaia, 2009).

Micro-semantic modeling of each type, functioning in the table **Excel** or **Access** programs, created by us (Sanaia, 2012), showed that each type corresponds to a certain number of sets of semes. Thus, a semic set of metaphor and its functional type became interrelated. This fact raised a doubt that the dialectic law - „ Quantitative changes move to the qualitative one“ also is revealed in the language.

As far as Arutyunova's theory about functional types of metaphor comes from the communication types of denotation units, we decided to bring this typology back to the common-

denotation frames, and we believe that the experiment was a success. We received the following result: 1. Identifier type includes only semantic nuclei – identifier and categorical sign, if this is hyperonomia, and differential is it is a usual member of semantic field. 2. Auxiliary – structural - only includes semantic nuclei. 3. Generalized - only nuclei. 4. Cognitive – only nuclei. 5. Expressive – nuclei, differential and internal form (Sometimes differential and internal form coincides with each other). 6. Figurative - Presupposition icon if it is sharply expressed. 7. Evaluative – included one more seme and is evaluative connotative seme. 8. Emotional type included all above-mentioned semes and emotional seme. 9. Connotative information about stylistic marking is added to stylistic. 10. Socio-cultural (national) type except all others includes one more seme that presents connotative socio-cultural information.

Thus, our research is an experiment to prove the fact that the typical classification of denotation units has been closely associated with their seme set and this connection may be reflected in the architectonics of macro-semantic modeling of language or units may be arranged according to the seme quantitative sets. Alternatively, this link may be reflected in the table of micro-semantic modeling added with a new zone of description which may be called a functional type.

As everything new, the model presented by us is in the process of development and improvement.

მულტივარიანტული პარალელური ტექსტები რუსულ ეროვნულ კორპუსში

დიმიტრი სიჭინავა
რუსული ენის ინსტიტუტი (რუსეთის ფედერაცია)
mitrius@gmail.com

პარალელურ კორპუსებში, რომლებიც რუსული ეროვნული კორპუსის (http://ruscorpora.ru) ნაწილს წარმოადგენენ, მალე შევა ე.წ. **მულტივარიანტული პარალელური ტექსტები**, თითოეულ დედნისეულ ტექსტზე რამდენიმე ალტერნატიული თარგმანით. ეს მულტივარიანტული კორპუსიც ხელმისაწვდომი იქნება ინტერნეტით. ამასთანავე, შექმნილია ფუნქციურად ეკვივალენტური ზმნური ლექსიკო-გრამატიკული ფორმების მონაცემთა ბაზა, რომლისთვისაც გამოყენებულ იქნა მულტივარიანტული რუსულ-ფრანგული კორპუსი. ამ მონაცემთა ბაზის ერთ-ერთი მთავარი მიზანი გახლდათ რუსული და ფრანგული ენების ზმნურ ფორმებს შორის არსებული სხვადასხვა ტიპის შესაბამისობების სტატისტიკის დადგენა და, კერძოდ კი, უპირო ზმნური კონსტრუქციებისა, როგორიცაა რუსული *мне кажется,* ფრანგული *il me semble*.

მულტივარიანტული კორპუსისა და მონაცემთა ბაზის შექმნის ტექნოლოგია უნდა ემსახურებოდეს შემდეგ ძირითად ფუნქციებს:

1) პარალელური ტექსტების შეთანადება თარგმანის რამდენიმე ვარიანტთან;

2) პარალელური ტექსტების მორფოლოგიური ანოტირება და ლემატიზაცია;
3) პარალელური ტექსტების ჩართვა კორპუსში თარგმანის რამდენიმე ვარიანტის თანხლებით;
4) ეკვივალენტური ლექსიკოგრამატიკული ზმნური ფორმების (რუსულ-ფრანგული) მონაცემთა ბაზის აგება;
5) ზმნურ ფორმებს შორის არსებული შესაბამისობების სტატისტიკის გამოთვლა.

პარალელური კორპუსის შესახებ არსებულ ნაშრომებში ჩვეულებრივ გამოყენებულია თარგმანის მხოლოდ ერთი ვარიანტი მოცემულ ენაზე. პარალელური კორპუსების ლექსიკისა და გრამატიკის გამოკვლევებში შეიძლება გამოყენებულ იქნეს ალტერნატიული თარგმანები ყველა შესაძლო შესაბამისობის გათვალისწინებით. მათი საშუალებით შეიძლება აისახოს ობიექტური ვარიანტულობა სამიზნე ენაში, რაც ფასეული რესურსი იქნება ორივე ენის შეპირისპირებითი გრამატიკული აღწერისათვის. მაგალითად, მოსალოდნელია მეტი ნაირსახეობის დაფიქსირება თარგმანებს შორის, როდესაც არსებობს სტრუქტურული განსხვავებები მოცემულ ორ ენას შორის (აღსანიშნმა, რომელსაც არ აქვს აშკარა შესაბამისობა სამიზნე ენის გრამატიკასა თუ ლექსიკაში, შესაძლოა მეტი ვარიანტულობა გამოავლინოს თარგმანებში).

შემუშავდა შეთანადებულ კორპუსებზე დამყარებული რაოდენობრივი მეთოდოლოგიები მონო, პოლი და ჰიპერშესაბამისობების ანალიზისათვის და, საზოგადოდ, შეპირისპირებითი გრამატიკული აღწერისათვის. რაოდენობრივი მეთოდები შეიძლება გამოვიყენოთ ვარიანტულობის ყველაზე ხშირი შემთხვევების გამოსავლენად, ასევე მოცემულ ორ ენას შორის არსებული რეგულარული შესატყვისობების ანალიზისათვის. როდესაც დაფიქსირდება თარგმანებს შორის არსებული ნაირსახეობა მოცემულ ლექსიკურ თუ გრამატიკულ თავისებურებასთან მიმართებით, სტატისტიკური ანალიზი საშუალებას მოგვცემს, აღმოვაჩინოთ, რომელი კონტექსტური მახასიათებელი კორელირებს თითოეულ ვარიანტთან. ამ კონტექსტური მახასიათებლების მეშვეობით აღიწერება თითოეული ვარიანტის თავისებურება. ფაქტორული ანალიზი და, კერძოდ, შესატყვისობების ანალიზი კარგად ესადაგება ვარიანტთა მრავალრიცხოვნობის პირობებში კონტექსტური მახასიათებლების სტაბილური ჯგუფების იდენტიფიკაციის ამოცანას. სტატისტიკური ანალიზი გამოყენებულია საკვლევი მრავალრიცხოვანი ტექსტების რეზიუმეს მისაღებად. მაშინ, როცა თითოეული კონტექსტი გამოუსადეგარია დასკვნების გამოსატანად და ასევე შეუძლებელია წარმოვიდგინოთ სრული სურათი, როდესაც საქმე გვაქვს თავმოყრილ მრავალრიცხოვან კონტექსტთან, სტატისტიკური ანალიზი სასარგებლოა ვარიანტების კარტოგრაფირებისათვის.

Multivariant Parallel Texts within the Russian National Corpus

Dmitry Sitchinava
Institute of the Russian language (Russian Federation)
mitrius@gmail.com

The parallel corpora as a part of the Russian National Corpus (http://ruscorpora.ru) will now include so-called **multivariant parallel texts** where for each original text a set of different alternative translations is provided. The multivariant corpus will likewise be available online. Concurrently, a Database of functionally equivalent lexico-grammatical verbal forms has been created using the multivariant Russian-French corpus. One of the main objectives of the database creation is to obtain statistical estimates of different types of equivalences between the Russian and French verbal forms, and, in particular, the impersonal verbal constructions of the kind of 'it seems to me' (Russian *mne kazhetsja,* French *il me semble*).

The technology for creating the multivariant corpus and the database should support the following basic functions:

- alignment of parallel texts with several variants of translation;
- morphological annotation and lemmatization of parallel texts;
- including parallel texts with several variants of translation into the corpus;
- constructing the database of equivalent lexico-grammatical verbal forms (Russian - French);
- calculating the statistical estimates of equivalences between the verbal forms.

The existing works on parallel corpora usually use only one option of translation per language. Lexical and grammatical studies performed on parallel corpora may use alternative options of translations, envisaging all possible correspondences. They can reflect objective variability in the target language and can be a valuable resource for contrastive grammatical description of both languages. For instance, one can expect to observe greater variability amongst the translations when there are structural differences between the two languages (a «signifié» with no clear correspondence in the grammar and/or lexicon of the target language is more likely to exhibit variation in its translations).

Quantitative methodologies for the analysis of mono-, poly- and hyperequivalences and, more generally, for contrastive grammatical description based on aligned corpora has been elaborated. Quantitative methods may be used for identifying the most frequent cases of variability, including an analysis of the regular correspondences between the two linguistic systems. When a variability of translations is observed for a given lexical or grammatical feature, statistical analysis will allow for discovering which contextual features are correlated with each variants. These contextual features help describing the values of each variant. Factorial analysis, and in particular Correspondences analysis, is well suited for this task of identifying the stable groups of contextual features across numerous instances of variants. Statistical analysis is used as a mean for providing a summary of numerous translations

(polyequivalence) under scrutiny. While each single context is of no use for drawing conclusions, and while it is also impossible to figure out the big picture when dealing with numerous collected contexts, statistical analysis is useful for mapping variants.

თათრული ენის ეროვნული კორპუსი „თუგან თელ“: გრამატიკული ანოტირების სტრუქტურა და მახასიათებლები

ჯავდეთ სულეიმანოვი, ოლგა ნევზოროვა, ბულატ ხაკიმოვი
თათრეთის მეცნიერებათა აკადემიის გამოყენებითი სემიოტიკის კვლევითი ინსტიტუტი, ყაზანი (რუსეთის ფედერაცია)
alsu_73@list.ru, nevzoro@gmail.com, khakeem@yandex.ru

დღესდღეობით საკმაოდ აქტუალურია თურქული ენების ფართოდ ხელმისაწვდომი ელექტრონული კორპუსების შედგენისაკენ მიმართული პროექტები. ყველაზე თვალსაჩინო პროექტებს შორის შეიძლება დასახელდეს თათრული, უიღურული, ბაშკირული, ხაკასური და თუვიური ენების კორპუსები, რომლებიც განხორციელების სხვადასხვა ფაზაშია.

თათრული ენის ეროვნული კორპუსის (თეეკ) „თუგან თელის“ შედგენის პროექტის ფარგლებში თავდაპირველად შეიქმნა თათრული სალიტერატურო ენის ახალი მორფოლოგიური ტიპის კორპუსი. ეს კორპუსი შეიცავს თანამედროვე თათრული სალიტერატურო ენის სხვადასხვა ჟანრისა და სტილის ტექსტებს. მხატვრული და არამხატვრული ტექსტების განაწილება თათრული ენის თანამედროვე კორპუსში (2012 წლის ბოლოს მდგომარეობით) წარმოდგენილია პირველ ცხრილში.

ცხრილი 1. ტექსტების განაწილება ჟანრების მიხედვით თეეკ–ში

ჟანრი	სიტყვათა რაოდენობა	ხვედრითი წილი კორპუსში, %
მხატვრული	19 279 033	71,45 %
არამხატვრული	7 703 258	28,55 %
მთლიანად	26 982 291	100 %

თათრულის აგლუტინაციური მორფოლოგიის ფორმალური რეპრეზენტაციისათვის გამოყენებულ მოდელში სიტყვაფორმა აიგება ფუძეზე რეგულარული დერივაციული და ფლექსიური მაწარმოებლების თანამიმდევრული დართვის გზით. მაშასადამე, იმისათვის, რომ მოვნიშნოთ სიტყვაფორმა, აუცილებელია მისი სტრუქტურის გაანალიზება აფიქსთა ჯაჭვის მიხედვით.

კორპუსში წარმოდგენილი მორფოლოგიური ინფორმაცია თათრულის სიტყვაფორმის შესახებ საერთოდ ორ ძირითად ველს მოიცავს: 1) მეტყველების ნაწილის თავისებურება; 2) მორფოლოგიური ნიშან–თვისებების (პარამეტრების) რაოდენობა.

თუ მივიღებთ მხედველობაში თათრული ენის მორფოტაქტიკის თავისებურებებს, გრამატიკული პარამეტრები დაიყოფა რთულ/მარტივ და სავალდებულო/ფაკულტატიურ პარამეტრებად. რთული პარამეტრები წარმოდგენილია ამა თუ იმ გრამატიკული კატეგორიის მნიშვნელობების მთელი წყებით, ხოლო მარტივი პარამეტრები მხოლოდ ერთი მნიშვნელობითაა წარმოდგენილი. სავალდებულო პარამეტრი ყოველთვის პოვნიერია ამა თუ იმ მეტყველების ნაწილის სიტყვაფორმის აღწერაში. განსხვავებით ზოგიერთი კონკრეტული მნიშვნელობისაგან, ფაკულტატიურ პარამეტრს შეუძლია ასევე გადმოსცეს მნიშვნელობის არაპოვნიერება (ანუ კუთვნილებითობის კატეგორიის მნიშვნელობა თათრული ენის არსებით სახელებში). კორპუსში შემავალი ტექსტების მორფოლოგიური მონიშვნა ხორციელდება თათრული ენის ორდონიანი მორფოლოგიური ანალიზის მოდულის მეშვეობით პროგრამულ ხელსაწყოში PC-KIMMO.

ქვემოთ წარმოგიდგენთ თათრული ენის წინადადების მორფოლოგიური მონიშვნის მაგალითს, რომლებსაც ახლავს გრამატიკული კომენტარები:

Bez uramnan ikää ütep barabız

ორი ჩვენგანი ქუჩას მიუყვებოდა

Bez	Uramnan	Ikää	ütep	barabız
Bez	uram-nan	ikä-ü	üt-ep	bar-a-bız
N/ Pro1_Plu	N+CASE_ABL(DAn)	Num+NUM_GROUP(AU)	V+PARTIC(Ip)	V+PRES(A)+ 1P_PLU(bIz)
Noun/ Pronoun_1sg Pl	Noun+ Ablative	Numeral+collective affix	Verb+Gerund	Verb+Present Tense + 1p, pl.
Awl/ We	down the street	the two of us	passing	go

შემუშავებულ იქნა აღნიშვნების სისტემა სათანადო მორფემების მიერ გამოხატული მორფოლოგიური კატეგორიების გადმოსაცემად, რომლის დროსაც გავითვალისწინეთ თანამედროვე ტიპოლოგიური და თურქოლოგიური გამოკვლევები.

თეეკ-ი განთავსებულია ინტერნეტში EANC–ის პლატფორმის გამოყენებით, რომელიც თავდაპირველად შეიქმნა აღმოსავლეთ სომხურის ეროვნული კორპუსისათვის. საძიებო სისტემა ახორციელებს ძიებას ზუსტი ფორმის, ლემისა და რამდენიმე გრამატიკული პარამეტრის მიხედვით.

თეეკ-ი შეიძლება ჩავთვალოთ თათრული ენის სხვადასხვა დონის კონცეპტუალური და ფუნქციური მოდელების კომპლექსად. როგორც ღია სისტემა, ეს კორპუსი იძლევა ახალი ენობრივი მოდელებისა და მათზე დამყარებული ანოტირების სისტემების დამატების საშუალებას (კერძოდ, დამატებით შემუშავდა თეეკ–ის სემანტიკური ანოტირების მოდელი).

National Corpus of the Tatar Language „Tugan Tel“: Structure and Features of Grammatical Annotation

Dzhavdet Suleymanov, Olga Nevzorova, Bulat Khakimov
Research Institute of Applied Semiotics of Tatarstan Academy of Sciences (Russian Federation)
alsu_73@list.ru, nevzoro@gmail.com, khakeem@yandex.ru

Nowadays the projects of designing generally accessible electronic corpura of Turkic languages are quite topical. Among the most prominent projects, the corpora of Tatar, Uyghur, Bashkir, Khakas and Tuvan languages can be mentioned, being at different phases of implementation.

Within the project of designing of the National Corpus of the Tatar language „Tugan Tel“[1] (TatNC), initially a new corpus of a morphological type was developed for the Tatar literary language. The corpus contains texts of different genres and styles of the contemporary Tatar literary language. The distribution of the texts in the TatNC into fiction and non-fiction at the end of 2012 is presented in Table 1.

Table 1. Distribution of texts according to their genre in TatNC

Genre	Amount of words	Share in the corpus, %
Fiction	19 279 033	71,45 %
Non-fiction	7 703 258	28,55 %
Total	26 982 291	100 %

The model used for the formal representation of the Tatar agglutinative morphology is that in which the wordform is built by consecutive adding to the stem of regular derivational and inflectional affixes. Thereby, in order to mark up a wordform, it is necessary to analyze the structure of its affix chain.

The morphological information about a Tatar wordform, contained in the corpus, generally consists of two main fields: 1) part-of-speech characteristics; 2) number of morphological features (parameters).

Taking into consideration the characteristics of Tatar morphotactics, grammatical parameters are divided into complex/simple and compulsory/optional. Complex parameters are represented by a range of meanings of a grammatical category while simple parameters are represented by a single meaning. A compulsory parameter is always present in the description of a wordform of a certain part of speech. An optional parameter, apart from some concrete meaning, can also convey the absence of meaning (i.e. the

[1] The project is carried out within the scope of the Program of fundamental research of the Presidium of the Russian Academy of Sciences „Corpus Linguistics“ 2012-2014, http://web-corpora.net/TatarCorpus/search/ ?interface language=ru

meaning of the category of possessiveness in Tatar nouns). Morphological mark-up of corpus texts is carried out using the module of the two-level morphological analysis of the Tatar language realized in the program tool PC-KIMMO [PC-KIMMO].

Below we present an example of the morphological mark-up of the Tatar sentence

Bez uramnan ikäü ütep barabız

The two of us walk down the street

Accompanied with grammatical comments:

Bez	uramnan	Ikäü	ütep	barabız
Bez	uram-nan	ikä-ü	üt-ep	bar-a-bız
N/ Pro1_Plu	N+CASE_ABL(DAn)	Num+NUM_GROUP(AU)	V+PARTIC(Ip)	V+PRES(A)+1P_PLU(bIz)
Noun/ Pronoun_1sg Pl	Noun+ Ablative	Numeral+collective affix	Verb+Gerund	Verb+Present Tense +1p, pl.
Awl/ We	down the street	the two of us	passing	go

A system of denotations has been developed to designate morphological categories, expressed by corresponding morphemes, which takes into consideration contemporary typological and Turkic researches [Mishar dialect, 2007].

The TatNC is placed in Internet using the platform EANC, originally developed by the company CorpusTechnologies for the East Armenian National Corpus [EANC]. The search system supports the search by the exact form, by lemma and by a number of grammatical parameters.

The TatNC can be considered as a complex of conceptual and functional models of different levels of the Tatar language. As an open system, the corpus allows adding of new linguistic models and annotation systems based on them (in particular, the model of semantic annotation of TatNC is being developed additionally).

References

[Mishar Dialect; 2007] The Mishar Dialect of the Tatar Language: Essays on Sintax and Semantics. Eds. E.A. Lutikova, K.I. Kazenina, V.D. Solovyeva, S.G. Tatevosova. – Kazan: Magarif, 2007. 383 p. (in Russian)

[PC-KIMMO] PC-KIMMO, http://www.sil.org/pckimmo/about_pc-kimmo.html

[EANC] Eastern Armenian National Corpus, http://www.eanc.net/

ქართული დიალექტური კორპუსი – შედგენილობა და ტექსტების უნიფიცირების პრობლემები

ნარგიზა სურმავა, ლია ბაკურაძე, მარინა ბერიძე
თსუ არნ. ჩიქობავას სახ. ენათმეცნიერების ინსტიტუტი (საქართველო)
nargizasurmava@yahoo.com; l.bakuradze@gmail.com; marine.beridze@gmail.com

ქართული დიალექტური კორპუსის სტრატეგიაა ტექსტური ფონდის მაქსიმალური სისრულე, რაც იმას ნიშნავს, რომ კორპუსში დიალექტები წარმოდგენილი არ არის გაწონასწორებულად, ბალანსირების პრინციპით, არამედ მასში მოცულობის შეუზღუდავად შედის ამა თუ იმ დიალექტის ამსახველი ყველა სანდო და ღირებული მასალობრივი რესურსი.

ქდკ-ს ტექსტური ბაზა შეიცავს სამი ტიპის ტექსტებს: 1. გამოცემული (ნაბეჭდი ტექსტები); 2. ხელნაწერი კოლექციები – ენათმეცნიერების ინსტიტუტის საარქივო მასალები და პირადი კოლექციები, 3. ტრანსკრიფტები – გაშიფრული ვიდეო და აუდიომასალები.

ადრინდელი გამოცემების პარალელურად კორპუსში შეგვაქვს, ავტორებთან შეთანხმების საფუძველზე, ჩვენი კოლეგების მიერ მომზადებული და ამჟამად გამოცემული სანდო მასალები (მაგალითად, შავშურ-იმერხეული მასალა მონოგრაფიიდან „შავშეთი“), რომელებიც კორპუსის მეშვეობით ხვდება აქტიურ სამეცნიერო მიმოქცევაში.

კორპუსი იქცა პირველადი გამოქვეყნების საშუალებად XX საუკუნის 30-იან, 50-იან წლებში (და შემდგომ) ჩაწერილი უნიკალური მასალებისათვის, როგორებიცაა: ქართველ ებრაელთა მეტყველების ნიმუშები ჩაწერილი როზა თავდიდიშვილის მიერ (1940), ალექსანდრე ჯიშიაშვილი – ქართლური ტექსტების კოლექცია (1938), მერი ცინცაძე – ქვემოაჭარული მასალები (1950), გ. ცოცანიძე – თუშური ტექსტები (1982)...

რაც შეეხება ტრანსკრიფტებს, ვიდეო და აუდიო მასალების აბსოლუტური უმრავლესობა გაშიფრული და ჩაწერილია კორპუსის შემდგენელთა ძალებით. ამ სახის მასალები გვაქვს აჭარული, მესხური, კახური, ქართლური, ფერეიდნული, ინგილოური და სამხრული დიალექტებისათვის. დღესდღეობით კორპუსში მათი მხოლოდ ნაწილია განთავსებული.

კორპუსის მომზადების პროცესში ერთ-ერთი მნიშვნელოვანი ეტაპია ტექსტების გრაფემული, პუნქტუაციური და პროსოდიული უნიფიკაცია. უნიფიკაციის წესების შემუშავებისას ორიენტირებულები ვიყავით მორფოლოგიური ანოტაციის მოთხოვნილებასა და კორპუსში ძიების სიმარჯვეზე. ვითვალისწინებდით ომონიმიის თავიდან აცილების შესაძლებლობასაც. ამავე დროს ვეყრდნობოდით ტექსტის ჩაწერის ქართულ დიალექტოლოგიაში მიღებულ ტრადიციას. თუმცა გარკვეულ სირთულეს გვიქმნიდა ის, რომ ამ მხრივ ერთგვაროვნება არც ადრინდელ გამოცემებშია დაცული – ეს განსაკუთრებით შეეხება პროსოდიულ აღნიშვნებს. ენკლიტიკებისა და პროკლიტიკების გამოყოფის სისტემურობას ზოგ გამოცემაში მეტად აქვს დათმობილი ყურადღება (მაგალითად, აღმოსავლეთ საქართველოს მთის დიალექტებში), ზოგში – ნაკლებად. გამოხატვის საშუალებებიც არაერთგვაროვანია (ზოგან დეფისი, ზოგან აპოსტროფი).

არასისტემურობა და არაერთგვაროვნება შეინიშნება ინტონაციის გამოხატვის მხრივაც (გამოყენებულია ხან მახვილის ნიშანი, ხან გრძელი ხმოვანი). შევეცადეთ, ყველა ეს სირთულე,

ცოცხალი დიალექტური მეტყველების ტრანსკრიფციასთან დაკავშირებულ სხვა წამოჭრილ პრობლემებთან ერთად (ე. წ. ჰეზიტაციის, ფალსტარტის აღნიშვნა), გაგვეთვალისწინებინა სპეციალურად კორპუსისათვის შემუშავებულ უნიფიკაციის წესებში; შეძლებისდაგვარად მკაცრად გაგვესაზღვრა თითოეული პუნქტუაციური სიმბოლოს (დეფისი, ტირე, აპოსტროფი, ორწერტილი, მრავალწერტილი, კვადრატული ფრჩხილები...) ფუნქციური გამოყენება კორპუსში და გაგვეტარებინა მთელ ტექსტურ მასივში.

მაგალითისთვის მოვიყვანთ ორიოდე პუნქტს ტექსტის მომზადების ინსტრუქციიდან:

- ტექსტის თანმხლები დამატებითი ინფორმაცია, რომელიც არ უნდა აისახოს ლექსიკონსა და სიტყვანში, ჩაისმება კვადრატულ ფრჩხილებში: ა) ჩამწერის რეპლიკა ან კითხვა, ბ) გამომცემლის კომენტარი (resp.; sic და ა.შ.);
- დეფისი მოხსნილია:
  ა) მავრცობხმოვნიან სახელებთან ენკლიტიკის დროს: **ამხანაგსა ჲყავ, ტოტსა ჲჭრის** და არა: ამხანაგსა-ჲყავ, ტოტსა-ჲჭრის;
  ბ) აღმოსავლეთ საქართველოს მთის დიალექტების ტექსტებში ენკლიტიკის დროს: **ქალი'დ** და არა: ქალი-დ.
  გ) ენკლიტიკის შედეგად მიღებულ ფორმებთან: **ხვარაფერი, არ ალი...** და არა: ხვ-არაფერიაო, არ-ალი...
- აუდიო და ვიდეოფაილების გაშიფვრისას თუ ჩანაწერი ალაგ–ალაგ არ ისმინება, ამას გრაფიკულად გამოვხატავთ კვადრატულ ფრჩხილებში ჩასმული სამწერტილით – **[…].**

აუდიო და ვიდეოტრანსკრიფტებში, რომლებშიც მეტი საშუალება არსებობს ხმოვან ჩანაწერთან ტექსტის დაახლოებისა, გადავწყვიტეთ გაგვეკეთებინა ტექსტის ორი ვარიანტი: ერთი, კორპუსის მორფოლოგიური ანოტაციის მოთხოვნილებიდან გამომდინარე, და მეორე, პროსოდიულ-ინტონაციური აღნიშვნებით, რემარკებითა და კომენტარებით, _ ტექსტების გამოსაცემი ვერსიებისათვის და აგრეთვე, მულტიმედიური კორპუსისათვის.

Georgian Dialect Corpus: Composition and Problems of Text Unification

Nargiza Surmava, Lia Bakuradze, Marina Beridze
Arn. Chikobava Institute of Linguistics, TSU (Georgia)
nargizasurmava@yahoo.com, l.bakuradze@gmail.com, marine.beridze@gmail.com

The strategy of the Georgian Dialect Corpus (GDC) is aimed at the maximum perfection of the text body, this implying that the corpus does not represent dialects with the principle of balance, rather, without the limitation of size, it includes any trustworthy and relevant data resource, reflecting any of the dialects.

The textual base of the GDC consists of three types of texts: 1. published (print texts); 2. manuscript collections – archival materials of the Institute of Linguistics and personal collections; 3.

transcripts – transcribed video and audio resources.

Alongside the earlier publications, the corpus, based on the authors' consent, the corpus will incorporate recently prepared and published trustworthy materials by our colleagues (for instance, Shavshian-Imerkhevian data from the monograph *Shavsheti*), being included in the actual scholarly circulation owing to the corpus.

The corpus became a means for the first publication of the unique texts, recorded in the 1930s, 1950s (and later), such as: samples of the Georgian Jews' speech, recorded by Roza Tavdidishvili (1940), collection of Kartlian texts by Alexandre Jishiashvili (1938), Lower Ajaran materials by Meri Tsintsadze (1950), Tush texts by Giorgi Tsotsanidze (1982).

As for transcripts, the overwhelming majority of video and audio materials were transcribed and recorded by their compilers. Such materials are available for the Ajaran, Meskhetian, Kakhetian, Kartlian, Fereidanian, Ingilo, and southern dialects. Only a portion of them has been placed in the corpus to date.

In the process of the development of the corpus, one of the important stages is the graphemic, punctuation and prosodic unification of texts. During establishing the unification rules, we paid attention to 1) requirement of the morphological annotation and 2) efficiency of corpus query. We also considered the possibility of the avoidance of homonymy. Meanwhile, we followed the tradition of text recording, accepted in Georgian dialectology. However, the fact, that homogeneity has not been present in earlier publications, made some complications; especially, this is the case with prosodic marking. The systemic character of the identification of enclitics and proclitics attracts more attention in some publications (for instance, in highland dialects of eastern Georgia), while it attracts less attention in others. Designations also vary (a hyphen in some of them, an apostrophe in others).

Non-systemic character and heterogeneity occur with respect to intonation as well (sometimes a stress marker is used, sometimes – a long vowel). We did our best to consider all these difficulties, alongside with other problems (so called hesitation, false start) occurring with transcribing of oral dialectal speech, within the unification rules of dialect texts, specially developed for the corpus, to determine, as strictly as possible, the functional uses of each punctuation symbol (hyphen, dash, apostrophe, colon, ellipsis, square brackets...) in the corpus and to sustain the principle throughout the text body.

Here are a couple of items from the guidelines for text preparation:

- complementary information, not to be shown in the dictionary and wordlist, will be inserted in square brackets: 1) recorder's cue or question. b) publisher's comment (*resp., sic*, etc.);
- a hyphen is removed:
  a) in substantives, having a perserverant vowel, in case of enclitic: **amxanagsa yq'av, t'ot'sa yč'ris** and not: **amxanagsa-yq'av, t'ot'sa-yč'ris**;
  b) in texts of highland dialects of eastern Georgia, in case of enclitic: **kali'd** and not: **kali-d**;
  c) in forms, yielded as a result of enclitic: **xvaraperi, ar ali...** and not: **xv-araperiao, ar-ali...**
- if, during the transcribing of audio and video files, a recording is inaudible at some places, it will be graphically marked as a bracketed ellipsis **[...]**.

In audio and video transcripts, providing better opportunities for the rapprochement of a text to a recording, we decided to make two versions of a text: one, in accordance with the requirement of the morphological annotation of the corpus, and another, accompanied with prosodic-intonation markers, remarks and comments – for publishing versions of texts and also for a multimedia corpus.

OLAT - ქართული ენის სწავლების ინოვაციური პლატფორმა ფრანკფურტის უნივერსიტეტში

მარიამ ყამარაული
ფრანკფურტის გოეთეს უნივერსიტეტი (გერმანია)
mariam_kamarauli@hotmail.de

ენების სწავლება დღესდღეობით წარმოუდგენელია ელექტრონული პროგრამების ონლაინრეჟიმში გამოყენების გარეშე. თანამედროვე ტექნოლოგიების გლობალური გამოყენების პირობებში ეს ერთგვარი გამოწვევაა ქართული ენისათვის.

ფრანკფურტის უნივერსიტეტის ემპირიული ენათმეცნიერების ინსტიტუტის კავკასიოლოგიის განხრის პროგრამაში ქართული ენის ელექტრონული სწავლების პროგრამა ერთ-ერთი პირველია კავკასიური ენებისათვის ელექტრონული სწავლების პლატფორმის შექმნის თვალსაზრისით.

ელექტრონული სწავლების დღესდღეობით ყველაზე გავრცელებულ E-learning-ის პროგრამას წარმოადგენს Moodle (Modular Object-Oriented Dynamic Learning Environment) – ონლაინსწავლების ტექნოლოგიების ისტორიაში ერთ-ერთი პიონერული პროგრამა, რომელიც ჯერ კიდევ ინარჩუნებს წამყვან პოზიციებს საერთაშორისო სამომხმარებლო ბაზარზე – ყველასათვის ხელმისაწვდომი და ადვილად ასათვისებელი უფასო პროგრამული პროდუქტი Moodle არის ტექნიკური ქამელეონი, რომელიც ერგება ყველა პროგრამულ ფორმატს და ჰყავს მომხმარებლების ფართო წრე.

ელექტრონული სწავლების პროგრამა OLAT (Online Learning And Training), რომელიც ფრანკფურტის გოეთეს სახ. უნივერსიტეტში გამოიყენება, ონლაინსწავლების ინოვაციურ პროგრამას წარმოადგენს და დიდაქტიკური სცენარის მოქნილობითა და დახვეწილობით გამოირჩევა – OLAT-ში ძირითადი აქცენტი გადატანილია არა სალექციო კურსების შინაარსობრივ მხარეზე, არამედ სტუდენტთა მხრიდან აქტიურ ჩართულობაზე – თვითგანათლებაზე, თვითკონტროლზე და თვითგანვითარებაზე.

Moodle-ისაგან განსხვავებით OLAT-ს აქვს ალმა-მატერის ძლიერი ინდივიდუალური რეპრეზენტაბელურობა, აღჭურვილია ინდივიდუალური მოხმარებისათვის საჭირო ძლიერი ტექნიკური ხელსაწყოებით, უზრუნველყოფს დღესდღეობით არსებული ნებისმიერი ფორმატის გამოყენებას და ხელს უწყობს ინტენსიური სწავლების ინტერაქციას ჯგუფებს შორის, გაცილე-

ბით მოქნილია სწავლების პროცესში გლობალური დაგეგმვისა და პარტიციპაციის თვალსაზრისით, იყენებს განსხვავებული კომუნიკაციის არხებს, იძლევა სასწავლო სცენარის ინდივიდუალური შედგენისა და განხორციელების საშუალებას.

ქართული ენის E-learning-ის კურსი, რომელიც ფრანკფურტის უნივერსიტეტშია არის დანერგილი, არის პირველი პლატფორმა, რომელმაც შეძლო ქართული ანბანის ადაპტაცია და ჩართვა ელექტრონული სწავლების ფორმატში. კურსის ხანგრძლივობა შეადგენს 2 სემესტრს და შეიცავს 28 ლექციას. კურსი შეიცავს ქართული ენის ფონეტიკის, მორფოლოგიისა და სინტაქსის საკითხებს. ასევე სპეციალურ პრობლემებს ქართული ენის პრაგმატიკიდან.

OLAT goes Georgian
(E-learning-platform of Georgian at the Frankfurt University)

Mariam Kamarauli
Goethe University Frankfurt (Germany)
mariam_kamarauli@hotmail.de

Today it is no longer possible to imagine a learning system that makes no use of online electronic programs. Especially the learning of foreign languages is a challenge which needs more resources than books. The present paper is about an electronic course of the Georgian language designed with the help of the E-learning group at Frankfurt University, which is represented in all its departments.

The most widespread e-learning program is Moodle, which stands for „Modular Object-Oriented Dynamic Learning Environment“. Moodle is the Pioneer of Online Learning Technologies and is open source. Because of its technical adaptivity to contents of all kinds, Moodle has a wide user community.

The second program to be dealt with here is OLAT, which stands for „Online Learning And Training“. It has strong qualities for online learning, online teaching and tutoring, with few educational restrictions. It is the most advanced program as far as online learning programs are concerned. The emphasis of OLAT is on self-education, self-control and self-development.

To draw a comparison of the two programs, it is important to highlight the following points: Moodle lacks individual tradition and the „corporate identity“ of a given alma mater; it is too global, separates technical learning tools and is too strongly concentrated on the evaluation of learning skills. In contrast to this, OLAT has a strong „corporate identity“ and provides technical security as well as powerful tools for individual care and support of all current content formats, intensive learning interaction between groups, global planning and implementation in the learning process, separate communication channels and initiation and implementation of different learning scenarios. It is exactly this reason which leads to the conclusion that OLAT is stronger and more advanced than Moodle.

The E-learning course of the Georgian language system presented here is the first platform which has adapted the original Georgian alphabet, and it is one of the first courses in Caucasian languages taught at Frankfurt University. The course lasts two semesters and includes 28 lectures (14 lectures each semester). The whole course includes phonetics, morphology, syntactic specialties of the Georgian language and some aspects of its semantics and pragmatics.

Each lecture is based on 10-12 slides, which have horizontal structure and stand for the thematic development of a given subject. There are two kinds of slides: major slides, which comprise basic information, and minor slides, which include far-reaching information and specifications of problems, specialties and exceptions. The use of this two-line system depends on how intensive a student wants to learn the content and also if this Georgian course is used for a major or minor subject.

Linguistic terms are linked internally to a glossary containing their definition and externally, to scientific literature which relates to the linguistic issue they stand for. A „Lernbar" contains two toolbars, one vertical and one horizontal. The vertical toolbar allows the student to see the working history and the course overview. It also allows to set a bookmark on a page, to write individual notes and to see the individual course development. The horizontal toolbar is for the navigation between the slides and also for the enlargement of the pages. On the last pages of every online lecture, a test is attached, which can have different forms: multiple choice, drag&drop-test, text-mark test, etc. At the end of the test, the students can see the results and can evaluate themselves. Since February 2013, a fourth version of „Lernbar" has been released, which is particularly designed for mobile devices like Smartphone and Tablet-PC.

ქართული წინადადების ანალიზის მონახაზი

გიორგი ჩიკოიძე
არჩილ ელიაშვილის სახ. მართვის სისტემების ინსტიტუტი, ტექნიკური უნივერსიტეტი (საქართველო)
gogichikoidze@yahoo.com

წინადადების ანალიზი წარმოადგენს „შინაარსი↔ტექსტი" მოდელის ფუნქციონირების ერთ-ერთ მიმართულებას. მისი საპირისპირო, ანუ სინთეზური მიმართულებისგან განსხვავებით, გრაფიკული (ან აკუსტიკური) წარმოდგენა, რომელიც უშუალო აღქმას ექვემდებარება, ანალიზური პროცესის საწყისი წერტილია, სახელდობრ, ტექსტი გვევლინება, როგორც სიმბოლოთა თანმიმდევრობა, სადაც ტექსტური სიტყვაფორმები წარმოდგენილია ერთმანეთისგან ხარვეზებით გამიჯნული ასოთა მონაკვეთების ჯაჭვის სახით, ზოგი მათგანი მონიშნულია სპეციფიკური სიმბოლოებით ანუ სასვენი ნიშნებით. სიტყვაფორმების ასეთი მკვეთრი გამიჯვნა, პირველ რიგში, ამარტივებს ტექსტის ანალიზს მეტყველებასთან შედარებით, სადაც ფონემათა უწყვეტი ჯაჭვები ხშირად რამდენიმე სიტყვის ჯგუფს გამოხატავს და არა ცალკეულ სიტყვაფორმას.

ტექსტური წარმოდგენის ეს თავისებურება თითოეული ტექსტური სიტყვაფორმის შესაბამის სალექსიკონო ერთეულთან უშუალო და ცალსახა იდენტიფიკაციის საშუალებას იძლევა, რასაც მოჰყვება სიტყვაფორმის გრამატიკული და სემანტიკური მახასიათებლების დადგენაც. რაც შეეხება გრამატიკულ მახასიათებლებს, მათი განსაზღვრა მაქსიმალურად მარტივდება იმ ლექსიკონის კონტექსტში, რომელიც მოიცავს ყოველი ერთეული მორფოლოგიური პარადიგმის გენერატორს, ანუ სისტემას, რომელიც წარმოშობს მოცემული ერთეულის პარადიგმის ფორმათა სრულ სიას, რომლის წევრები განლაგებულია მათი გრამატიკული მახასიათებლების შესაბამისად. ამის შედეგად სიის რომელიმე წევრთან შესავალი ტექსტური სიტყვაფორმის აბსოლუტური იდენტობა ამ ფორმის გრამატიკული მახასიათებლების დადგენის ტოლფასი იქნება (სწორედ ასეთია 2009-2011 წლებში მართვის სისტემების ინსტიტუტში დამუშავებული ქართული კომპიუტერული ლექსიკონი).

მორფოლოგიური ანოტირება ლინგვისტიკური ანალიზის მეტად მნიშვნელოვანი ასპექტია, რაც უზრუნველყოფს ანალიზური პროცესის შემდგომ წინსვლას მისი იდეალური (თუმც ჯერ ვერმიღწეული) მიზნისკენ, ანუ „შინაარსის“ უშუალოდ გამოხატულებისაკენ. ამ მიმართულებით შემდეგი მნიშვნელოვანი ნაბიჯია მთლიანი წინადადების სტრუქტურის, მისი აგებულების დადგენა უკვე სტრუქტურის ცალკეული ელემენტების (სიტყვაფორმების) შესახებ მოპოვებული მორფოლოგიური ინფორმაციის საფუძველზე.

სტრუქტურის აგება და ასახვა სხვადასხვა სქემით ხდება (მაგალითად, უშუალო შემადგენელების ხისებრი გრაფით).

ჩვენ არ მივყვებით როლებრივი ორგანიზაციის სქემას, რომელიც ეყრდნობა, ერთი მხრივ, სემანტიკური როლების და, მეორე მხრივ, ზმნური სუპერპარადიგმის ცნებას.

სუპერპარადიგმა წარმოადგენს ერთი და იმავე ზმნური ლექსემიდან ნაწარმოები პარადიგმების ერთობლიობას, რომელიც ჯამში შეიძლება განხილულ იქნეს როგორც გარკვეული ვირტუალური („გლობალური“) პროცესის ასახვა; სემანტიკური როლები (CS, AG, OB, AD) კი ამ თვალსაზრისით პროცესის განვითარების ფარგლებში მონაწილეთა სტაბილურ ფუნქციას გამოხატავს.

სემანტიკური როლების შინაარსის სტაბილურობის მიუხედავად, ტექსტებში მათი გამომხატველი აქტანტები გარკვეულად იცვლება სხვადასხვა პარადიგმის ფარგლებში, მაგრამ მათი მორფოლოგიური გაფორმების ვარირება გარკვეულ კანონზომიერებას ექვემდებარება. ვარირების ხასიათი მჭიდროდაა დაკავშირებული სუპერპარადიგმის ტიპთან, აგრეთვე მის კონკრეტულ წევრ პარადიგმასთან, რომელსაც ზმნა განეკუთვნება და საბოლოოდ განისაზღვრება იმ სერიით, რომელსაც ზმნის მწკრივი ეკუთვნის.

ამ კანონზომიერების გათვალისწინებითა და უკვე არსებული მორფოლოგიური ანოტირების საფუძველზე, ანალიზის პროცესს შეუძლია იმ აქტანტების (თუნდაც ჰიპოთეტური) გამოყოფა, რომლებიც მოცემული ზმნის კონტექსტში შესაბამის სემანტიკურ როლებს გამოხატავენ. ეს განაპირობებს მარტივი წინადადების ცენტრალური სტრუქტურის _ ე.წ. წინადადების „ბირთვის“ (Core structure) დადგენასა და მის გამიჯვნას „პერიფერიული“ (periphery) კომპონენტისგან, რაც საბოლოოდ მეტად მნიშვნელოვანი ეტაპია წინადადების სტრუქტურის დასადგენად. ამგვარად, აქტანტთა გაფორმების ეს კანონზომიერება და ქართული მორფოლოგიის სპეციფიკური სიმდიდრე უზრუნველყოფს გამონათქვამის მორფოლოგიური ანოტირების უშუალო გადა-

სვლას როლებრივ სტრუქტურაზე, რომლის ჩამოყალიბება შინაარსის სფეროში უფრო ღრმა „შეჭრად“ გვესახება, ვიდრე ის, რაც ჩვეულებრივ სინტაქსურ წარმოდგენას ახასიათებს.

მორფოლოგიურ ინფორმაციასა და როლებრივ სტრუქტურას შორის ასეთი მიმართების არსებობა მიგვანიშნებს ანალიზური პროცესის პარალელური წარმართვის შესაძლებლობაზეც. სახელდობრ, შეიძლება ვიფიქროთ პროცესის განვითარების ისეთ სქემაზე, რომლის თანახმად, მომავალ როლებრივ სტრუქტურაში ყოველი სიტყვის მორფოლოგიურ ანალიზს ჰიპოთეზების ჩამოყალიბება მოჰყვება. აშკარაა, რომ ასეთი მიდგომა უფრო ახლოა ბუნებრივი ენობრივი სისტემის გამოყენებასთან (მეტყველების მოსმენა ან ტექსტის წაკითხვა) და ამგვარად, ენის მოდელირების ძირითად, ფუნდამენტურ ამოცანასთანაც.

The Pattern of the Analysis of Georgian Sentence

George Chikoidze

Institute of Control Systems, Technical University (Georgia)
gogichikoidze@yahoo.com

The analysis of a sentence represents one of the directions of the meaning↔text model functioning. As different from the synthetic direction, the initial point of the process of analysis is represented by a graphical (or acoustic) representation subordinated to immediate perception. Specifically, a text is shown as a coordination of symbols where textual word forms are represented in a chain separated by space bars, some of them being highlighted by specific symbols or punctuation marks. Such a sharp separation of word forms first of all simplifies the analysis of a text in comparison with speech in which continuous chains of phonemes often represent a group of some words and not separate word form.

Such kind of a peculiarity provides an opportunity for an immediate and unequivocal identification of each textual word form with a lexical unit, followed by certain grammatical and semantic characteristics. As for grammar characteristics, their determination is maximally simplified in the context of the dictionary, including the generator of a morphological paradigm of each unit, or the system, generating a list of paradigm forms of given unit members, coordinated in accordance with their grammatical characteristics, as a result of which absolute identity of input textual word form with a member of the list will be equal to establishment of grammar characteristics of the form (that is what Georgian Computer Dictionary is like, created in 2009-2011 in the Institute of Control Systems, supported by Rustaveli Foundation).

Morphological annotation is a rather important aspect of analysis, provides a basis for furthering of the process of analysis to its ideal (yet not achieved) aim or immediate representation of „meaning“. The next important step is to establish the whole structure of a sentence on the basis of already procured morphologic information which is characteristic to separate elements (word forms).

Structure building and representation is made differently (for instance, by the graph of immediate constituents).

We do not follow the scheme of the role organization, depending, on the one hand, on the semantic roles and on the notion of the verb super-paradigm, on the other. The latter (super-paradigm) represents the sum of paradigms which are derived from the one and the same verbal lexeme that can be considered as a reflectance of a certain virtual ('global') process developing.

In spite of stability of the content of semantic roles in the texts, actants expressing them experience some changes in the context of various paradigms. However, variation of their morphologic forming depends on some kind of regularity. The characteristics of varying is connected with the type of a super-paradigm, then with the exact member paradigm to which the verb belongs to and finally it is defined by the series to which the verb row belongs to.

That kind of relations between the morphological information and the role structure indicates to the opportunity of parallel building of the process of analysis. Specifically, we may think about the scheme of process developing according to which in the further role structure the morphological analysis of each word will be followed by formation of a hypothesis in its probable place. It is obvious that such an approach is closer to the usage of the system of a natural language (listen to discourse or read a text) and, thus, to the basic fundamental task of language modeling.

კორპუსის შედგენა – წარსული, აწმყო და სამომავლო პერსპექტივები

ჯინ ჰადსონი
მალმეს უნივერსიტეტი (შვედეთი)
jean.hudson@mah.se

„რეალურად არსებული ენის" ელექტრონულად შენახული მონაცემების გამოყენება სანათმეცნიერო კვლევაში ჯერ კიდევ საყმაწვილო ასაკშია. მე-20 საუკუნის 80-იან წლებში – კომპიუტერული ეპოქის გარიჟრაჟზე – ჩვენ გვაოცებდა ახალი ხედვა: ლექსიკონების შემდგენლებს განსაკუთრებით ხიბლავდა კორპუსების გამოყენების შესაძლებლობა სიტყვებისა და ფრაზების უფრო ზუსტი განმარტების ჩამოსაყალიბებლად ნებისმიერ ენაზე (თუმცა იმ დროს ეს ძირითადად ინგლისურს ეხებოდა). ეს კარგი დასაწყისი იყო, მაგრამ ამჟამად ჩვენ უფრო დახვეწილად უნდა გამოვიყენოთ კორპუსის შეუფასებელი შესაძლებლობები.

რადგანაც ამ კონფერენციის მთავარი მიზანია იდეებისა და გამოცდილების გაცვლა–გამოცვლა ქართული ენის კომპიუტერული რესურსების განვითარებაზე განსაკუთრებული აქცენტით, მე მინდა, გაგიზიაროთ ზოგიერთი მოსაზრება კორპუსის შედგენის გზების შესახებ და იმ ძიების შესახებ, რომელიც ემყარებოდა ადრეული კორპუსული კოლექციების შექმნისა და კვლევის წარმატებებსა თუ წარუმატებლობებს ინგლისურ ენასთან მიმართებით.

განსაკუთრებულ ყურადღებას მივაქცევთ სამეტყველო მასალის შეგროვების მნიშვნელობას. საშუალო მოქალაქეს ყოველდღიურად აქვს ენის ცოცხალი განცდა, რამდენადაც ის უფრო მეტად მეტყველებს ან ისმენს, ვიდრე წერს ან კითხულობს. ამას ის ფაქტიც უნდა დავუმატოთ, რომ ენის ევოლუცია მიმდინარეობს ზეპირმეტყველების და არა წერის გზით (ეს უკანასკნელი Homo Sapiens-ის ისტორიაში შედარებით გვიანდელი გამოგონებაა) და აშკარა ხდება, რომ ზეპირმეტყველებითი მასალის შემცველი სულ უფრო და უფრო მოცულობითი კორპუსები გვჭირდება. ზეპირმეტყველებითი მასალის კორპუსული კვლევა გვიჩვენებს არა მარტო იმას, როგორ არის აგებული ესა თუ ის ენა, არამედ ასევე გვაწვდის ფასეულ ინფორმაციას ენისა და შემეცნების ურთიერთობის შესახებ.

მსურს, გაგიზიაროთ გამოცდილება, რომელიც მივიღე, როგორც კორპუსის შემდგენელმა (განსაკუთრებით ეს ეხება კორპუსს CANCODE – ინგლისური დისკურსის კემბრიჯული და ნოტინგემური კორპუსი) და როგორც სამეტყველო მასალის კოგნიტიური თვალსაზრისით მკვლევარმა.

Corpus Compilation - Past, Present, and Future Perspectives

Jean Hudson
Malmö University (Sweden)
jean.hudson@mah.se

The use of electronically stored 'real language' data for linguistic investigation is still in its infancy. In the 1980s – the dawn of the computer era - we were amazed by the insights that were pouring in: dictionary-makers were especially attracted by the way in which corpora could be put to use in a more accurate description of what words and phrases mean, in whatever language (although mostly the English language, in those days). This was a good beginning, but now we need to make more sophisticated use of this invaluable tool.

Since the main aim of this conference is to exchange ideas and experiences, with particular regard to the development of computational resources for the Georgian language, I would like to share some ideas regarding future pathways for corpus compilation and investigation based on the successes and failures of earlier corpus collection and research in the English language.

In particular, I will be highlighting the importance of the collection of speech data. Most language that is experienced by the average citizen daily is spoken or heard – not written or read. Add to this the fact that language evolution happens through speech, not writing (which is a relatively new invention in the history of Homo Sapiens) and it becomes evident that we need more and larger corpora of spoken data. Corpus investigation of spoken data shows us not only how a particular language is structured, it also gives us valuable insight into the relationship between language and cognition.

I would like to share some of the experience that I have gained as a corpus compiler (in particular CANCODE – Cambridge and Nottingham Corpus of Discourse in English) and as a researcher of speech data from a cognitive perspective.

პირთა საძიებელი

25. ზაალ კიკვიძე
zaalk@yahoo.com
გვ. 71
26. გიორგი კილაძე
giorgi_x2000@yahoo.com
გვ. 73
27. მანანა კობაიძე
kobaidze@comhem.se
გვ. 74
28. მიხაილ კოპოტევი
mihail.kopotev@helsinki.fi
გვ. 77
29. ირინა ლობჟანიძე
irina_lobzhanidze@iliauni.edu.ge
გვ. 82
30. ნანა ლოლაძე
nanaloladze@yahoo.com
გვ. 40
31. მაია ლომია
maia-lomia@mail.ru
გვ. 78
32. ლიანა ლორთქიფანიძე
l.lordkipanidze@yahoo.com
გვ. 21
33. თინათინ მარგალიტაძე
tinatin@margaliti.ge
გვ. 83, 85
34. მარიამ მანჯგალაძე
mariam@ice.ge
გვ. 37, 50
35. თამარ მახარობლიძე
ateni777@yahoo.com
გვ. 88
36. დავით ნადარაია
dnad@itex.ge
გვ.21
37. ელენე ნაპირელი
elene.napireli@gmail.com
გვ. 91
38. ოლგა ნევზოროვა
nevzoro@gmail.com
გვ. 105
39. მაია ნინიძე
maianinidze@yahoo.com
გვ. 94
40. სერგეი პოტიომკინი
prolexprim@gmail.com
გვ. 97
41. ნინო სანაია
nsanaia@yahoo.com
გვ. 99
42. დიმიტრი სიჭინავა
mitrius@gmail.com
გვ. 102
43. ჯავდეთ სულეიმანოვი
alsu_73@list.ru
გვ. 105
44. ნარგიზა სურმავა
nargizasurmava@yahoo.com
გვ. 109
45. რატი სხირტლაძე
rati2008@gmail.com
გვ. 31, 50
46. თედო უთურგაიძე
mariam@ice.ge
გვ. 50
47. ნინო ფანცხავა
Nino.p99@gmail.com
გვ. 29
48. ზაქარია ფურცხვანიძე
pourtskhvanidze@em.uni-frankfurt.de
გვ. 60
49. გიორგი ქერეჭაშვილი
hello@dictionary.ge
გვ. 83, 85
50. ირაკლი ღარიბაშვილი
igar@hotmail.com
გვ. 29

51. მარიამ ყამარაული
mariam_kamarauli@hotmail.de
გვ. 112
52. ნათია შენგელაია
nshengelaia@yandex.ru
გვ. 73
53. გიორგი ჩიკოიძე
gogichikoidze@yahoo.com
გვ. 114
54. ბულატ ხაკიმოვი
khakeem@yandex.ru
გვ. 105
55. ჯინ ჰადსონი
jean.hudson@mah.se
გვ. 117

List of Participants

1. Lia Bakuradze
l.bakuradze@gmail.com
p. 110
2. Maia Barikhashvili
maiahereti@yahoo.com
p. 17
3. Larisa Beliaeva
Iauranbel@gmail.com
p. 20
4. Khatuna Beridze
beridze@illinois.edu
p. 27
5. Marina Beridze
marine.beridze@gmail.com
p. 23, 110
6. Manana Bukia
manbuki@rambler.ru
p. 30
7. George Chikoidze
gogichikoidze@yahoo.com
p. 116
8. Sophiko Daraselia
sopod@yahoo.com
p. 46
9. Ketevan Datukishvili
k_datukishvili@yahoo.com
p. 42
10. Nino Doborjginidze
nino_doborjginidze@iliauni.edu.ge
p. 49
11. Lali Ezugbaia
ezugbaia@ice.ge
p. 52
12. Kakha Gabunia
kgabunia@cciir.ge
p. 33
13. Irakli Garibashvili
igar@hotmail.com
p. 30
14. Rusudan Gersamia
rgersamia@iliauni.edu.ge
p. 80
15. Jost Gippert
gippert@em.uni-frankfurt.de
p. 36
16. Ketevan Gochitashvili
ketevan.gochitashvili@tsu.ge
p. 39
17. Natia Gorgadze
ngiorgadze@cciir.ge
p. 33
18. Jean Hudson
jean.hudson@mah.se
p. 118
19. Grigol Kakhiani
gkakhiani@gmail.com
p. 27
20. Mariam Kamarauli
mariam_kamarauli@hotmail.de
p. 113
21. Vidas Kavaliauskas
vk1119@gmail.com
p. 66
22. Giorgi Keretchashvili
hello@dictionary.ge
p. 84, 87
23. Bulat Khakimov
khakeem@yandex.ru
p. 107
24. Zaal Kikvidze
zaalk@yahoo.com
p. 72

25. George Kiladze
giorgi_x2000@yahoo.com
p. 74
26. Manana Kobaidze
kobaidze@comhem.se
p. 76
27. Mikhail Kopotev
mihail.kopotev@helsinki.fi
p. 78
28. Tsitsino Kvantaliani
tsitsino.nino.kvantaliani@gmail.com
p. 69
29. Rusudan Landia
landiarus@gmail.com
p. 69
30. Irina Lobzhanidze
irina_lobzhanidze@iliauni.edu.ge
p. 82
31. Nana Loladze
nanaloladze@yahoo.com
p. 42
32. Maia Lomia
maia-lomia@mail.ru
p. 80
33. Liana Lortkipanidze
l.lordkipanidze@yahoo.com
p. 23
34. Tamar Makharoblidze
ateni777@yahoo.com
p. 90
35. Mariam Manjgaladze
mariam@ice.ge
p. 39, 52
36. Tinatin Margalitadze
tinatin@margaliti.ge
p. 84, 87
37. David Nadaraia
dnad@itex.ge
p. 23
38. Elene Napireli
elene.napireli@gmail.com
p. 92
39. Olga Nevzorova
onevzoro@gmail.com
p. 107
40. Maia Ninidze
maianinidze@yahoo.com
p. 95
41. Nino Pantskhava
Nino.p99@gmail.com
p. 30
42. Zakharia Pourtskhvanidze
pourtskhvanidze@em.uni-frankfurt.de
p. 62
43. Sergey Potemkin
prolexprim@gmail.com
p. 98
44. Nino Sanaia
nsanaia@yahoo.com
p. 101
45. Natia Shengelaia
nshengelaia@yandex.ru
p. 74
46. Dmitry Sitchinava
mitrius@gmail.com
p. 104
47. Rati Skhirtladze
rati2008@gmail.com
p. 33, 52
48. Dzhavdet Suleymanov
alsu_73@list.ru
p. 107
49. Nargiza Surmava
nargizasurmava@yahoo.com
p. 110
50. Manana Tandaschwili
tandaschwili@em.uni-frankfurt.de
p. 62

51. Tedo Uturgaidze
mariam@ice.ge
p. 52
52. Karina Vamling
karina.vamling@mah.se
p. 57
53. Ruprecht von Waldenfels
waldenfels@issl.unibe.ch
p. 55
54. Jolanta Zabarskaite
jolanta.zabarskaite@lki.lt
p.59
55. Merab Zakalashvili
GILCE@Wanex.ge
p.42

THE INTERNATIONAL CONFERENCE

GEORGIAN LANGUAGE AND MODERN TECHNOLOGIES – 2013

საერთაშორისო კონფერენცია

ქართული ენა და თანამედროვე ტექნოლოგიები – 2013

THE INTERNATIONAL CONFERENCE

GEORGIAN LANGUAGE AND MODERN TECHNOLOGIES – 2013

საერთაშორისო კონფერენცია

ქართული ენა და თანამედროვე ტექნოლოგიები – 2013

THE INTERNATIONAL CONFERENCE

GEORGIAN LANGUAGE AND MODERN TECHNOLOGIES – 2013

საერთაშორისო კონფერენცია

ქართული ენა და თანამედროვე ტექნოლოგიები – 2013