
LANGUAGE AND MODERN
TECHNOLOGIES

საორგანიზაციო კომიტეტი:

ავთანდილ არაბული – თსუ არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველო)

მანანა თანდაშვილი – გოეთეს სახელობის ფრანკფურტის უნივერსიტეტი (გერმანია)

მარინა ბერიძე – თსუ არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველო)

დარეჯან თვალთვაძე – ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)

რუსუდან პაპიაშვილი – თსუ არნ. ჩიქობავას სახელობის ენათმეცნი ერების ინსტიტუტი (საქართველო)

მარიამ ყამარაული – გოეთეს სახელობის ფრანკფურტის უნივერსიტეტი (გერმანია)

სარედაქციო კოლეგია:

ლია ბაკურაძე – თსუ არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველო)

ზაალ კიკვიძე – თსუ არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველო)

თინათინ მარგალიტაძე – ივ. ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)

ჯინ ჰადსონი – მალმეს უნივერსიტეტი, ენისა და ლინგვისტიკის დეპარტამენტი (შვედეთი)

ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი
გოეთეს სახელობის ფრანკფურტის უნივერსიტეტი
არნოლდ ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი

საერთაშორისო კონფერენცია

ენა და თანამედროვე ტექნოლოგიები – 2015

თბილისი
2015

Organizing Committee:

Avtandil Arabuli - Arnold Chikobava Institute of Linguistics, TSU, Georgia

Manana Tandashvili - Goethe University Frankfurt am Main, Germany

Marina Beridze - Arnold Chikobava Institute of Linguistics, TSU, Georgia

Darejan Tvaltvadze - Ivane Javakhishvili Tbilisi State University, Georgia

Rusudan Papiashvili - Arnold Chikobava Institute of Linguistics, TSU, Georgia

Mariam Kamarauli - Goethe University Frankfurt am Main, Germany

Editorial Board:

Lia Bakuradze - Arnold Chikobava Institute of Linguistics, TSU, Georgia

Jean Hudson – Malmö University, Department of Language and Linguistics, Sweden

Zaal Kikvidze - Arnold Chikobava Institute of Linguistics, TSU, Georgia

Tinatin Margalitadze - Ivane Javakhishvili Tbilisi State University, Georgia

Ivane Javakhishvili Tbilisi State University
Goethe University Frankfurt an Main
Arnold Chikobava Institute of Linguistics

International Conference

Language and Modern Technologies – 2015

Tbilisi
2015

მუშაობის განრიგი

10 სექტემბერი

საქართველოს პარლამენტი

თბილისი, რუსთაველის 8

09.00-10.00 – მონაწილეთა რეგისტრაცია
10.00-11.30 _ კონფერენციის გახსნა
11.30-12.00 _ შესვენება
12.00-14.45 – I პლენარული სხდომა
14.45-15.30 _ შესვენება
15.30-18.10 _ II პლენარული სხდომა

11 სექტემბერი

საქართველოს პარლამენტის ეროვნული ბიბლიოთეკის საგამოფენო დარბაზი

თბილისი, გუდიაშვილის 7

10.00-11.15 –სექცია 1
11.15-11.45 – შესვენება
11.45-13.25 _ სექცია 2
13.25-13.45 _ შესვენება
13.45-14.45 _ საჯარო ლექცია
14.45-15.45 _ შესვენება
15.45-17.15 _ სექცია 3 – პრეზენტაციები
17.15 _ შეხვედრა უცხოელ მეცნიერებთან

12 სექტემბერი

საქართველოს პარლამენტის ეროვნული ბიბლიოთეკის საგამოფენო დარბაზი

თბილისი, გუდიაშვილის 7

10.00-11.15 –სექცია 4
11.15-11.45 – შესვენება
11.45-13.50 –სექცია 5
13.50-14.50 – შესვენება
14.50-17.20 – სექცია 6 – პრეზენტაციები
17. 20-17.45 – შესვენება
17.45 – მრგვალი მაგიდა

რეგლამენტი
პლენარული მოხსენება – 40 წუთი
პრეზენტაცია – 30 წუთი
მოხსენება – 25 წუთი
მსჯელობა –10 წუთი

Conference Schedule

10 September

Parliament of Georgia

Tbilisi, Rustaveli str.8

09.00-10.00 – Registration of participants
10.00-11.30 – Opening of the conference
11.30-12.00 – Break
12.00-14.45 – Plenary Session 1
14.45-15.30 – Lunch
15.30-18.10 – Plenary Session 2

11 September

Exhibition Hall, National Parliamentary Library of Georgia

Tbilisi, Gudiashvili str. 7

10.00-11.15 – Session 1
11.15-11.45 – Break
11.45-13.25 – Session 2
13.25-13.45 – Break
13.45-14.45 – Public lecture
14.45-15.45 – Break
15.45-17.15 – Session 3 – Presentations
17.15 – Meeting with international scholars

12 September

Exhibition Hall, National Parliamentary Library of Georgia

Tbilisi, Gudiashvili str. 7

10.00-11.15 – Session 4
11.15-11.45 – Break
11.45-13.50 – Session 5
13.50-14.50 – Break
14.50-17.20 – Session 6 – Presentations
17.20-17.45 – Break
17.45 – Round table

Time-limit
Plenary session – 40 minutes
Presentation – 30 minutes
Paper – 25 minutes
Discussion – 10 minutes

პროგრამა

PROGRAM

მუშაობის განრიგი

10 სექტემბერი

საქართველოს პარლამენტი

ილია ჭავჭავაძის დარბაზი

თბილისი, რუსთაველის 8

09.00-10.00	**კონფერენციის მონაწილეთა რეგისტრაცია**
10.00-11.30	**კონფერენციის გახსნა** სხდომის თავმჯდომარე: **ივანე კიღურაძე** – საქართველოს პარლამენტის განათლების, მეცნიერებისა და კულტურის კომიტეტის თავმჯდომარე
10.00-10.45	მისალმებები: **დავით უსუფაშვილი** – საქართველოს პარლამენტის თავმჯდომარე **თამარ სანიკიძე** – საქართველოს განათლებისა და მეცნიერების მინისტრი **იოსტ გიპერტი** – ფრანკფურტის გოეთეს უნივერსიტეტის ემპირიული ენათმეცნიერების ინსტიტუტის დირექტორი **მარინა ჩიტაშვილი** – შოთა რუსთაველის ეროვნული სამეცნიერო ფონდის გენერალური დირექტორი
10.45-11.10	**მ. თანდაშვილი** – ქართული ეროვნული კორპუსი (ახალი ვერსიის პრეზენტაცია)
11.10-11.30	**მ. ბერიძე** – ქართული დიალექტური კორპუსი (ახალი ვერსიის პრეზენტაცია)
11.30-12.00	შესვენება
12.00-14.45	**I პლენარული სხდომა** სხდომის თავმჯდომარეები: **ი. გურევიჩი, ა. არაბული**
12.00-12.40	**ი. გიპერტი** – ქართველოლოგია დიგიტალური ჰუმანიტარიის კონტექსტში
12.40-13.20	**ტ. მაკენერი** – კოლოკაციები და კონტექსტი – კოლოკაცია და ქსელები
13.20-13.50	**ზ. კირტავა** – საკანონმდებლო ტერმინების ელექტრონული ლექსიკონის (თეზაურუსის) შექმნა საქართველოს პარლამენტის საკანონმდებლო ინფორმაციის მართვის სისტემის განვითარების პროექტის ფარგლებში
13.50-14.20	**დიტერ ვან აიტფანკი** – CLARIN - ისტორია და პერსპექტივები
14.20-14.45	მსჯელობა

14.45-15.30	შესვენება
15.30-18.10	**II პლენარული სხდომა** სხდომის თავმჯდომარეები: **ი. გიპერტი, თ. მარგალიტაძე**
15.30-16.10	**ი. გურევიჩი** – თანამედროვე ტენდენციების კვალდაკვალ: ენობრივი რესურსები და ინსტრუმენტები რესურსებით ნაკლებად უზრუნველყოფილი ენებისათვის.
16.10-16.50	**ე. ჰარდი** – მარკირება მეტყველების ნაწილთა მიხედვით სხვადასხვა ტიპის ენაში: რამდენიმე თეორიული საფუძველი მორფოსინტაქსური ანოტირების სქემებისათვის
16.50-17.30	**ს. შაროვი** – ვარირების შესწავლა დიდ ინტერნეტკორპუსებში ლექსიკოგრაფიული აღწერის ამოცანის გადასაჭრელად
17.30-18.10	**პ. მოირერი** – ქართული ენის მორფოსინტაქსური ანალიზი - მეთოდები და გამოწვევები

11 სექტემბერი

საქართველოს პარლამენტის ეროვნული ბიბლიოთეკის საგამოფენო დარბაზი

თბილისი, გუდიაშვილის 7

10.00-11.15	**სექცია 1** სხდომის თავმჯდომარეები: **ს. შაროვი, ლ. ეზუგბაია**
10.00-10.25	**თ. მარგალიტაძე, ი. ორმოცაძე** – ინგლისურ-ქართული სამეცნიერო ტექსტების პარალელური კორპუსის პლატფორმა და დარგობრივი ლექსიკოგრაფია
10.25-10.50	**ქ. დათუკიშვილი, ნ. ლოლაძე, მ. ზაკალაშვილი** – ქართული ენის ელექტრონული ლექსიკონის შედგენის პრინციპებისათვის
10.50-11.15	**კ. გაბუნია** – ქართულ-ინგლისური მანქანური თარგმანის ერთი საკვანძო საკითხისათვის (ზმნური სიტყვაფორმების შესატყვისობა ქართულსა და ინგლისურში)
11.15-11.45	შესვენება
11.45-13.25	**სექცია 2** სხდომის თავმჯდომარეები: **ე. ჰარდი, ნ. დობორჯგინიძე**
11.45-12.10	**ს. დარასელია, ს. შაროვი** – ქართული ენის მორფოლოგიური ანოტირებისას გამოვლენილი შეცდომების ანალიზი.
12.10-12. 35	**ი. ლობჟანიძე** – ქართული ენის ანალიზატორის გაუმჯობესებისა და განვითარების პერსპექტივები ქართული ენის კორპუსის საფუძველზე
12.35-13.00	**მ. ყამარაული** – დეტერმინატორები და მოდიფიკატორები ძველ და ახალ ქართულში

13.00-13.25	**მ. ბერიძე, ლ. ლორთქიფანიძე, დ. წადარაია** – ქართული დიალექტური კორპუსის მორფოლოგიური ანოტირების კონცეფციისათვის
13.25-13.45	შესვენება
13.45-14.45	საჯარო ლექცია **ტ. მაკენერი** – Graphcoll
14.45-15.45	შესვენება
15.45-17.45	**სექცია 3** – პრეზენტაციები
15.45-16.15	**ლ. ეზუგბაია, ლ. ბაკურაძე, ნ. სურმავა, მ. ბარიხაშვილი** – პროექტი „ქართული ენა საზღვარგარეთ – ქართული დიალექტები და ლაზური თურქეთში, ირანსა და აზერბაიჯანში“
16.15-16.45	**ნ. დოხორჯგინიძე** – ქართული ენის კორპუსის კონცეფცია
16.45-17.15	**მ. მანჯგალაძე, ქ. გოჩიტაშვილი** – ქართული ენის ელექტრონული სწავლების კურსი - ახალი ვერსია (A1 – B2 დონეები) და პროგრამის განვითარების პერსპექტივა
17.15	**მ. ჩიტაშვილი** – შეხვედრა უცხოელ მეცნიერებთან (შოთა რუსთაველის ეროვნული სამეცნიერო ფონდი)

12 სექტემბერი

საქართველოს პარლამენტის ეროვნული ბიბლიოთეკის საგამოფენო დარბაზი

თბილისი, გუდიაშვილის 7

10.00-11.15	**სექცია 4** სხდომის თავჯდომარეები: **პ. მოირერი, მ. მანჯგალაძე**
10.00-10.25	**ნ. სურმავა, ც. კვანტალიანი, მ. კიკონიშვილი, მ. ბერიძე** – ქდკ-ს ავტომატური ანოტირების შედეგების ანალიზი (იმერული, აჭარული, კახური დიალექტების მასალის მიხედვით)
10.25-10.50	**ნ. სანაია** – მეტაფორული შესიტყვებების მიკროსემანტიკური მოდელირების პრობლემები (ზმნური მეტაფორული შესიტყვებების მასალაზე დაყრდნობით)
10.50-11.15	**ლ. ბაკურაძე, მ. ბერიძე** – „დიალექტური კუნძულის“ ლინგვოკულტურული სივრცის მოდელირება და პრეზენტაცია ქდკ–ში (ფერეიდნული დიალექტი)
11.15-11.45	შესვენება

11.45-13.50	**სექცია 5** სხდომის თავჯდომარეები: **ნ. ლოლაძე, ნ. სანაია**
11.45-12.10	**ც. ხახვიაშვილი, ნ. ბილანიშვილი** – „ქართლის ცხოვრების“ პარალელური (ქართულ-სომხური) კორპუსი
12.10-12.35	**მ. ბარიხაშვილი, ე. ნაპირელი, რ. პაპიაშვილი** – ქდკ-ს ინგილოური ლექსიკონის მიმართება ლექსიკოგრაფიულ წყაროებთან
12.35-13.00	**დ. ანფიმიადი** – ქართული დიალექტური კორპუსი, როგორც სასწავლო-საგანმანათლებლო რესურსი სასკოლო ჰუმანიტარული სწავლებისათვის
13.00-13.25	**თ. კალხიტაშვილი** – საქართველოს ეპიგრაფიკული ძეგლების კორპუსი (კორპუსის შედგენილობა და ელექტრონული გამოცემის სტანდარტი)
13.25-13.50	**ნენა ნვოსუ-ნვორუ** – Linguascript: მეცნიერების გამოყენება ენის განვითარებისათვის
13.50-14.50	შესვენება
14.50-17.20	**სექცია 6** – პრეზენტაციები
14.50-15.20	**კ. ფხაკაძე, მ. ჩიქვინიძე, გ. ჩიჩუა, ი. ბერიაშვილი, დ. კურცხალია** – „კიდევ ერთი ნაბიჯი მოსაუბრე ქართული თვითგანვითარებადი ინტელექტუალური კორპუსისაკენ“ - მიზნები და პირველი შედეგები
15.20-15.50	**ნ. ამირეზაშვილი, რ. ერემიანი, ლ. ლორთქიფანიძე, ლ. სამსონაძე, გ. ჩიკოიძე, ა. ჩუტკერაშვილი, ნ. ჯავაშვილი** – ქართული ენის კომპიუტერული მოდელები
15.50-16.20	**თ. მახარობლიძე** – ქართული ჟესტური ენის დოკუმენტირება
16.20-16.50	**მ. ტურაშვილი** – ქართული ხალხური ცხოველთა ზღაპრების ელექტრონული კატალოგი
16.50-17.20	**დ. თვალთვაძე, მ. მადუაშვილი, ე. კვირკველია** – ელექტრონული კურსებისა და ტექსტური ბაზების გამოყენება სწავლების პროცესში (თსუ ჰუმანიტარულ მეცნიერებათა ფაკულტეტის გამოცდილება)
17.20-17.45	შესვენება
17.45	მრგვალი მაგიდა – ენის ტექნოლოგიები და დიგიტალური ჰუმანიტარია მოდერატორი : **იოსტ გიპერტი**

სემინარი „კორპუსის შექმნის მეთოდოლოგიური საფუძვლები“

საქართველოს პარლამენტის ეროვნული ბიბლიოთეკის საგამოფენო დარბაზი

თბილისი, გუდიაშვილის 7

13 სექტემბერი

14.00-14.30	სემინარის გახსნა (ი. გიპერტი, მ. თანდაშვილი)
14.30-16.00	**ი. გურევიჩი** – ანოტაციის მეთოდები
16.00-17.00	**ე. ჰარდი** – მორფოსინტაქსური ანოტირების სქემები
17.00	**მ. თანდაშვილი** – ქართული ენის ეროვნული კორპუსი

14 სექტემბერი

10.00-12.00	**ნ. ბუბენჰოფერი** – თემატური კორპუსები
12.00-13.00	შესვენება
13.00 -15.00	**ე. ტაიხი** – კორპუსების შექმნა
15.00-16.00	შესვენება
16.00	**ი. გიპერტი** – კორპუსის შექმნის საფუძვლები

15 სექტემბერი

10.00-12.00	**უ. კვასტჰოფი** – ვრცელი ტექსტური კორპუსები და მათი გამოყენება ლექსიკოგრაფიაში
12.00-13.00	შესვენება
13.00-15.00	**პ. მოირერი** – ლემატიზაცია და ომონიმიის მოხსნა
15.00-16.00	შესვენება
16.00	შემაჯამებელი დისკუსია

Conference Schedule

September 10

Parliament of Georgia
Ilia Chavchavadze Chamber

Tbilisi, Rustaveli st. 8

09.00-10.00	**Registration of conference participants**
10.00-11.30	**Opening**
	Chairperson: **Ivane Kiguradze** – Chair of the Science, Education and Culture Committee of the Parliament of Georgia
10.00-10-45	Welcoming speeches: **Davit Usupashvili** – Chairman of the Parliament of Georgia **Tamar Sanikidze** – Minister of Education and Science of Georgia **Jost Gippert** – Director, Goethe University Frankfurt/Main, Institute for Empirical Linguistics **Marina Chitashvili** – Director General of Shota Rustaveli National Science Foundation
10.45-11.10	**M. Tandashvili** – Georgian National Corpus (New version)
11.10-11.30	**M. Beridze** – Georgian Dialect Corpus (New version)
11.30-12.00	Break
12.00-14.45	**Plenary Session 1** Chairpersons: **I. Gurevych, A. Arabuli**
12.00-12.40	**J. Gippert** – Kartvelology in the Age of Digital Humanities
12.40-13.20	**T. McEnery** – Collocations and Context – Collocation and Collocation Networks
13.20-13.50	**Z. Kirtava** – Compilation of the Electronic Dictionary (Thesaurus) of Legislative Terms within the Framework of the Project "Development of Legislative Information Management System for the Parliament of Georgia"
13.50-14.20	**D. van Uytvanck** – CLARIN
14.20-14.45	Discussion
14.45-15.30	Break
15.30-18.10	**Plenary Session 2** Chairpersons: **J. Gippert**, **T. Margalitadze**
15.30-16.10	**I. Gurevych** – Catching up with the trends: language resources and tools for less-resourced languages

16.10-16.50	**A. Hardie** – Part-of-speech tagging in different kinds of language: some theoretical bases for morphosyntactic annotation schemata
16.50-17.30	**S. Sharoff** – Studying variation in large web-corpora for the task of lexicographic description
17.30 -18.10	**P. Meurer** – Morphosyntactical analyses of Georgian – methods and challenges

September 11

Exhibition Hall, National Parliamentary Library of Georgia

Tbilisi, Gudiashvili st. 7

10.00-11.15	**Session 1** Chairpersons: **S. Sharoff, L. Ezugbaia**
10.00-10.25	**T. Margalitadze, I. Ormotsadze** – The Platform of the English-Georgian Parallel Corpus of Scientific Texts and Specialized Lexicography
10.25-10.50	**K. Datukishvili, N. Loladze, M. Zakalashvili** – On the Principles of Compilation of the Electronic Dictionary of Georgian
10.50-11.15	**K. Gabunia** – One Key Issue of Georgian-English Machine Translation (Equivalence of verbal word-forms in Georgian and English)
11.15-11.45	Break
11.45-13.00	**Session 2** Chairpersons: **A. Hardie, N. Doborjginidze**
11.45-12.10	**S. Daraselia, S. Sharoff** – Error Analyses in Part-of-Speech Tagging in Georgian
12. 10-12. 35	**I. Lobzhanidze** – Improvement of the Georgian Morphological Analyzer on the basis of the Corpus of Modern Georgian Language
12.35-13.00	**M. Kamarauli** – Determiners and Modifiers in Old and Modern Georgian
13.00-13.25	**M. Beridze, L. Lordkipanidze, D. Nadaraia** – On the Morphological Concept of the Morphological Annotation of the Georgian Dialect Corpus
13.25-13.45	Break
13.45-14.45	Public lecture **T. McEnery** – Graphcoll
14.45-15.45	Break
15.45-17.15	**Session 3** – Presentations
15.45-16.15	**L. Ezugbaia, L. Bakuradze, N. Surmava, M. Barikhashvili** – Project: The Georgian Language Abroad – Georgian Dialects and Laz in Turkey, Iran and Azerbaijan

16.15-16.45	**N. Doborjginidze** – Georgian Language Corpus: Concept and Methodology
16.45-17.15	**M. Manjgaladze, K. Gochitashvili** – ELearning Course of Georgian –New Version (A1-B2 Levels) and Perspectives for Course Development
17.15	**Marina Chitashvili** – Meeting with international scholars (Shota Rustaveli National Science Foundation)

September 12

Exhibition Hall, National Parliamentary Library of Georgia

Tbilisi, Gudiashvili st. 7

10.00-11.15	**Session 4** Chairpersons: **P. Meurer, M. Manjgaladze**
10.00-10.25	**N. Surmava, Ts. Kvantaliani, M. Kikonishvili, M. Beridze** –Analysis of the Results of the Automated Annotation of GDC (Based on the data of Imeretian, Acharan, Kakhetian dialects)
10.25-10.50	**N. Sanaia** – The Challenges of Micro-Semantic Modeling of Metaphoric Collocations (Based on the Data of Verbal Metaphoric Collocations)
10.50-11.15	**L. Bakuradze, M. Beridze** – Modeling and Presenting of the Lingua-cultural Area of a "Dialect Island" in GDC (Fereidanian Dialect)
11.15-11.45	Break
11.45-13.50	**Session 5** Chairpersons: **N. Loladze, N. Sanaia**
11.45-12.10	**Ts. Khakhviashvili, N. Bilanishvili** – Parallel (Georgian-Armenian) Corpus of Kartlis Tskhovreba (Life of Kartli)
12.10-12.35	**M. Barikhashvili, E. Napireli, R. Papiashvili** – The Relationship of the Ingiloan Dictionary of GDC to Lexicographic Sources
12.35-13.00	**D. Anphimiadi** – Georgian Dialect Corpus as an Educational Resource for Teaching the Humanities at School
13.00-13.25	**T. Kalkhitashvili** – Epigraphic Corpora of Georgia's Inscriptions (Corpus Structure and the Standard of Electronic Edition)
13.25-13.50	**N. Nwosu-Nworuh** – Linguascript: Applying Science in Language Development
13.50-14.50	Break
14.50-17.20	**Session 6** - Presentations:
14.50-15.20	**K. Pkhakadze, M. Chikvinidze, G. Chichua, I. Beriashvili, D. Kurtskhalia** – The Aims and First Results of the Project "One More Step Towards Georgian Talking Self-Developing Intellectual Corpus"

15.20-15.50	**N. Amirezashvili, R. Eremyan, L. Lortkipanidze, L. Samsonadze, G. Chikoidze, A. Chutkerashvili, N. Javashvili** – Computer Models of the Georgian language
15.50-16.20	**Tamar Makharoblidze** – Documentation of GESL (Georgian Sign Language)
16.20-16.50	**M. Turashvili** –Electronic Catalogue of the Georgian Animal Folktales
16.50-17.20	**D. Tvaltvadze, M. Maduashvili, E. Kvirkvelia** –The Use of Electronic Courses and Textual Data in the Process of Teaching (The Experience of the Faculty of Humanities at Tbilisi State University)
17.20-17.45	Break
17.45	Round table – Language Technologies and Digital Humanities
	Moderator: **Jost Gippert**

Workshop on "Methodological Foundations of Corpus Building":

Exhibition Hall, National Parliamentary Library of Georgia

Tbilisi, Gudiashvili st. 7

September 13

14.00-14.30	Opening of the Workshop (J. Gippert, M. Tandashvili)
14.30-16.00	**I. Gurevych** – Annotation methods
16.00-17.00	**A. Hardie** – Morphosyntactic Annotation Schemata
17.00	**M. Tandashvili** – Georgian National Corpus

September 14

10.00-12.00	**N. Bubenhofer** – Thematical Corpora
12.00-13.00	Break
13.00 -15.00	**E. Teich** – Comparing corpora
15.00-16.00	Break
16.00	**J. Gippert** – Foundations of corpus building

September 15

10.00-12.00	**U. Quasthoff** – Large text corpora and their application in lexicography
12.00-13.00	Break
13.00-15.00	**P. Meurer** – Lemmatization and Disambiguation
15.00-16.00	Break
16.00	Plenary discussion and farewell

ქართული ენის კომპიუტერული მოდელები

**ნინო ამირეზაშვილი, რუდოლფ ერემიანი, ლიანა ლორთქიფანიძე,
ლია სამსონაძე, გიორგი ჩიკოიძე, ანა ჩუტკერაშვილი, ნინო ჯავაშვილი**
საქართველოს ტექნიკური უნივერსიტეტის
არჩილ ელიაშვილის მართვის სისტემების ინსტიტუტი (საქართველო)
l_lordkipanidze@yahoo.com, gogichikoidze@yahoo.com

საქართველოს ტექნიკური უნივერსიტეტის არჩილ ელიაშვილის მართვის სისტემების ინსტიტუტში გასული საუკუნის 50–იანი წლების ბოლოდან დღემდე მიმდინარეობს კომპიუტერული ლინგვისტიკის მიმართულებით მეცნიერული ფუნდამენტური კვლევები.

2009–2010 წლებში შესრულდა რუსთაველის ფონდის მიერ დაფინანსებული ორი პროექტი:

- ქართული ენის კომპიუტერული სუფლიორი უნარდაქვეითებულ პირთათვის (პროექტის ხელმძღვანელი ლიანა ლორთქიფანიძე);
- ავტომატური განმარტებით-კომბინატორული ლექსიკონი, როგორც ქართული ენის მოდელირების საფუძველი (პროექტის ხელმძღვანელი გიორგი ჩიკოიძე).

პროექტების ფარგლებში ქართული ენისათვის დამუშავდა სხვადასხვა კომბინაციური მეთოდიკა (ლექსიკური ფუნქციები, სინონიმური მწკრივები, სემანტიკური როლები და სუპერპარადიგმები); მთავარი შედეგი იყო ქართული ენის კომპიუტერული ლექსიკონის შექმნა, რომელიც, ამავე დროს, მორფოლოგიური გენერატორის ფუნქციას ასრულებს ანუ თითოეული სალექსიკონო ერთეულისგან აწარმოებს შესაბამის სრულ პარადიგმას. ლექსიკონი შეიძლება გამოყენებული იქნეს ენობრივ ავტომატურ სისტემებში (ლექსიკური თარგმანი, ენობრივი დიალოგი კომპიუტერთან, ტექსტური კორპუსების ავტომატური ანოტირება და სხვა).

წლების განმავლობაში მუშავდებოდა ენის მოდელირების ფუნდამენტური საკითხები. კერძოდ, შეიქმნა ენობრივი ალგორითმების წარმოდგენის ისეთი საშუალება, რომელიც, ერთი მხრივ, ასახავს ენობრივი სისტემის ფუნქციონირების ფუნდამენტურ თვისებებს (ორმიმართულებიანობა და პარალელურობა), მეორე მხრივ, ხელსაყრელია კომპიუტერული რეალიზაციისათვის. დამუშავდა ქსელური მეთოდი, რომლის საშუალებით შესაძლებელია ორმიმართულებიანი (ანალიზი/სინთეზი) ერთობლივი პროცესორის ფორმულირება. ქსელური მეთოდი ნათლად და მარტივად წარმოადგენს ალგორითმის სტრუქტურას, დასაშვებს ხდის მრავალდონიანი, ანუ პარალელურად მოქმედი სისტემების აგებას, ქსელის ყოველმა რკალმა შეიძლება ასახოს ენის ფუნდამენტური ნიშნური ხასიათი და ა.შ. ყველაზე გამოსადეგად კომპიუტერის და ენის სტრუქტურებს შორის არსებული წინააღმდეგობის გადასალახავად ქსელური მიდგომაა მიჩნეული (გ. ჩიკოიძე).

2011 წელს შესრულდა საქართველოს ტექნიკური უნივერსიტეტის მიერ დაფინანსებული პროექტი „ენის სწავლების კომპიუტერული მხარდაჭერა“ (პროექტის ხელმძღვანელი ლიანა ლორთქიფანიძე).

ამჟამად ჯგუფი მუშაობს რუსთაველის ფონდის საგრანტო პროექტზე „ქართული ენის კორპუსის სრული (მორფოლოგიური, სინტაქსური, სემანტიკური) ანოტირების სისტემა“ (2013–

2015). მორფოლოგიური ანოტირების გარდა, რომელიც არსებითად ეყრდნობა წინა პროექტში შექმნილ გრამატიკულ ლექსიკონს, იგეგმება კორპუსის შემადგენლობის მონიშვნა სინტაქსურ–სემანტიკური მახასიათებლებითაც (ლექსიკური ფუნქციები, სემანტიკური როლები და მათი მიმართებები, ზმნური სუპერ-პარადიგმები).

ჯგუფის მიერ რეალიზებულია და ამჟამად ნაყოფიერად გამოიყენება შემდეგი პროგრამული პროდუქტები:

1. ქართული ენის კომპიუტერული სუფლიორი უნარდაქვეითებულ პირთათვის (**Prophet_Geo)**;
2. თანამედროვე ქართული ენის მორფოლოგიური ლექსიკონი თანდართული პროცესორით: **GeoTrans;**
3. ენის მორფოლოგიის მულტიენობრივი კომპაილერი: **MuMoCom**;
4. კონკორდანსების შედგენის ნახევრად ავტომატური სისტემა: **MultiLingConc**;
5. მულტიენობრივი ლექსიკური მთარგმნელის კომპაილერი: **MuLexTranCom**.

ქართული ენის კომპიუტერული სუფლიორი (**Prophet_Geo)** არის ბეჭდვისას სიტყვების მოკარნახე პროგრამა, რომელიც მომხმარებელს სთავაზობს სიტყვებს მათი საწყისი ასოების მიხედვით. საკარნახებელი სიტყვა იძებნება ლექსიკონში და ეკრანზე გამოდის ერთი და იმავე ასოებით დაწყებული სიტყვების სია. სიტყვათა თანმიმდევრობას სიაში განსაზღვრავს მოცემულ ენაში მათი გამოყენების სიხშირე, რომლის რიგი მოცემული ენისათვის დამახასიათებელი ამ სიტყვების ხმარების სიხშირის მიხედვითაა შედგენილი. სუფლიორი თავსებადია ვინდოუსის ყველა პროგრამასთან. პროგრამა სასარგებლო აღმოჩნდა ყველა იმ პირთათვის, რომელთაც აქვთ სხვადასხვა სახის მოტორული და ლექსიკური დარღვევები. მათ შესაძლებლობა მიეცათ ნაკლები ძალისხმევით გაემარტივებინათ წერის პროცესი. გარდა ამისა, კომპიუტერში ტექსტის აკრეფის პროცესის დასაჩქარებლად „სუფლიორის“ გამოყენება მიზანშეწონილია ნებისმიერი რიგითი მომხმარებლისთვისაც. პროგრამა დაინერგა თბილისის 203-ე ყრუ და სმენადაქვეითებულთა საჯარო სკოლა-პანსიონში.

ცნობილია, რომ მორფოლოგიურ ლექსიკონში სალექსიკონო ერთეული დახასიათებული უნდა იყოს იმ გრამატიკული ნიშნებით, რომელთაც არსებითი მნიშვნელობა აქვთ გრამატიკულად სწორი სიტყვაფორმების ასაგებად. ლექსიკონის ელექტრონული ვერსია – **GeoTrans**–ი მომხმარებლის მიერ მოწოდებულ ამოსავალ ერთეულს პასუხობს მისი სრული მორფოლოგიური პარადიგმით. უნდა აღინიშნოს, რომ სისტემა წარმოქმნის მხოლოდ ისეთ სიტყვაფორმებს, რომლებიც თანამედროვე ლიტერატურულ ნორმებს შეესაბამება. გარდა ამისა, ლექსიკონი ავტომატური მორფოლოგიური ანალიზისთვისაც გამოიყენება. ამოსავალ სიტყვათა და წესთა ლექსიკონი შევსებულია 100000 ერთეულით. აღნიშნულ ლექსიკონზე დაყრდნობით შესაძლებელია დაახლოებით 24 400 000 სიტყვაფორმის სინთეზ/ანალიზის დემონსტრირება. პროგრამა **GeoTrans**–ი ამჟამად წარმატებით გამოიყენება ტექსტური კორპუსის მორფოლოგიური ანოტირებისთვის.

დროთა განმავლობაში ყველა ენა იცვლება. იცვლება მისი ლექსიკაც. ისევე, როგორც შეუძლებელია გადაითვალოს რიცხვების უსასრულო სიმრავლე ან დასახელდეს მსოფლიოში არსებული ყველა საკუთარი სახელი, შეუძლებელია ბუნებრივი ენების სრული ლექსიკონების შე-

დგენა. მაგრამ ეს ეხება მხოლოდ ტრადიციული, სტაციონალური ლექსიკონების ბეჭდურ ვერსიებს. ინტერნეტსივრცეში, სადაც აისახა თანამედროვე ცხოვრების მრავალი სფერო, ბუნებრივი ლექსიკის ყოველი ახალი გამოვლინება ფიქსირდება ეგრეთ წოდებული „სიტყვის კარიბჭის“ დახმარებით. მსგავსი სისტემა ავტომატურად გამოავლენს ახალ სიტყვაფორმას, რომელიც ლინგვისტური დამუშავების შემდეგ შევა ენის ავტომატურ გრამატიკულ ლექსიკონში. რიგი ენებისთვის ლექსიკონების შევსება-გაფართოების პროცესი გამარტივებულია გრამატიკული კომპაილერის დახმარებით, რომელიც ენის ფორმალური მოდელის ავტომატური რეალიზაციის ყველაზე თანამედროვე ინსტრუმენტია.

ენის მორფოლოგიის მულტიენობრივი კომპაილერი (**MuMoCom)** ცოდნის დაგროვების ექსპერტული სისტემაა. მასში ენის მორფოლოგიის ფორმალური მოდელი ავტომატურადაა რეალიზებული ე.წ. ცოდნის შეძენის და მისი შინაარსის შინაგანი წარმოდგენის პროცესით. სხვადასხვა ენის მორფოლოგიის შესახებ **MuMoCom**-ის დახმარებით შეძენილი ცოდნა გამოიყენება ავტომატურ პროცესორში. სისტემა ინვარიანტულია ენის მიმართ. პროგრამული აპლიკაცია საკმაოდ მოხერხებულია და მნიშვნელოვნად აადვილებს ლინგვისტ-ექსპერტის შრომას. მისი ინსტრუმენტებით ენის მორფოლოგიურ პროცესორზე მუშაობა შესაძლებელია პარალელურ რეჟიმში ნებისმიერი რაოდენობის ექსპერტისთვის; მას აქვს პოპულარული მაიკროსოფტ-ვორდთან თავსებადი ინტერფეისი; მასში სამუშაოდ არაა აუცილებელი სპეციალიზებული ენის შესწავლა (Xerox Finite State Tool-ისგან განსხვავებით); ასევე გამარტივებულია ენის კორპუსისთვის მორფოლოგიური ანოტირების სტანდარტების შემუშავება და შემდგომში ამ სტანდარტებში ნებისმიერი ცვლილების შეტანა; შესაძლებელია ნებისმიერი ენის სხვადასხვა ვარიაციისთვის (დროის, სივრცის, წარმომავლობის, ჟანრის და ა.შ. მიხედვით) ცალკეულ ენათა მორფოლოგიური პროცესორების ბიბლიოთეკების კომპილირება და სხვ.

სწორედ **MuMorCom**-ის გამოყენებით შეიქმნა როგორც ქართული ენის კომპიუტერული სუფლიორი (**Prophet_Geo)**, ისე თანამედროვე ქართული ენის მორფოლოგიური ლექსიკონი თანდართული პროცესორით (**GeoTrans)**. ამჟამად მიმდინარეობს კორპუსის ტექსტების მორფოლოგიური ანოტირება **MuMorCom**-ზე დაყრდნობით, შოთა რუსთაველის ეროვნული სამეცნიერო ფონდის მიერ დაფინანსებული პროექტის – „ქართული ენის კორპუსის სრული (მორფოლოგიური, სინტაქსური, სემანტიკური) ანოტირების სისტემა“ – ფარგლებში.

კონკორდანსს ცენტრალური ადგილი უჭირავს კორპუსის ლინგვისტიკაში, ვინაიდან ტექსტში მრავალი მნიშვნელოვანი ენობრივი ნიმუშის მიგნების საშუალებას იძლევა. კონკორდანსების შედგენის ნახევრად ავტომატური სისტემა (**MultiLingConc)** დაეხმარება ენის სპეციალისტებს პრაქტიკულად ნებისმიერ ენაზე ჩაწერილი ტექსტების კვლევაში, რაც შემდეგში მდგომარეობს: 1) ენათა კონტრასტული, სინტაქსური, მორფოლოგიური, ლექსიკოლოგიური აღწერა, 2) ენათა ურთიერთშედარება, 3) ენათა შორის გრამატიკული და ლექსიკოლოგიური პარალელების გამოვლენა.

ალგორითმების სისტემა ქმნის პროგრამულ დანართს, რომლის საშუალებითაც ხდება შერჩეული ტექსტის დამუშავება, რის შედეგადაც სხვადასხვა ენის სპეციალისტები მიიღებენ კონკრეტული ტექსტის კონკორდანსისა და მრავალენოვანი ლექსიკონის როგორც კომპიუტერულ, ასევე პოლიგრაფიულ ვერსიას.

სისტემაში გათვალისწინებულია მრავალენოვანი თვითშევსებადი კომპიუტერული ლექსიკონის ალგორითმის რეალიზაცია, რაც გულისხმობს ყოველ ჯერზე (ახალი ენის, ახალი ტექსტის ჩაწერისა და დამუშავების დროს) სალექსიკონო ბაზის შევსება-გამდიდრებას, ეს კი, თავის მხრივ, იწვევს ერთხელ უკვე გამოყენებული ფორმისთვის სალექსიკონო ინფორმაციის ავტომატურ მიწერას. საბოლოოდ, გარკვეული ენისთვის სხვადასხვა მკვლევრის მიერ სხვადასხვა ტექსტის დამუშავების შემდეგ ჩაწერილი და შევსებული ლექსიკონი უკვე შეიძლება ენის კორპუსში განთავსდეს.

კომპიუტერული ლინგვისტიკის ყველაზე დიდ მიღწევად და, გარკვეულწილად, საბოლოო მიზნად, სამართლიანად შეიძლება მოვიაზროთ ავტომატური მანქანური თარგმანი. ამ მეტად რთული ამოცანის ბოლომდე გადასაწყვეტად აუცილებელია უპირველესად ლექსიკური მთარგმნელის შექმნა. ენის რიგითი მომხმარებლისათვის ასეთი სისტემაც საკმაოდ ღირებულია, რადგან ის უადვილებს მომხმარებელს უცხო ტექსტის ინტენსიურ გაცნობას და ხშირად უფრო სასარგებლოცაა ტექსტის შედგენისას.

ლექსიკურ მთარგმნელში შეპირაპირებულია ორენოვანი ლექსიკონები და ენის პროცესორები მომხმარებლის მიერ ტექსტის დამუშავების შესაბამისად. უცხო ტექსტის წაკითხვისას ენის მორფოლოგიური პროცესორის ანალიზატორი მომხმარებელს მოუძებნის საძიებო სიტყვის ბაზისურ, ანუ სალექსიკონო ფორმას იმ შემთხვევაშიც, როცა ეს სატექსტო ფორმა მკვეთრადაა განსხვავებული სალექსიკონო ფორმისაგან (მაგალითად, ინგლისური catch-caught, mouse-mice, ქართული დაგლეჯს-დაგლიჯა, გათლის-გათალა. წევს-დაწვა-დაწოლილა, ზის-სხედან). შემდეგ კი შესთავაზებს ყველა შესაძლო შესატყვისს მშობლიურ ენაზე. იგივე გამეორდება საპირისპირო მიმართულებით უცხო ენაზე თარგმნის დროს.

მულტიენობრივი ლექსიკური მთარგმნელის კომპაილერი (**MuLexTranCom**) ძირითადად ეყრდნობა **MuMoCom** სისტემას, რომლის ენის მიმართ ინვარიანტულობა ენის მიმართ მომხმარებელ ლინგვისტს ეხმარება მისთვის სასურველი ორენოვანი მორფოლოგიური ავტომატური ლექსიკონის კომპილირებაში. კომპიუტერული დანართის მოქნილი უტილიტების გამოყენებით იქმნება პროგრამული პროდუქტი დაპროგრამების გარეშე და პრაქტიკულად ავტომატური მთარგმნელის ლექსიკური მარაგის ზომა შემოსაზღვრული არ არის.

ჩვენი ჯგუფი ბოლო წლების განმავლობაში მოქმედებს იმ სტრატეგიით, რომ ცალკეულ პროექტებში დამუშავებული საკვანძო თეორიული და პრაქტიკული საკითხები მომავალში ენის ტექნოლოგიზების საიმედო დასაყრდენად იქცეს. ჩვენ მიერ შექმნილი კომპიუტერული პროდუქტები მრავალი ლინგვისტური მიმართულებით გამოიყენება. ლინგვისტებისთვის ფართო ასპარეზი იქმნება ქართული ენის კომპიუტერული მოდელის რეალიზაციისათვის მისი მრავალი გამოვლინებისა და ცვლილებების გათვალისწინებით. აღსანიშნავია, რომ ჩვენ მიერ პროგრამული პროდუქტების სახით შექმნილია ნაციონალური კორპუსის მენეჯერის კომპილაციისათვის ძირითადი კომპონენტები.

Computer Models of the Georgian Language

Nino Amirezashvili, Rudolf Eremyan, Liana Lortkipanidze, Lia Samsonadze, Giorgi Chikoidze, Ana Chutkerashvili, Nino Javashvili
Archil Eliashvili Institute of Control Systems, Georgian Technical University (Georgia)
l_lordkipanidze@yahoo.com, gogichikoidze@yahoo.com

Fundamental scientific researches in Computational Linguistics have been carried out at the Archil Eliashvili Institute of Control Systems, Georgian Technical University, since the late 1950s. Two projects, supported by Rustaveli Foundation, were developed in 2009-2010:

- Georgian Computer Prompter for Disabled Persons (Research Director – Liana Lortkipanidze);
- Automatic Explanatory-combinatorial Dictionary as a Basis of Georgian Language Modeling (Research Director – George Chikoidze).

Within the framework of the project, various combination methods for the Georgian language have been developed (lexical functions, synonymic series, semantic roles and super-paradigms); the main result was the creation of the Georgian language computer dictionary, which, at the same time, carries out functions of a morphological generator; in other words, it produces the full paradigm for each lexical unit. The dictionary can be used in language automatic systems (lexical translation, lingual dialogue with the computer, automatic annotation of text corpora, etc.).

Fundamental problems of language modeling have been developed for many years. Namely, the means of presentation of language algorithms was developed, which reflects fundamental characteristics of functioning of language algorithms, on the one hand (bi-directionality and parallelism), and is very convenient for computer realization, on the other.

A network method has been developed, allowing formulating of a bi-directional (analysis/ synthesis) combined processor. The network method represents the structure of the algorithm clearly and simply, makes it possible to build multilevel, i.e. parallel operating systems; each arc of the network may reflect the fundamental signed character, etc. The network approach is considered as the best way to overcome the resistance between the computer and the language structures (G. Chikoidze).

In 2011 the project "Computer Support of Language Learning," supported by the Georgian Technical University, was developed (Research Director – Liana Lortkipanidze).

The team is currently working on Rustaveli Foundation project "The Full (Morphological, Syntactical, Semantical) Annotation System of Georgian Language Corpora" (2013-2015). Besides the morphological annotation which is essentially based on the dictionary created in the previous project, it is planned to mark the corpus components according to their syntactic/semantic characteristics (lexical functions, semantic roles and their relations, verbal super-paradigms).

The following software products have been implemented by the team and are being used productively at present:

1. Georgian Computer Prompter for Disabled Persons (**Prophet_Geo**);
2. Modern Georgian Morphological Dictionary with Attached Processor: **GeoTrans**;
3. Multilingual Language Morphology Compiler: **MuMoCom**;

4. Semi-automatic System of Compiling Concordances: **MultiLingConc**;
5. Multilingual Lexical Translator Compiler: **MuLexTranCom**.

1. Georgian Language Computer Prompter (**Prophet_Geo**) is a program, which dictates words while printing. It offers a user words by their initial letters. A word is searched in the dictionary and a list of words, beginning with the same letters, appears on the screen. The line of offered words is compiled by the frequency of its usage in the language. Prophet_Geo is compatible with all Windows applications.

The program became beneficial for all the people who have different types of motor and lexical disorders. They are able to simplify the writing process with less effort. In addition, using the "Prompter" is advisable to accelerate the process of text printing for any regular consumer. The program is implemented in Public School 203 for the deaf and diminished hearing children.

2. It is known that units in the morphological dictionary should be characterized by the grammatical features, which are essential for building grammatically correct word forms. Electronic version of the dictionary – **GeoTrans** gives a full morphological paradigm as an answer to the respond of the initial form given by a user.

It should be noted that the system generates only word forms, which correspond to the modern literary standards. In addition, the dictionary is used for automatic morphological analysis. The dictionary of the initial words and the rules are supplemented with 100,000 units. Based of the mentioned dictionary, it is possible to demonstrate synthesis/analysis of 24 400 000 word forms. The **GeoTrans** program is currently successfully used for the morphological annotation of the text corpus.

3. Languages change through time; vocabulary changes as well. Just as it is impossible to count the infinite set of numbers, or nominate all proper names which exist in the world, it is impossible to create complete dictionaries of natural languages. But it is only about the traditional, stationary versions of print dictionaries.

In the Internet, where many areas of modern life are reflected, the appearance of each new vocabulary is fixed by using so-called "Gate of the word". A similar system automatically detects a new word form, which fits into the automatic grammar vocabulary of the language after linguistic processing. For some languages, the filling-widening process of dictionaries is simplified with the help of the grammar compilers which is the modernist tool of the automatic realization of a formal language model.

Language morphology multilingual compiler (MuMoCom) is a knowledge accumulation expert system. The morphology of the formal language model is automatically realized in it by the process of the so-called Knowledge acquisition and its contents submission. Knowledge about morphology of multiple languages acquired with the help of **MuMoCom** – is used in an automatic processor. The system is invariant towards the language. A Software application is very convenient and greatly simplifies the work of linguists and experts.

Working on the language morphologic processor with its instruments is possible in parallel mode for any number of experts. It has a popular Microsoft – Word compatible interface. For working with the program it is not necessary to learn a specialized language (in contrast to – Xerox Finite State Tool); It is also simplified developing morphological annotate standards for a language corpus and then making any changes as well. It is possible to compile libraries of any language morphological processors for different (by time, space, origin, genre, etc.) variations, etc.

By means of applying of MuMorCom, Georgian Language Computer Prompter (Prophet_Geo) was designed as well as Modern Georgian Morphological Dictionary with attached processor: (**GeoTrans**); Nowadays the morphological annotation of text corpora on the basis of MuMorCom is conducted within the framework of the project "The Full (Morphological, Syntactical, Sematical) Annotation System of Georgian Language Corpora", financed by Shota Rustaveli National Scientific Foundation.

4. Concordance takes the central place in Corpus Linguistic, as it enables us to find many important language patterns in the text. Semi-automatic system of compiling concordances (**MultiLingConc**) will help language professionals to research practically in any language written texts, as follows: 1) Contrast, syntactic, morphological, lexicological description of languages; 2) Comparison of languages; 3) revealing grammatical and lexicological parallels between the languages.

The system of algorithms creates software which processes the selected text and as a result specialists of different languages get computer and printing versions of a particular language concordance as well as a multilingual dictionary.

The system provides algorithm realization of a self-filling computer dictionary which means that every time (while recording or processing a new language or a new text) the base of the dictionary gets filled. If the form has already been used, the program provides automatic writing of its dictionary information. Finally, after processing many different texts of certain language by variety of different researchers, the recorded and filled dictionary may be placed in the corpus of the language.

Automatic machine translation can fairly be considered as the main achievement and, in some ways, the final goal of Computational Linguistics. In order to solve this too complicated task completely first of all it is necessary to create lexical translator. This type of system is rather valued among the ordinary users, as it makes it easier for them to learn foreign-language texts intensively thus is much more useful while composing text.

5. The lexical translator integrates bilingual dictionaries and language processors according to the text that is processed by the user. While reading foreign text the analyzer of language morphological processor will find the basic dictionary form of a required word even if the text form is significantly different from its dictionary form (e.g. English: catch-caught, mouse-mice; Georgian: daglejs-daglija, gatlis-gatala, c'evs-dac'va-dac'olila, zis-sxedan). Then the compiler offers the user probable all corresponding words in his/her native language. The same is about translating into a foreign language.

The Compiler of the Multilingual Lexical Translator (**MuLexTranCom**) is mainly based on **MuMoCom** system, which is invariant towards language, helps a user-linguist to compile desired automatic bilingual morphologic dictionary. Program product is created by using convenient utilities of computational supplement without programming and the number of the vocabulary units of automatic translator is not limited.

The strategy of our team is to provide reliable support for future language technologies by the theoretical and practical key issues that have been worked up in separate projects for the past years. The computer products created by our team are used in various linguistic areas. It is a challenge for linguists to create computational model of a language, taking into account its multi detections and changes. It should be mentioned that as a program product we have created main components to compile a national corpus manager.

ქართული დიალექტური კორპუსი, როგორც სასწავლო-საგანმანათლებლო რესურსი სასკოლო ჰუმანიტარული სწავლებისათვის

დიანა ანფიმიადი
თსუ არნ. ჩიქობავას სახ. ენათმეცნიერების ინსტიტუტი (საქართველო)
Diana@saba.com.ge

კორპუსოლოგია დიდი ხანია გასცდა მხოლოდ ვიწრო სამეცნიერო ინტერესებს და ფართო კულტურული თუ საგანმანათლებლო მნიშვნელობა შეიძინა. ერთიან ელექტრონულ სივრცეში განთავსებული ერთგვარი ენობრივი არტეფაქტები განსაკუთრებულ სიცოცხლეს სწორედ კულტურულ და საგანმანათლებლო არეალში მოხვედრისას იძენენ. განსაკუთრებით მნიშვნელოვანი და მეტიც, სასიცოცხლოდ აუცილებელი სასკოლო განათლების პროცესში კორპუსული კვლევების გამოყენების პრეცედენტია, ვინაიდან სკოლამ უნდა გაზარდოს 21-ე საუკუნის ელექტრონული კორპუსების შემქმნელი თუ მომხმარებელი.

ჩვენს მოხსენებაში გვსურს წარმოვადგინოთ სასკოლო განათლების რამდენიმე მოდული, სადაც დიალექტური კორპუსის, როგორც საგანმანათლებლო რესურსის სათანადო გამოყენება საინტერესო სასწავლო პროცესის წარმართვისა და, რაც მთავარია, მნიშვნელოვანი შედეგების მიღწევის საშუალებას მოგვცემს.

ერთენოვანი, პარალელური თუ სხვადასხვა სპეციფიკის კორპუსები მთელ მსოფლიოში აქტიურად გამოიყენება როგორც ენების სწავლების, აგრეთვე სხვადასხვა დისციპლინის დამუშავების პროცესში. გამომდინარე იქიდან, რომ ელექტრონული კორპუსი გვთავაზობს ენობრივ მასალას სტუქტურირებულად, სისტემურად, იგი ამარტივებს კვლევისა და სწავლის პროცესებს.

მოხსენებაში განვიხილავთ ეროვნული სასწავლო გეგმით გათვალისწინებული ქართული ენისა და ლიტერატურის სწავლების პროცესში დიალექტური კორპუსის გამოყენების პერსპექტივებს, კონკრეტულად კი, მიმოვიხილავთ შემდეგ საკითხებს:

1. დიალექტური კორპუსი ქართული ენისა და ლიტერატურის სწავლების პროცესში:
    - ქართული ენის გაკვეთილები – გრამატიკული კატეგორიები, ენობრივი ფორმები, დიალექტური ლექსიკა, ეთნონიმები;
    - ქართული ლიტერატურის გაკვეთილები – ლიტერატურული სიუჟეტები, ლიტერატურული პერსონალიები;
    - ქართული ფოლკლორი – მითი, ლეგენდა, ხალხური პოეზია, ზღაპრები;
    - ხალხური „ვეფხისტყაოსანი“;
    - ფაქტები, მეფეები, ისტორიული პირები, მიგრაციები, შემოსევები, ომები, შინამრეწველობა, ეთნოგრაფია;
    - ტოპონიმები, გეოგრაფიული სახელები და ა.შ.;
    - სამოქალაქო განათლება – ქალთა უფლებები, კლასობრივი ამბები, დამოუკიდებლობისა და სახელმწიფოებრიობის საკითხები.

წარმოვადგენთ გაკვეთილის საჩვენებელ გეგმებს, მასალებს, რესურსებს, როგორც ტექნოლოგიებით გამდიდრებული პროექტებით სწავლების, აგრეთვე, სწავლების სხვა ტიპით მუშაობისათვის – თეორიული თუ პრაქტიკული, საკლასო ან საშინაო, ინდივიდუალური თუ ჯგუფური აქტივობების მაგალითზე. აგრეთვე, წარმოვადგენთ დავალებების, საკლასო თუ კლასგარეშე აქტივობების სხვადასხვა ტიპსა და მაგალითს, ამასთანავე, გაკვეთილებს რამდენიმე თემაზე, როგორებიცაა: „ადამიანთა პორტრეტები“, „ჩაიწერე დიალექტი“, „ნაცნობი და უცნობი მწერლები“ და სხვა.

Georgian Dialect Corpus as an Educational Resource for Teaching the Humanities at School

Diana Anphimiadi
Arn. Chikobava Institute of Linguistics, TSU (Georgia)
Diana@saba.com.ge

Corpus Studies has long gone beyond narrow scholarly interests and acquired a broad cultural and/or educational significance. Certain linguistic artifacts, loaded in the common electronic space, become particularly vital following their introduction into the cultural and educational realms. Particularly important and, moreover, vitally necessary is the precedent of the application of corpus studies in the process of school education as far as school has to bring up developers and/or users of electronic corpora of the 21st century.

The paper will present several modules of school education within which an appropriate application of a dialect corpus, as of an educational resource, allows for conducting of an interesting teaching/learning resource and, what is most important, for achieving of significant results.

Monolingual, parallel and various specialized corpora are actively applied all over the world in the process of both language teaching and developing of various disciplines. With a view to the fact that electronic corpora provide linguistic data in a structured, systematized way, it simplifies the process of research and teaching.

The paper discusses perspectives of applications of a dialect corpus in the process of teaching of the Georgian language and literature, as envisaged in the National Curriculum; specifically, the following issues will be dealt with:

1. Dialect corpus in the process of teaching of the Georgian language and literature:
   I. Lessons in Georgian language: grammatical categories, linguistic forms, dialect vocabulary, ethnonyms;
   II. Lessons in Georgian literature: literary plots, literary personalia;
   III. Georgian folklore: myth, legend, folk poetry, fairy tales;
   IV. Folk versions of *The Knight in the Panther's Skin*;

V. Facts, kings, historical personalities, invasions, wars, cottage industry, ethnology;
VI. Place-names, geographic names, etc.;
VII. Civic education: women's rights, class relations, issues of sovereignty and statehood.

The paper will present illustrative plans, materials, and resources of lessons, exemplified both by technologically enriched project-based teaching and by other teaching activities, theoretical or practical, classroom or home, individual or group. In addition, it will present various tasks, classroom and extra-curricular activities, besides, lessons on various topics, such as: "Human portraits," "Record a dialect," "Known and unknown writers," etc.

„დიალექტური კუნძულის“ ლინგვოკულტურული სივრცის მოდელირება და პრეზენტაცია ქდკ-ში (ფერეიდნული დიალექტი)

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

ქართული დიალექტური კორპუსის კონცეფცია იმთავითვე გულისხმობდა ქართული ენობრივი სივრცის მოდელის შექმნას. კორპუსის მეტაანოტირების სისტემა ისეა შედგენილი, რომ მისი საშუალებით შესაძლებელია ქართული ენის ნებისმიერი ქვესისტემის ან სხვა ქართველური ენების მონაცემების შენახვა, აღწერა, კვლევა. კორპუსის მოდელი შეიძლება გამოყენებული იქნეს ნებისმიერი მულტიენობრივი არეალის დოკუმენტირებისათვის.

განსაკუთრებული აქცენტი კორპუსის შექმნისას გაკეთდა საზღვარგარეთ გავრცელებული ქართული დიალექტების ადეკვატური „ენობრივი პორტრეტის“ შესაქმნელად. ამისთვის მოეწყო რამდენიმე ექსპედიცია ირანის, აზერბაიჯანისა და თურქეთის ტერიტორიაზე კომპაქტურად მოსახლე ეთნიკური ქართველების მეტყველების ნიმუშების მოსაპოვებლად. ასევე გაიშიფრა და გამოსაცემად მომზადდა წინამორბედი ექსპედიციის (1998) მოპოვებული აუდიომასალა. გაციფრებულია და კორპუსში ინტეგრირებული ის მცირერიცხოვანი ტექსტები, რომლებიც სხვადასხვა დროს გამოქვეყნდა ქართველ მეცნიერთა მიერ (არნ. ჩიქობავა, ვ. თოფურია, მ. თოდუა, დ. ჩხუბიანიშვილი, თ. უთურგაიძე...).

ფერეიდნული ირანის ისლამურ რესპუბლიკაში, ისპაჰანის მახლობლად, ფერეიდანის რეგიონში მცხოვრებ ქართველთა დიალექტია. ის ვითარდებოდა განსაკუთრებული იზოლაციის პირობებში, შორს ბირთვული კულტურული და ენობრივი არეალიდან, შესაბამისად, აქვს ყველა ნიშანი ტიპური დიალექტური კუნძულისა.

ფერეიდნულმა დიასპორამ ოთხი საუკუნის განმავლობაში შეინარჩუნა თითქმის სრული „კუნძულური იზოლაცია“ და მხოლოდ გასული საუკუნის 60-იანი წლებიდან იწყება იზოლაციის რღვევა მისთვის თანმხლები, დაჩქარებული ტემპით მიმდინარე ასიმილაციური პროცესებით.

ფერეიდნულ დიალექტში ბირთვულ არეალთან კონტაქტის აღდგენის შემდეგ კულტურული და ენობრივი რეინტეგრაციის პროცესი დაიწყო, რაც, ბუნებრივია, აისახა ენობრივ პროცესებზეც. ქართულ დიალექტურ კორპუსში წარმოდგენილი ფერეიდნული მასალა სწორედ ამ მოვლენების ენობრივი სარკეა.

ჩვენ მიზანდასახულად ვცადეთ ფერეიდნული ლინგვოკულტურული სივრცის მოდელირება ქდკ-ში ტექსტური და ლექსიკოგრაფიული მასალის მოპოვების პროცესშივე. რეპრეზენტაციულობის ხარისხის გასაზრდელად ჩვენ ვიყენებდით დიალექტური მასალის მოპოვების ტრადიციულ მეთოდებს: თემატური თხრობის მოტივირება, სპონტანური სამეტყველო კომუნიკაციის ინიცირება, სპეციალურად შედგენილი დარგობრივი და იდეოგრაფიული კითხვარების გამოყენება და სხვ. მასალას ვიწერდით ისე, რომ დაცული ყოფილიყო ასაკობრივი და გენდერული ბალანსი, თემატური მრავალფეროვნება, რეგიონალური ვარიაციულობა და სხვ.

ჩვენ მოვახერხეთ და ჩავიწერეთ ფერეიდნული ენობრივი კოდის ყველაზე „კაშკაშა“ მატარებლები. მათ შორის ისეთები, რომელთა მეტყველების ნიმუში ადრეც იყო დაფიქსირებული. გვაქვს „ერთი მთქმელის“ სუბკორპუსისთვის ძალიან საინტერესო ტექსტური და ლექსიკოგრაფიული მასალა...

კორპუსის მეტაანოტირებისა და ლინგვისტური ანოტირების სისტემა იმგვარადაა ორგანიზებული, რომ დიალექტურ კუნძულში მშობლიურ (ბირთვულ) კულტურულ არეალთან ინტეგრირების თანამდევი პროცესების თვალის გადევნებაც იყოს შესაძლებელი. მეტიც, კორპუსში „ჩაშენებული“ ფერეიდნული ლექსიკონი შესაძლებელია, სასწავლო ფუნქციითაც იყოს გამოყენებული.

ლექსიკონის მარკირების განსაკუთრებული სისტემა საშუალებას გვაძლევს ახალი, სალიტერატურო ქართულის გავლენით გაჩენილი ლექსიკა გამოვყოთ საერთო ლექსიკიდან. ტრადიციული დიალექტოლოგია არ ითვალისწინებს დიალექტში ასეთი მონაცემის გამოყოფას და არ აღიარებს მას დიალექტური სტრატის აღწერისას რელევანტურად.

დიალექტურ კუნძულში ენობრივი პროცესების განვითარება, როგორც ეს ლიტერატურაშია აღწერილი, სწორედ შიდადიალექტურ ვარიაციებსა და სალიტერატურო ენისადმი მათ მიმართებაზეა დამოკიდებული. ასეთ ენობრივ კოლექტივში ორი ფაქტორი აქტიურდება: სალიტერატურო ენის გავლენა და სხვადასხვა დიალექტის თანაარსებობისას (იგულისხმება, რომ გადასახლებულები სხვადასხვა დიალექტის წარმომადგენელი იყვნენ) გაჩენილი ერთგვარი „გაწონასწორების“ პროცესი, რაც მიმართულია საერთო დიალექტის – კოინეს ჩამოსაყალიბებლად (ჟირმუნსკი 1929, 492).

ქართული დიალექტური კუნძულებიდან ფერეიდნული გამოირჩევა იმით, რომ მას სალიტერატურო ენასთან კონტაქტი და, შესაბამისად, მისი გავლენა არ ჰქონია საუკუნეთა განმავლობაში. რაც შეეხება დიალექტთაშორის მიმართებასა და „გაწონასწორებას“ – ეს პროცესი ბუნებრივად საგულვებელია ფერეიდანში, რადგან, როგორც ცნობილია, გადასახლება რამდენიმე ეთნოგრაფიული კუთხიდან მოხდა (კახეთი, ქართლი, ივრისპირი...). ყოველივე ამის გამო ჩვენი მიზა-

ნი იყო არა მხოლოდ ფერეიდნული მეტყველების სტატიკური დიალექტური პროფილის, არამედ მისი განვითარებისა და განსაკუთრებით, ბირთვულ ენობრივ არეალთან კავშირის აღდგენის შემდგომ მისი განვითარების დინამიური პროცესის ჩვენება, აგრეთვე, შიდადიალექტური განსხვავებების მაქსიმალურად გამოვლენა და მონიშვნა.

გრამატიკული მარკირების სისტემაში ჩვენ დავამატეთ სპეციალური ნიშნები ფსევდოლიტერატურული (pseudo) და ინოვაციური (New) ფორმებისთვის.

ეს მახასიათებლები საშუალებას გვაძლევს 400 წლის წინანდელი ფერეიდნული დიალექტური მონაცემიდან გამოვარჩიოთ მცდარი _ ფსევდოლიტერატურული ფორმები (საკვე / საკვები...); სალიტერატურო ენიდან შესული ფორმები (ქალიშვილი / ყორი...); სალიტერატურო ენიდან შესული და დიალექტის მორფონოლოგიურ "კანონიკაზე" გაწყობილი ფორმები (ბავშივები < ბავშვები; გადიცვალა < გარდაიცვალა...). ლექსიკონის რედაქტორში გამოყოფილია სპეციალური ველი ილუსტრაციის რეგიონალური ვარიაციულობის მოსანიშნად (რომელიც სპეციფიკური ფორმის ჩაწერის ადგილზე მიუთითებს).

მოხსენებაში დეტალურად იქნება აღწერილი, როგორ შეიძლება კორპუსის მრავალდონიანი ანოტირების სისტემის გამოყენება ფერეიდნული დიალექტური კუნძულის შესახებ მაქსიმალური ინფორმაციის მისაღებად და როგორ ახდენს კორპუსი ფერეიდნული კოლექტიური და ინდივიდუალური ენობრივი მსოფლხატის რეპრეზენტაციას.

ლიტერატურა:

ჟირმუნსკი – Жирмунский, В. М. Проблемы переселенческой диалектологии // Жирмунский В. М. общее и германское языкознание. Избр. труды. Л., 1976. с. 492-516.

ბერიძე, ბაკურაძე – М. М. Беридзе, Л. Д. Бакурадзе, Грузинский диалектный остров в Иране, მინსკი, 2014.

ბერიძე, სურმავა, ნადარაია – ქართული დიალექტური კორპუსი და საქართველოს ეთნოსოციალური სურათი, საერთაშორის კონფერენცია – МЕГАЛИНГ – 2009 – შრომები, კიევი 2009.

Modeling and Presenting of the Lingua-cultural Area of a „Dialect Island“ in GDC (Fereidanian Dialect)

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

The concept of the Georgian Dialect Corpus initially implied the creation of the model of the Georgian linguistic area. The meta-annotating system of the corpus is designed so that data of any of the varieties of Georgian or of Kartvelian languages can be stored, described, studied. The corpus model can be used for documenting of any multi-linguistic area.

A special emphasis was made on the creation of an adequate “linguistic portrait” of the Georgian dialects abroad. For the sake of this, several expeditions were organized in order to collect speech samples of ethnic Georgians, compactly settled in Iran, Azerbaijan, and Turkey. The audio-data, collected by the previous expedition (1998), were deciphered and prepared for publication. The few texts, published by Georgian scholars (Arn. Chikobava, V. Topuria, M. Todua, D. Chkhubianishvili, T. Uturgaidze…) at various periods of time, were digitalized and integrated into the corpus.

Fereidanian is a dialect spoken by ethnic Georgians inhabiting the region of Fereidan, near Isfahan, Islamic Republic of Iran. It developed within a particularly isolated environment, distant from the core cultural and linguistic area; hence, it has all properties of a typical dialect island.

Fore four centuries, the Fereidanian diaspora maintained the almost complete “island isolation”, and, only since the 1960s, the isolation started to deteriorate, accompanied with peculiar increased assimilation processes.

Following the re-establishment of the contact of Fereidanians with the core area, a process of cultural and linguistic reintegration began, which, naturally enough, was reflected in linguistic processes. The Fereidanian data, presented in the Georgian Dialect Corpus, are a linguistic mirror of those phenomena.

We purposefully tried to model the Fereidanian lingua-cultural area in the GDC during the process of collecting of textual and lexicographic data. In order to increase the degree of representativeness, we applied traditional methods of collecting of dialect data: motivation of thematic narration, initiation of spontaneous speech communication, specially designed sectoral and ideographic questionnaires, etc. The data were recorded so as to maintain the age and gender balance, thematic diversity, regional variation, etc.

We managed to record “the brightest” speakers of the Fereidanian linguistic code, among them, those who had been recorded earlier. We possess very interesting textual and lexicographic data for a sub-corpus of “one speaker”…

The meta-annotation of the corpus and the linguistic annotation system are organized so as to allow for observing of the concurrent processes with the integration of the dialect island with the core cultural area. Moreover, the Fereidanian dictionary may be used for teaching/learning purposes.

The peculiar marking system of the dictionary allows for identifying of words, having emerged as a result of the influence of Standard Georgian, from the common vocabulary. Traditional dialectology does not consider identification of such a datum and does not acknowledge it relevant in the description of a dialect stratum.

Development of linguistic processes within a dialect island, as it has been described in literature, depends on intra-dialect variations and their attitude to a standard language. Two factors become activated in such a language community: influence of a standard language and, emerged within the co-existence of dialects (provided that re-settlers were speakers of various dialects), a certain "balancing" process directed towards the establishment of a common dialect, a koiné (Zhirmunsky 1929, 492).

Fereidanian is peculiar among Georgian dialect islands with respect to the fact that it has not had any link with the standard language and, hence, was not influence by it for centuries. As for inter-dialect relations and "balancing", this process is essentially conceivable in Fereidan as far as, as it is known, their exile took place from several ethnographic provinces (Kakheti, Kartli, Ivrispiri…). Owing to these, we were aimed at presenting of the static dialect profile of Fereidanian and of the dynamic process of its development following the restoration of its link with the core linguistic area; in addition, to maximally reveal and markup intra-dialect differences.

In the system of the grammatical markup, special tags were added for pseudo-standard and new forms.

These features allow for the identification of false, pseudo-standard forms from the 400-year-old Fereidanian dialect data (sak've=sak'vebi 'food'); forms, originating from the standard language (kališvili / q'ori 'daughter'); forms, originating from the standard language and accommodated to the morphonological rules of the dialect (bavšivebi< bavšvebi 'children'; gadicvala<gardaicvala 's/he passed away'). The dictionary editor has a special field to mark regional variations of illustrations (indicating to the location of recording of a specific form).

The paper will present a detailed description how a multi-level annotation system of the corpus can be used for collection of maximum information about the Fereidanian dialect island and how the corpus represents the Fereidanian collective and individual linguistic worldview.

ქდკ-ს ინგილოური ლექსიკონის მიმართება ლექსიკოგრაფიულ წყაროებთან

მაია ბარიხაშვილი, ელენე ნაპირელი, რუსუდან პაპიაშვილი
თსუ არნ. ჩიქობავას სახ. ენათმეცნიერების ინსტიტუტი (საქართველო)
maiahereti@yahoo.com, elene.napireli@tsu.ge, ruso_papiashvili@hotmail.com

ინგილოური ონლაინლექსიკონი იქმნება ქართული დიალექტური კორპუსის (ქდკ-ს) წიაღში. ლექსიკონში ინტეგრირდება არსებული ბეჭდური ლექსიკონები; სამეცნიერო შრომებსა და ტექსტებში დადასტურებული, აგრეთვე 1935 წლის „შინამრეწველობის მასალებიდან“ და სხვა დიალექტოლოგიური ექსპედიციების არქივებიდან ამოკრებილი ლექსიკა (მ. ჯანაშვილი, ზ. ედილი, რ. ღამბაშიძე, გრ. იმნაიშვილი, ვ. აბაშვილი, ნ. როსტიაშვილი...).

ონლაინლექსიკონი არ წარმოადგენს ბეჭდური წყაროს ზუსტ ასლს. იგი სხვადასხვა სტრუქტურულ ველს მოიცავს: ლემა, შესატყვისი, განმარტება, გრამატიკული და ფონეტიკური ვარიაციები, ილუსტრაცია, შიდა ბმულები (იხილე; იგივეა რაც) და სხვ. ამ ველებში სტრუქტურირებული ბეჭდური წყაროს მონაცემები გარკვეული სახის ცვლილებას განიცდის. ეს ცვლილებები ძირითადად ლექსიკოგრაფიული რედაქტორის სტრუქტურითაა შეპირობებული, თუმცა ხშირად აუცილებელი ხდება უფრო არსებითი ხასიათის ჩარევა და ავტორისეული სიტყვა-სტატია ზოგჯერ საკმაოდ იცვლის სახეს.

ძირითადად საჭირო ხდება შემდეგი ტიპის ცვლილებები:

1.სწორდება, ზუსტდება ან ემატება მნიშვნელობა; მაგ.:

- ღüნეჲ – **ბუნტ ღüნეჲ** „ამბოხი ღვინო“ (რ. ღამბაშიძე, 1988); ამ განმარტების ბუნდოვანების გამო გადავამოწმეთ ამ სიტყვის მნიშვნელობა ინგილოური დიალექტის მატარებლებთან, აღმოჩნდა, რომ ეს არის ამღვრეული ღვინო, რაც შესაბამისად აისახა ონლაინლექსიკონში;
- წოდ – **წოდ-წოდ**: წოდ-წოდ დაგჭრი „ნაკუწ-ნაკუწად, ასო-ასო დაგჭრი, აგკუწავ“ (რ. ღამბაშიძე, 1988 – განმარტება არ მოჰყვებოდა), ეს ლემა ილუსტრაციის მიხედვით განვმარტეთ „ნაკუწ-ნაკუწად, ასო-ასო“;
- ლმ **გათეთრევაჲ**: თუთაჲ გათეთრევაჲ ნ თუთაჲ (რ. ღამბაშიძე, 1988 – განმარტება არ მოჰყვებოდა); დავადგინეთ მისი მნიშვნელობები: ქდკ. გათეთრევაჲ 1. „გათეთრება /შეღებვა, შეფერვა/“ 2. ავადმყოფობის დროს ფერის დაკარგვა;
- და მრავალი სხვა.

2. ზოგჯერ ლექსიკოგრაფიულ წყაროში არ არის გატარებული ლექსიკური ერთეულის წარმოდგენის ერთი პრინციპი. ონლაინლექსიკონში ვცდილობთ ამ ხარვეზის გასწორებას:

- ცალკეულ შემთხვევებში ლექსიკურ ერთეულად გამოყოფილი მესამე პირის ზმნური ფორმები ჩვენს ლექსიკონში აისახა გრამატიკული ვარიაციის ველში, ხოლო ლემად ჩაიწერა ასეთი ზმნების მასდარი; მაგ.: ლმ. გარაქოვს, განმ. ვარგა, საჭიროა (რ. ღამბაშიძე).

ქდკ. – გარაქოვა ქნაჲ „საჭიროდ ყოფნა, გამოდგომა“ **გარაქოვს** ფორმა შესაბამისი განმარტებით აისახა ილუსტრაციის ველში, ზმნური ფორმა კი გრამატიკული ვარიაციის შესაბამის გრაფაში.

- ცალკე სალექსიკონო ერთეულად გამოტანილი ყველა მიმღეობა (თუ ის გასუბსტანტივებული არაა) და ზმნიზედა (ზმნიზედად კვალიფიცირებული ვითარებითი ბრუნვის ფორმა) გავაერთიანეთ მასდარის ბუდეში, როგორც გრამატიკული ვარიაციები, ხოლო დიფერენცირებული მნიშვნელობები გავიტანეთ ილუსტრაციის ველში.
- როდესაც ზმნისწინს არ ჰქონდა სიტყვაწარმოებითი ფუნქცია, მთავარ ლემად გავიტანეთ უზმნისწინო ფორმა, ხოლო ზმნისწინიანი ვარიანტები შევიტანეთ ამ ლემის გრამატიკული ვარიაციის ველში (იხ. წიგნში მოკრბცაჲ, ხოლო ელექტრონულ გამოცემაში კრბცაჲ). ისეთი ფორმა, რომელიც არ არის მთავარი ლემის არც გრამატიკული და არც ფონეტიკური ვარიაცია, მნიშვნელობით კი ერთია, გაგვაქვს სინონიმად (იხ. ჟაგი ფაყლაჲ /ჟაგურ ფაყლაჲ)

3. ბეჭდური წყაროს გარკვეული ინფორმაცია გამოტოვებულია ქდკ-ს ელექტრონულ ლექსიკონში – ეს არის ბმულები, რომლებიც სხვა ლექსიკონების ამა თუ იმ ფორმასთან გვაგზავნის, თუმცა აღნიშნული ინფორმაცია ინახება სალექსიკონო სტატიის სამუშაო ველში და მომავალში, როცა ბაზას დაემატება ყველა სხვა ლექსიკონი, ივარაუდება შიდა ბმულებით ამ ინფორმაციის გააქტიურება.

4. ქდკ-ს ლექსიკონში არ აისახება ის ბმულები, რომლებიც ილუსტრაციაში გამოყენებულ მასალასთან გვაგზავნის. ბმულით „იხილე“ ჩვენს ლექსიკონში აღნიშნულია სალექსიკონო ერთეულთან დაკავშირებული ფორმები.

ქდკ-ს ლექსიკოგრაფიული რედაქტორში შესაძლებელია ისეთი შიდა ბმულების გააქტიურება, რომლებიც წყაროებში არაა მოცემული.

გარდა ამისა, რედაქტორი იძლევა იმის საშუალებას, რომ ნებისმიერ დროს დაემატოს ახალი ლექსიკოგრაფიული წყაროს მასალა ყველა საჭირო საავტორო მითითებით. მიმდინარეობს ქდკ-ს ინგილოური ლექსიკონის შედარება სხვა ლექსიკოგრაფიულ წყაროებთან, რომლებიც დღეისათვის უკვე ელექტრონულ რესურსად არის ქცეული. შედარების შედეგად აღმოჩენილი ყველა სახის ცვლილება ავტომატურად აისახება ონლაინლექსიკონში:

- იდენტურ ლემებს, რომელთაც იდენტური განმარტება აქვთ, დაემატება ყველა ის წყარო, რომელშიც დადასტურებული იქნება მსგავსი ფორმები.
- იდენტურ ლემებს, რომელთაც აქვთ განსხვავებული განმარტება, ემატება ახალი მნიშვნელობა ცალკე სალექსიკონო სტატიად შესაბამისი წყაროს მითითებით;
- განსხვავებული ლემები, რომლებიც მხოლოდ ამ ავტორთან დასტურდება, ინტეგრირდება ლექსიკონში ცალკე სალექსიკონო სტატიის სახით.

მოხსენებაში დეტალურად იქნება განხილული ქდკ-ს ინგილოური ონლაინლექსიკონის წყაროები და მათი ასახვის თავისებურებები.

The Relationship of the Ingiloan Dictionary of GDC to Lexicographic Sources

Maia Barikhashvili, Elene Napireli, Rusudan Papiashvili
Arn. Chikobava Institute of Linguistics, TSU (Georgia)
maiahereti@yahoo.com, elene.napireli@tsu.ge, ruso_papiashvili@hotmail.com

The online Ingiloan dictionary is developed within the Georgian Dialect Corpus (GDC). The dictionary will incorporate available print dictionaries, vocabulary from scholarly works and recorded texts, also from the 1935 *Cottage Industry Materials*, and from archives of other dialectological expeditions (M. Janashvili, Z. Edil, RF. Gambashidze, Gr. Imnaishvili, V. Abashvili, N. Rostiashvili, etc.).

An online dictionary is not an exact copy of its print source. It comprises various structural fields: lemma, equivalent, definition, grammatical and phonetic variations, illustrations, cross-references (*see; same as*), etc. Structured within these fields, data from a print source undergo certain changes. These changes are mostly conditioned by the structure of a lexicographic editor; however, frequently it is necessary to intervene in a more essential way and an original entry becomes rather modified.

Mostly the following kinds of changes become necessary:

1. A sense is either corrected, or specified or added.

2. Sometimes, a lexicographic source does not maintain a single principle for presenting of a lexical item. We try to reclaim such shortcomings:

- In individual cases, 3rd person verb forms, identified as lexical items, were reflected in a field of grammatical variation, and masdars of such verbs were recorded as lemmas;
- All participles (provided that they are not substantivized) and adverbs (adverbial case forms qualified as adverbs), having appeared as individual entries, were integrated into a field of a masdar as grammatical variations, and differential meanings were placed in a field of illustrations;
- Whenever a preverb did not have a word-building function, a preverbless form was drawn as a principal lemma, while preverbed variants were included in a field of grammatical variation of a lemma in question. A form, being neither a grammatical or phonetic variation of a principal lemma but identical in meaning, is drawn as a synonym.

3. Certain information of a print source are omitted in the GDC electronic dictionary – these are references to certain forms of other dictionaries; however, the information is stored in a working field of an entry, and, in the future, when the base incorporates all other dictionaries, the information will be activated by means of cross-referencing.

4. The GDC dictionary does not reflect the references to data used in illustrations. In our dictionary, the reference “See” refers to forms associated with an entry.

Within the lexicographic editor of GDC, cross-references, not appearing in sources, can be activated.

Besides, the editor provides an opportunity to eventually add data from a new lexicographic source referring to all necessary author information. The Ingiloan Dictionary of GDC is compared to other lexicographic sources, having already been made an electronic resource by now. All kinds of changes, detected as a result of comparison, are automatically reflected in the online dictionary:

- Identical lemmas, having identical definitions, will take on all the sources evidencing similar forms;
- Identical lemmas, having distinct definitions, will take on a new meaning as an individual entry, referring to its source;
- Distinct lemmas, occurring with only a certain author, will be integrated into the dictionary as individual entries.

The talk will provide a detailed discussion of the sources of the Online Ingiloan Dictionary of GDC and peculiarities of their appearance in it.

ქართული დიალექტური კორპუსის მორფოლოგიური ანოტირების კონცეფციისათვის

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

ქართული დიალექტური კორპუსი (ქდკ) შეიქმნა და თანმიმდევრულად ვითარდება უკვე რამდენიმე წელია. მიმდინარე ეტაპზე ვმუშაობთ მორფოლოგიური ანოტირების პრობლემებზე. კორპუსის მორფოლოგიური ანოტირების კონცეფცია მუშავდება კორპუსზე მუშაობის დაწყების დღიდან (ბერიძე, ნადარაია 2009). ბუნებრივია, სამუშაო პროცესი იწვევს ამ კონცეფციის განვითარების აუცილებლობას.

თავიდანვე დაიგეგმა, რომ მორფოლოგიური ანალიზის პროცესში გამოყენებული იქნებოდა ორი ძირითადი რესურსი: **სალიტერატურო ენის მორფოლოგიური ანალიზატორი Geo Trans-**ი და სპეციალურად დამუშავებული **ქართული დიალექტური ლექსიკონები**, როგორც ჩვენ მათ ვუწოდეთ, **ნახევრადავტომატური** ლექსიკონები.

ნახევრადავტომატური ლექსიკონების შექმნა გულისხმობს: არსებული ბეჭდური ლექსიკონების ელექტრონული ვერსიის შექმნას, უნიფიცირებას, მეთაური სიტყვისთვის – ლემისთვის პირველადი გრამატიკული ინფორმაციის (გრამატიკული ჯგუფის მარკერების) მიწერას. ასევე, ლექსიკონის სხვადასხვა სტრუქტურულ ველში (განმარტება, ილუსტრაცია, შენიშვნა...) დადასტურებული ფონეტიკური, გრამატიკული ან სიტყვაწარმოებითი ვარიაციების ლემასთან „მიბმას“ და მათთვის როგორც გრამატიკული, ისე პარადიგმული მახასიათებლების მიწერას. ამგვა-

რად „აღჭურვილი“ ლექსიკონი გამოიყენება ტექსტური სიტყვანის გარკვეული (ლექსიკონში ასახული) ნაწილის იდენტიფიცირებისა და მარკირებისათვის Geo trans-ის მეშვეობით (ბერიძე, ლორთქიფანიძე, ნადარაია 2015-1, 2015-2, 2014).

მორფოლოგიური ანალიზისთვის მნიშვნელოვან ინსტრუმენტს წარმოადგენს სისტემა Geo Trans-ი, რომლის საშუალებითაც იდენტიფიცირებული და გაანალიზებულია დიალექტური ტექსტების ის ნაწილი, რომელიც სალიტერატურო ენობრივ ინვენტართან საერთო ლექსიკურ და ფორმაწარმოებით ენობრივ მასალას მოიცავს (ამოცნობის დიაპაზონი 20% _ 45%).

დიალექტური კორპუსის ლინგვისტური ანოტირება ამ ეტაპზე გულისხმობს მორფოსინტაქსურ მახასიათებელთა ორსაფეხურიან იერარქიას. ესენია:

1. გრამატიკული ჯგუფის მახასიათებლები (არსებითი სახელი, ზედსართავი სახელი, ზმნა...);
2. ფორმაწარმოებითი მახასიათებლები (რიცხვი, ბრუნვა, პირი, მწკრივი და ა. შ.).

მეორე საფეხურის იერარქიაში გაერთიანებულია აგრეთვე ზოგი სიტყვაწარმოებითი, სემანტიკური, სინტაქსური ინფორმაციის მატარებელი მახასიათებელი. მოხსენებაში დეტალურად იქნება აღწერილი ქდკ-ს მორფოსინტაქსური მარკერების სისტემა.

Geo Trans-ის საშუალებით ანოტირების პროცესი ასეთი თანმიმდევრობით წარიმართება:

- ტექსტური მასალიდან იქმნება **სრული სიტყვანი** ცალკეული დიალექტებისათვის;
- სიტყვანი მუშავდება Geo Trans-ის საშუალებით _ იდენტიფიცირებულ სიტყვებს მიეწერება ორი დონის მორფოლოგიური მარკერი: გრამატიკული ჯგუფის მარკერი და ფორმაწარმოებითი მარკერი (სიტყვათა ნაწილს მიეწერება სიტყვაწარმოებითი ან სემანტიკური მარკერებიც);
- საერთო ტექსტური სიტყვანიდან გამოიყოფა ორი სია: ამოცნობილი (იდენტიფიცირებული) და ამოუცნობი (არაიდენტიფიცირებული) სიტყვების სიები;
- იდენტიფიცირებული სიტყვების სიაში გამოიყოფა ომონიმური და არაომონიმური სიტყვების სიები.

ამ პროცედურების შემდეგ ანოტირების პროცესი გრძელდება ქდკ-ს ანოტირების რედაქტორში.

ანოტირების რედაქტორი წარმოადგენს სპეციალურად შექმნილ პროგრამულ ხელსაწყოს, კორპუსის დამოუკიდებელ, მრავალფუნქციურ სტრუქტურულ ჩანართს, რომლის საშუალებითაც ხორციელდება:

- Geo Trans-ის მიერ შემოთავაზებული ”ამოცნობილი” სიების ატვირთვა რედაქტორში;
- სიების კავშირი **კონტექსტების ბაზასთან** (რაც იძლევა სიტყვის ყველა კონტექსტური რეალიზაციის გადამოწმების საშუალებას);
- სიების კავშირი **გრამატიკული მარკერების ბლოკთან** (რაც იძლევა არასწორი ანოტირების გასწორების საშუალებას);
- Geo Trans-ის მიერ შემოთავაზებული ანალიზის ტესტირება;
- ანალიზში დაშვებული შეცდომის გასწორება;

- ტესტირებული (გასწორებული) სიის ავტომატურად მინიჭება კონტექსტებისთვის;
- რედაქტორს აქვს დამატებითი ხელსაწყოები, რომლებიც საშუალებას გვაძლევს: დავაგენერიროთ ნებისმიერი დიალექტის სიტყვების სია ნებისმიერ დროს; ჩამოვტვირთოთ მიმდინარე ხედი; ავტვირთოთ ანოტირებული სიტყვების სია; აგრეთვე მივიღოთ სტატისტიკური ინფორმაცია, კონკრეტული სიის ან მისი ნაწილის შესახებ (თუ რამდენი სიტყვაფორმაა, რამდენია ანოტირებული, ამათგან რამდენია ომონიმური და რამდენი – არაომონიმური).

ანოტირების რედაქტორში მუშაობას აადვილებს მრავალდონიანი საძიებო სისტემა, რომელიც სიის „გაფილტვრის“ საშუალებას იძლევა სხვადასხვა მახასიათებლის მიხედვით, მაგალითად, ანოტირებაზე მომუშავე ოპერატორს შეუძლია თავად შექმნას მისთვის საჭირო სია შემდეგი მახასიათებლების მიხედვით:

- მხოლოდ ლემა, ლემა და სიტყვაფორმა, სრული სია;
- გრამატიკული ჯგუფის მარკერი (სრული სია, მხოლოდ არსებითი სახელი, მხოლოდ ზმნა, მხოლოდ ზედსართავი სახელი და ა. შ.);
- სრული სიტყვა, სიტყვის დასაწყისი, სიტყვის სეგმენტი ან სიტყვის დაბოლოება;
- სრული სია, ომონიმური (ანოტირებული), არაომონიმური (ანოტირებული) და არაანოტირებული სიტყვა.

ამავე დროს, რედაქტორში არის ცვლილებათა კონტროლის საშუალება: საერთო სიაში შეცვლილი მარკერი განსხვავებული ფერის ფონით არის წარმოდგენილი, შესაძლებელია ცალკე სიად წარმოდგენა ყველა იმ სიტყვისა, რომელთა მარკირებაშიც ოპერატორმა გარკვეული, თუნდაც სულ უმნიშვნელო ცვლილება შეიტანა.

ანოტირების პროცესის ეფექტურობის გასაზრდელად მივმართეთ მაღალი სიხშირის სიტყვების ანალიზის ექსპერიმენტს: გამოიყო და ცალკე დამუშავდა 1000-ზე მეტ კონტექსტში რეალიზებული სიტყვების სია. გამოვლინდა 140 ყველაზე ხშირად რეალიზებული სიტყვაფორმა. ეს სიტყვები 500 000-მდე კონტექსტშია დამოწმებული. ამათგან თითქმის ნახევარი არაომონიმურია და ავტომატურად შეიძლება მოინიშნოს და აღიწეროს ყველა დიალექტში ან დიალექტთა ნაწილში. ასეთი ავტომატური ოპერაციის შედეგად 200 000-მდე კონტექსტი ავტომატურად ანოტირდა კორპუსში.

„ხშირი სიტყვების“ მეორე ნახევარს შეადგენენ სიტყვები, რომლებიც ომონიმურები არიან ამოსავალ სისტემაშივე, ან წარმოადგენენ დიალექტის შიგნით, ან დიალექტსა და სალიტერატურო ენას შორის გამოვლენილ გრამატიკულ ომონიმებს. ანოტირებისას სიტყვანის ეს ნაწილი უკვე დიფერენცირებულ მიდგომას საჭიროებს.

ე.წ. „ამოუცნობი“ სიტყვების სიების შემდგომი დამუშავების ავტომატიზაციისათვის იქმნება მორფოლოგიური ანალიზატორის დიალექტური მოდული. ეს სამუშაო გულისხმობს სალიტერატურო ენის ანალიზატორში იმ წესების დამატებას, რომლებიც გამოვლინდა და შემუშავდა დიალექტთა დონეზე. წესები იქმნება ცალკეული დიალექტებისთვის და ანალიზიც ცალკეული დიალექტის სიტყვანში ხორციელდება.

ლიტერატურა:

Beridze Marina, Liana lordkipanidze, David Nadaraia – Georgian Dialect Corpus: Problems and Prospects"; Jost Gippet/ Ralf Gehrke (eds): Historical Corpora. Challenges and Perspectives. The Georgian Dialect Corpus: problems and procpects. (CLIP),vol.5 Tubingen (Narr), 2015.

Beridze Marina, Liana lordkipanidze, David Nadaraia – lexicographic conteption of Georgian Dialect Corpus and problems of its morphological annotation, Applied linguistics in ressearch and education _ Proceedings of the VII)-th International Bianual Conference, Saint-Peterburg 2014.

Beridze Marina, Liana lordkipanidze, David Nadaraia – The Corpus of Georgian Dialects Dialect Dictionaries with the Functions of Representativeness and Morphological Annotation in Georgian Dialect Corpus, Logic, Language, and Computation. *Proceedings of 10th International Tbilisi Symposium on Logic, Language, and Computation*, TbiLLC 2013, Gudauri, 2013.

Beridze Marina, Liana lordkipanidze – The Corpus of Georgian Dialects, NLP, Corpus Linguistics, Corpus Based Grammar Research Proceedings of Fifth International Conference Smolenice, Slovakia, 25–27 November, Bratislava, Slovakia 2009.

On the Concept of the Morphological Annotation of the Georgian Dialect Corpus

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

The Georgian Dialect Corpus (GDC) has been designed and has been consistently developing for several years. At the current stage, we work on the problems of the morphological annotation. The concept of the morphological annotation of the corpus is being developed since the inception of the corpus work (Beridze, Nadaraia 2009). Naturally enough, the working process causes the necessity of the development of the concept.

It was initially planned that two resources would be used in the process of the morphological analysis: **Geo Trans**, **Morphological Parser of the Standard Language**, and specially processed **Georgian dialect dictionaries**, as we referred to them, **semi-automated** dictionaries.

The creation of **semi-automated** dictionaries implies: creation of electronic versions of available paper dictionaries, their unification, assigning of primary grammatical information (grammatical group markers) to a head-word – a lemma; in addition, "to attach" phonetic, grammatical, and/or word-formation variations, occurring in various structural fields of a dictionary (definition, illustration, note…) to a lemma and to assign both grammatical and paradigmatic features to them. A dictionary, "equipped" in such a way, is used for identifying and marking of a certain portion of a textual word list

(occurring in a dictionary) by means of **Geo Trans** (Beridze, Lordkipanidze, Nadaraia 2015-1, 2015-2, 2014).

A significant tool for morphological analysis is the system Geo Trans, having analyzed a part of the dialect texts with common lexical and inflectional data with the standard language inventory (recognition accuracy 20%-45%).

Currently, the linguistic annotation of the Georgian Dialect Corpus implies a two-stage hierarchy of morphosyntactic features; they are:

1. Grammatical group features (noun, adjective, verb…);
2. Inflectional features (number, declension, person, screeve, etc.).

The hierarchy of the second stage incorporates features with some word-formation, semantic, syntactic information. The paper will provide a detailed description of the set of morphosyntactic markers (tags) of the GDC.

By means of Geo Trans, the annotation process will proceed in the following sequence:

- a complete word list for individual dialects will be established from the textual data;
- the word list will be established by means of Geo Trans – identified words will be assigned bi-level morphological markers: a grammatical group marker and an inflection marker (some words will be assigned either word-formation or semantic markers as well);
- two lists will be identified from the common word list: identified and unidentified word lists;
- lists of ambiguous and non-ambiguous words will be identified from the list of identified words.

Following the procedure, the annotation process will be continued within the annotation editor of GDC.

The annotation editor is a specially developed software tool, an independent multi-functional structural insertion of the corpus, by means of which

- "identified" lists, provided by Geo Trans, are uploaded to the editor;
- a link of lists to the base of contexts is established (providing for checking of all contextual occurrences);
- a link of lists to the block of grammatical markers is established (providing for revising of incorrect annotations);
- provides for testing of the analysis by Geo Trans;
- provides for revising of committed mistakes in analysis;
- provides for automated assigning of a tests list to contexts;
- the editor has additional tools, providing for: generation of a word list of any dialect at any time; download a current view; upload a list of annotated words; besides, receive statistical information about an individual list or its part (number of tokens, number of annotated ones, number of ambiguous and non-ambiguous ones).

Work within the annotation editor becomes easier owing to the multi-level query system, providing for filtering a list according to various features; for instance, an operator, working on annotation, can generate a necessary list in accordance with the following features:

- only a lemma, a lemma and a token, a complete list;
- grammatical group marker (complete list, only a noun, a verb, only an adjective, etc.);
- a complete word, beginning of a word, a segment of a word, an ending of a word;
- a complete list, ambiguous (annotated), non-ambiguous (annotated) and non-annotated word.

Meanwhile, the editor allows for the control of changes: a changed marker in the common list is presented with a distinct color background. All the words, in the marking of which an operator made a certain, even a slight change, can be presented as an individual list.

In order to increase the effectiveness of the annotation process, we conducted an experiment with the analysis of high-frequency words: a list of words, occurring in more than a thousand contexts, was identified and processed. 140 most frequently occurring words were identified. Almost half of them are non-ambiguous and can be automatically marked and described in all dialects or in a part of dialects. As a result of such an automated operation, up to 200 000 contexts were automatically annotated in the corpus.

Another part of “frequent words” are those which are ambiguous either in the initial system or within a dialect or between a dialect and the standard language. This part should be treated differentially in annotation.

For the sake of the automation of the further processing of so called “unidentified” words, a dialect module of the morphological parser is developed. The work involves adding of the rules to the standard language parser, which were identified and processed at the dialectal level. Rules are established for individual dialects and analysis is carried out within a word list of an individual dialect.

პარალელური და შედარებითი ტექსტური კორპუსები ლექსიკოგრაფიაში

ლარისა ბელიაევა
გერცენის სახელობის სანკტ პეტერბურგის სახელმწიფო პედაგოგიური უნივერსიტეტი (რუსეთი)
lauranbel@gmail.com

მეცნიერებასა და ტექნიკაში მომხდარი დრამატული ცვლილებების, კვლევის ახალი მიმართულებების წარმოშობისა და ცოდნის ახალი სფეროების გაჩენის შედეგად მივიღეთ მკვეთრი ჩამორჩენა იმ სპეციალური ლექსიკოგრაფიული რესურსების მხრივ, რომლებიც ხელს უნდა

უწყობდეს მთარგმნელობით საქმიანობას. სპეციალური ლექსიკოგრაფიული სისტემები, რომლებიც გამიზნულია თარგმნითი ლექსიკონების შესაქმნელად და სამართავად, რეალურ სიტუაციაში მოითხოვს ახალი ტერმინების სისტემატურ დაფიქსირებას, რისთვისაც, თავის მხრივ, საჭიროა შემუშავდეს და გამოყენებულ იქნეს ორივე მეთოდი – ლექსიკური ერთეულების (ერთსიტყვიანი ტერმინები და ტერმინოლოგიური შესიტყვებები, რომლებიც სპეციალურ სამეცნიერო ტექსტებში გვხვდება) წმინდა ლინგვისტური და სტატისტიკურ-ლინგვისტური ანალიზი და მათი ავტომატური ამოკრების მეტრიკა. ტექსტებიდან ავტომატურად ამოსაკრებ ლექსიკურ ერთეულებს ტერმინოლოგიური კანდიდატები ეწოდება. ტერმინთა ამოკრების პროცედურები და სპეციალური მეტრიკა საშუალებას გვაძლევს შევაფასოთ:

- ლექსიკურ ერთეულთა თავსებადობის პრეფერენციები (unithood), მათი სინტაგმატური სიახლოვე;
- იმ ამოკრებილ შესიტყვებათა ტერმინობა (termhood), რომლებიც ტერმინოლოგიურ კანდიდატებად მიიჩნევა;
- გარკვეული შესიტყვების მახასიათებლები სპეციალურ ტექსტურ კორპუსებში ან სპეციალური ენის ტერმინოლოგიაში.

უფრო მეტიც, მუშავდება სპეციალური მეთოდები, რომლებიც საშუალებას გვაძლევს, შევაფასოთ ტექსტის გარკვეული ლექსიკური ერთეულების პოტენციალი, რამდენად ფუნქციონირებენ ისინი, როგორც საკვანძო სიტყვები.

თარგმნითი ლექსიკონების შექმნის თანამედროვე მიდგომა გულისხმობს რეალური ტექსტების კორპუსის ფორმირებასა და გამოყენებას, რადაც შეიძლება მივიჩნიოთ მონაცემთა ბაზები, რომლებიც არა მარტო კვლევით, არამედ პრაქტიკულ ლექსიკოგრაფიულ ამოცანებსაც წყვეტენ. როგორც წესი, წერილობითი ტექსტების კორპუსები მოიცავენ ისეთ ტექსტებს, რომლებიც გაირჩევა სფეროს, ავტორის, ფუნქციისა და ა.შ. მიხედვით, ასევე დამატებული აქვთ მარკირება კორპუსში პოვნიერი წინადადებების ფორმატისა და სინტაქსური სტრუქტურის მიხედვით. ეს მარკირება შეიძლება ეფუძნებოდეს ანალიზის შედეგებს და იგი განსაზღვრავს ლექსიკურ ერთეულთა მარკერების ნაწილს მეტყველების ნაწილთა მიხედვით.

კორპუსის ნაწილად გამოყენებული სამეცნიერო ტექსტი წარმოადგენს გარკვეული სუბიექტის სამეცნიერო სფეროში კოგნიტური და საკომუნიკაციო მოღვაწეობის შედეგს, რომელიც მიმართულია სპეციფიკური ობიექტისაკენ – რეალობის ფაქტებისა და პროცესებისაკენ. სამეცნიერო ტექსტის სტრუქტურა ასახავს ადამიანის კოგნიტური მოღვაწეობის ძირითად კომპონენტს და შეიცავს სამეცნიერო ცოდნას, როგორც თავის პროდუქტს; ამგვარი ტექსტი შეიძლება გამოვიყენოთ მონაცემთა და ინფორმაციის მოპოვების პროცესებში. ტექსტური მონაცემები არ არის სტრუქტურირებული, სამეცნიერო ტექსტსაც აქვს ძალიან რთული შინაგანი სტრუქტურა და იგი წარმოადგენს ცოდნის აღმოჩენის წყაროს ტექსტებში. გარკვეული ტექსტის ყალიბები თანხვდება იმ ნომინაციებს, რომლებიც გამოიყენება ობიექტისათვის მოცემულ გამოკვლევაში, ხოლო რეალური სამყაროსა და სუბიექტური რეალობის რეპრეზენტაცია დაკავშირებულია ცოდნის რეპრეზენტაციასთან.

პარალელური ტექსტების კორპუსებიდან ტერმინების ამოკრების ავტომატური სისტემების შექმნის კონცეფციას 20 წელზე მეტია, რაც იკვლევენ და იგი ნაწილობრივ რეალიზებულია სხვადასხვა ტერმინოლოგიურ პროექტში. ამ იდეის რეალიზაციაში ჩვენ ვხედავთ არა მარტო იმ

სირთულეებს, რომლებიც დაკავშირებულია იმასთან, რომ სხვადასხვა ენის ტერმინოლოგიური სისტემები არ არის სიმეტრიული, არამედ სპეციფიკურ პრობლემასაც, რომელიც მდგომარეობს მართებული თარგმანების შერჩევაში პარალელური ტექსტების კორპუსებისათვის, რადგანაც თარგმანის ხარისხი (როგორც მანქანური, ისე ადამიანური) საკმაოდ ხშირად არ არის ადეკვატური. ამიტომაც, სავსებით ბუნებრივია შედარებით ახალი იდეა შედარებითი ტექსტების კორპუსების შესაქმნელად და გამოსაყენებლად, რომლის კონსტრუქცია შეიძლება ეფუძნებოდეს სპეციალისტთა შეფასებას შესადარებელი ენების ტექსტების შესახებ.

შესადარებელი ტექსტების სპეციფიკურ წყაროს წარმოადგენს კონფერენციების მასალები, რომლებიც ერთსა და იმავე სამეცნიერო პრობლემატიკას ეხება, მაგრამ ორგანიზებულია სხვადასხვა ქვეყანაში და იმართება სხვადასხვა სამუშაო ენაზე. რაც შეეხება საერთაშორისო სამეცნიერო კონფერენციებს, ძირითადი სამუშაო ენა ინგლისურია; ამგვარად, ინგლისურ-რუსული და რუსულ-ინგლისური თარგმნითი ლექსიკონების პრობლემის მოსაგვარებლად აუცილებელია ერთი საკითხისადმი მიძღვნილი, რუსულ და ინგლისურ ენებზე ჩატარებული სამეცნიერო კონფერენციების მასალების შემცველი სპეციალური საკვლევი კორპუსების შექმნა.

შედარებით კორპუსში სამეცნიერო ტექსტი უნდა განვიხილოთ, როგორც შედეგი ინფორმაციის გადაცემისა და როგორც წყარო (ამოსავალი წერტილი) ინფორმაციის მოპოვებისა და ამოკრებისა. ამის კვალად სამეცნიერო ტექსტის შინაარსი, განსაკუთრებით მისი მნიშვნელობა, რომელიც უნივერსალურია და შეიძლება ამოვკრიბოთ ავტორისა და მიმღების თეზაურუსების მინიმალური დამთხვევის შემთხვევაში, ძირითადად განისაზღვრება ინფორმაციით განსახილველ ობიექტთა შესახებ, რაც ენათმეცნიერულად მათი სახელწოდებებით აღიწერება – სახელები და სახელური ჯგუფები.

ამგვარი ტექსტური კორპუსების შედარებითი ანალიზი კიდევ რამდენიმე მახეს წარმოაჩენს. „ინგლისურენოვანი“ საკონფერენციო ტექსტები ძირითადად გლობალურ ინგლისურზეა შესრულებული, რაც სინტაქსურ სტრუქტურებში ხშირი შეცდომების პოვნიერებას ნიშნავს, რომელთაც უშვებენ ავტორები თავიანთი დედაენის გავლენით; იქ ასევე არ გვაქვს ტერმინოლოგიური ჰარმონიზაცია, რომლის ფარგლებშიც ტერმინები წარმოადგენენ ავტორის დედაენის სათანადო ლექსიკური ერთეულების თარგმანებს, ნაცვლად იმისა, რომ გამოყენებულ იქნეს სტანდარტიზებული ნომინაციები. თავის მხრივ, „რუსულენოვანი“ ტექსტები „დამძიმებულია“ შემზარავი ოფიციალური სტილით, თავკიდურა ობიექტიანი სინტაქსური კონსტრუქციებით და მკაფიო საზღვრების არქონით იმ სახელურ ჯგუფებს შორის, რომლებიც ახდენენ ტერმინთა სახელდებას და ასრულებენ სხვადასხვა როლს წინადადებაში. ეს კი რუსული ენის სახელთა ლექსიკისთვის დამახასიათებელ ვითარებაში, როდესაც ბრუნვის ფორმები მეტისმეტად მრავალმნიშვნელიანია, იწვევს იმას, რომ შეუძლებელი ხდება ტერმინთა საზღვრებისა და სტრუქტურის მართებულად განსაზღვრა. მაგრამ სამეცნიერო ტექსტის შინაარსი, განსაკუთრებით მისი მნიშვნელობა, განსაკუთრებით მისი ნაწილი, რომელიც უნივერსალურია და შეიძლება ამოვკრიბოთ ავტორისა და მიმღების თეზაურუსების მინიმალური დამთხვევის შემთხვევაში, ძირითადად განისაზღვრება ინფორმაციით განსახილველ ობიექტთა შესახებ, რაც ენათმეცნიერულად მათი სახელწოდებებით აღიწერება – სახელები და სახელური ჯგუფები. მაგრამ ამ ჯგუფების სტრუქტურა არ ემთხვევა ერთმანეთს ინგლისურ და რუსულ ენებში.

ტერმინოლოგიური კანდიდატების ამოკრების პროცესი პარალელური კორპუსის შემთხვევაში ეფუძნება წინადადებათა მიხედვით შეთანადებასა და ფრაზების მიხედვით კოორდინაციას წინასწარ შეთანადებულ ორ წინადადებას შორის, რაც ემყარება წინადადების საზღვართა და ნაწილთა ფორმალურ მახასიათებლებს, მოცულობის შესაბამისობასა და ტექსტის პრაგმატიკულ სტრუქტურებს. მიუხედავად ყველა თავჩენილი ტექნიკური და ენობრივი სირთულისა, ეს პროცესი სავსებით რეალიზებადია. ტექსტების შედარებითი კორპუსის გამოყენებისას ოდენ ტერმინოლოგიური შეთანადებაა შესაძლებელი, რაც ემყარება, პირველი, ტერმინოლოგიური ერთეულების გამოვლენილ ერთენოვან დოკუმენტებს, რომელიც ახასიათებს კორპუსის ორივე ნაწილს, და, მეორე, მათ შედარებას, როგორც თარგმანის კანდიდატებისას; ასევე, სტაბილური შესიტყვებების ძიებას, რომლებშიც ეს ერთენოვანი ტერმინები ბირთვების როლში გამოდიან. შემდგომი შედარებითი ანალიზი მოითხოვს ცოდნას ავტომატური თარგმნითი ლექსიკონებიდან და მანქანური თარგმნის სისტემებიდან, რაც საშუალებას მოგვცემს, შევამოწმოთ შერჩეული ტერმინოლოგიური წყვილები.

სპეციალური ენების წყვილთათვის პარალელური ან შედარებითი კორპუსების შექმნა და რთული ენობრივი და სტატისტიკური პარამეტრების იდენტიფიკაცია წარმოადგენს ურთიერთდაკავშირებულ პრობლემებს და წარმოაჩენს ახალ ტერმინებს და მათ თარგმანს.

Parallel and Comparable Text Corpora in Lexicography

Larisa Beliaeva
"Herzen" State Pedagogical University of Russia St. Petersburg (Russia)
lauranbel@gmail.com

Dramatic changing of the situation in science and technology, origin of new lines of investigation and, and what is more, new knowledge domains results in sharp backlog of specialized lexicographic resources, supporting translation work. Special lexicographic systems, meant for creation and management of translation dictionaries, in actual condition require constant "tracking" of new terms, that in turn assumes development and use both methods of purely linguistic and statistic-linguistic analysis and metrics of automatic extraction of lexical units – one-word terms and terminological collocations found in real special scientific texts. Lexical units to be automatically extracted from the texts are called term candidates (TC). Term extraction procedures and special metrics permit to evaluate:

- compatibility preferences of lexical units (unithood), their syntagmatic proximity,
- termhood of the extracted collocations considered as TC,
- charactiristics (salience) of a certain collocation in a specialized text corpora or in the terminology of a language for special purposes (LSP).

Furthermore special methods enabling to evaluate potential of certain lexical units (LE) of the text to function as key words (keyness) are under development.

Modern approach to creation of translation dictionaries assumes formation and using real text corpora, which can be considered as databases for solving not only research, but also practical lexicographic tasks. As a rule, written text corpora include texts which are specialized according to a domain, author, function etc, as well as added with tagging according to format and syntactic structure of sentences in the corpus. This tagging could be based on the parsing results and determines the part-of-speech tags of lexical units.

A scientific text used as a corpora constituent is the result of cognitive and communicative activities of the certain subject in a scientific domain aimed at a specific object – facts and processes of actual reality. The structure of a scientific text reflects the main components of a human scientific cognitive activity and contains the scientific knowledge as its product, such text can be used in the processes of data and information mining. The textual data are not unstructured and a scientific text has a very complex implicit structure and is the source for knowledge discovery in texts. The patterns of a particular text follows the nominations used for the object of the study, a representation of real world and subjective reality is related to knowledge representation.

The conception of creation of automated systems for term extraction from parallel text corpora has been under study for more than 20 years and is partially realized in various terminological projects. In realization of this idea we can see not only complexities, related to the fact that terminological systems of different languages are not symmetrical, a specific problem is selection of proper translations for parallel text corpora, as the quality of translations (both machine and human) every so often is not adequate. Therefore the relatively new idea to build and use comparable text corpora, construction of which could be based on expert evaluation of texts on compared languages, is quite natural.

A specific source of comparable texts is proceedings of conferences, devoted to the same scientific problems, but organized in different countries and with different working languages. As for international scientific conferences the main working language is English, thus for solving the problem of English-Russian and Russian-English translation dictionaries it is necessary to build specialized research corpora of conference proceedings with the same subject in English and Russian languages is expedient.

Scientific text in a comparable corpus is to be considered as the result of information transfer and the source (starting point) of information mining and extraction. Thereafter the scientific text content, especially the part of its sense which is universal and can be extracted in case of minimum coincidence of the author's and recipient's thesauri, is mainly determined by information on the objects under consideration, linguistically described by their names – nouns and noun phrases.

At comparative analysis of such text corpora shows several additional pitfalls. "English" conference texts for the most part are written in global English, that means frequent infringement of syntactic sentence structures, caused by the authors' native language influence, and absence of terminology harmonization, due to which terms represent translations of appropriate lexical units of native author language, instead of using the standardized nominations. "Russian" texts, in turn, are "weighed down" by dreadful scientific officialese, frequent use of syntactic structures with object in the first position of sentence and absence of explicit boundaries between noun phrases that nominate the terms and perform different role in a sentence. In the situation of the overdeveloped case ambiguity, which is characteristic for noun lexicon of Russian language, it results in impossibility of proper

determination of term boundaries and structure. But the scientific text content, especially the part of its sense which is universal and can be extracted in case of minimum coincidence of the author's and recipient's thesauri, is mainly determined by information on the objects under consideration, linguistically described by their names – nouns and noun phrases. But the structures of these phrases do not coincide in English and Russian languages.

The process of TC extraction in case of parallel corpora is based on sentence-by-sentence alignment and phrase-by phrase coordination between two sentences previously aligned which relies on formal characteristics of boundaries and parts of sentences, conformity of volume and pragmatic text structures. Under all arising technical and linguistic complexities this process is quite realizable. When using comparable text corpora only terminological alignment is possible, which relies on first, revealing monolingual files of terminological units characteristic for both parts of a corpus and second, their comparison as pair of translation candidates, as well as search of stable collocation with these monolingual terms as nuclei. Further comparative analysis requires knowledge from automated translation dictionaries and machine translation systems, enabling to verify the term pairs chosen.

Building a parallel or comparable text corpora for a certain pairs of languages for special purposes and identification of complex of linguistic and statistical parameters represent correlated problems and find new terms and their translations

References:

Belyaeva, L. Scientific Text Corpora as a Lexicographic Source // SLOVKO 2009. NLP, Corpus Linguistics, Corpus Based Grammar Research, Proc. from the Intern.Conference, November 25 – 27 2009, Smolenice, Slovakia. – pp. 19-25

Beliaeva L. Applied Lexicography and Scientific Text Corpora // Transcations on Business and Engineering Intelligent Applications. Galina Setlak, Kassimir Markov (ed.). Rzeshov, Poland: ITHEA, 2014. – pp. 55-63

Delpech E., Daille B. Dealing with lexicon acquired from comparable corpora: validation and exchange // Proceedings, 9th Conference on Terminology and Knowledge Engineering (TKE). – Fiontar, Dublin City University, 2010. Pp. 229-223.conformity.

Delgado, M., Martin-Bautista, M.J., Sanchez, D., Vila, M.A. Mining Text Data: Special Features and Patterns // Lecture Notes In Computer Science, Vol. 2442, Springer-Verlag GmbH, 2002. – pp. 140-151

Feldman, R., Dagan, I. Knowledge discovery in textual databases (KDT) // Proceedings of the 1st Int. Conference on Knowledge Discovery and Data Mining (KDD-95), AAAI Press, 1995. – pp. 112–117

TTC Project: Terminology Extraction, Translation Tools and Comparable Corpora //URL: http://www.ttc-project.eu/about-ttc/concept-and-objectives

ქართულ-ინგლისური მანქანური თარგმანის ერთი საკვანძო საკითხისათვის (ზმნური სიტყვაფორმების შესატყვისობა ქართულსა და ინგლისურში)

კახა გაბუნია
ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)
kgabunia@cciir.ge

1. ორმხრივი მანქანური თარგმანი რამდენიმე კომპონენტს მოიცავს, რომელთა ეტაპობრივი გავლის გარეშე, ადეკვატური შედეგის მიღება შეუძლებელია; რასაკვირველია, ეს კომპონენტები ურთიერთკავშირშია და ნებისმიერ ეტაპზე წინა ეტაპის მონაცემთა კორექტირება აუცილებელია:

ა. ლექსიკურ ერთეულთა დამუშავება ორივე ენაში და ზუსტი შესატყვისობის დადგენა (ომონიმიის, სინონიმიის, პოლისემიისა და სხვა პრობლემების გათვალისწინებით);
ბ. სიტყვაფორმათა ანალიზი და სალექსიკონო (სახელური თუ ზმნური ფორმაწარმოების საყრდენი ფუძის გამოყოფა) ერთეულების გამოყოფა;
გ. სინტაქსური ანალიზატორის შექმნა, რომელიც წინადადების ფარგლებში დასრულებული აზრის სრულფასოვანი თარგმანის ამოცანას გადაჭრის.

წინამდებარე მოხსენება სიტყვაფორმათა ანალიზს ეხება (მორფოლოგიის სფერო), თუმცა, საუბარი, ასევე, იქნება ამ ტიპის ანალიზის მნიშვნელობაზე სინტაქსური ანალიზატორის შესაქმნელად.

2. ქართული ტექსტების მორფოლოგიური, სინტაქსური და სემანტიკური ანალიზის გარეშე შეუძლებელია ისეთი პრობლემის გადაჭრა, როგორიცაა მანქანური თარგმანი ქართული ენიდან სხვა ენაზე (ამ შემთხვევაში საუბარია ქართულ-ინგლისურ / ინგლისურ-ქართულ სათარგმნ პროგრამაზე).

ქართულ ენაში წინადადების ხერხემალია ზმნა; ზმნური ფორმის აგების კანონზომიერებებს განსაკუთრებული მნიშვნელობა ენიჭება თარგმანის განხორციელების პროცესში.

ზმნური ფორმა წარმოადგენს მორფემათა კანონზომიერ თანამიმდევრობას, რომლის ცენტრშიც დგას ზმნური პარადიგმის (იგულისხმება უღლების პარადიგმა) ფუძე. პარადიგმის ფუძის მარცხნივაც და მარჯვნივაც განლაგებულია როგორც წარმოქმის, ისე – უღლების კატეგორიების გამომხატველი აფიქსები (მარცხნივ – გვარის, ვერსიის, პირისა და ობიექტის რიცხვის მარკერები, ხოლო მარჯვნივ – სუფიქსები: თემის ნიშანი, რომელიც მწკრივის მარკერია (უთურგაიძე, ეზუგბაია, 2011...), სავრცობი, კონკრეტული დრო-კილოს მაწარმოებელი სუფიქსი, პირის ნიშანი, რიცხვის ნიშანი). ამ უკიდურესად მწყობრი და განუმეორებელი მორფოლოგიური სტრუქტურის (ცალკე განხილვის საგანია მორფოლოგიური ომონიმიის პრობლემა) ანალიზის საფუძველზე განისაზღვრება ამა თუ იმ კონკრეტული ზმნის სემანტიკურ კატეგორიათა „ჯაჭვი“, რომელიც, თავის მხრივ, განსაზღვრავს მწკრივის / ნაკვთის სტატუსს. მაგალითად, ზმნური სიტყვაფორმა

„ააშენა“ დროის მიხედვით – წარსულია, კილოს მიხედვით – თხრობითი, ასპექტი – სრულია. მხოლოდ ერთი მახასიათებლის მიხედვით განსხვავდება ამ ფორმისგან ზმნა *„აშენებდა“* (დრო – წარსული, კილო – თხრობითი, ასპექტი – უსრული)... ხშირად ერთი მწკრივის ფორმა მეორისგან რამდენიმე მახასიათებლით განსხვავდება, თუმცა მწკრივის (ამ შემთხვევაში უმჯობესი იქნებოდა ტერმინ *ნაკვთის* გამოყენება) სტატუსი უცვლელი რჩება: შდრ. მეორე პირის წყვეტილის თხრობითი კილოს ფორმა *„ააშენე“* (კონტექსტი: *„შარშან შენ სახლი ააშენე“* – წარსული დრო, თხრობითი კილო, სრული ასპექტი...) და ბრძანებითი კილოს გამომხატველი *„ააშენე!“* (კონტექსტი: *„მომავალ წელს სახლი ააშენე!“* – მომავალი დრო, ბრძანებითი კილო, სრული ასპექტი...).

მწკრივი უნდა განვიხილოთ, როგორც შედგენილი სემანტიკური კატეგორია, რომელიც რამდენიმე სემანტიკურ მახასიათებელს / კომპონენტს მოიცავს:

დრო	კილო	ასპექტი	გზისობა	აქტი	თანამდევრობა

აქვე უნდა აღინიშნოს, რომ მწკრივის სემანტიკური სტრუქტურა უბრალო მექანიკური „ჯაჭვი“ კი არ არის, არამედ ერთმანეთთან შეთავსებადი კომპონენტების თვისებრივი ერთობლიობა; მიუხედავად იმისა, რომ შეიძლება ამ კომბინაციაში ერთი (ან რამდენიმე) კომპონენტი შეიცვალოს, თითოეულ მწკრივთან დაკავშირებულ კომბინაციათა რიცხვი სასრულია და სხვა მწკრივი ამავე კომბინაციას არასოდეს გაიმეორებს...

მოხსენებაში მოცემულია ცდა, აღიწეროს ქართული ზმნის მწკრივებთან დაკავშირებული ყველა შესაძლო კომბინაცია (კონტექსტების გათვალისწინებით) სემანტიკური შემადგენლების კომბინატორიკის თვალსაზრისით.

3. ქართულ-ინგლისური ავტომატური სათარმნი პროგრამისთვის უმნიშვნელოვანესი ამოცანაა ქართული და ინგლისური ზმნური სიტყვაფორმების შესაბამისობაში მოყვანა (იგულისხმება მწკრივთა შესატყვისობა): სხვაგვარად, ყოველი კონკრეტული ქართული მწკრივის შესაბამისი ინგლისური ზმნური კონსტრუქციის აგება.

აღსანიშნავია, რომ ყოველ კონსტრუქციას აქვს თავისი სემასიოლოგიური სტრუქტურა; დასადგენია შესაბამისი სემასიოლოგიური კონსტრუქციების ინგლისურენოვანი შესატყვისობების დადგენა.

ზემოთ წარმოდგენილი მწკრივის სემანტიკური კომპონენტები უნივერსალურ ხასიათს ატარებს და საკმაოდ იდენტურ სურათს გვაძლევს ქართული და ინგლისური „მწკრივების“ ურთიერთმიმართების თვალსაზრისით. შესაძლებლად მიგვაჩნია ამ კომპონენტების კომბინაციათა შესატყვისობის დადგენა ორივე ენისათვის, რაც მეტ-ნაკლები სიზუსტით მოგვცემს ზმნური ფორმის ადეკვატურ თარგმანს (რასაკვირველია, არ გამოვრიცხავთ ხელოვნური, არაბუნებრივი კონსტრუქციების აგებას, რომელთა თავიდან აცილებაც მხოლოდ მრავალფეროვანი კონტექსტების ანალიზის შედეგად არის შესაძლებელი).

One Key Issue of Georgian-English Machine Translation(Equivalence of Verbal Word-forms in Georgian and English)

Kakha Gabunia
Ivane Javakhishvili Tbilisi State University (Georgia)
kgabunia@cciir.ge

1. Machine translation from Georgian into another language consists of several components which must be done stage-by-stage to get an adequate result; of course, these components are interconnected and it is necessary to gradually correct data from a previous stage:

a. To process lexical items in both languages and to determine exact equivalents (with a view to homonymy, synonymy, polysemy, and other problems);
b. To analyze word-forms and to identify entries (to identify base stems for nominal and verbal inflection);
c. To create a syntactic analyzer to solve the task of complete translation of a complete idea within a sentence.

The present paper deals with word-form analysis (domain of morphology); however, it will also address the significance of such an analysis for the creation of a syntactic analyzer.

2. It is impossible to solve a problem like machine translation from Georgian into another language (in this case we are dealing with Georgian-English/English Georgian translation software) without initial analyses of a Georgian text morphologically, syntactically, and semantically.

The backbone of a Georgian sentence is the verb; the patterns of verb-form building are especially important in the process of translation.

A verbal form is a regular sequence of morphemes, in the center of which there is a stem of a verbal paradigm (viz. conjugation paradigm). On both, left and right side of a paradigm stem there are affixes marking both the derivational and inflectional categories (to the left – markers of voice, version, person and number object; to the right – suffixes: thematic suffix which is a screeve marker (Uturgaidze, Ezugbaia, 2011), an extension marker, a suffix inflecting specific tense/mood, a person marker, a number marker). The analysis of this extremely ordered and unique morphological structure (the problem of morphological ambiguity is a subject of another research) determines the "chain" of semantic categories of a certain verb, which in its turn determines a status of a screeve. For example, verbal word-form ***აშენა*** (s/he built it) is Past according to tense, Indicative – according to mood, Perfect – according to aspect. The form ***აშენებდა*** (Past, Indicative, Imperfect) differs from the last one in terms of only one feature… Frequently, a form of one screeve differs from another in terms of several features; however, the status of a screeve remains unchanged: cf. Indicative of 2nd person Aorist ***აშენე*** (context: "Last year you **built** this house" – past tense, indicative mood, perfect aspect…) to imperative ***აუშენე!*** (Context: "Build a house next year!" – Future Tense, Imperative Mood, Perfect Aspect…).

A screeve should be viewed as a compound semantic category encompassing several semantic features/components:

Tense	Mood	Aspect	Iterative	Act	Coherence

It should also be noted here that a semantic structure of a screeve is not a mere mechanical "chain", but a qualitative unity of compatible components; despite the fact that one (or several) components in this combination can be changed, combinations linked to each screeve are finite and no other screeve ever repeats the same combination…

The paper is an attempt to describe all possible combinations related to verb screeves (with respect to context) with a view to combinability of semantic constituents.

3. The most important task for Georgian-English automatic translation software is to bring Georgian and English verbal word-forms in correspondence to each other (that is screeve correspondences): in other words, to build a corresponding English verbal construction for each Georgian screeve.

It should be noted that each construction has its own semantic structure; English equivalents for each semantic construction should be established. The above presented semantic components of a screeve have a universal character and present a rather identical picture in terms of the relationship between Georgian and English "screeves". I believe that it is possible to establish the equivalence of these component combinations for both languages, which, with more or less accuracy, will yield in adequate translations of verbal forms (of course, I do not exclude the possibility of occurrence of artificial, unnatural constructions, which can be avoided only as a result of analyses of diverse contexts).

თანამედროვე ტენდენციების კვალდაკვალ: ენობრივი რესურსები და ინსტრუმენტები რესურსებით ნაკლებად უზრუნველყოფილი ენებისათვის

ირინა გურევიჩი
დარმშტადტის ტექნიკური უნივერსიტეტი (გერმანია)
gurevych@ukp.informatik.tu-darmstadt.de

გასულ წლებში მნიშვნელოვანი გამოკვლევები ჩატარდა მსხვილმასშტაბიანი ენობრივი რესურსებისა და ინსტრუმენტების შესაქმნელად და სხვადასხვა ენებში მათ ასამოქმედებლად. პირველ რიგში, შემოთავაზებული იქნა მნიშვნელობის დონეზე და სხვადასხვა ენების მიხედვით ერთმანეთთან შესაბამისობაში მოყვანილი მსხვილმასშტაბიანი ლექსიკურ-სემანტიკური რესურსები, მაგალითად UBY ან BabelNet-ი. შემდგომ ეტაპზე გამოჩნდა ტექსტის ანალიზის რთული სისტემების შემუშავების მხარდამჭერი მრავალკომპონენტიანი ჩარჩო-პროგრამები, მა-

გალითად NLTK, GATE ან DKPro. დაბოლოს, დაინერგა ინტერნეტზე დაფუძნებული საანოტაციო ინსტრუმენტები დიდი კორპუსების სელექციური და დისტრიბუციული ანოტირებისთვის ლინგვისტური ანალიზის სხვადასხვა დონეზე. ამგვარ ინსტრუმენტებში აგრეთვე გამოყენებულია მანქანური სწავლების მეთოდი. წინამდებარე მოხსენებაში მსმენელებს გავაცნობთ ზემოხსენებულ პროგრამულ რესურსებს და ზოგადად დავახასიათებთ მათ შესაძლო გამოყენებას რესურსებით ნაკლებად უზრუნველყოფილი ენებისათვის, ქართული ენის კონკრეტულ მაგალითზე.

Catching up with the trends: language resources and tools for less-resourced languages

Irina Gurevich
Technical University Darmstadt (Germany)
gurevych@ukp.informatik.tu-darmstadt.de

In the past years, significant research efforts have gone into the production of large-scale language resources and tools as well as making them interoperable across languages. First, large-scale lexical-semantic resources aligned at the sense level and across languages have been proposed, for example UBY, or BabelNet. Second, multi-component frameworks to support the development of complex text analysis systems have emerged, e.g. NLTK, GATE, or DKPro. Finally, web-based annotation tools for selective and distributed annotation of large corpora at different levels of linguistic analysis have been implemented. Such tools also make use of machine learning. This talk will introduce the above described developments and outline their possible applications to less-resourced languages, based on Georgian as one specific example.

ქართული ენის ელექტრონული ლექსიკონის შედგენის პრინციპებისათვის

ქეთევან დათუკიშვილი, მერაბ ზაკალაშვილი
ლინგვისტური ტექნოლოგიების ჯგუფი (საქართველო)
k_datukishvili@yahoo.com, GILC@Wanex.ge
ნანა ლოლაძე
თსუ არნ. ჩიქობავას სახ. ენათმეცნიერების ინსტიტუტი (საქართველო)
ლინგვისტური ტექნოლოგიების ჯგუფი (საქართველო)
nanaloladze@yahoo.com,

თანამედროვე ლექსიკოგრაფიაში სულ უფრო ფართოდ გამოიყენება კომპიუტერული ტექნოლოგიები. ელექტრონული ლექსიკონები საშუალებას აძლევს მომხმარებელს მიიღოს დიდი მოცულობის ინფორმაცია. ამასთან, ისინი გამოირჩევა ძიების მრავალმხრივი საშუალებებით. ელექტრონული ლექსიკონების ხარისხი დამოკიდებულია არა მხოლოდ წყაროების მოცულობასა და მრავალფეროვნებაზე, არამედ სრულყოფილ ლექსიკოგრაფიულ სერვისზე.

ჩვენი მიზანია შევქმნათ ქართული ენის ელექტრონული განმარტებითი ლექსიკონი, რომელიც აღჭურვილი იქნება მრავალმხრივი საძიებო სისტემით. იგი მომზადდება ქართული ენის მორფოლოგიური პროცესორისა და პროგრამა „ლექსიკოგრაფის“ ბაზაზე.

ლექსიკონის მომხმარებელს საშუალება ექნება მისთვის საინტერესო სალექსიკონო ერთეული აირჩიოს ანბანური სიიდან ან აკრიფოს შესაბამის უჯრაში. ანბანურ სიაში სიტყვები მოცემული იქნება სალექსიკონო ერთეულის სახით (სახელებისათვის – სახელობითი ბრუნვისა და ზმნებისათვის – საწყისის ფორმით). არჩეული სიტყვის შესახებ მომხმარებელი მიიღებს შესაბამის განმარტებას. ამასთან, ლექსიკონს ექნება ძიების ორიგინალური სისტემა: აკრეფისას განმარტების მოძიება შესაძლებელი იქნება არა მარტო სალექსიკონო ერთეულის, არამედ ნებისმიერი სიტყვაფორმის მიხედვით. თუ მომხმარებელი აკრეფს სიტყვას, რომელიც არ არის სალექსიკონო ერთეული, ანუ სახელი არ არის სახელობით ბრუნვაში ან ზმნა – საწყისის ფორმით (სახლმა, კაცთან, ვწერდით, გამიკეთებია და ა.შ.), მაშინ ლექსიკონი მომხმარებელს შესთავაზებს ამ სიტყვის სალექსიკონო ფორმას (სახლი, კაცი, წერა, გაკეთება და ა.შ.) შესაბამისი განმარტებით. ერთ სალექსიკონო ერთეულს შეიძლება დაუკავშირდეს ასობით და ათასობით სიტყვაფორმა. მაგალითად, სიტყვას „**დაწერა**“ უკავშირდება შემდეგი ფორმები: დავწერ, დაწერ, დაწერს, დავწერთ, დავწერდი, დავწერე, დამიწერია, დაიწერა, დაწერილა და ა.შ. ყველა ეს სიტყვაფორმა წარმოადგენს ერთ ლექსიკურ ერთეულს: **დაწერა.** ამგვარი ბმით ჩვენს ლექსიკონში შესაძლებელი იქნება განმარტების მიღება მილიონობით სიტყვაფორმის შესახებ.

პროგრამა „ლექსიკოგრაფში“ სალექსიკონო სტატიები წარმოდგენილია სტრუქტურირებული სახით. აქ ცალკე ველებადაა განთავსებული სიტყვის ფუნქციონირების სფერო, დარგი, მეტყველების ნაწილი და ა.შ., გარდა ამისა, მითითებულია სიტყვათა სემანტიკური ველი: მცენა-

რეები, ცხოველები, ავეჯი, ნივთიერებები და მისთ. ეს საშუალებას მისცემს მომხმარებელს ინ-ფორმაცია მოიძიოს აღნიშნული მახასიათებლების მიხედვითაც.

On the Principles of Compilation of the Electronic Dictionary of Georgian

Ketevan Datukishvili, Merab Zakalashvili
Linguistic Technologies Group (Georgia)
k_datukishvili@yahoo.com, GILCE@Wanex.ge
Nana Loladze
Arn. Chikobava Institute of Linguistics, TSU (Georgia)
Linguistic Technologies Group (Georgia)
nanaloladze@yahoo.com

Computer technologies are widely applied in the modern lexicography. Electronic dictionaries enable users to get vast amount of the appropriate information. In addition, they are remarkable by their diversified means of searching. The quality of the electronic dictionaries depends not only on the volume and diversity of the sources used, but on the perfect lexicographical service as well.

We are aiming at compiling of the electronical explanatory dictionary of the Georgian language supplemented with the diversified search system. This dictionary will be compiled on the basis of the Georgian language morphological processor and the programm "Lexicographer".

The user of the Dictionary will be able to select a needed dictionary entry out of the alphabetic list or type it in the corresponding box. Words in the alphabetic list will be given in the form of the dictionary entries (the nominative case for nouns and the infinitive form for verbs). The user will receive the corresponding explanation of the selected word. Besides, the dictionary will be appended with the original search system: typing into it makes it possible to find a definition not only according to a dictionary entry but also according to any possible word form. If a user types a word which does not represent a dictionary entry (i.e. a noun not in the nominative case and a verb not in the infinitive form), (e.g. saxl**ma**, kac**tan**, **vcerdit, gamiketebia**, etc.), then the electronic dictionary offers the dictionary forms of the words (**saxli, kaci, cera, gaketeba**, etc.), provided with corresponding explanations. Hundreds of thousands of the word forms are expected to be associated with one dictionary entry. For example, the word "dacera" (write) is associated with the forms *davcer, dacer, dacers, davcert, damiceria, davcerdi*, etc. All word forms represent the forms of a single lexical unit – **dacera.** Such linkage makes it possible to find even millions of forms and their definitions in our dictionary.

Dictionary entries are represented in the structurized form in the program "Lexicographer". In it the functional area of the word, its usage domain, part of speech, etc. appear in the individual fields. Besides, the semantic fields are also denoted: plants, animals, furniture, substances and others. This will help the user to search for the information in according with those characteristics as well.

ქართული ენის მორფოლოგიური ანოტირებისას გამოვლენილი შეცდომების ანალიზი

სოფიკო დარასელია
ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)
sopod@yahoo.com
სერგეი შაროვი
ლიდსის უნივერსიტეტი (დიდი ბრიტანეთი)
s.sharoff@leeds.ac.uk

წინამდებარე ნაშრომში განხილულია ქართულის ენის ვებკორპუსის – KaWaC-ის მორფოლოგიური ანოტირება და ანოტირებისას გამოვლენილი შეცდომების ანალიზი. KaWaC წარმოადგენს თანამედროვე ქართული ენის ვებკორპუსს, რომელიც შეიქმნა ლიდსის უნივერსიტეტში და განთავსებულია აღნიშნული უნივერსიტეტის ვებგვერდზე. ვებკორპუსი ასახავს თანამედროვე ქართულ ენას, კერძოდ, ენის განვითარების ბოლო 10-15 წელიწადს და მოიცავს სხვადასხვა ჟანრის, რეგიონისა და ა.შ. ტექსტებს (Daraselia, Sharoff, 2014) ინტერნეტსივრციდან.

ქართული ენის ვებკორპუსის ტექსტური მასალა ჩამოიტვირთა 697 ვებსაიტის 618468 ვებგვერდიდან და მოიცავს 150 მილიონზე მეტ სიტყვას, კერძოდ, კორპუსში არის:

- 182845341სიტყვა და 2865042 სიტყვის უნიკალური ხმარების შემთხვევა;
- 230597937ტოკენი, რაც მოიცავს პუნქტუაციის ნიშნებს სიტყვებთან ერთად.

მორფოლოგიური ანოტირება დაიწყო ქართული ენისათვის მორფოლოგიური მახასიათებლების სისტემის შექმნით, რომელიც რამდენიმე ეტაპად განხორციელდა. ტაგსეტები შემუშავდა სპეციალურად შექმნილი უნივერსალური სქემის საფუძველზე (მულტიტექსტური მორფოსინტაქსური პარამეტრების საფუძველზე MULTEXT-East Morphosyntactic Specification). მორფოლოგიური მახასიათებლების სისტემის შემუშავების პროცესში მაქსიმალურად გათვალისწინებულია მულტიტექსტური მორფოსინტაქსური პარამეტრების მახასიათებლები და ახალი მახასიათებლები წარმოდგენილია ქართული ენის მორფოლოგიური თავისებურებების ასახვის მიზნით (Daraselia, Sharoff, 2014).

მულტიტექსტური მორფოსინტაქსური პარამეტრების ბაზაზე შექმნილი ქართული ენის მორფოლოგიური მახასიათებლების სისტემა მოიცავს 15 ძირითად კატეგორიას, ესენია: არსებითი სახელი, ზედსართავი სახელი, ნაცვალსახელი, რიცხვითი სახელი, ზმნა, ზმნიზედა, თანდებული, კავშირი, ნაწილაკი, შორისდებული, მასდარი, მიმღეობა, შედგენილი შემასმენელი, აბრევიატურა და უცნობი. ზემოთ წარმოდგენილ თითოეულ კატეგორიას მინიჭებული აქვს შესაბამისი წყვილები, სულ 331 წყვილი.

მულტიტექსტური მორფოსინტაქსური პარამეტრები განსაზღვრავს ძირითად მორფოლოგიურ კატეგორიებს და მათთან დაკავშირებულ ქვეკატეგორიის წყვილებს, რასაც „მორფოსინტაქსური აღწერა“ – (Morphosyntactic Description (MSD)) ეწოდება. მაგალითად:

არსებითი სახელი, ტიპი = საზოგადო, რიცხვი = მხოლობითი, ბრუნვა = სახელობითი

ზემოთ წარმოდგენილი ქვეკატეგორიების წყვილების სტრუქტურას შეესაბამება „მორფოსინტაქსური აღწერის“ (ე.ი. MSD) მახასიათებელი: *Ncsn (Noun, Type= Common, Number = singular, Case = Nominative),* რაც გამოყენებულია მორფოლოგიური ანოტირებისას (Sharoff, etal, 2008).

მორფოლოგიური ანოტირების პირველ საფეხურზე ქართული ენის ვებკორპუსიდან ამოღებული იქნა სიხშირული სია – ყველაზე ხშირად ხმარებული 5 000 ქართული სიტყვა და შესრულდა ამ მასალის ხელით ანოტირება. მეორე ეტაპზე კი დაახლოებით ერთი მილიონი ანოტირებული წინადადებების ბაზაზე მომზადდა საწვრთნელი მასალა კორპუსის ავტომატური მორფოლოგიური ანოტირებისათვის და მიღებული საწვრთნელი მასალით კი განხორციელდა მთლიანი ვებკორპუსის მორფოლოგიური ანოტირება.

ამგვარად, ჩატარებული კვლევის საფუძველზე მოვახდინეთ მთლიანი ვებკორპუსის ავტომატური მორფოლოგიური ანოტირება და შედეგად კი მივიღეთ ანოტირებული ვებკორპუსის პირველი ვერსია.

მორფოლოგიური ანოტირების შეცდომების გამოსავლენად გამოვიკვლიეთ ავტომატურად ანოტირებული მასალა (დაახლოებით 40 000 ავტომატურად ანოტირებული წინადადება) და შეცდომების ანალიზი წარმოდგენილია ქვემოთ. ქართული ენის ვებკორპუსის მორფოლოგიური ანოტირებისას გამოვლენილი შეცდომების ტიპებია:

ა) **ლექსიკური უზუსტობა:** შემთხვევა, როდესაც ესა თუ ის სიტყვა მოცემულია საწვრთნელ მასალაში, მაგრამ არა იმ მახასიათებლით, რას მას უნდა ჰქონდეს მოცემულ კონტექსტში. ქვემოთ წარმოდგენილ მაგალითში ***შეიძლება*** არის ზმნა, საწვრთნელ მასალაში კი იგი წარმოდგენილია მხოლოდ როგორც ნაწილაკი და ავტომატური ანოტირებისას მოცემული სიტყვას აქვს არასწორი მახასიათებელი: “Q” = ნაწილაკი:

დარჩა [Vi3sa] / რამე [Psn] /, რისი [Psn] /ჭამაც [Wsn] /შეიძლება [Vi3sp]?

ამ სიტყვის სწორი მახასიათებელი მოცემულ კონტექსტში არის *Vi3sp = Verb, 3rd person, Singular, PresentTense (ზმნა, მესამე პირი, მხოლობითი რიცხვი, აწმყო დრო).*

ბ) **უცნობი სიტყვა:** ეს არის სიტყვა, რომელიც არ არის მოცემულ საწვრთნელ მასალაში და მისი ტაგირება ხდება ავტომატურად კონტექსტიდან გამომდინარე. ზოგიერთ შემთხვევაში ტაგერი ერთმანეთისაგან ვერ არჩევს საზოგადო და საკუთარ არსებით სახელებს, განსაკუთრებით მაშინ, როდესაც საუბარია ისტორიულ ადგილებზე. მაგალითად:

აწყურის ღვთისმშობლისა და ზარზმის ღვთისმშობლის...

მოცემულ მაგალითში ***აწყურის*** და ***ზარზმის*** მონიშნულია როგორც საზოგადო არსებითი სახელები (*Ncsg*), მაშინ როდესაც მოცემული მაგალითისათვის სწორი მახასიათებელია *Npsg (Noun/proper/singular/genitivecase*).

გ) **ბუნდოვანი სიტყვა**: მორფოლოგიური ანოტირებისას ერთ-ერთ დიდ სირთულეს წარმოადგენს მასდარი და მიმღეობა, ვინაიდან კონტექსტი არასაკმარისია მასდარისა და მიმღეობის სწორად მონიშვნისათვის. მაგალითად:

ჩვენ არ გვეგების ამის ცოდნა.

მოცემულ კონტექსტში **ცოდნა** არის მასდარი (მახასიათებელი: *Wsn*), მაშინ, როდესაც იგი ზემოთ წარმოდგენილ მაგალითში შეცდომით მონიშნულია როგორც არსებითი სახელი (*Ncsn*).

გარდა მასდარისა და არსებითი სახელისა, ბუნდოვანი შემთხვევები ასევე გამოვლენილია ზმნასა და მასდარს შორის, მაგალითად:

და [C] შექმნის [Vi3sf] მათგან [Ppg] ისეთ [Psd] ძალას [Ncsd]

ამ მაგალითში **შექმნის** არის ზმნა (*Vi3sf*), მაგრამ მონიშნულია შეცდომით როგორც მასდარი ნათესაობით ბრუნვაში (*Wsg= Masdar / singular/gentivecase*). მოცემული სიტყვის სწორი მახასიათებელი კი უნდა იყოს *Vi3sf" (Verb / 3rd person, singular/futuretense*): ზმნა, მესამე პირი, მხოლობითი რიცხვი, მყოფადი დრო.

მეორე მაგალითში კი, პირიქით, მასდარი მონიშნულია როგორც ზმნა:

ოჯახის [Ncsg] / შექმნის [Wsg]/ გადაწყვეტილების [Ncsg] / მიღება [Wsn]

ამ კონტექსტში **შექმნის** არის მასდარი, რომელსაც შეესაბამება მახასიათებელი: *Wsg(= Masdar, number = Singular, Case = Genitive*): მასდარი, მხოლობითი რიცხვი, ნათესაობითი ბრუნვა.

ამგვარად, ნაშრომში განვიხილეთ ქართული ენის ვებკორპუსის პირველი ავტომატური მორფოლოგიური ანოტირება და ანოტირების პროცესში გამოვლენილი შეცდომები. უნდა აღინიშნოს, რომ მოცემული კვლევით არ დასრულებულა მორფოლოგიური ანოტირების სამუშაო. ანოტირების შემდგომ ეტაპზე დაგეგმილია არსებული შეცდომების გასწორება და მორფოლოგიური ანოტირების ინსტრუმენტის დახვეწა და უფრო დეტალური მორფოლოგიური ანოტირების მოდელის წარმოდგენა.

Error Analyses in Part-of-Speech Tagging in Georgian

Sophiko Daraselia
Ivane Javakhishvili Tbilisi State University (Georgia)
sopod@yahoo.com
Serge Sharoff
The University of Leeds (United Kingdom)
s.sharoff@leeds.ac.uk

The paper discusses the Georgian web corpus KaWaC and part-of-speech tagging of the corpus. KaWac is a large web corpus of contemporary Georgian built at the University of Leeds. The corpus represents contemporary Georgian language within the period of the last 10-15 years and contains wide range of text types, topics and regions (Daraselia, Sharoff, 2014) from the Internet.

The corpus texts were crawled from 618468 web pages from 697 websites. It contains over 150 million words, in particular:

- Words: 182845341, 2865042 Type
- Lemmas: 182845341, 1427952 Types
- Tokens: 23550807, 3447266 Type

In this paper, we discuss the corpus composition and the process of creation of a computational model of Georgian morphology for the purposes of corpus annotation, creating the Georgian tagset using the MULTEXT-East Morphosyntactic Specifications.

The Part of Speech Tagging started with the designing of the Georgian tagset carried out in several steps. The tagset is designed according to MULTEXT-East Morphosyntactic Specifications. The MTE specifications of several corpora were directly taken from the MULTEXT-East resources, We created new MSDs for specific Georgian cases, such as Masdar, Participle, etc.

The tagset contains 15 main categories: noun, verb, adjective, pronoun, adverb, adposition, conjunction, numeral, particle, interjection, masdar, participle, compound verb, abbreviation and residual. Each category has attribute value pairs, in total resulting 331 attribute value pairs.

The MULTEXT-East Morphosyntactic Specifications defines main morphosyntactic categories and their attribute value pairs and describes morphosyntactic properties of words that is called Morphosyntactic Descriptions (MSDs). For instance:

Noun, Type= Common, Number = singular, Case = Nominative

The specifications of the feature structure above correspond to a single MSD tag *Ncsn,* which is used in automatic morphological analysis and disambiguation (Sharoff, et al, 2008).

The part-of-speech tagging started with manual annotation. At the first stage, frequent list from the corpus (about 5000 most frequent Georgian words) were extracted and tagged manually. At the second stage, we have extracted about one million sentence-to-sentence tagged data and corrected the data (about one hundred thousand words) manually. At the third stage, manually tagged and corrected training data were applied to the whole corpus resulting in the first version of the annotated corpus for Georgian. At the final stage, we have extracted automatically tagged sentence-to-sentence tagged data, examined, and analysed the errors of the part-of-speech tagging.

What is the accuracy of the part-of speech tagging for Georgian and to what extent taggers get wrong? We have looked at the automatically annotated data (about forty thousand sentence-to-sentence tagged words), examined, and evaluated the errors in part-of-speech tagging, such as:

a) **Lexicon gap:** Here, the word appears in the training data, but never with the tag, which it has in this context. For example, below, *შეიძლება* is a verb, but in the training data, it occurs as a particle.
   დარჩა [Vi3sa]/ რამე [Psn]/, რისი [Psn]/ ჭამაც [Wsn]/ შეიძლება [Vi3sp]?
   შეიძლება is with a wrong tag "Q" = particle

   The correct tag for the word in this context is: *Vi3sp = Verb, 3rd person, Singular, Present Tense*

b) **Ambiguous word**: a word not occurring in the training data and the tagger has to rely on the context, here the tagger in some cases cannot differentiate proper and common nouns, in particular, old Georgian /historical proper names, For example, *აწყურის ღვთისმშობლისა და ზარზმის ღვთისმშობლის*, in this example *აწყურის* and *ზარზმის* are tagged as common nouns whilst the correct tag for the words *აწყურის* and *ზარზმის* is *Npsg: Noun/proper/singular/genitive case.*

c) **Difficult linguistics: Needs much syntax**: sometimes it is necessary to have a broad contextual knowledge to determine the right tag. For example, below, a tagger cannot correctly choose between the Present (Vi3sp) and Future (Vi3sf) tag for *შეაკავშირებს* and chooses the tag wrongly, the right tags for this example are:
   რომელიც [C] / შეაკავშირებს [Vi3sf] / ჩვენს [Ppd] / მეურნეთ [Ncpd]/ ამა [Pse] / თუ [C] / იმ [Pse] / სახით [Ncsi]

d) **Underspecified /unclear:** Masdars and Participles are a bit problematic cases in part-of-speech tagging in Georgian. The given context not sufficient to distinguish these categories. For example, below, ცოდნა is a Masdar:

ჩვენ არ გვეგების ამის ცოდნა

Here, ცოდნა is tagged as a Noun (Ncsn), but in this context it is a Masdar with MSD tag: *Wsn*.

Unclear cases can occur with verbs and masdars, as well. For example, in the first example, შექმნის is clearly a verb, but it also can be masdar in genitive case as shown in the second example:
Example 1:

და [C] შექმნის [Vi3sf] მათგან [Ppg] ისეთ [Psd] ძალას [Ncsd]

Here, it is tagged wrongly with "*Wsg" = Masdar / singular/gentive case*, the right tag in this context is *"Vi3sf" = Verb / 3rd person, singular/future tense*.

Example 2:

ოჯახის [Ncsg]/ შექმნის [Wsg] / გადაწყვეტილების [Ncsg] / მიღება [Wsn]

In this example, შექმნის is Masdar in Genitive case with MSD tag: Wsg = Masdar, number = Singular, Case = Genitive.

In both examples, შექმნის has the same spelling and phonetics, whether it is a verb (*Vi3sf= Verb / 3rd person, singular/future tense*) or a masdar (*Wsg = Masdar, number = Singular, Case = Genitive*) depends on the context.

Thus, the paper discusses a large web-corpus of Georgian, focusing on part-of-speech tagging and error analyses in POS tagging. We think that the research undertaken at the University of Leeds resulting in a Georgian web-corpus, part-of-speech tagging and producing first version of annotated KaWaC corpus, can be used for morphological annotation and developing corpus resources not only for Georgian, but for other Kartvelian languages as well.

ქართული ენის კორპუსის კონცეფცია

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

საპრეზენტაციოდ წარმოდგენილია 2009-2014 წლებში ილიას სახელმწიფო უნივერსიტეტის ლინგვისტურ კვლევათა ინსტიტუტის მიერ მომზადებული ქართული ენის კორპუსი (GLC), რომელიც ამ ეტაპზე 100 000 000-ზე მეტ სიტყვაფორმას მოიცავს (corpora.iliauni.edu.ge). იგი შედგება მონოლინგვური და ბილინგვური კორპუსებისგან. მონოლინგვური კორპუსში წარმოდგენილია ა. ძველი და საშუალი ქართული ენის კორპუსი, ბ. ახალი და თანამედროვე ქართული ენის კორპუსი.

ძველი ქართული ენის კორპუსში ცალკეა გამოყოფილი მთარგმნელობითი კორპუსი, რომელიც, თავის მხრივ, დაყოფილია როგორც ქრონოლოგიურ-სტილური, ანუ მთარგმნელობითი სკოლების (წინაათონური, ათონური, ანტიოქიური და ა. შ.), ასევე სათარგმნი დედნების ენის მიხედვით (შდრ. ბერძნული – ქართული, სირიული – ქართული, ქრისტიანული არაბული – ქართული, სომხური – ქართული). ბილინგვურ ნაწილში წარმოდგენილია "ქართლის ცხოვრებისა" (ძველი ქართული – ძველი სომხური) და "ვეფხისტყაოსნის" პარალელური (ქართულ-ინგლისური) კორპუსები. სრულდება მუშაობა "შუშანიკის წამების" ბილინგვურ (ძველი ქართული – ძველი სომხური) კორპუსზე.

სადისკუსიოდ წარმოვადგენთ კონცეფციას, რომელიც ქართული ენის კორპუსს უდევს საფუძვლად და რომელიც არა მარტო ისტორიული გრამატიკის რაკურსით წარმოაჩენს ამ ენაზე შემონახულ უმდიდრეს მემკვიდრეობას, არამედ მისი სოციალური და საზოგადოებრივი ფუნქციების განვითარების, სხვა ენებთან და კულტურებთან ურთიერთობის თვალსაზრისითაც.

მოხსენებაში ასევე შევეხებით საქართველოს ეპიგრაფიკის კორპუსს და ქართული ენის ტექსტური მემკვიდრეობის დოკუმენტირებისა და ციფრული ჰუმანიტარიის სხვა პროექტებსაც, რომლებიც ლინგვისტურ კვლევათა ინსტიტუტში ხორციელდება.

Georgian Language Corpus: Concept and Methodology

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

I aim to present the Georgian Language Corpus (GLC) developed at the Institute of Linguistic Studies of Ilia State University during 2009-2014 (corpora.iliauni.edu.ge). At present the corpus contains over 100 000 000 word forms and has two main sections, monolingual and bilingual. The monolingual section consists of a) Old and Middle Georgian Corpus, and b) New and Modern Georgian Corpus. The Old Georgian Corpus on its part contains a translation corpus structured according to translation schools, i.e. the chronological and stylistic principle (pre-Athonite, Athonite, Antiochian, etc.) and source texts (cf. Greek – Georgian, Syriac – Georgian Christian Arabic – Georgian, Armenian – Georgian). The bilingual section includes parallel corpora of *Kartlis Tskhovreba* (*The Georgian Chronicle*, a Georgian-Armenian corpus) and *Vepkhistkaosani* (*The Knight in the Panther's Skin,* a Georgian-English corpus). The work on the Old Georgian-Armenian bilingual corpus of *The Martyrdom of Holy Queen Shushanik* will soon be completed.

I offer for discussion the conceptual framework underlying the GLC. The rich language legacy is presented not only from the perspective of historical grammar but also of social and public functions and relations with other languages and cultures.

I will also dwell on the Georgian epigraphic corpus and other projects of digital humanities and Georgian text documentation implemented at the Institute of Linguistic Studies.

პროექტი „ქართული ენა საზღვარგარეთ – ქართული დიალექტები და ლაზური თურქეთში, ირანსა და აზერბაიჯანში“

ლალი ეზუგბაია
თბილისის თავისუფალი უნივერსიტეტი (საქართველო)
l.ezugbaia@freeuni.edu.ge
ლია ბაკურაძე, ნარგიზა სურმავა, მაია ბარიხაშვილი
თსუ არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველო)
l.bakuradze@gmail.com; nargizasurmava@yahoo.com; maiahereti@yahoo.com

პროექტის მიზანი იყო საზღვაგარეთ (თურქეთის, ირანისა და აზერბაიჯანის ტერიტორიაზე) მცხოვრები ეთნიკური ქართველების მეტყველების ნიმუშების მოპოვება, თანამედროვე ტექნოლოგიებით დოკუმენტირება და კვლევა, მეორე მხრივ, ისტორიულ სამშობლოსთან ენობრივი და კულტურული რეინტეგრაციისათვის ხელის შეწყობა.

პროექტის ფარგლებში შესრულდა შემდეგი სამუშაო:

- შეიქმნა ქართული დიალექტების (ფერეიდნულის, ინგილოურის, ჩვენებურების ქართულის) და ლაზურის ნიმუშების ვიდეო კოლექცია;
- გაიშიფრა და გამოსაცემად მომზადდა აღნიშნული კოლექციის დიდი ნაწილი;
- ქართული დიალექტებისა და ლაზური კილოკავების ტექსტური კოლექციისა და ბეჭდური მასალის საფუძველზე მომზადდა ელექტრონული ლექსიკონები.

დიალექტური ტექსტებისა და ელექტრონული ლექსიკონების დამუშავება მოხდა ქართული დიალექტური კორპუსის ლექსიკოგრაფიული კონცეფციის მიხედვით (მ. ბერიძე). ფერეიდნულის, ინგილოურის, ჩვენებურების ქართულისა და ლაზურის ელექტრონული ლექსიკონები პირველი ნაბიჯია ამ კონცეფციის განხორციელების გზაზე.

პროექტის ფარგლებში მომზადებული ტექსტები და ლექსიკონები ინტეგრირებულია ქართულ დიალექტურ კორპუსში (ქდკ).

ტექსტებისა და ლექსიკონების მომზადებაზე მუშაობდნენ: ფერეიდნულზე – მარინა ბერიძე, ლია ბაკურაძე, ინგილოურზე – მაია ბარიხაშვილი, ელენე ნაპირელი. ჩვენებურების ქართულზე – ნარგიზა სურმავა, ლაზურზე – ლალი ეზუგბაია, მანანა ბუკია, მარინა ჯღარკავა, სოფიო ბერულავა. პროგრამული უზრუნველყოფა ეკუთვნის დავით ნადარაიას.

ელექტრონული ლექსიკონებით სარგებლობა შესაძლებელია როგორც ქართული დიალექტების კორპუსის, ასევე თავისუფალი უნივერსიტეტის ვებგვერდების საშუალებით შემდეგ მისამართებზე: http://corpora.co/#/dictionaries, http://freeuni.edu.ge/node/1004

პროექტი დაფინანსებული იყო 2011-2014 წლებში შოთა რუსთაველის ეროვნული სამეცნიერო ფონდის მიერ და განხორციელდა თბილისის თავისუფალი უნივერსიტეტის ბაზაზე არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტის მეცნიერი თანამშრომლების მონაწილეობით.

მოხსენებაში დეტალურად იქნება წარმოდგენილი პროექტის შედეგები და განვითარების პერსპექტივა.

Project: The Georgian Language Abroad – Georgian Dialects and Laz in Turkey, Iran and Azerbaijan

Lali Ezugbaia
Tbilisi Free University (Georgia)
l.ezugbaia@freeuni.edu.ge
Lia Bakuradze, Nargiza Surmava, Maya Barikhashvili
Arn. Chikobava Institute of Linguistics at TSU (Georgia)
l.bakuradze@gmail.com; nargizasurmava@yahoo.com; maiahereti@yahoo.com

On the one hand, the goal of the project was to collect speech samples of ethnic Georgians living abroad (Turkey, Iran, and Azerbaijan), to document and study them by means of modern technologies and, on the other hand, to promote the linguistic and cultural reintegration to the historical homeland.

The following work was performed within the project:

1. The video collection of Georgian dialects (Fereidanian, Ingiloan, Georgian of Chveneburebi) and Laz samples have been produced;
2. A large part of the collection has been deciphered and prepared for publication;
3. The electronic dictionaries, based on the text collection and the printed material of Georgian dialects and the local dialects of Laz, were compiled.

Dialectal texts and electronic dictionaries have been processed according to the lexicographical concept of the Georgian Dialect Corpus (M. Beridze). Fereidanian, Ingiloan, Georgian of Chveneburebi and Laz electronic dictionaries are the first step in the implementation of the concept.

The texts and the dictionaries, prepared within the project, are integrated into the Georgian Dialect Corpus (GDC).

The texts and the dictionaries were prepared by Marina Beridze and Lia Bakuradze (Fereidanian), Maya Barikhashvili and Elene Napireli (Ingiloan), Nargiza Surmava (Georgian of Chveneburebi), Lali Ezugbaia, Manana Bukia, Marina Jgharkava and Sophie Belurava (Laz). The software was developed by David Nadaraia.

The electronic dictionaries are available on the following websites of the Georgian Dialect Corpus and Tbilisi Free University: http://corpora.co/#/dictionaries, http://freeuni.edu.ge/node/1004.

The project was financed by Shota Rustaveli National Science Foundation in 2011-2014 and was implemented on the bases of Tbilisi Free University by participating the researchers of Arn. Chikobava Institute of Linguistics.

The results and the development prospects of the project will be presented in detail in the talk.

ელექტრონული კურსებისა და ტექსტური ბაზების გამოყენება სწავლების პროცესში (თსუ ჰუმანიტარულ მეცნიერებათა ფაკულტეტის გამოცდილება)

დარეჯან თვალთვაძე, მაია მადუაშვილი, ეკა კვირკველია
ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)
darejan.tvaltvadze@tsu.ge, eka.kvirkvelia@tsu.ge

დღეს, როდესაც ციფრული ჰუმანიტარია მძლავრად იკიდებს ფეხს საქართველოში და აქტიურად მიმდინარეობს მუშაობა ქართული ენის ეროვნული კორპუსის შესაქმნელად, ელექტრონული სწავლების დანერგვასთან ერთად მნიშვნელოვანად მიგვაჩნია სწავლების პროცესში არსებული ელექტრონული ბაზების მაქსიმალურად გამოყენება, რათა სტუდენტებმა შეძლონ კორპუსში მუშაობისათვის აუცილებელი უნარ-ჩვევების გამომუშავება და შემდგომში საკუთარი ინტერესისამებრ გამოყენება.

უნივერსიტეტის სასწავლო პროცესში ბოლო დროს აქტიურად დაინერგა ელექტრონული სწავლების პორტალი (MOODLE), სადაც განთავსებულია ელექტრონული კურსები სხვადასხვა საგნისათვის. მათ შორის არის „ქართული პალეოგრაფიისა“ და „ქართული სალიტერატურო ენის ისტორიის“ კურსები. ელექტრონული კურსის გამოყენებით სტუდენტებს შეუძლიათ სალექციო თემებისა და პრეზენტაციების, სასწავლო მასალისა და დავალებების ნახვა, კურსთან დაკავშირებული სიახლეებისათვის თვალის მიდევნება. ამგვარი მიდგომა აიოლებს მასალაზე წვდომას, ეფექტურს ხდის როგორც სწავლების, ისე სწავლის პროცესს და, ამასთანავე, არის დროის ეკონომიის მოქნილი საშუალება.

„ქართული პალეოგრაფიის“ ელექტრონულ კურსზე შესაძლებლობის ფარგლებში განთავსებულია საინტერესო ვიზუალური მასალა: საქართველოსა და მის ფარგლებს გარეთ დაცული ხელნაწერებისა და ცნობილი ქართველი მწიგნობრების ავტოგრაფების ფოტოები, ბმულები ქართული სამწიგნობრო კერებისა და უცხოური ლიტერატურისა ბერძნული პალეოგრაფიის შესახებ. ატვირთულია დოკუმენტური ფილმები და ა. შ.

„ქართული სალიტერატურო ენის ისტორიის“ კურსის ფარგლებში სილაბუსით გათვალისწინებული მასალის ათვისების შემდეგ სტუდენტებს მოეთხოვებათ რეფერატისა და პრეზენტაციის მომზადება. რეფერატისა და პრეზენტაციის მომზადებისას აუცილებელი პირობაა ემპირიული მასალის მოსაძიებლად ელექტრონული ბაზებისა და კორპუსების გამოყენება, რაც იძლევა საკითხთა როგორც სინქრონიული, ისე დიაქრონიული კუთხით კვლევის საშუალებას. ძირითადი ელექტრონული რესურსია ქართული ენის ეროვნული კორპუსი (GNC) და ის ქვეკორპუსები, რომლებიც საფუძვლად დაედო მის შექმნას: TITUS-ის ელექტრონული ტექსტების ბაზა, ARMAZI-ს ელექტრონული ტექსტების ბაზა, GEKKO – ქართული კორპუსი და ქართული დიალექტური კორპუსი.

აღნიშნულ ტექსტურ ბაზებსა და კორპუსებზე მუშაობა სტუდენტებს უფართოვებს თვალსაწიერს, აძლევს საშუალებას აღიქვან ენობრივი პროცესები ისტორიულ ჭრილში და ამავე

დროს დააკვირდნენ ენის განვითარების თანამედროვე ვითარებას, თავად შეამოწმონ ესა თუ ის თეორიული დებულება და მასალის საშუალებით მოახდინონ მისი ვერიფიკაცია. მოქნილი საძიებო სისტემები ამარტივებს კვლევას, იძლევა სიტყვაფორმათა ანალიზის, სტატისტიკის გაკეთებისა და სანდო დასკვნების გამოტანის საშუალებას. შესაბამისად, სამუშაო პროცესიც უფრო მარტივი და სახალისო ხდება სტუდენტებისათვის.

უნდა აღინიშნოს, რომ თსუ ჰუმანიტარულ მეცნიერებათა ფაკულტეტი არა მხოლოდ მომხმარებელია თანამედროვე კორპუსებისა, არამედ აქტიურად არის ჩართული ქართული ენის ეროვნული კორპუსის შექმნაში. საერთაშორისო პროექტში **„ქართული ენის ეროვნული კორპუსი – ტექნოლოგიური ჩარჩოს შექმნა“**, რომელიც Volkswagen Stiftung-ის მიერ ფინანსდება, მონაწილეობს სტუდენტთა ორი ჯგუფი. ერთი, რომელშიც 16 სტუდენტია ჩართული, ახორციელებს ძველი ქართული ლიტერატურის ძეგლების დიგიტალიზირებას, ხოლო მეორე ჯგუფი – უკვე არსებული ტექსტების ანოტირებას.

ამგვარად, მსოფლიოში წარმოუდგენელი სისწრაფით მიმდინარე ტექნიკური პროგრესის პირობებში ძალზე მნიშვნელოვანია არსებული ელექტრონული რესურსების უფრო აქტიურად გამოყენება და დანერგვა სწავლების პროცესში. ამგვარი მიდგომა საშუალებას აძლევს სტუდენტებს წარმართონ თავიანთი სამეცნიერო კვლევები თანამედროვე მეთოდებისა და ტექნოლოგიების გამოყენებით.

Application of Electronic Courses and Textual Data in the Process of Teaching (Know-how of the Faculty of Humanities, Tbilisi State University)

Darejan Tvaltvadze, Maya Maduashvili, Eka Kvirkvelia
Ivane Javakhishvili Tbilisi State University (Georgia)
darejan.tvaltvadze@tsu.ge, eka.kvirkvelia@tsu.ge

In our days, when digital humanities is gaining strength in Georgia and activities are under way to develop the national corpus of the Georgian Language, application of digital databases in the process of teaching is considered to be very significant together with establishing of the practice of Elearning, in order to enable students to acquire and develop necessary skills for working with the corpus and to apply them afterwards in accordance with their own interests.

Recently, the academic process at University Curriculum saw the implementation of the electronic teaching portal MOODLE, which offers various digital courses. Among them there are the courses: “Georgian Paleography” and “The History of the Georgian Literary Language”. Students can browse lecture topics, presentations, teaching materials and other homework as well as follow the course

updates by means of using the electronic resources. This approach allows an easy access to the materials and makes both teaching and learning effective. It also provides a flexible tool for saving time.

The electronic course "Georgian Paleography" offers as much access as possible to rather interesting visual materials: photos picturing the manuscripts preserved both within and outside Georgia and famous Georgian scholars; it makes available links to Georgian literary schools, foreign literary sources, Greek paleography, documentaries and other materials.

According to the guidelines, students are required to prepare an essay and a presentation after digesting the materials offered within the framework of the syllabus for "the History of the Georgian Literary Language". The key point in preparing an essay and a presentation is the application of electronic databases and corpora in their empirical research that allows the synchronic as well as diachronic approach in their study. The main electronic resource is the Georgian National Corpus (GNC) and the sub-corpora, having laid a foundation for its development. The electronic text database TITUS, another database for electronic texts – ARMAZI, the Georgian Corpus (GEKKO) and the Georgian dialect Corpus (GDC).

Work on the above mentioned textual databases and corpora broadens the students' outlook and enables them to perceive the linguistic processes within a historical context. It also allows them to observe the present-day state of its development, to verify theoretical tenets based on the materials provided. Flexible search engines simplify investigation, allow analyses of word-forms, statistic survey, leading to credible conclusions. Therefore, the working process becomes easy and engaging for students.

It should be noted that Faculty of Humanities, TSU, is not only a user of present-day corpora but it is actively involved in the development of National Corpus of the Georgian Language. The International project "Georgian National Speech Corpus – Creation of Technological Framework," funded by Volkswagen Stiftung, involves two groups of students. One group includes 16 members and implements a digitalization of the ancient monuments of Georgian literary sources. The other group carries out the annotation of the already-existing texts.

Thus, with respect to the incredible speed of technical progress all over the world, it is of immensely important to more actively apply available electronic resources and to implement them in the teaching/learning process. This approach allows students to conduct their research applying up-to-date methods and technologies.

საქართველოს ეპიგრაფიკული ძეგლების კორპუსი (კორპუსის შედგენილობა და ელექტრონული გამოცემის სტანდარტი)

თამარ კალხიტაშვილი
ილიას სახელმწიფო უნივერსიტეტი (საქართველო)
kalkhitashvili.tamar@gmail.com

მოხსენებაში წარმოდგენილია *საქართველოს ეპიგრაფიკული ძეგლების კორპუსის* პირველი ქვეპროექტი – ***III-X საუკუნეების აღმოსავლეთ საქართველოს ქართული, არამეული, ბერძნული და სომხური წარწერების კორპუსი***, რომელიც EpiDoc-ის სტანდარტის მიხედვით ხორციელდება ილიას სახელმწიფო უნივერსიტეტის ლინგვისტურ კვლევათა ინსტიტუტში. ახალი ელექტრონული გამოცემის შესასრულებლად წინასწარ შეირჩა, დამუშავდა და ქართულად ითარგმნა EpiDoc-ის აღწერილობა, არსებული კრიტიკული და დიპლომატიური გამოცემების მიხედვით მომზადდა წარწერების XML და XSLT ფაილები, ასევე წარწერათა მაღალი რეზოლუციის ფოტოდოკუმენტაცია; KML ფორმატში შეიქმნა წარწერების რუკები, დამუშავდა ვებგვერდის დიზაინი და მისი ინგლისურენოვანი თარგმანი.

არმაზის ბილინგვა, რომელიც ყველა არსებულ ნაბეჭდ გამოცემასა და ონლაინ რესურსში წარმოდგენილი იყო ებრაული ტრანსკრიფციით (ალეფ-ბეთ ივრითით), აღნიშნულ კორპუსში დოკუმენტირებული იქნა 2007 წელს მ. ევერსონის მიერ (Everson 2007) ინიცირებული სტანდარტული უნიკოდით. შეიქმნა ფონტი, რომელიც ზუსტად გაიმეორებს არმაზული არამეულის მოხაზულობას.

თანამედროვე ელექტრონული გამოცემის სტანდარტით (EpiDoc) დაძლეული იქნა ეპიგრაფიკული ძეგლის გამოცემის მანამდე არსებული, ლაიდენის კონვენციის (Van Groningen 1932) მთავარი ნაკლი. მიუხედავად იმისა, რომ ეს უკანასკნელი იძლეოდა სემანტიკური კოდირების საშუალებას (Roued-Cunliffe 2010), ძირითადი აქცენტი გადატანილი იყო ეპიგრაფიკის, როგორც არქეოლოგიური და ისტორიული დისციპლინის როლზე და ჯეროვანი ყურადღება არ ექცეოდა წარწერის ენობრივ შესწავლას (Cayless, Roueché, et al. 2009). ელექტრონული გამოცემის სტანდარტი წარწერების ახლებური აღქმის შესაძლებლობას იძლევა.

ერთი მხრივ, ტექსტის კოდირება და მეორე მხრივ, ბმულებით გამდიდრებული აღწერის მონაცემები წარწერას სრულფასოვნად წარმოადგენს – როგორც ტექსტს და როგორც არქეოლოგიურ ობიექტს (Bodard 2008).

Epigraphic Corpora of Georgia's Inscriptions (Corpus Structure and the Standard of Electronic Edition)

Tamar Kalkhitashvili
Ilia State University (Georgia)
kalkhitashvili.tamar@gmail.com

The talk is a presentation of the first sub-project of the corpus of the epigraphic monuments. The corpus project of Georgian, Aramaic, Greek and Armenian inscriptions from 3rd-10th centuries is carried out in accordance with the EpiDoc standard at ISU Linguistic Research Institute. To carry out the new electronic edition, EpiDoc guidelines has been selected, processed and translated into Georgian in advance the diplomatic and critical editions were prepared by the XML and XSLT files, as well as high resolution photo documentation; web site design was developed and its English translation was prepared.

Armazi Bilingual, in all printed and online publications, was presented with Hebrew transcriptions (Hebrew aleph-beth); in our corpora, the Aramaic text of Armazi Bilingual was documented by Everson in 2007 (Everson 2007) by means of initiated standard Unicode characters. In addition, the font with exact shapes of "Armaz Aramaic" was created.

Up-to-date electronic publishing standards (EpiDoc) allowed to overcome the main shortcoming associated with publication of epigraphic monuments in the past, of the Leiden Convention (Van Groningen 1932). Even though the Leiden Convention offered the semantic encoding method (Roued-Cunliffe 2010), the main emphasis was made on epigraphy, as the role of the archaeological and historical disciplines, and due attention was not paid to the linguistic study of inscriptions (Cayless, Roueché, et al. 2009). The standard of electronic publishing allows for a novel perception of inscriptions.

Both text encodings and the description, enriched with link data perfectly present the inscription both as a text and as an archaeological artifact (Bodard 2008).

საკანონმდებლო ტერმინების ელექტრონული ლექსიკონის (თეზაურუსის) შექმნა საქართველოს პარლამენტის საკანონმდებლო ინფორმაციის მართვის სისტემის განვითარების პროექტის ფარგლებში

ზვიად კირტავა
საქართველოს პარლამენტი (საქართველო)
zkirtava@parliament.ge

სამართლებრივი სფეროს განვითარება მისი მეტაენის განვითარების გარეშე შეუძლებელია.

ენის გამოყენებით ხდება ურთიერთობების დამყარება და მოწესრიგება, მათ შორის იმ სამართლებრივი ურთიერთობების ჩამოყალიბებაც, რომელზეც დგას საკანონმდებლო, საკონსტიტუციო, სასამართლო, სამოქალაქო პროცესები. ამ სფეროს ნორმალური განვითარება მოითხოვს დარგის სპეციფიკური ენის პარალელურად განვითარებას, ცნებითი და ტერმინოლოგიური სისტემის ჰარმონიზაციას.

საქართველოს პარლამენტის საკანონმდებლო ინფორმაციის მართვის სისტემის განვითარების პროექტი მომზადდა 2014 წელს და განხორციელდა იმავე წლის აგვისტო-ოქტომბერში. პროექტის მონაწილეები გაერთიანებულნი იყვნენ შემდეგ 5 სამუშაო ჯგუფში:

- საკანონმდებლო აქტების სემანტიკური მოწესრიგების
- საკანონმდებლო აქტების სტრუქტურიზაციის
- საკანონმდებლო პროცესების მოწესრიგების
- კლასიფიკატორებისა და ელექტრონული თეზაურუსის
- ინფორმაციული ტექნოლოგიების.

პროექტის ერთ-ერთი მნიშვნელოვანი ნაწილია საკანონმდებლო ტერმინთა ელექტრონული ლექსიკონის (თეზაურუსის) ტექნოლოგიური ჩარჩოს საპილოტე ვერსიის განხორციელება და სრულმასშტაბიანი სამომავლო სამუშაოების დაგეგმვა.

საკანონმდებლო ტერმინთა ელექტრონული ლექსიკონი – თეზაურუსი წარმოადგენს საქართველოს კანონმდებლობაში გამოყენებულ ტერმინთა სისტემატიზებულ ელექტრონულ ბაზას. მასში მოცემული იქნება თითოეული ტერმინის ერთი ან რამდენიმე მნიშვნელობა, რომლითაც იგი გამოიყენება კანონმდებლობის პრაქტიკაში.

საკანონმდებლო თეზაურუსი დაინტერესებულ პირს საშუალებას მისცემს:

- გაეცნოს კონკრეტული საკანონმდებლო ტერმინის პრაქტიკაში გამოყენებულ ყველა მნიშვნელობას;
- განასხვაოს გამოყენების სფეროთა მიხედვით ტერმინის სხვადასხვა ინტერპრეტაცია;
- სწრაფად მოიძიოს კონკრეტული ტერმინები
- უზრუნველყოს ტერმინთა ერთგვაროვნება

- ჩამოაყალიბოს ტერმინთა განსაზღვრება-განმარტებები

თეზაურუსის სამუშაოები – პროექტის ფარგლებში შემუშავდა:

- კანონმდებლობაში გამოყენებულ ტერმინთა ერთიანი ელექტრონული კრებულის შექმნის მეთოდოლოგია
- საძიებო სისტემა, რომელიც დაეფუძნება ანბანურ კლასიფიკაციას
- ტერმინის ყველა განმარტების ასახვის შესაძლებლობა
- ტერმინის სამართლებრივ წყაროზე (აქტზე) გადასვლის (ბმულის) შესაძლებლობა
- ტერმინის განმარტებაში შემავალი სხვა ტერმინების განმარტებაზე გადასვლის შესაძლებლობაც.
- საპილოტე სისტემაში ინტეგრირდა 6 კანონი და პრაქტიკულად შეიქმნა ინსტრუმენტი, რომლის გამოყენებითაც უკვე შესაძლებელია ვრცელი თეზაურუსის შექმნა.

ერთი მხრივ – ტერმინების განმარტებისა და მეორე მხრივ – მათი გამოყენების აქტების გაერთიანებით მომზადდა ქართული პოლიტიკური კორპუსის ჩამოყალიბების საფუძველი.

თეზაურუსში წარმოდგენილი ტერმინის შესახებ ინფორმაცია ორ სტრუქტურულ ველში იქნება წარმოდგენილი. ერთში მოცემულია დეფინიცია, მეორეში კი მითითებულია ინფორმაცია იმ კონკრეტულ საკანონმდებლო აქტზე, რომელშიც გამოყენებულია/ციტირებულია მოცემული ტერმინი. განმარტებისას უპირატესობა მიენიჭება იმ მნიშვნელობას, რომელსაც მას კონკრეტული საკანონმდებლო აქტი აძლევს.

ელექტრონული თეზაურუსში ერთი ტერმინის ყველა მნიშვნელობა პარალელურად იქნება წარმოდგენილი. განსხვავებული მნიშვნელობები ერთმანეთთან და სათანადო საკანონმდებლო აქტებთან – ჰიპერტექსტული ბმულებით იქნება დაკავშირებული.

საკანონმდებლო ტერმინების განმარტებითი ელექტრონული ლექსიკონის განვითარება უზრუნველყოფს საზოგადოების ფართო ფენებისათვის საკანონმდებლო ანბანის ხელმისაწვდომობას, რაც ხელს შეუწყობს სამოქალაქო საზოგადოების შენებას და ასევე მნიშვნელოვანი ნაბიჯი იქნება ქართული ენის დღემდე არასაკმარისად განვითარებული სეგმენტის – საკანონმდებლო-იურიდიული მეტაენის გაძლიერებისა და ქართული ენის პოლიტიკური კორპუსის ჩამოყალიბებისკენ.

ეს არის ის უნიკალური შემთხვევა, როცა თანამედროვე ტექნოლოგიები წარსულში დაკარგული საგანძურისა და ენის გამდიდრებისა და გაძლიერების შესაძლებლობას იძლევა. ამ შესაძლებლობის განხორციელებისათვის ახლო მომავალში სერიოზული სამუშაოა საჭირო, რომლის მხოლოდ მცირე ნაწილია ჯერ შესრულებული. თუმცა პერსპექტივა გარკვეულია და ვიცით, საით და როგორ უნდა ვიაროთ.

Compilation of the Electronic Dictionary (Thesaurus) of Legislative Terms within the Framework of the Project "Development of Legislative Information Management System for the Parliament of Georgia"

Zviad Kirtava
Parliament of Georgia (Georgia)
zkirtava@parliament.ge

Development of legal sphere is impossible without the development of its respective metalanguage.

Use of language enables the establishment and normalization of relations, including the relations upon which legislative, constitutional, judicial and civil processes are based. Normal development of this sphere requires the simultaneous development of its specific language and the harmonization of its conceptual and terminological systems.

The project "Development of Legislative Information Management System for the Parliament of Georgia" was prepared in 2014 and was implemented in August to October of the same year. The participants of the Project were included in five Working Groups, namely those of:

- Semantic Normalization of Legislative Acts
- Structuralization of Legal acts
- Normalization of Legislative Processes
- Classifiers and Electronic Thesaurus; and of
- Information Technologies

One of the important components of the Project is the effort to implement a pilot version of the technological framework for the Electronic Dictionary (Thesaurus) of Legislative Terms and to draw up plans for full-scale future activities.

The Electronic Dictionary – Thesaurus of Legislative Terms is a systematized electronic database of the terms used in Georgian legislation. It will include one or more meanings of each particular term, in which it is used in the legislative practice.

The Legislative Thesaurus will enable an interested person to:

- know all meanings of each particular legislative term as used in practice
- distinguish between various interpretations of a term as per spheres of application
- quickly look up a concrete term
- guarantee the uniformity of terms
- formulate definitions and explanations of terms

Thesaurus-related work – Within the framework of the Project there were developed:

- The methodology for the setting up of an integrated electronic collection of terms used in the sphere of legislation
- A search system, which will be based on alphabetic classification
- The possibility to reflect all explanations of a term
- The ability to move (via hyperlink) to the legislative source (act) of the term in question
- The ability to move also to the explanations of all other terms included in the explanation of the term in question itself
- The pilot system has integrated 6 laws, thus creating in effect a tool, which can be used for the compilation of an inclusive thesaurus.

By combining the explanations of terms on the one hand, and the respective acts of their application on the other hand, there was formed the basis for the building of a Georgian political corpus.

The information about each term included in the thesaurus will be represented in two structural fields. One field contains definition, while the other one provides the information on the specific legislative act, wherein the term in question is used / quoted. In the process of explanation, preference will be given to the meaning, in which it is used by the concrete legislative act.

In the electronic thesaurus, all meanings of each particular term will be represented in parallel. Different meanings will be connected with one another as well as with respective legislative acts by means of hypertext links.

The development of explanatory electronic dictionary of legislative terms will provide wide circles of the society with the access to the basics of legislation, will strengthen civil society building effort, and will also be a major step forward towards the reinforcement of the hitherto underdeveloped segment of the Georgian language – the metalanguage of legislative and judicial fields, as well as towards the building of the Georgian political text corpus.

This is the unique case when modern technologies provide the possibility to enrich and enhance the language and the treasure lost in the past. In order to convert this possibility into reality, serious work must be carried out in the near future, of which only a small part is done as yet. However, the future prospects are clear and we know where to go and how.

ქართული ენის ანალიზატორის გაუმჯობესებისა და განვითარების პერსპექტივები ქართული ენის კორპუსის საფუძველზე

ირინა ლობჟანიძე
ილიას სახელმწიფო უნივერსიტეტი (საქართველო)
irina_lobzhanidze@iliauni.edu.ge

ბუნებრივი ენის დამუშავებისას არსებობს სხვადასხვა ტიპის კომპიუტერული უზრუნველყოფა, რომლითაც ხდება სიტყვის ანოტირება მორფოლოგიის დონეზე, კერძოდ, სიტყვაში შესულ სხვადასხვა სტრუქტურულ ერთეულს ენიჭება სხვადასხვა მორფოლოგიური ტაგი. მსგავს შემთხვევებში აგლუტინაციური ტიპის ენებისათვის, ძირითადად გამოიყენება ქსეროქსის სასრული პოზიციის საშუალებანი. შესაბამისად, ქართული ენისათვის შექმნილი ანალიზატორის საფუძველს ქსეროქსის კალკულუსი წარმოადგენს.

აღნიშნული მოხსენებით გვინდა ყურადღების გამახვილება იმ სისტემაზე, რომელშიც ხდება ანალიზატორის შედეგების გადამოწმება და დამატებითი ერთეულების შემატება ქართული ენის კორპუსის გამოყენებით.

მორფოლოგიური ანალიზატორი შეიქმნა შოთა რუსთაველის ეროვნული სამეცნიერო ფონდის მხარდაჭერით (AR/320/4-105/11). მითითებული პროექტის მთავარი მიზანი იყო ქართული ენის კორპუსის შექმნა და, შესაბამისად, კვლევის შედეგები ზოგადად ემსახურებოდა შემდეგი ამოცანების გადაწყვეტას:

ა) ანალიზატორის/გენერატორის შექმნა თანამედროვე ქართული ენისთვის ქსეროქსის სასრული პოზიციის საშუალებების გამოყენებით;

ბ) მორფოსინტაქსისა და მორფონოლოგიური წესების გაწერა ისე, რომ არ მოხდეს არასწორი ქართული სიტყვების კლასიფიცირება და გენერაცია;

გ) თანამედროვე ქართული ენის კორპუსის ანოტირება მორფოლოგიის დონეზე ზემოხსენებული ანალიზატორის გამოყენებით.

შესაბამისად, მოხსენებაში ძირითადი ყურადღება გამახვილდება ანოტირების შედეგების გადამოწმებაზე ქართული ენის კორპუსის გამოყენებით.

1. ქართული ენის კორპუსის ანოტირება მორფოლოგიური ანალიზატორის გამოყენების საფუძველზე

ქართული ენა განეკუთვნება მორფოლოგიური კატეგორიებით მდიდარ ენათა რიგს. ქართული ენის მორფოლოგიური სტრუქტურის აღწერა მოიცავს შემდეგ თავისებურებებს: ფლექსიურ კატეგორიათა მთელი რიგების აღწერას; იმ ერთეულთა ასახვას, რომლებსაც ზმნური და სახელური პარადიგმები მოიცავენ; სხვადასხვა ელემენტთა ურთიერთდამოკიდებულებათა აღწერას და რეგულარული, ნახევრად-რეგულარული და ირეგულარული ნიმუშების აღწერას. ბუნებრივია, რომ ენის ყველა ზემოხსენებული თავისებურება ართულებს ქართული ენის მოდელირებას.

ქართული ენის მორფოლოგიური ანალიზატორი შეიქმნა სასრული პოზიციის ავტომატების გამოყენებით. სასრული პოზიციის კომპიუტერული საშუალებანი ფართოდ გამოიყენება მსოფლიოში კომპიუტერული ფონოლოგიული და მორფოლოგიური ანალიზისათვის სხვადასხვა ენაში. ანალიზატორი შეიქმნა შოთა რუსთაველის სამეცნიერო ფონდის დაფინანსებით AR/320/4-105/11 პროექტის ფარგლებში. სისტემის თავისებურებიდან გამომდინარე მორფოსინტაქსი კოდირებულია ლექსიკონებში, ხოლო შენაცვლების წესები კოდირებულია რეგულარული გამოსახულებების სახით. გარდა ძირითადი თავისებურებებისა, გასათვალისწინებელი იყო ისიც, რომ თანამედროვე ქართული ენა ბოლომდე აგლუტინაციური არ არის. ფართოდ მიღებული ცნებების ფარგლებში აგლუტინაციური ენა გულისხმობს იმას, რომ სიტყვის ძირი არ იცვლება, ხოლო ნებისმიერი სახის აფიქსი, რომელსაც კონკრეტული გრამატიკული ფუნქცია ახლავს, ძირს პირდაპირ ემატება. ქართული ენა, ამ მხრივ, ნარევი ხასიათისაა. ამ მხრივ, განსაკუთრებით საინტერესოა ქართული ზმნური პარადიგმა, რომელშიც ვხვდებით უამრავ არაკონკატენციურ პროცესს, რომელიც გენერაციის თვალსაზრისით ბევრად უფრო რთულია.

ანალიზატორის შედეგებზე დაყრდნობით ხდება ქართული ენის კორპუსში შესული ტექსტების ანოტირება მორფოლოგიის დონეზე და სამომავლოდ ანალიზატორში შეტანილი კატეგორიების საფუძველზე კორპუსის საძიებო სისტემის გამართვა, რომელიც ამ ეტაპზე საშუალოდ 150-მდე გრამატიკული ერთეულის მოძებნის საშუალებას გვთავაზობს.

კორპუსში შესული ტექსტების მრავალფეროვნება ითხოვს ანალიზატორის მუდმივ განახლებას და მის განახლებაზე კონტროლს. შესაბამისად, ამ ძირითადი პრობლემის გადასაწყვეტად ქართული ენის კორპუსში ამ ეტაპზე შექმნილია დამატებითი შიდა პანელი შემდეგი ბლოკებით:

ა) ტექსტის ატვირთვა – აღნიშნული ბლოკიდან ხდება ქართული ენის კორპუსში სხვადასხვა ჟანრის ტექსტის ატვირთვა და ატვირთული ტექსტების ანოტირება;

ბ) ტექსტები – აღნიშნული ბლოკი ორი ნაწილისგან შედგება: პირველი ნაწილი უზრუნველყოფს მეტაინფორმაციის გადამოწმებისა და კორექტირების საშუალებას, ხოლო მეორე ნაწილი გვაძლევს ომონიმიის მოხსნის შესაძლებლობას, ანალიზატორის შედეგების გადამოწმებას და, საჭიროების შემთხვევაში, კორექტურას;

გ) სიტყვების კორექტირება – ბუნებრივია, რომ მიუხედავად კორექტურისა, ტექსტებში შიგადაშიგ გვხვდება ისეთი სიტყვები, რომლებიც ენაში არ დასტურდება და წარმოადგენს მხოლოდ ორთოგრაფიულ შეცდომას. ბუნებრივია, რომ ანალიზატორისათვის მსგავსი სიტყვების ამოცნობა შეუძლებელია. ამიტომაც მსგავსი სიტყვების დადასტურების შემთხვევაში საჭირო ხდება მათი კორექტირება;

დ) უცნობი სიტყვები – სხვადასხვა ჟანრის ტექსტში შეიძლება შეგვხვდეს ან ტერმინოლოგიური ერთეული ან ნასესხობა ანდა რაიმე სხვა ტიპის სიტყვა, რომელიც ანალიზატორის ლექსიკონში შესული არ არის. შესაბამისად, სისტემა ინახავს და აგროვებს მსგავსი ტიპის ერთეულებს. ამ ბლოკიდან შესაძლებელი ხდება მსგავსი ტიპის ერთეულებისათვის კლასის მიჩენა, რაც გვაძლევს კონკრეტული სიტყვის კონკრეტულ ლექსიკონში მოთავსების საშუალებას.

2. ტესტირება და სამომავლო განვითარების პერსპექტივა

სასრული პოზიციის გარდამქმნელები გვაძლევენ იმის ნახვის საშუალებას, თუ რომელი სიტყვა შეემატა ან გამოაკლდა საერთო ქსელს. მორფოლოგიური ანალიზატორის შემოწმება

მუდმივად მიმდინარეობს. ქართული ენის ანალიზატორის მთავარი დანიშნულებაა სხვადასხვა ტიპის სიტყვების როგორც გარჩევა, ასევე გენერაციაც. შესაბამისად, მისი მუშაობის შედეგი მოწმდება კორპუსში შესულ სხვადასხვა ჟანრის ტექსტებში არსებულ სიტყვებზე. ამ ეტაპზე მორფოლოგიურ ანალიზატორში შესული ბლოკებიდან ყველაზე მეტი ფორმა დასტურდება ზმნური პარადიგმის ბლოკში, რომელიც აგენერირებს 9 მილიონამდე ფლექსიურ ფორმას საშუალოდ, მათ შორის 80%–ზე მეტი დასტურდება კორპუსში. ამ ეტაპზე მიმდინარეობს ანალიზატორის გამდიდრება კორპუსში დადასტურებული სიტყვებით. ანალიზატორის განვითარების მთავარ ამოცანად გვესახება ურთიერთდამოკიდებულებათა ასახვა სინტაქსის დონეზე და ამის საფუძველზე კონკრეტული ტიპის ლინგვისტური ომონიმიის მოხსნა.

ლიტერატურა:

Beesley, K., Karttunen, L. *Finite State Morphology.* Stanford: CSLI Publications, 2003.

Gurevich, O. "A Finite-State Model of Georgian Verbal Morphology." *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL.* NY: Association for Computational Linguistics, 2006. 45-48.

Jurafsky, D., Martin, H. J. *Speech and Language Processing.* New Jersey: Pearson Education International, 2009.

Kapanadze, O. "Describing Georgian Morphology with a Finite-State System." *Lecture Notes in Computer Science*, 2010: 114-122.

Meurer, P. "A Computational Grammar for Georgian." *Lecture Notes in Computer Science*, 2009: 1-15.

Stump, G. T. *Inflectional Morphology: a Theory of Paradigm Structure.* NY: Cambridge University Press, 2001.

მელიქიშვილი, დ. *ქართული ზმნის უღლების სისტემა.* თბილისი: ლოგოს პრესი, 2001.

შანიძე, ა. *ქართული ენის გრამატიკის საფუძვლები.* თბილისი: თბილისის უნივერსიტეტის გამომცემლობა, 1973.

SOFTWARE:

Xerox Finite-State Tools (tools: lexc, xfst, lookup; operating system: Windows).

Improvement of the Georgian Morphological Analyzer on the Basis of the Corpus of Modern Georgian Language

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

There are a lot of applications of Natural Language Processing, which give appropriate annotation at the level of morphology, especially, each structural unit of a word is equipped with concrete morphological tags. For the cases of agglutinating languages, generally, Xerox tools are used. Thus, the morphological analyzer of Georgian Language is compiled on the basis of Xerox calculus.

In this paper, we present the system which provides assistance in the testing of the morphological analyzer and its enrichment with new words using the Corpus of Modern Georgian Language.

The morphological analyzer was developed within the framework of the project (AR/320/4-105/11)[1] financed by the Shota Rustaveli National Science Foundation. The main aim of the above-mentioned project was to create the Corpus of Modern Georgian Language, and, appropriately, the results of research were used to reach the following targets:

a) to build an analyzer/generator for Modern Georgian by means of the Xerox finite state tools;

b) to accurately model both morphotactics and morphophonological alternation rules, so that only well-formed Georgian words are classified and generated;

c) to provide annotation of the corpus of Modern Georgian language using the above-mentioned analyzer.

Thus, the main attention is paid to the testing of the annotation results using the corpus of Modern Georgian language.

1. Annotation of the Corpus of Modern Georgian Language on the basis of Morphological Analyzer

Modern Georgian language belongs to a morphologically rich languages. Descriptions of Georgian morphological structure emphasize the large number of inflectional categories; the large number of elements that a verb or a noun paradigms can contain; the interdependence in the occurrence of various elements and the large number of regular, semi-regular and irregular patterns. All the above-mentioned peculiarities make computational model of Georgian morphology a rather difficult task.

The Morphological analyzer of Modern Georgian language has been developed using finite state automata. Such kind of tools has been applied to the analysis of phonology and morphology in different languages. The analyzer was developed within the framework of the project AR/320/4-105/11 financed

[1] The opinion represented in the paper belongs to the author and may be does not correspond to the opinion of the Shota Rustaveli National Science Foundation

by the Shota Rustaveli National Science Foundation. The morphotactics is encoded in the lexicons and alternation rules are encoded in regular expressions. In addition to the above-mentioned peculiarities we had to take into account the fact that Modern Georgian language can't be considered as completely agglutinating. According to the existing definitions, the main peculiarity of agglutinating languages is that the root of a word doesn't change and each affix added directly to the root has its own grammatical function. From this point of view Georgian language is of mixed nature; especially, the paradigm of Georgian verb undergoes non-concatenative processes, which are more difficult from the viewpoint of computer generation.

The results of analyzer are used for the annotation of texts in the Corpus of Modern Georgian language at the level of morphology and for the organization of query system, which, at this stage of development, is able to find about 150 grammatical categories.

The diversity of texts in the corpus requires permanent renovation of the analyzer and control for this renovation. Thus, to reach the above-mentioned task, the inner panel of the Corpus of Modern Georgian Language, at this stage, is equipped with the following blocks:

a) Text Upload – the mentioned block provides the upload of texts of different genres to the Corpus of Modern Georgian language and the annotation of text already uploaded;

b) Texts – the mentioned block is subdivided into three parts: the first part provides checking of meta-annotation and its correction (in case of need), the second part allows us to remove homonymy, to check the results of analyzer and to correct them (in case of need);

c) Word Correction – in spite of correction of texts, we meet some words, which don't exist and can be considered as an orthographic mistake. The analyzer can't provide their annotation. Thus, such kind of words can be corrected directly on-line;

d) Unknown words – the texts of different genres have different type of words, e.g. terminological words, borrowings etc. These types of words are not represented in the lexicon of analyzer. Thus, the system collects such kind of word. This block allows the editor to give appropriate class to the word. The above-mentioned information allows us to put the word directly to the lexicon.

2. Testing and Future Development

The xerox finite state transducers allows us to see which words were added or removed from network. Thus, the check is performed permanently. The analyzer is designed for broad coverage. Thus, the output is compared against the words extracted from texts of different genres. At present the verbal paradigm generates over 9000000 inflected forms and recognizes over 80% of the forms frequently used in the corpus. The further stage of development is to equip a model with syntax block, which will allow us to avoid some kind of linguistic homonymy.

კოლოკაციები და კონტექსტი – კოლოკაცია და ქსელები

ტონი მაკენერი
ლანკასტერის უნივერსიტეტი (დიდი ბრიტანეთი)
a.mcenery@lancaster.ac.uk

სიტყვათშეხამებათა შესახებ არსებული საენათმეცნიერო კვლევა-ძიება საკმაოდ მოცულობითია. ფერთის (Firth 1957: 6) წინადადება იმის თაობაზე, რომ განეხილათ „რა გარემოცვაში იმყოფება სიტყვები“, განხორციელდა სხვადასხვაგვარად (იხ. Evert 2004, 2008, 2010) და გამოკვლეულ იქნა სხვადასხვა კონტექსტში (მაგ., Baker et al. 2008, Xiao&McEnery 2006, Syanova&Schmitt 2008). და მაინც, კოლოკაციების შესწავლის ორმოცდაათი წლის განმავლობაში ამ კვლევა-ძიების შედეგად შეთვისებული ცოდნის დიდი ნაწილი ჯერ კიდევ შესაფასებელია სისტემური სახით და დასანერგია იმ ინსტრუმენტებში, რომლებსაც იყენებენ კორპუსის სფეროში მომუშავე ლინგვისტები (იხ. Gries 2013, სადაც წარმოდგენილია ნიშანდობლივი დისკუსია იმის შესახებ, თუ როგორ უნდა გავაუმჯობესოთ კოლოკაციების კვლევა). ტრადიციულად შემოთავაზებულია კოლოკაციების გამოყოფის სამი კრიტერიუმი: (I) დისტანცია, (II) სიხშირე, და (III) ექსკლუზიურობა. დისტანცია აზუსტებს დიაპაზონს ბირთვის (ჩვენთვის საინეტერესო სიტყვის) ირგვლივ, რომლის ფარგლებშიც ჩვენ ვეძებთ კოლოკატებს. ამ დიაპაზონს ეწოდება „კოლოკაციის ფანჯარა.“ დისტანცია კოლოკატსა და ბირთვს შორის შეიძლება იყოს არაუმეტეს ერთი სიტყვისა, თუ ჩვენ გვაინტერესებს, მაგალითად, ზედსართავები, რომლებიც უშუალოდ უძღვიან არსებით სახელს ინგლისურ ენაში, ან არაუმეტეს ოთხი ან ხუთი სიტყვისა ბირთვის ორივე მხარეს, თუ ჩვენ გვაინტერესებს უფრო ზოგადი შეხამებები (კოლოკაციისა და დისტანციის თაობაზე პოლემიკის თაობაზე იხ. Sinclair et al, 2004 [1970]: 42-48). მეორე კრიტერიუმი – გამოყენების სიხშირე წარმოადგენს სიტყვათშეხამების ტიპურობის მნიშვნელოვან ინდიკატორს. მაგალითად, არსებითი სახელი love ხშირად გვხვდება წინდებულთან in და, მაშასადამე, in love მნიშვნელოვანი „ნაჭერია“ ინგლისურში. მიუხედავად ამისა, in ასევე შეიძლება შეგვხვდეს ბევრი სხვა არსებითი სახელის წინ, როგორებიცაა case, fact, ან school. შესაბამისად, მათ შორის (love და in) არსებული ურთიერთობა არ არის ექსკლუზიური. მეორე მხრივ, love უფრო ძლიერად და ექსკლუზიურად დაკავშირებულია არსებით სახელთან affair; როდესაც სიტყვა affair ჩნდება ტექსტში, დიდი ალბათობაა იმისა, რომ წინამავალი სიტყვა იყოს love. ზემო განხილულ სამ კრიტერიუმთან ერთად შ. გრიზი (Gries 2013) გამოყოფს კიდევ სამ კრიტერიუმს, რომლებიც უნდა გავითვალისწინოთ; (IV) მიმართულება, (V) დისპერსია, და (VI) ტიპისა (უნიკალური სიტყვაფორმა) და ტოკენის დისტრიბუცია კოლოკატებს შორის.

მიმართულება არ ნიშნავს იმას, რომ ორ სიტყვას შორის არსებული მიმართების ძალა იშვიათადაა სიმეტრიული. მაგალითად, სიტყვას affair უფრო ძლიერი ურთიერთობა აქვს სიტყვასთან love, ვიდრე სიტყვას love აქვს affair-სთან, რადგანაც love სხვა სიტყვებთან უფრო მეტადაა თანაპოვნიერი, ვიდრე პირიქით. სიტყვათშეხამების ტრადიციული საზომები მაინც ვერ წვდებიან ამ განსხვავებას, რადგან კორპუსის ლინგვისტიკაში ფართოდ გამოყენებული მათი

უმრავლესობა სიმეტრიულ საზომებს წარმოადგენს. 1 ამიტომაა, რომ გრიზი (Gries 2013) გვთავაზობს გამოვიყენოთ Delta P, როგორც საზომი, რომელიც ითვალისწინებს მიმართულებას ისე, რომ გვაძლევს კოლოკაციის ძალის ორ სხვადასხვა სიდიდეს სიტყვათა ნებისმიერი წყვილისათვის. დისპერსია წარმოადგენს ბირთვისა და კოლოკატების დისტრიბუციას კორპუსში. მაგალითად, ბრიტანულ ეროვნულ კორპუსში სიტყვა affair ეხამება სიტყვას love 189 შემთხვევაში, რაც გადანაწილებულია 151 ტექსტში. ეს შედარებით თანაბარი დისტრიბუციაა კიდევ ერთ პოტენციურ კოლოკატთან agape-სთან (ბერძნული სიტყვა, რომელიც არარომანტიკულ სიყვარულს აღნიშნავს) შედარებით, რომელიც ცხრაჯერ გვხვდება მხოლოდ ორ ტექსტში. დაბოლოს, გრიზი (Gries 2013) წარმოგვიდგენს ტიპისა (უნიკალური სიტყვაფორმა) და ტოკენის დისტრიბუციას, როგორც სასურველ კრიტერიუმს, რომელიც ნაწილობრივ ამოქმედებულია ლექსიკური მიზიდულობის G საზომში (Daudaravičius&Marcinkevičienė 2004). ეს კრიტერიუმი ითვალისწინებს არამარტო მოცემული კოლოკაციური ურთიერთობის ძალას (ვთქვათ, სიტყვებს love და affair შორის), არამედ ასევე კონკურენციის დონეს ბირთვული სიტყვის ირგვლივ არსებულ კოლოკატთა სხვადასხვა ტიპს შორის. ბრიტანულ ეროვნულ კორპუსში გვაქვს კოლოკატთა ცამეტი ათასი სხვადასხვა ტიპი, რომლებიც სიტყვას affair კონკურენციას უწევენ სიტყვასთან love ახლომდებარე ადგილისათვის.

ამ კრიტერიუმებს უნდა მივამატოთ მეშვიდე მახასიათებელი: კავშირუნარიანობა ცალკეულ კოლოკატს შორის. სიტყვათა კოლოკატები გვხვდება არა იზოლირებული სახით, არამედ წარმოადგენენ სემანტიკური ურთიერთობების რთული ქსელის ნაწილს, რომლებიც საბოლოო ჯამში წარმოაჩენენ თავიანთ მნიშვნელობებს და ტექსტის ან კორპუსის სემანტიკურ სტრუქტურას. მაგალითად, ბრიტანულ ეროვნულ კორპუსში სიტყვა affair არ ეხამება ისეთ სიტყვებს, როგორებიცაა unrequited, undying ან madly, მაგრამ ასეთებთან დაკავშირებულია სიტყვის love მეშვეობით, რომელიც ეხამება როგორც სიტყვას affair, ასევე ზემო ნახსენებ სამს (მსგავსად სხვებისა).

წინამდებარე მოხსენებაში ვასაბუთებთ, რომ კოლოკატები უნდა განვიხილოთ არა იზოლირებული სახით, არამედ უფრო ფართო კოლოკაციური ქსელების ნაწილად.

Collocations and Context – Collocation and Collocation Networks

Tony Mcenery
Lancaster University (United Kingdom)
a.mcenery@lancaster.ac.uk

The linguistic research on word associations is vast. Firth's (1957: 6) suggestion to look at the "company that words keep" has been operationalised in a number of different ways (see Evert 2004, 2008, 2010) and has been explored in a number of different contexts (e.g. Baker et al. 2008, Xiao & McEnery 2006, Syanova & Schmitt 2008). However, more than fifty years into the research on

collocations, many of the lessons learned from this research have yet to be systematically evaluated and fully implemented in the tools that corpus linguists use (see Gries 2013 for an important discussion about how research into collocations can be improved). Traditionally, three criteria for identifying collocations have been proposed. These are: (i) distance, (ii) frequency, and (iii) exclusivity. The distance specifies the span around a node word (the word we are interested in) where we look for collocates. This span is called the 'collocation window'. The distance of the collocate from the node can be as little as one word if we are interested, for instance, in the adjectives immediately preceding a noun in English, or as much as a span of four or five words on each side of the node, if we are interested in more general associations (for a debate on collocational distance, see Sinclair et al. 2004 [1970]: 42-48). The second criterion, frequency of use, is an important indicator of the typicality of a word association. For instance, the noun *love* occurs frequently with the preposition *in* and therefore *in love* is an important 'chunk' in the English language. However, *in* can also appear in front of many other nouns, such as *case*, *fact*, or *school*. Consequently, the relationship between *love* and *in* is not exclusive. On the other hand, *love* is much more strongly and exclusively connected with the noun *affair*; when the word *affair* appears in text, there is a large probability that the preceding word is *love*. In addition to the three criteria discussed above, Gries (2013) points out three other criteria that should be considered: (iv) directionality, (v) dispersion and (vi) type-token distribution among collocates.

Directionality refers to the fact that the strength of the attraction between two words is rarely symmetrical. For example, the word *affair* has a stronger relationship with the word *love* than *love* with the word *affair* because *love* co-occurs with other words than *affair* more often than vice versa. Yet the traditional association measures do not capture this difference because the majority of those commonly used in corpus linguistics are symmetrical measures.[1] Gries (2013) therefore suggests using Delta P as a measure that takes directionality into account by producing two different values of collocational strength for any pair of words. Dispersion is the distribution of the node and the collocates in the corpus. For example, in a general corpus of British English such as the BNC the word *affair* collocates with *love* in 189 cases distributed across 151 texts. This is a relatively even distribution compared to another potential collocate *agape* (a Greek term for non-romantic love), which occurs 9 times but only in 2 texts. Finally, Gries (2013) raises type-token distribution as a desirable criterion which has been partly operationalized through the lexical gravity *G* measure in Daudaravičius & Marcinkevičienė (2004). This criterion takes into account not only the strength of a given collocational relationship (say between *love* and *affair*), but also the level of competition for the slot(s) around the node word from other collocate types. In the BNC, there are about 13 thousand different collocate types which compete with *affair* for a slot near the word *love*.

To these criteria we should add a seventh feature: the connectivity between individual collocates. Collocates of words do not occur in isolation, but are part of a complex network of semantic relationships which ultimately reveals their meaning and the semantic structure of a text or corpus. For example, in the BNC the word *affair* does not collocate with words such as *unrequited*, *undying* or *madly* but is connected with these through the word *love* which collocates with both *affair* and the three terms mentioned above (among others).

As I argue in this presentation, collocates should not be considered in isolation but rather as part of larger collocation networks.

References:

Baker, P., Gabrielatos, C., Khosravnik, M., Kryzanowski, M., McEnery, T. and Wodak, R. (2008) 'A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press', Discourse and Society, Volume 19, Issue 3, pp 273-306.

Daudaravičius, V. & R. Marcinkevičienė. (2004). Gravity counts for the boundaries of collocations. International Journal of Corpus Linguistics, 9(2), 321–48.

Evert, S. (2004). The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart. Veröffentlicht.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling, & M. Kytö (Eds.), Corpus Linguistics. An International Handbook (pp. 223-233). Chapter 58. Berlin, Germany: de Gruyter.

Evert, S. (2010). Computational Approaches to Collocations. Retrieved from http://www.collocations.de/ (last accessed March 2015).

Firth, J.R. 1957. Papers in Linguistics. Oxford: Oxford University Press.

Gries, S. Th. (2013). 50-something years of work on collocations: What is or should be next…. International Journal of Corpus Linguistics, 18(1), 137-166.

Sinclair, J., Jones, S., & Daley, R. (2004). English Collocation Studies: The OSTI Report. London, UK: Continuum.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. Canadian Modern Language Review/La Revue Canadienne des Langues Vivantes, 64(3), 429-458.

Xiao, Z. and McEnery, T. (2006) 'Collocation, semantic prosody and near synonymy: a cross-linguistic perspective' Applied Linguistics, Volume 27, Issue 1, pp. 103–129.

ქართული ენის ელექტრონული სწავლების კურსი – ახალი ვერსია (A1 – B2 დონეები) და პროგრამის განვითარების პერსპექტივა

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

XXI საუკუნის საგანმანათლებლო სივრცე წარმოუდგენელია ინოვაციური ტექნოლოგიებისა და თანამედროვე სასწავლო პროგრამების გარეშე, რომლებშიც ინტეგრირებულია ინოვაციური ტექნოლოგიების გამოყენებით შექმნილი სასწავლო კურსები და აქტივობები. თანამედროვე თუ ტრადიციული უნივერსიტეტების უმეტესობა ცდილობს სტუდენტებს შესთავაზოს განათლების მიღების ალტერნატიული გზები, როგორებიცაა: დისტანციური, ელექტრონული თუ შერეული ტიპის სწავლება. ამ ტიპის სასწავლო პროგრამები მოქნილია, გამდიდრებულია ვიდეო, აუდიო, გრაფიკული მედიასაშუალებებით, ახლავს სხვადასხვა ტიპის ინტერაქტიული სავარჯიშოები, ითვალისწინებს თვითტესტირებისა და შემაჯამებელი, ფინალური ტესტირების სისტემას და სხვ. რაც მთავარია, არის ხარისხიანი განათლების მიღებაზე ორიენტირებული.

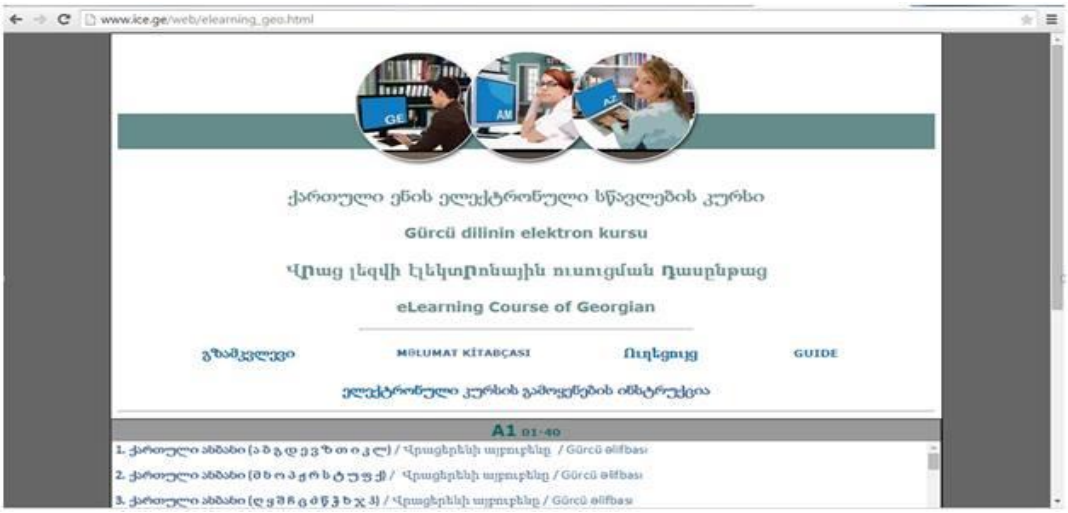
პოპულარულ დისტანციურ თუ ელექტრონულ სასწავლო კურსებს შორის, რა თქმა უნდა, ენის სასწავლო კურსები განსაკუთრებულ ადგილს იკავებს, მაგალითისათვის, თუ ერთ-ერთი ე.წ. "ღია რესურსის" მონაცემებს გადავხედავთ, სადაც სხვადასხვა ენის კურსია განთავსებული, ვნახავთ, რომ 2011 წლის მონაცემებით, ვებგვერდს 300 000 მომხმარებელი ჰყავდა, ხოლო 2014 წელს მათი რიცხვი 25 მილიონამდე გაიზარდა, აქედან 12, 5 მილიონი აქტიური შემსწავლელია (https://ru.wikipedia.org/wiki/Duolingo).

პროექტი, რომელსაც ჩვენ წარმოვადგენთ, დაფინანსდა ფონდ „ღია საზოგადოება – საქართველოს" მიერ სამოქალაქო საზოგადოების მხარდაჭერის პროგრამის ფარგლებში და ითვალისწინებდა ეთნიკური უმცირესობების წარმომადგენელთა მხარდაჭერას სახელმწიფო ენის შესწავლის თვალსაზრისით. იგი განხორციელდა არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტში და, შესაბამისად, განთავსებულია ინსტიტუტის ვებგვერდზე: http://www.ice.ge/web/elearning_geo.html

თანამედროვე ტექნოლოგიების გამოყენებით შეიქმნა ქართული ენის შემსწავლელი ელექტრონული კურსი, რომელიც ხელმისაწვდომია ნებისმიერი მსურველისათვის, ვინც კი დაინტერესდება შეისწავლოს სახელმწიფო ენა. ელექტრონული კურსი არ არის კომერციული დანიშნულების; მომხმარებელთა რაოდენობისა და არეალის მიხედვით თუ ვიმსჯელებთ, იგი დღეს ერთ-ერთი წარმატებული პროექტია.

ქართული ენის ელექტრონული სწავლების კურსი ორ ეტაპად დაფინანსდა:

1. „ქართული ენის ელექტრონული სწავლების კურსი (A1, A2, B1 დონეები)“ და 2. ,,ქართული ენის ელექტრონული სწავლების კურსის დანერგვა და განვითარება საგანმანათლებლო სივრცეში (A1-B2 დონეები)”, შესაბამისად, კურსის პირველი ვერსია ინტერნეტსივრცეში განთავსდა 2012 წელს, ხოლო მეორე, განახლებული და შევსებული ვერსია ხელმისაწვდომი გახდა 2014 წლის სექტემბრიდან.



პროექტი, ასევე, ითვალისწინებდა საქართველოს რეგიონებში ქართულის, როგორც მეორე ენის მასწავლებელთა ტრენინგებს, რათა მიღებული ელექტრონული რესურსით სარგებლობის წესებს გასცნობოდნენ ადგილობრივი პედაგოგები და, როგორც მულტიპლიკატორებს, დაენერგათ ახალი ტიპის სწავლება; გაევრცელებინათ თანამედროვე მეთოდებით ენის სწავლების რესურსის შესახებ ინფორმაცია.

დღეისათვის“**ქართული ენის ელექტრონული სწავლების კურსი“** არის გაკვეთილების ციკლი(140 თემა/გაკვეთილი), რომელიც საშუალებას აძლევს კომპიუტერის ნებისმიერ მომხმარებელს: ეტაპობრივად, საფუძვლიანად შეისწავლოს ქართული ენა (გრამატიკა, ლექსიკა, ფრაზეოლოგია) ევროპის საბჭოს მიერ განსაზღვრულ ზოგადევროპული სტანდარტის A1, A2, B1 და B2 დონეებზე; გამართოს მეტყველება; განივითაროს წერის უნარი და დახვეწოს წერის კულტურა.

ქართული ენის ელექტრონული სწავლების კურსი ისეთ პლატფორმაზეა დაფუძნებული, რომელიც უპრობლემოდ გულისხმობს პროექტის გაფართოვებას, დახვეწასა და სხვა ენების დამატებას. უმნიშვნელოვანესია ინგლისურენოვანი, თურქულენოვანი, რუსულენოვანი ვერსიების მომზადება, რადგან ჩვენი დიასპორის წარმომადგენელთა თუ ემიგრანტთა უმეტესობა ამ ენების გავრცელების არეალშია მოქცეული.

- ვგეგმავთ ახალი სასწავლო კომპონენტების შემოტანას, მაგ., სასწავლო კურსის ინტეგრირება სხვადასხვა ტიპის კორპუსთან, რაც შემსწავლელს საშუალებას მისცემს, ჰქონდეს უშუალო კონტაქტი ნებისმიერი ტიპის ტექსტთან და, თემატიკიდან გამომდინარე, შეასრულოს კონკრეტული დავალება.

- ვგეგმავთ კურსის შინაარსობრივ გაფართოებას. კერძოდ, კურსის ფარგლებში ვვარაუდობთ პრაგმატული კომპეტენციის განმავითარებელი აქტივობების დამატებას და სპეციალური დავალებების შეთავაზებას, რომლებიც ქართულ ენაში არსებულ დისკურსულ მოდელებს გააცნობს და მათი სწორად გამოყენების შესაძლებლობას მისცემს სტუდენტებს.
- ენის სხვადასხვა რეგისტრის მახასიათებლების გაცნობა შემსწავლელებისაგან მოითხოვს ენის მნიშვნელოვან ფონურ ცოდნას. ვფიქრობთ, რომ ამ ეტაპზე გამართლებული იქნება შემსწავლელებს შესაბამისი მასალის (როგორც ტექსტების, ისე დავალებების) სახით ამ ტიპის ინფორმაციაც შევთავაზოთ სრულფასოვანი კომუნიკაციური უნარების გასავითარებლად.
- თანამედროვე კვლევები აჩვენებს, რომ არავერბალური ლინგვისტური სტრუქტურები კომუნიკაციის უმნიშვნელოვანესი ნაწილია. ამდენად, მათი გაცნობა შემსწავლელებისათვის ერთ-ერთი მნიშვნელოვანი ამოცანაა კურსის ავტორებისათვის. მით უმეტეს, რომ ელექტრონული კურსი იძლევა შესაძლებლობას, თეორიული ინსტრუქციების ნაცვლად, სწორედ ვიზუალურად გამდიდრებული, ავთენტური მასალით შევთავაზოთ ამ ტიპის ინფორმაცია.

პროექტის შესრულებული და შესასრულებელი ამოცანები პირობითად სამ ეტაპად შეიძლება დავყოთ:

1. პროექტის პირველი ვერსია (A1-B1დონეები);
2. პროექტის ახალი ვერსია (A1-B2 დონეები);
3. პროექტის განვითარება – განვრცობა ახალი ენებით, დამატებითი სავარჯიშოებითა და ვიზუალური, ვიდეო- თუ აუდიომასალით; კორპუსთან ინტეგრირებული სასწავლო კომპონენტის შემოტანა; კურსის შინაარსობრივი გაფართოება; პრაგმატული კომპეტენციის განმავითარებელი აქტივობების დამატება; სპეციალური დავალებების შეთავაზება, რომლებიც ქართულ ენაში არსებულ დისკურსულ მოდელებს წარმოაჩენს.

ELearning Course of Georgian – New Version (A1-B2 Levels) and Perspectives for Course Development

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

Contemporary education makes a wide use of innovative technologies. Up-to-date technologies are integrated in the educational programs as well. Both modern and traditional universities offer students alternative ways of education: Distant, Electronic or Blended ones. These types of programs are flexible, enriched with video, audio and graphic media, involve interactive tasks, self-assessment and final tests. Most importantly, they are focused on high-quality education.

Language courses are among the most popular distant or Electronic courses. For example, Duolingo (https://ru.wikipedia.org/wiki/Duolingo), a free language-learning platform, offering different language courses, had 300 000 users in 2011. The number of users increased to 25 million by 2014, 12. 5 million of whom are active learners.

The present project was financed by Open Society – Georgia Foundation within the framework of the program supporting civil society. The aim of the program was to support national minorities in learning the state language. The project was carried out at Arnold Chikobava Institute of Linguistic. The product is uploaded on the web-site of the institute: http://www.ice.ge/web/elearning_geo.html.

ELearning Course of Georgian, which was developed in the framework of the above mentioned program, is based on a new methodology and up-to-date technologies. The course is accessible for everyone who is willing to learn the Georgian Language and is a non-commercial product. If we take the number of users and the areas of use into account, we can consider it as a successful one.

Elearning Course of Georgian was implemented as 2 projects:

“Elearning Course of Georgian (A1, A2, B1 levels)

Implementation and Development of Elearning Course of Georgian in Educational Space (A1-B2 levels). The first version of the course was displayed on the Internet in 2012 and the second one, the updated version has been available since September of 2014.

Georgian as a second language teachers’ training was a part of the project. The aim of the trainings was to get teachers familiar with the course and the principles of its use. Afterwards they would disseminate information about the new resource of language teaching among their colleagues.

Elearning Course of Georgian is a collection of lessons (140 topics/lessons), which allows learners to: gradually and fundamentally acquire the Georgian language (grammar, vocabulary, phraseological units) according to the standards of the Common European Framework of Reference for Languages:

learning, teaching, assessment (CEFR) (A1, A2, B1 and B2 levels); develop speaking, listening and writing skills.

The platform which is used for Elearning Course of Georgian Language allows course authors to make changes, complete and add new languages as well.

Preparation of English, Turkish, Russian versions of the course is very important since the majority of members of Georgian diaspora or immigrants live in the areas where those languages are widespread.

- We plan to add new teaching components, e.g. to integrate the course with different language corpora, which will allow learners to have direct access to various kinds of texts and, depending on the topic, complete some tasks.
- We intend to extend the contents of our course by adding pragmatic competence development activities as well offering learners tasks which make them familiar with the discourse patterns of Georgian Language.
- Considerable background knowledge of language is a prerequisite for acquiring different language registers. For development of the comprehensive communicative skills it would be useful to offer materials (texts and tasks) to advanced (B1-B2 levels) learners with such knowledge.
- According to the current researches, non-verbal linguistic structures are important part of communication. Hence, offering them to learners must be the most important task for course developers. One of the benefits of the Ecourse is a possibility to give learners visually enriched, authentic materials instead of theoretical instructions.

Already fulfilled and future goals of our project could be divided into 3 groups:

1. The first version of the course (A1-B1 levels);
2. The updated version of the course (A1-B2 levels);
3. The course development – to add new languages, tasks and visuals, video and audio materials; to add a new teaching component integrated with language corpora to extend the content; to add activities for development of pragmatic competence and tasks involving Georgian discourse patterns.

ინგლისურ-ქართული სამეცნიერო ტექსტების პარალელური კორპუსის პლატფორმა და დარგობრივი ლექსიკოგრაფია

თინათინ მარგალიტაძე, ია ორმოცაძე
ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)
tinatin@margaliti.ge; iaormotsadze@yahoo.com

მოხსენებაში განხილული იქნება ინგლისურ-ქართული სამეცნიერო ტექსტების პარალელური კორპუსის პლატფორმის გამოყენების შესაძლებლობა დარგობრივ ლექსიკოგრაფიაში, კერძოდ კი, ინგლისურ-ქართული დარგობრივი ლექსიკონების შესადგენად. პარალელური კორპუსის პლატფორმა შემუშავდა ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტის ლექსიკოგრაფიულ ცენტრში და აღნიშნული პლატფორმა პირველად 2013 წელს წარმოვადგინეთ კონფერენციაზე „ქართული ენა და თანამედროვე ტექნოლოგიები“. პარალელური კორპუსის გამოყენების პერსპექტივები მრავალმხრივია როგორც ზოგად, ისე სპეციალიზებულ ლექსიკოგრაფიულ პროექტებში. თუ ამჟამად ლექსიკოგრაფიული ცენტრის მიერ შექმნილი პარალელური კორპუსი მხოლოდ ინგლისურ-ქართულ ენებზე არსებულ სამეცნიერო ტექსტებს ეფუძნება, მომავალში მისი შექმნის პრინციპები და მეთოდოლოგია თავისუფლად შეიძლება იქნეს გამოყენებული სხვა ევროპულ-ქართული პარალელური კორპუსების შესაქმნელად.

2013 წელს გამართულ კონფერენციაზე ჩვენ წარმოვაჩინეთ კორპუსის ბაზების მოწყობის ძირითადი პრინციპები. ეს პრინციპები ხანგრძლივი ფიქრისა და განსჯის შედეგად ჩამოყალიბდა და მიზნად ისახავდა ბაზაში ტექსტების იმგვარ განაწილებას, რაც მომავალში საშუალებას მოგვცემდა პარალელური კორპუსი გამოგვეყენებინა ინგლისურ-ქართულ დარგობრივ ლექსიკონებზე სამუშაოდ. ამ მოსაზრებებიდან გამომდინარე კორპუსში შეიქმნა ჯგუფები, რომლებიც დაიყო კრებულებად, ხოლო კრებულები ტექსტურ წყვილებად. მაგ., საქართველოს მეცნიერებათა აკადემიის „მოამბე“ კორპუსში ერთ ჯგუფს წარმოადგენს, რომელიც იყოფა კრებულებად ტომებისა და დარგების მიხედვით: ტომი 65, ექსპერიმენტული მედიცინა; ტომი 65, მათემატიკა; ტომი 65, ენტომოლოგია; ტომი 66, მათემატიკა და ა.შ. კრებულები მოიცავს სამეცნიერო ჟურნალში გამოქვეყნებული სტატიების ინგლისურ-ქართულ თეზისებს. კორპუსის მეორე ჯგუფია არქეოლოგია, რომლის კრებულები მოიცავს: დმანისის გათხრების შესახებ გამოქვეყნებულ სტატიებს ინგლისური თარგმანებითურთ, ვანის გათხრებს და სხვა. კრებულებში შეტანილი ტექსტები ავტომატურად იშლება ტექსტურ წყვილებად, ინგლისურ-ქართულ წინადადებებად, რომლებიც რედაქტირდება ხელით, რათა ავტომატურად დანაწევრებული წინადადებები გათანაბრდეს და ძიების დროს მივიღოთ იდეალური ინგლისურ-ქართული წინადადებების წყვილები.

კორპუსში შეტანილი ტექსტები მეცნიერების თითქმის ყველა დარგს მოიცავს და ამ ტექსტებისა და დარგების რაოდენობა ყოველდღიურად იზრდება. ეს დარგებია: მათემატიკა, მექანიკა, გეოფიზიკა, ქიმია, ჰიდროლოგია, გეოლოგია, პალეონტოლოგია, სამშენებლო მექანიკა, მანქანათმცოდნეობა, ჰიდროტექნიკა, ელექტროტექნიკა, ბოტანიკა, გენეტიკა, ფიზიოლოგია, ბიოფიზიკა, ბიოქიმია, ენტომოლოგია, ექსპერიმენტული მორფოლოგია, ექსპერიმენტური მედიცინა, ფინანსები, არქეოლოგია, ეთნოგრაფია, ქართველოლოგია და სხვა. დღევანდელი მდგომარე-

ობით კორპუსში არის 1908 კრებული, 28 000 ხელით გათანაბრებული ტექსტური წყვილი და 1 მილიონამდე სიტყვაფორმა.

კორპუსის მართვის პანელის ფუნქციებია:

- კორპუსების ჯგუფების შექმნა და რედაქტირება;
- კრებულების შექმნა და რედაქტირება;
- ტექსტური წყვილების დამატება და რედაქტირება;
- ტექსტის ავტომატური დანაწევრება, გათანაბრება და ტექსტური წყვილების გენერირება შემდგომი რედაქტირების შესაძლებლობით.

კორპუსზე მუშაობის შემდეგ ეტაპზე მიზნად დავისახეთ შეგვემუშავებინა კორპუსში დარგების მიხედვით ტერმინოლოგიის ტაგირების / მონიშვნისა და კორპუსიდან აღნიშნული ტერმინოლოგიის ამოღების მეთოდოლოგია. შემუშავდა პარალელური კორპუსის პროგრამის სათანადო მოდული, სპეციალური პლატფორმა, რომელიც საშუალებას იძლევა კორპუსში არსებული მასალიდან ამოვიღოთ წინასწარ მონიშნული დარგობრივი ტერმინოლოგია. აღნიშნული მოდულის დამუშავების შემდეგ კორპუსის მართვის პანელის უკვე არსებულ ფუნქციებს დაემატა ახალი ფუნქცია, კერძოდ, **ტექსტში ტაგირებული / მონიშნული ტერმინების აღქმა და ძიება კორპუსში.**

აღსანიშნავია, რომ ტერმინთა ტაგირებისათვის არ გახდა საჭირო სხვადასხვა დარგობრივი ტაგების რთული სისტემის შექმნა. ჩვენ მიერ ზემოთ აღწერილმა ბაზების მოწყობის პრინციპმა, კერძოდ, მასალის კრებულებად, სწორედ დარგების მიხედვით დალაგების პრინციპმა, საშუალება მოგვცა შეგვემუშავებინა ტერმინთა ტაგირების / მონიშვნის უნიფიცირებული სისტემა – ვარსკვლავით ინიშნება სხვადასხვა დარგის ტერმინები, შემდეგ კი კორპუსიდან ახალშექმნილი მოდულის საშუალებით შესაძლებელია ტაგირებული ტერმინების დარგების მიხედვით სორტირება და ამოღება შემდგომი ლექსიკოგრაფიული დამუშავებისათვის. ამა თუ იმ დარგის ტერმინებს კორპუსიდან ვიღებთ ტერმინთა ინგლისურ ეკვივალენტებთან ერთად, აგრეთვე, რაც მეტად მნიშვნელოვანია, კორპუსიდან ვიღებთ ამა თუ იმ ტერმინის შესიტყვებებს, ასევე ინგლისური თარგმანებითურთ.

პარალელური კორპუსის აღნიშნული მეთოდოლოგიისა და ახალი მოდულის კვლევა წარიმართა კორპუსის ერთ-ერთი ჯგუფის, კერძოდ, ფინანსების მასალაზე. თიბისი ბანკმა ჩვენი კვლევისათვის მოგვაწოდა სხვადასხვა საბანკო დანიშნულების ინგლისურ-ქართული პარალელური ტექსტები: წლიური ანგარიშები, საბანკო პროდუქტები, ხელშეკრულებები და სხვა. აღნიშნული ტექსტები შევიდა პარალელური კორპუსის ცალკე ჯგუფად, დამუშავდა, წინადადებები გათანაბრდა, მოინიშნა ტერმინოლოგია. პარალელური კორპუსის ახალი მოდულის დახმარებით ტაგირებული ტერმინოლოგია ამოვიღეთ კორპუსიდან და დავამუშავეთ. ჩვენი მოხსენება სწორედ ამ კვლევის ეტაპებს და შედეგებს წარმოადგენს კონფერენციაზე. მომავალში კორპუსის ეს ჯგუფი შეივსება ისეთი ვრცელი ინგლისურ-ქართული ტექსტებით, როგორებიცაა: აუდიტორული და საფინანსო აღრიცხვების მსოფლიო სტანდარტების სახელმძღვანელოები და მათი ქართული თარგმანები.

აქვე უნდა აღინიშნოს, რომ სამეცნიერო ტექსტების ინგლისურ-ქართულ პარალელურ კორპუსს უკვე აქვს საძიებო ინტერფეისი, რომელიც განკუთვნილია კორპუსის მონაცემთა ბაზაში სასურველი ქართული ან ინგლისური სიტყვის ან შესიტყვების ძიებისა და შესაბამისი შედე-

გების შემცველი პარალელური ტექსტური წყვილების გამოტანისთვის. ძიება ხორციელდება ორმხრივად – როგორც ნაწარმოების ორიგინალ, ასევე ნათარგმნ ტექსტში. ძიება შესაძლებელია როგორც მთლიან მასალაში, ასევე სხვადასხვა პარამეტრების მიხედვით, მათ შორის ნაწარმოების ჯგუფების, კრებულების, ტიპების, ჟანრების, ავტორების, წლების და ა.შ. ფილტრაციით. ძიების შედეგებში ასევე ხელმისაწვდომია ინფორმაცია ნაწარმოების ავტორის, თარგმანის ავტორის, გამოცემის წლისა შესახებ და ა.შ.

The Platform of the English-Georgian Parallel Corpus of Scientific Texts and Specialized Lexicography

Tinatin Margalitadze, Ia Ormotsadze
Ivane Javakhishvili Tbilisi State University (Georgia)
tinatin@margaliti.ge, iaormotsadze@yahoo.com

The paper will discuss the possibility of application of the platform of the English-Georgian parallel corpus of scientific texts in specialized lexicography, in particular, in compiling English-Georgian specialized dictionaries. The platform of the parallel corpus was developed at the Lexicographic Centre of Iv. Javakhishvili Tbilisi State University and was initially presented at the international conference 'The Georgian language and Modern Technologies' in 2013. Prospects of application of the parallel corpus platform are multi-faceted in both: general as well as specialized lexicography. At present, the mentioned parallel corpus is mainly based on English-Georgian scientific texts, although the platform and the methodology can be used in future for the creation of similar parallel corpora with respect to other European languages.

In 2013 we mainly focused on the principles of arrangement of data in corpus databases. These principles were worked out after a long period of deliberation and aimed at the arrangement of texts in databases in a way that would enable the application of the corpus in specialized lexicography in future. Proceeding from these considerations text groups, text sets and text pairs were introduced in the corpus. For example, 'The Bulletin of the Academy of Sciences of Georgia' forms one text group in the corpus. It is further subdivided into text sets according to volumes and domains: volume 65, experimental medicine; volume 65, mathematics; volume 65, entomology; volume 66, mathematics etc. The text sets comprise English-Georgian abstracts of the papers published in the 'Bulletin'. The second text group of the corpus is archaeology and its text sets comprise: articles published about Dmanisi excavations, excavations in Vani *etc* with English translations. Texts uploaded in text sets are automatically broken down into text pairs, i.e. English-Georgian sentence pairs, which are edited manually in order to align automatically broken down sentences and get the ideal pairs of English-Georgian sentences in search results.

Texts uploaded in the corpus comprise all fields of knowledge: mathematics, mechanics, geophysics, chemistry, hydrology, geology, paleontology, machine building science, hydraulic engineering, electrotechnics, botany, genetics, physiology, biophysics, biochemistry, entomology, experimental morphology, experimental medicine, financing, archaeology, ethnography, Kartvelology etc. At present the corpus contains 1908 text sets, 28 000 manually aligned text pairs and up to a million tokens. New texts are added to the corpus on the daily basis.

The next stage of working on the parallel corpus platform was to elaborate the methodology of tagging and extracting specialized terminology from the corpus. A special module of the parallel corpus program was developed that enables us to extract the previously tagged terminology from the corpus. After the development of this module, to the already existing functions of the corpus control panel, namely:

- Management functionalities of text groups
- Management functionalities of text sets
- Management functionalities of text pairs
- Automatic breakdown of texts by sentences, sentence alignment, generation of pairs and further manual alignment options,

A new function was added, in particular:

- Recognition of and search for the tagged terms in the corpus.

It should be mentioned that the tagging process of terminology did not require creation of a complicated tagging system. The principles of the arrangement of corpus databases, described above, enabled us to develop a unified system of tagging. Specialized terminology is tagged by means of an asterisk and then, by applying the newly developed module, it is sorted according to domains and extracted from the parallel corpus for the further lexicographic processing. Specialized lexicography is extracted from the corpus alongside with its English equivalents and what is also important, collocations of terms, with their respective English translations, can be extracted as well.

The study of the mentioned methodology and the new module was conducted on the basis of one of the text groups of the parallel corpus, namely financing. TBC bank provided us with English-Georgian texts of annual reports, agreements, bank products and so on. These texts were uploaded, processed, sentence-aligned and terminology-tagged in the corpus. By means of the new module the tagged terminology was extracted from the corpus and processed. Our paper will present this case study with the detailed account of the stages, as well as the results of the research. To the financing text group of the corpus in future will be added substantial texts of manuals of international financial and audit reporting standards and their Georgian translations.

It should also be mentioned that English-Georgian parallel corpus of scientific texts already has an interface for searching Georgian or English words or collocations and displaying the proper text pairs containing the search results on the screen. Parallel Corpus User Interface has search functions against original, as well as translated material and in whole text. Search may be conducted according to the different parameters such as text groups, text sets, types, genres, authors, years etc. The information about the work, author, translator, publishing year etc. is also available in search results.

ქართული ჟესტური ენის დოკუმენტირება

თამარ მახარობლიძე
ილიას სახელმწიფო უნივერსიტეტი (საქართველო)
ateni777@yahoo.com

ქართული ჟესტური ენა არის საქართველოს ტერიტორიაზე მცხოვრებ ყრუთა ენა. საქართველოში არ არსებობს ამ ენაზე რაიმე ტიპის მასალა ან ტექსტი, არ გვაქვს არანაირი ფორმის ჩანაწერი. არსად არასოდეს ჩატარებულა ამ ენის ასეთი კვლევები. მსოფლიოში არ არსებობს არც ერთი დოკუმენტი ქართულ ჟესტურ ენაზე. ამრიგად, ძალზე დროულია და ფასეულია ამ ენის დოკუმენტირების დაწყება. საერთოდ, ნებისმიერი ენის დოკუმენტირება მრავალი თვალსაზრისით არის საჭირო და მნიშვნელოვანი.

ყრუთა თემის ინტერესებიდან გამომდინარე, ადგილობრივ ბაზარზე ჩვენი პროდუქტის მიმართ დიდი მოლოდინია. ქართული ჟესტური ენის დოკუმენტირება მნიშვნელოვანია შემდეგი მოსაზრებების გამო:

- ქართული ჟესტური ენა არის ნაკლებად შესწავლილი ენა და ნებისმიერი სამომავლო კვლევებისათვის, უპირველეს ყოვლისა, უნდა ჩატარდეს ენის დოკუმენტირება. ენის დოკუმენტირებული მასალა გამოიყენება როგორც საყრდენი მასალა სხვადასხვა მიმართულებით ჩასატარებელი ლინგვისტური და არა მხოლოდ ლინგვისტური კვლევებისათვის. შექმნილი პროდუქტი გამოიყენება როგორც ფუნდამენტური, ასევე გამოყენებითი კვლევებისათვის;
- ქართული ჟესტური ენა, ისევე როგორც ნებისმიერი ჟესტური ენა, არ არის სამწერლობო ენა და ამდენად, რთულია ამ ენის განვითარებაზე დაკვირვება;
- ენის დოკუმენტირებული მასალა კი გამოიყენება დიაქრონიული დასკვნებისათვისაც და ეს განსაკუთრებით ფასეული იქნება უფრო შორეული მომავლის თვალსაზრისით. ამდენად, ეს პროდუქტი არასოდეს არ დაკარგავს თავის ღირებულებას;
- მსოფლიოში წამყვანი ტექნოლოგიური კომპანიების მიერ მუშავდება ჟესტურ ენათა კომპიუტერული თარგმანის ახალი ტექნოლოგიური მიდგომები, ნებისმიერ შემთხვევაში, იმისათვის რომ ქართველმა ყრუებმა შეძლონ ამ სფეროში მიღწეული ინოვაციებით სარგებლობა, საჭიროა ქართული ჟესტური ენის კორპუსის შექმნა, ენის დოკუმენტირება კი გახლავთ ამ მიმართულებით გადადგმული პირველი ნაბიჯი;
- ქართული ჟესტური ენის დოკუმენტირებას უფრო ახლო პლანის პრაგმატული მნიშვნელობაც შეიძლება ჰქონდეს: ფაქტობრივად, აქ იქნება პირველი ჩაწერილი ტექსტები, რომელთა გამოყენებაც შესაძლებელი იქნება სასკოლო და საუნივერსიტეტო სწავლებისათვის. აღსანიშნავია, რომ, სამწუხაროდ, საქართველოში დღემდე არ არსებობს ამ ტიპის ტექსტები;
- წარმოდგენილი პროექტის ფარგლებში შექმნილი პროდუქტი გამოიყენება ენის პრაქტიკული სწავლებისათვისაც.

ქართული ჟესტური ენის დოკუმენტირება ძალზე აქტუალურია ყრუთა – როგორც ლინგვისტური უმცირესობების პრობლემების გადასაჭრელად, მათი უფლებების დასაცავად და მა-

თი სამოქალაქო საზოგადოებაში ფართო ინტეგრაციისათვის, ქართული ჟესტური ენის დოკუმენტირების პროცესი მოიაზრებს ამ ენაზე ტექსტებისა და სიტყვა-გამოთქმების ინოვაციური პროგრამული სისტემით გადაღებასა და ჩაწერას ორ- და სამგანზომილებიანი ახალი ტექნოლოგიებით Kinect-ითა და Leap Motion-ით, ასევე სტანდარტული ვიდეო ფორმატით. ასეთი კომბინირებული ტიპის გადაღება-ჩაწერა არის ჩვენი ინოვაციური ფორმატის მიდგომა და აადვილებს შემდგომში ენის სრული კორპუსის შექმნასა და შესაბამის მორფოლოგიურ ანალიზატორზე მუშაობას. ამ დროს პარალელურ რეჟიმში იქმნება ენობრივი ბაზა, რომელიც ყველა არსებულ ტექნოლოგიურ სისტემას მოერგება. მნიშვნელოვანია ამ ტიპის ბაზის შექმნა და არა მხოლოდ ვიდეო გადაღებები, რამდენადაც ასეთი მასალა სხვადასხვა პროგრამული მიდგომისათვის არის ხელმისაწვდომი და მარტივად წასაკითხი. ამ ტიპის ჩანაწერებს აღარ დასჭირდებათ ადაპტირება და გადაწერა პროგრამული კომპიუტერული თარგმანის განსხვავებული სისტემებისათვის. ფაქტობრივად, ერთდროულად იქმნება ქართული ჟესტური ენის ონლაინ არქივი და ბიბლიოთეკა, სასწავლო მასალები და ამ ენის კორპუსის საყრდენი მასალაც. ჩატარდება საველე სამუშაოები ყრუთა თემთან, ჩატარდება გადაღებები, მასალის ლინგვისტური ანალიზი და დესკრიფცია საერთაშორისო სტანდარტების მიხედვით, შეიქმნება ენის ე. წ. მინიკორპუსი, საბიბლიოთეკო ვირტუალური არქივი და ჟესტური ენის მანქანური თარგმანისათვის გამზადდება მონაცემთა ბაზა. ჩაწერილი მასალა განთავსდება ილიას სახელმწიფო უნივერსიტეტის სერვერზე. მოიაზრება ქართული ჟესტური ენის ლექსიკონის პროგრამული გადამუშავება და დოკუმენტირება, ასევე ქართული ჟესტური ენის ტექსტების დოკუმენტირება. შექმნილი მასალები დამუშავებული იქნება როგორც ლინგვისტური, ასევე პროგრამული თვალსაზრისით და გამოყენებადი იქნება სხვადასხვა პრაგმატული მიდგომისთვის. ჟესტური ენის დოკუმენტირებისთვის ძალზე მნიშვნელოვანია ევროპის ქვეყნებისა და ამერიკის შეერთებული შტატების გამოცდილების გაზიარება. ამ მიზნით ვაპირებთ უცხოელ კოლეგებთან მჭიდრო თანამშრომლობას.

Documentation of GESL (Georgian Sign Language)

Tamar Makharoblidze
Ilia State University (Georgia)
ateni777@yahoo.com

The goal of the project is the documentation of Georgian Sign Language – GESL. This is a language of Deaf and Hard of Hearing people (DHH) in Georgia. There are no documented materials or texts of GESL in Georgia or anywhere else. That is why the project is so timely and so highly valuable. Besides the expectations from DHH, the presented project is important because of the following factors: GESL is an unknown language and first of all the language documentation should be performed. This documented material can be used for any future linguistic and other kind of investigations in fundamental and applied sciences for the interested specialists worldwide. GESL, like other sign languages, is not a written language and it is difficult to follow internal processes of the language

development. The documented materials could be used for diachronic comparisons and this will be highly appreciated in terms of future investigations. The final product will never lose its meaning and value. The world's leading technological companies develop software for computer translation for sign into spoken languages and vice versa. In order to have the access and to be able to use such technological achievements, Georgian DHH will need GESL corpora and the project of GESL documentation is the first real step to this goal. GESL documentation project has the additional concrete pragmatic meanings. These recorded texts will be very useful for practical language teaching at schools and at the Universities. GESL documentation is instrumental for the protection of human and linguistic rights of DHH in Georgia and for their successful integration into the civil society. GESL documentation project intends to film the texts, sign phrases and signs with 3D and 2D technologies with video camera, Kinect and Leap-Motion. Such a combined method is our innovation and it makes easy to create the sign language corpora and it also simplifies the process of elaborating a specific tag-analyzing system for the GESL corpora. During this process we create the language base which will be easy to access for any type of technological system. It is important to make such kind of universal documentation and not to make just a video filming, because such documents will be easy accessible for any kind of software systems /engines, and there will not be any need to adopt these texts in future, as they can meet the different level technological demands. The presented project factually creates the online GESL archive and library, the learning material and GESL corpora basic materials. With close collaboration with DHH in Georgia, we'll film the data and provide the descriptive linguistic analysis on the existed international standards. The final documented texts, sign phrases and signs will be ready for any kind of pragmatic usage. The recorded material will be placed on the server of Ilia State University. For the successful achievement of the goal I intend to have a close collaboration with my foreign colleagues and to share their experience.

Linguascript: მეცნიერების გამოყენება ენის განვითარებისათვის

ნენა ნვოსუ-ნვორუ
ნდუფუ ალიკე იკვოს ფედერალური უნივერსიტეტი (ნიგერია)
nneabel@yahoo.com

ცნობარი *Ethnologue*[1] გვთავაზობს ენის განვითარების ჩარლზ ფერგიუსონისეულ (1968) განმარტებას, რომლის მიხედვითაც იგი სამ სფეროს მოიცავს: *გრაფიზაცია* – დამწერლობის სისტემის განვითარება; *სტანდარტიზაცია* – ნორმების ჩამოყალიბება, რომლებიც უპირატესი იქნება ტერიტორიულ და სოციალურ დიალექტებთან მიმართებით; *მოდერნიზაცია* – შესაძლებელი იყოს თარგმნა და განხორციელდეს დისკურსი იმ ფართო თემატიკის ირგვლივ, რაც ახასიათებს „ინდუსტრიულ, [...] მოდერნიზებულ საზოგადოებებს." უფრო ვრცლად, *Ethnologue* განმარტავს

[1] http://www.ethnologue.com/language-development

ენის განვითარებას, *როგორც მიმდინარე დაგეგმილი ღონისძიებების წყების შედეგს, რომელსაც ახორციელებენ ენობრივი თემები იმის უზრუნველსაყოფად, რომ ეფექტურად გამოიყენონ თავიანთი ენები თავიანთი სოციალური, კულტურული, პოლიტიკური, ეკონომიკური და სულიერი მიზნების მისაღწევად.*[1] წინამდებარე ნაშრომი ეხება ენის განვითარების შეფასების მეხუთე ინდიკატორს: „ენის გამოყენება ახალ მედიასაშუალებებში, როგორებიცაა ვებგვერდები, ჩეთრუმები (chat rooms), პოდკასტები (podcasts) და MP3 ჩამოტვირთვა, და მოკლე ტექსტური შეტყობინებების ასაკრებად მობილურ ტელეფონებსა და სხვა ელექტრონულ ხელსაწყოებზე".

- **მიზანი: რა პრობლემები გადაწყდება?**

ნიგერიის ძირითადი ენების ენობრივი სიცოცხლისუნარიანობა გაუმჯობესდება.

- **თეორიული ჩარჩო**

გრძელვადიანი პოტენციაციის კვლევა ხორციელდება ჰიპოკამპის სხვადასხვა ნაწილში; ჰიპოკამპი არის ტვინის მნიშვნელოვანი ნაწილი, რომელიც პასუხისმგებელია **სწავლასა** და **მეხსიერებაზე**. ეს ხდება სინაპტიკური ძალის გაზრდის შედეგად, რასაც სხვაგვარად გრძელვადიანი პოტენციაცია ეწოდება [...].

კვლევამ დაადასტურა, რომ არსებობს სინაფსის სამი ფუნქცია: ინჰიბიციური, მეხსიერებითი და სივრცითი. წინამდებარე ნაშრომში განვიხილავთ მეორე ფუნქციას, რათა დავამტკიცოთ, რომ გრძელვადიანი პოტენციაციისა და ქცევითი სწავლის კავშირს შეუძლია უწიგნურობა გადაქციოს წიგნიერებად.

მეთოდოლოგია

- გრძელვადიანი პოტენციაციის მეთოდის გამოყენება გრძელვადიანი პოტენციაციისა და ქცევითი სწავლების კავშირის წარმოსაჩენად მოსახლეობის სამიზნე ჯგუფისათვის.
- არსებული Uniskript-ის ანბანური სისტემის გამოყენება მოდელის (**Linguascript**) შესამუშავებლად იმ ენისათვის, რომელიც გამოყენებული იქნება ნიგერიის აქტიურ მოსახლეობაში.

- **სპეციფიკური ამოცანები**

სწრაფი ფუნქციონალური წიგნიერების ეფექტის მიღწევა ერთიანი ანბანური სისტემის წყალობით, რომელიც გამოყენებული იქნება ნიგერიის აქტიურ მოსახლეობაში.

- **გრძელვადიანი გამოყენება**

დროთა განმავლობაში საინფორმაციო და საკომუნიკაციო ტექნოლოგიების ექსპერტებთან თანამშრომლობით შეიძლება შემუშავდეს სამუშაო მოდელი, რომელიც საბოლოო ჯამში შესაძლოა დააპატენტონ მწარმოებლებმა, რათა მათ გამოუშვან სპეციალური კლავიატურა შესაბამისი მომხმარებლებისათვის.

[1] Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2014. *Ethnologue: Languages of the World, Seventeenth edition.* Dallas, Texas: SIL International. Online version: http://www.ethnologue.com

Linguascript: Applying Science in Language Development

Nnenna Nwosu-Nworuh
Federal University Ndufu Alike Ikwo (Nigeria)
nneabel@yahoo.com

Language development as defined by *Ethnologue*[1] with reference to the definition of Charles Ferguson (1968) as primarily dealing with three areas of concern: *graphization* – the development of a system of writing; *standardization* – the development of a norm that overrides regional and social dialects and *Modernization* –the development of the ability to translate and carry on discourse about a broad range of topics in ways characteristic of "industrialized, {…} modernized societies". More broadly, *Ethnologue* defines language development as *the result of the series of on-going planned actions that language communities take to ensure that they can effectively use their languages to achieve their social, cultural, political, economic, and spiritual goals*[2]. This research concerns the 5th indicator on evaluation of language development: "the use of the language in new media such as on web pages, in chat rooms, podcasts and MP3 downloads, and for SMS texting on mobile phones or other electronic devices".

- **Objective: What problem will this solve?**

To improve the linguistic vitality of the major Nigerian languages.

- **Theoretical framework**

Studies of LTP[3] are done in slices of hippocampus, an important organ of the brain for **learning** and **memory**. This occurs as a result of increase in synaptic strength, otherwise known as potentiation[…].

Research has proven that there are 3 functions of synapses: *inhibitive, memory and spatial*. The second function is the one to be discussed in this study to prove that the link between LTP and behavioural learning may turn illiteracy into functional literacy.

Methodology

- Applying the LTP theory to show connection between LTP and behavioural learning for target population
- Applying an existing Uniskript alphabet system to develop a model (**Linguascript**) for a language usable in Nigeria among the active population.

- **Specific objectives**

Achieving rapid functional literacy through a 'singular' alphabet system, usable by the active adult population in Nigeria.

- **Long term application**

Eventually, collaborate with ICT experts to develop it into a usable model which eventually could be patented so GSM manufacturers can produce special keyboard brands for the end users.

[1] http://www.ethnologue.com/language-development
[2] Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2014. *Ethnologue: Languages of the World, Seventeenth edition.* Dallas, Texas: SIL International. Online version: http://www.ethnologue.com
[3] Long Term Potentiation

სინტაქსზე დამყარებული მეთოდი ლექსიკური ომონიმიის მოსახსნელად სიტყვათა სპეციალური ჯგუფებისათვის

ოლგა ნევზოროვა
თათრეთის მეცნიერებათა აკადემიის გამოყენებითი სემიოტიკის
სამეცნიერო-კვლევითი ინსტიტუტი (რუსეთი)
ყაზანის ფედერალური უნივერსიტეტი (რუსეთი)
onevzoro@gmail.com
ალფია გალიევა
თათრეთის მეცნიერებათა აკადემიის გამოყენებითი სემიოტიკის
სამეცნიერო-კვლევითი ინსტიტუტი (რუსეთი)
ვლადიმირ ნევზოროვი
ყაზანის ეროვნული კვლევითი ტექნიკური უნივერსიტეტი (რუსეთი)

ტექსტებში სიტყვაფორმების განსხვავებული ყალიბების ძიებისა და მნიშვნელობის გამოხატვის შესაძლო საშუალებების აღწერასთან ერთად პოლისემიური სიტყვის მნიშვნელობის მოდელირების საკითხი კვლავაც იპყრობს თანამედროვე ენათმეცნიერთა და ბუნებრივ ენათა დამუშავების სისტემების შემქმნელთა ყურადღებას (Franz, 1996). ომონიმიის მოხსნა (Word Sense Disambiguation (WSD)) წარმოადგენს კომპიუტერული საშუალებებით კონტექსტებში სიტყვათა მნიშვნელობის იდენტიფიცირების უნარს. WSD მიიჩნევა ხელოვნური ინტელექტის სფეროს სრულფასოვან პრობლემად ანუ ამოცანად, რომლის გადაწყვეტა ისეთივე დიდ სირთულეს წარმოადგენს, როგორსაც ხელოვნური ინტელექტის ყველაზე რთული საკითხები. WSD შეიძლება განვიხილოთ როგორც საკლასიფიკაციო ამოცანა: სიტყვათა მნიშვნელობები წარმოადგენენ კლასებს, ხოლო ავტომატური კლასიფიკაციის მეთოდი გამოიყენება სიტყვის თითოეული პოვნიერების მისაკუთვნებლად ერთი ან რამდენიმე კლასისათვის, რაც დაფუძნებული იქნება კონტექსტიდან და გარეგანი ცოდნის რესურსებიდან მომდინარე ფაქტობრივ მასალაზე (Klapaftis & Manandhar, 2013). კლასტერიზაციის მიდგომას საფუძვლად უდევს ჰიპოთეზა იმის შესახებ, რომ სიტყვები ერთმანეთს ჰგვანან სემანტიკურად, თუკი ისინი გვხვდებიან ანალოგიურ დოკუმენტებში, ანალოგიურ კონტექსტუალურ ჩარჩოებში და ანალოგიურ სინტაქსურ კონტექსტში (Van de Cruys, 2010; Dorow & Widdows, 2003).

საკორპუსე მონაცემები კარგ ექსპერიმენტულ მასალას გვაწვდის ტექსტში პოვნიერი ოკაზიური ფორმების გამოსაკვლევად. არსებობს ოკაზიური ერთეულების ორი ტიპი: 1) ჩვეულებრივი სიტყვები, რომლებიც იღებენ ოკაზიურ მნიშვნელობებს, და 2) თავისთავად ოკაზიური სიტყვები. პირველი ტიპის ლექსიკურ ერთეულებს გადატანითი მნიშვნელობები აქვთ მეტაფორული და მეტონიმიური გადასვლების სახით. ამ მოვლენას ეწოდება სემანტიკური დერივაცია. მეორე ტიპის ლექსიკურ ერთეულებს აქვთ ფუძეთშეერთების ფორმა, რომელთაც სემანტიკური გადახრის მქონე ვალენტობის მოდელები ახასიათებს (Dean, 1988).

შემუშავებული ძირითადი მეთოდი გულისხმობს ბუნებრივ მოვლენათა ამსახველი პოლისემიური ერთეულების დისტრიბუციული კონტექსტუალური მოდელის კორპუსულ შესწავლას. ჩვენ შევარჩევთ კონტექსტთა ტიპურ კომპონენტებს ამ არსებითი სახელების (ტიპური პრედიკატები და მსაზღვრელები, მათთან დაკავშირებული სიტყვები, რომლებიც სხვადასხვა კლასს განეკუთვნებიან) ჩვეულებრივი და იდიომატური მნიშვნელობებისათვის. გრამატიკულად შეთანხმებული მსაზღვრელები (პრეპოზიციული ზედსართავები და მიმღეობები) და შეუთანხმებელი მსაზღვრელები (პოსტპოზიციური გენიტივები) ჩვენთვის განსაკუთრებით საინტერესოა, რადგანაც ამგვარი შესიტყვებები ხელს უწყობენ სემანტიკური გადასვლების იდენტიფიცირებასა და ტექსტში ახალი მნიშვნელობების კონსტრუირების მექანიზმების აღწერას.

ჩვენი მეთოდი ემყარება სინტაქსური დამოკიდებულების სტატისტიკას სიტყვათა შორის, რომლებიც დაჩნდებიან წინადადებაში სიტყვათა მახასიათებლების სიმრავლეთა წარმოსაქმნელად სამიზნე სიტყვის ორი მნიშვნელობის გასარჩევად (პირდაპირი და გადატანითი).

ამ მიზნით ჩვენ ვიყენებთ სპეციალიზებულ პროგრამულ სისტემას, რომელიც შევიმუშავეთ წინამდებარე მოხსენების ავტორებმა (Nevzorova & Nevzorov, 2009). ჩვენ დავნერგეთ კორპუსულ მონაცემთა ნახევრადავტომატური დამუშავება „ონტოინტეგრატორის“ სპეციალიზებული სისტემის მეშვეობით. ჩვენ შევქმენით სპეციალური პროგრამა, რომელიც აფასებს კონტექსტთა სინტაქსურ მსგავსებას პოლისემიური არსებითი სახელის თითოეული მნიშვნელობისათვის, თავს უყრის სტატისტიკურ ინფორმაციას კონტექსტთა შედგენილობის შესახებ და იღებს და სტატისტიკურად აზუსტებს მათში შემავალი პოლისემიური სიტყვების რაოდენობას.

ჩვენ შევადგინეთ სამიზნე არსებითების პირდაპირი და გადატანითი მნიშვნელობების კონტექსტური მახასიათებლების სია და დავადგინეთ მათი წონა თითოეული სამიზნე არსებითი სახელის ექსპერიმენტული ნიმუშისათვის. გამოკვლევის შედეგად ჩვენ შევადგინეთ ლექსიკური და გრამატიკული ბინარული მახასიათებლების სია, რომელიც საშუალებას იძლევა, კონტექსტურად გავარჩიოთ პოლისემიური სიტყვების მნიშვნელობა ექსპერიმენტული ნიმუშიდან. ომონიმიის მოსახსნელად ჩვენ ჩამოვაყალიბეთ ორი დამატებითი კრიტერიუმი პირდაპირი და გადატანითი მნიშვნელობების გასარჩევად, რომლებშიც გათვალისწინებულია რელევანტური ფორმალური მახასიათებლების წონა. ლექსიკური ომონიმიის მოხსნის პროცედურა შემუშავებული კრიტერიუმების გამოყენებით გამოითვლის Wdir და Wfig ფუნქციათა სიდიდეებს. თუკი Wdir (Wfi) ფუნქციის სიდიდეები აჭარბებს ზღვრულ სიდიდეებს, რომლებიც დაფიქსირებულია ექსპერიმენტებში, მაშინ სამიზნე სიტყვის ომონიმიის მოხსნა ხდება მოცემულ კონტექსტში.

ჩვენ ჩავატარეთ ექსპერიმენტები რუსული ენის პოლისემიური სიტყვების ლექსიკური ომონიმიის მოსახსნელად და მასში გამოვიყენეთ ჩვენ მიერ შექმნილი მეთოდი ლექსიკური ომონიმიის მოხსნისა ბუნებრივი მოვლენების აღმნიშვნელი არსებით სახელებში. საკორპუსე მასალა შეირჩა რუსული ენის ეროვნული კორპუსიდან.

ბუნებრივი მოვლენების აღმნიშვნელი პოლისემიური სიტყვების შემცველი კონტექსტების აღსაწერად და პირდაპირი და გადატანითი მნიშვნელობების გასარჩევად რელევანტურ მახასიათებლებზე დაყრდნობით შემუშავებული ფორმალური კრიტერიუმების სისტემა მთლიანობაში საკმაოდ ზუსტია ომონიმიის მოხსნის თვალსაზრისით.

ჩვენ გამოვავლინეთ, რომ სემანტიკურ მახასიათებლებს ლომის წილი უდევთ ომონიმიის მოხსნაში. სემანტიკური მახასიათებლების გასარჩევად საჭიროა სათანადო სემანტიკური რესურ-

სები. პირველი ნაბიჯები ამგვარი სემანტიკური რესურსების შესაქმნელად წინამდებარე გამოკვლევაში გადაიდგა, სახელდობრ, შევადგინეთ ურთიერთდაკავშირებული სიტყვების სპეციალური სიტყვანები სხვადასხვა მეტყველების ნაწილისათვის, რომელიც გამოყენებულ იქნებოდა ომონიმიის მოხსნის ალგორითმებში.

კონტექსტის მოცულობის განსაზღვრა, რაც საკმარისი იქნებოდა ომონიმიის მოსახსნელად, რთულ პრობლემას წარმოადგენს ავტომატური მეთოდებისათვის. ჩვენს ნაშრომში გამოყენებული მეთოდები მორგებულია კონტექსტის ფიქსირებულ მოცულობაზე ანუ სრულ წინადადებაზე. ბევრ შემთხვევაში ომონიმიის მოსახსნელად საჭიროა უფრო ფართო კონტექსტი: აბზაცი ან იქნებ მთელი ტექსტიც კი, რაც ძნელად თუ მოხერხდება კვლევის ამ ეტაპზე. ომონიმიის მოხსნის მეორე პრინციპულ სირთულეს წარმოადგენს ის, რომ შეიძლება დამატებითი ცოდნა ჩავრთოთ იმ დასკვნების ჩამოსაყალიბებლად, რომლებიც აუცილებელია ომონიმიის მოსახსნელად კონტექსტში.

The Syntax-based Method of Resolving Lexical Ambiguity for Special Groups of Words

Olga Nevzorova
Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences
Kazan Federal University (Russia)
onevzoro@gmail.com
Alfiya Galieva
Kazan Federal University, Kazan, Russia (Russia)
Vladimir Nevzorov
Kazan National Research Technical University (Russia)

The issue of the polysemic word sense modelling, together with searching for distinctive patterns of word use in texts and describing possible ways of making sense, continues to attract the attention of modern linguists and NLP systems developers (Franz, 1996). Word Sense Disambiguation (WSD) is the ability to identify the meaning of words in contexts in a computational manner. WSD is considered an artificial intelligence-complete problem, that is, a task whose solution is at least as hard as the most difficult issues of artificial intelligence. WSD can be viewed as a classification task: word senses are classes, and an automatic classification method is used to assign each occurrence of a word to one or more classes based on the evidence from the context and from external knowledge sources (Klapaftis and Manandhar, 2013). The underlying hypothesis of clustering approach is that words are semantically

similar if they appear in similar documents, within similar context windows, or in similar syntactic contexts (Van de Cruys, 2010; Dorow & Widdows, 2003).

Corpus data give us good experimental material for exploring the semantics of occasional forms in the text. There are two types of occasional units: 1) usual words that obtain occasional meanings and 2) occasional words as such. The first type lexical units have figurative senses in the forms of metaphoric or metonymic shifts. This phenomenon is called semantic derivation. The second type lexical units have the form of combinations of stems with semantic *deviant valency* patterns (Dean, 1988).

The main method developed implies corpus study of the distributive contextual model of polysemic units denoting natural phenomena. We select typical components of contexts for literal and figurative senses of these nouns (typical predicates and modifiers, different classes of associated words). Grammatically coordinated modifiers (adjectives and participles as premodifiers) and uncoordinated modifiers (postmodifiers in the Genitive case) are of particular interest for us since such collocations facilitate identification of semantic shifts and description of mechanisms of new senses construction in the text.

Our method is based on syntactic dependency statistics between words that occur in a sentence to produce sets of word feature vectors for resolving two sense (direct or figurative) of a target word.

For this purpose we use the specialized program system developed by the authors of this paper (Nevzorova & Nevzorov, 2009). We implemented semi-automatic processing of corpus data by means of specialized tools of “OntoIntegrator” program system. We created specialized software that assesses the syntactic resemblance of contexts for each sense of a polysemic noun, collects statistical information on contexts composition, and obtains and statistically evaluates collocations containing polysemic words.

We made up a list of contextual characteristics for direct and figurative senses of target nouns and evaluated their weights for the experimental sample for each target noun. As a result of the study we compiled a list of lexical and grammatical binary characteristics that enable contextual word sense disambiguation of polysemic words from the experimental sample. For disambiguation we formulated two additive criteria for discerning direct and figurative senses that take into account the weights of relevant formal features. The procedure of lexical disambiguation using the developed criteria computes values of function *Wdir* and *Wfig.* If values of function *Wdir (Wfig)* exceed the threshold values fixed in the experiments then the target word is disambiguated in a given context.

We carried out experiments on lexical disambiguation of polysemic words in Russian using our method of lexical disambiguation for nouns denoting natural phenomena. Corpus data have been selected from the National Corpus of the Russian Language.

The developed system of formal criteria for describing contexts containing polysemic words that denote natural phenomena and for disambiguating literal and figurative senses on the basis of relevant features, in general give sufficient accuracy in disambiguation.

We revealed that semantic characteristics made the main contribution to disambiguation. To discern semantic features we need appropriate semantic resources. The first steps towards creating such semantic resources were made in this study, namely specialized vocabularies of associated words for different parts of speech were compiled to be used in word-sense disambiguation algorithms.

Defining the context size, sufficient for disambiguation, is a difficult problem for automatic methods. The methods used in our work are attached to a fixed context size – that of the full sentence. In

many cases, disambiguation requires a larger context – the paragraph or maybe even the whole text, which was hardly solvable at this stage of work. Another principal difficulty in disambiguation is the fact that additional knowledge may be involved for building inferences that are necessary for disambiguation in a context.

References:

Klapaftis, Ioannis P. & Manandhar, Suresh (2013). *Evaluating Word Sense Induction and Disambiguation Methods*. Language Resources and Evaluation. 47: 579-605.

Dorow, B., & Widdows, D. (2003). Discovering corpus-specific word senses. In *Proceedings of the 10th conference of the European chapter of the ACL* (pp. 79–82). Budapest, Hungary: ACL.

Dean, P. (1988) Polysemy and cognition. *Lingua* 75: 325-361.

Nevzorova, O., Nevzorov, V. (2009) The Development Support System "OntoIntegrator" for Linguistic Applications // International Book Series "Information Science and ComputingG". Number 13. Intelligent Information and Engineering Systems. Supplement to the International Journal "Information Technologies & Knowledge". Volume 3. ITHEA, Rzeszow-Sofia: 78-84.

Franz, Alexander. (1996). *Automatic Ambiguity Resolution in Natural Language Processing: An Empirical Approach.* Berlin: Springer-Verlag.

Van De Cruys, T. (2010). *Mining for Meaning – the Extraction of Lexico-Semantic Knowledge from Text*. PhD Thesis. University of Groningen, The Netherlands: 12–18.

მეტაფორული შესიტყვებების მიკროსემანტიკური მოდელირების პრობლემები (ზმნური მეტაფორული შესიტყვებების მასალაზე დაყრდნობით)

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

ჩვენი მოხსენება ეძღვნება ფრანგულ და ქართულ ენებში არსებული „აზროვნების“ აღმნიშვნელი ისეთი ზმნური მეტაფორული შესიტყვებების მიკროსემანტიკური მოდელირების პრობლემებს, როგორებიცაა : Une idée traverse [l'esprit], les pensées assiègent qqn., chaser les pensées ... (ფრანგ.) აზრმა გაუელვა (თავში), ფიქრმა შეიპყრო, ... აზრები განდევნა (ქართ.) – რაც გულისხმობს მათი თავისებურებების განხილვას სემიოტიკურ და სემანტიკურ ჭრილში.

ამ ტიპის შესიტყვებები შეიძლება ჩაითვალოს ფრაზეოლოგიურ ერთეულად (ფე), იმ შემთხვევაში თუ ამ ტერმინს ფართო გაგებით გამოვიყენებთ, როგორც ეს ესმოდათ შ. ბალის, ვ. თელიას და ა. შ., მაგრამ ასეთ მიდგომას მივყავართ ფრაზეოლოგიის, როგორც ნებისმიერი ფიქსირებული ლექსიკის შემსწავლელი დისციპლინის განსაზღვრებამდე, რომლის საზღვრებიც „ზოგად სინტაქსურ კომბინატორიკამდე“ (Мокиенко, 2001) ფართოვდება. ჩვენ არ ვიზიარებთ ამ აზრს, იმის გამო, რომ იგი კვლევის ჩვენ მიერ არჩეული შესიტყვებების სპეციფიკას ვერ წარმოაჩენს და ვემხრობით ი. ანიჩკოვის პოზიციას, რომელიც წერს, რომ „ყველა მყარი შესიტყვება ვერ იქნება ფრაზეოლოგიზმი, მაგრამ ყველა ფრაზეოლოგიზმი მყარი შესიტყვებაა“ (Аничков, 1964).

მეორე მხრივ, მეტაფორულ შესიტყვებაში შემავალი სრული მნიშვნელობის მქონე სიტყვის (სემანტიკური საყრდენის) გვერდით ხმარებული, ირიბი ნომინაციის გზით და სემანტიკური ტრანსპოზიციის შედეგად მიღებული, ავტონომიურ ნომინაციას უნარმოკლებული სატელიტი კომპონენტი ფუნქციურად დამხმარე სიტყვას უტოლდება და ამ ნომინაციურ ერთეულებს ანალიტიკურ შესიტყვებებს ამსგავსებს. ამ მოსაზრებას ლექსიკური ანალიტიზმის მკვლევრის, ნ. დიმიტრიევას შემდეგი განმარტებაც ამყარებს: „ერთი სრული მნიშვნელობის მქონე და მეორე, დამხმარე სიტყვასთან გათანაბრებული, ავტონომიურ ნომინაციას უნარმოკლებული წევრისაგან შემდგარი მყარი შესიტყვება ანალიტიკური შესიტყვებაა“ (Дмитриева, 1971). ასეთ წარმონაქმნს ე. ეროფეევა ანალიტიკურ იდიომებს უწოდებს (Ерофеева, 1990).

სემიოტიკურ ასპექტში ორი ელემენტისაგან (სიტყვისაგან) შემდგარ ამ ენობრივ საშუალებებს ერთ სემიოტიკურ ნიშნად განვიხილავთ, იქიდან გამომდინარე, რომ ისინი ერთ რეფერენტზე დაიყვანება. მაგალითად: les pensées assiègent qqn., ანალოგიურადაა ქართულში: ფიქრმა შეიპყრო – აღნიშნავს ფიქრის პროცესს, რომელიც დახასიათებულია, როგორც შემაწუხებელი, ანუ დაიყვანება შემდეგ ლოგიკურ ინფორმაციაზე: „ფიქრობს“, ხოლო ის, რომ ეს ფიქრები შემაწუხებელია წარმოდგენილია მეტაფორული სახელდების საშუალებით უარყოფით-შეფასებითი პრესუპოზიციური ხასიათის მქონე კონოტატური ინფორმაციით. ამ ტიპის შესიტყვების ერთრეფერენტიანობა საშუალებას იძლევა მისი სემანტიკური შემადგენლობა განვიხილოთ ერთი სიტყვით აღნიშნული ერთი მნიშვნელობის მსგავსად სემურ-კომპონენტურ ანალიზზე დაყრდნობით. ეს მიდგომა ასევე იმთავითვე მოხსნის თითოეული შემადგენელი სიტყვის გრამატიკული ტაგირების აუცილებლობას და ეს უკანასკნელი, ვფიქრობთ, შინაარსიდან (მნიშვნელობიდან) გამომდინარე უნდა განხორციელდეს;

სემანტიკურ ასპექტში, ენის კორპუსისთვის ამ ტიპის შესიტყვებების სემანტიკური მოდელირება შესაძლებელია ჩვეულებრივი ცხრილის სახით, რომელიც **Excel** პროგრამაში ფუნქციონირებს. ამგვარი მნიშვნელობის ამსახველი ცხრილი ორი მაკროკომპონენტისაგან შედგება: დენოტატურისა და კონოტატურისაგან. დენოტატურ პარამეტრებში შეყვანილი გვაქვს შემდეგი ნიშნები: არქისემა და კატეგორიალური ნიშანი სემანტიკური ველის ჰიპერონიმისათვის და დიფერენციალური ნიშანი (ნიშნები) ველის ჩვეულებრივი წევრისათვის. ასევე ამ ნაწილში წარმოდგენილია მეტაფორის შიდაფორმა, როგორც დენოტატის ლოგიკურად თუ ემოციურად საგანგებოდ გამოკვეთილი სემა, თუ იგი სახეზეა.

კონოტატურ მაკროკომპონენტში სემების თანმიმდევრობა ჩვენ მიერ ჩატარებული კვლევიდან გამომდინარე ავაგეთ, რის არგუმენტაციასაც მოხსენებაში წარმოვადგენთ. ეს თანმიმდევ-

რობა შემდეგნაირია: მეტაფორული ხატი, შეფასება, ემოტივი, სტილისტური მარკერი. აგრეთვე შესაძლებელია კულტურულ-ეროვნული მარკერის აღნიშვნა, თუკი იგი სახეზეა;

ამრიგად, ჩვენ მიერ შედგენილი სავარაუდო ცხრილი შემდეგნაირად გამოიყურება:

სახელდებითი ერთეული	დენოტატური მაკროკომპონენტი				კონოტატური მაკროკომპონენტი			
	არქისემა	კატეგ. ნიშანი	დიფერენც. ნიშანი	შიდა ფორმა	ხატი	შეფასება	ემოტივი	სტილი

მოცემულ ცხრილში პარამეტრების რაოდენობა და ნაირსახეობა როგორც მაკროკომპონენტურ დონეზე, ასევე მიკროკომპონენტურ დონეზე შეზღუდული არ არის. რაც მეტი იქნება პარამეტრების რაოდენობა, მით უფრო ამომწურავად იქნება აღწერილი სახელდებითი ერთეულით მოწოდებული ინფორმაცია.

The Challenges of Micro-Semantic Modeling of Metaphoric Collocations (Based on the Data of Verbal Metaphoric Collocations)

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

The paper provides a deeper view into the challenges of micro-semantic modeling of metaphoric collocations, occurring both in Georgian and French and denoting concept of "thinking", namely: Une idée traverse [l'esprit], les pensées assiègent qqn., chaser les pensées . . . (French) – An idea ran to his head, was lost in thoughts, got rid of thoughts (English). We will discuss their peculiarities within the frameworks of semiotics and semantics.

Linguistic items consisting of two elements (words) are considered as a single semiotic sign as far as they are limited to a single meaning. For example, the analogue of les pensées assiègent qqn., in Georgian "was thinking" denotes only process of thinking, while the negative connotation of the phrase is revealed through the metaphoric collocation. The collocation with one meaning can be treated similarly as word-based componential analyses. The approach described above automatically dissolves the need of grammatically examining of each component and leaves it to concept (meaning) connotation analyses.

Semantic modeling of the same kind of collocations can be done through Excel sheets consisting of two macro-components: denotative and connotative. The denotative parameters include an arhiseme and a categorical feature for semantic hyperonyms and differentiating one for an ordinary component. An internal form of a metaphor might also be presented in the same part as a seme of a logically or emotionally salient denotative.

In the connotative macro-component, the sequence of semes was based on the research (argumentation will be provided in the talk). The sequence is the following: metamorphic image; evaluation, emotive, stylistic marker. It is also possible to integrate a cultural maker in case it exists.

In the list, the number and variety of the parameters both at the macro-componential as well as the micro-component levels are not limited. The more parameters are included, the more exhaustive description will be provided on what is presented by a nominating item.

ქდკ-ს ავტომატური ანოტირების შედეგების ანალიზი (იმერული, აჭარული, კახური დიალექტების მასალის მიხედვით)

ნარგიზა სურმავა, ციცინო კვანტალიანი, მარინა კიკონიშვილი, მარინა ბერიძე
არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველო)
nargizasurmava@yahoo.com; tsitsino.nino.kvantaliani@gmail.com

ქდკ-ს ავტომატური ანოტირების მთავარი ინსტრუმენტია სისტემა Geo Trans-ი, რომელიც შექმნილია ქართული სალიტერატურო ენის მორფოლოგიური ანალიზისთვის.

მორფოლოგიური ავტომატური ანალიზის ტესტირებისა და საბოლოო გრამატიკული ანოტირების პროცესი წარიმართება სპეციალური ინსტრუმენტით – ანოტირების რედაქტორით, რომელშიც წინასწარაა ატვირთული გაანალიზებული სიტყვების სიები – არაომონიმური და ომონიმური. ამ ეტაპზე მიმდინარეობს რამდენიმე დიალექტის მასალის ავტომატური ანალიზის ტესტირება (ქართლური, იმერული, აჭარული, კახური) და კონტექსტებისთვის მარკერების ავტომატურად მიწერა. პარალელურად მუშავდება რეკომენდაციები ანალიზატორის სრულყოფისა და მისი დიალექტური მოდულის შექმნისათვის.

ტესტირებისა და ანოტირების პროცესში მუშაობა რამდენიმე მიმართულებით წარიმართება:

- ლემატიზაციის შედეგების ტესტირება;
- სისტემათა შორის (სალიტერატურო ენასა და დიალექტს ან სხვადასხვა დიალექტს შორის) ომონიმიის პროგნოზირება, გამოვლენა, მოხსნა;
- ტრანსლაციის შედეგად გაჩენილი ომონიმიის პროგნოზირება, გამოვლენა, მოხსნა;
- სალიტერატურო ანალიზატორში გამორჩენილი სიტყვების ან ანალიზში დაშვებული შეცდომების გამოვლენა;

- ქდკ-ს კონცეფციის შესაბამისად ანალიზის კორექტირება.

მოკლედ აღვწერთ იმ ტიპურ პრობლემებს, რომელთა დაძლევაც უხდება მკვლევართა ჯგუფს ამ სამუშაოს შესრულებისას.

ყველაზე მეტი შეცდომა **ლემატიზაციის** საფეხურზე შეინიშნება. შეცდომები სხვადასხვა მიზეზითაა გამოწვეული: სალიტერატურო მორფოლოგიურ ლექსიკონში გამორჩენილია შესაბამისი ლექსიკური ერთეული; სალიტერატურო ანალიზატორის ლემატიზაციის კონცეფცია არ თანხვდება დიალექტურისას, სალიტერატურო ენის ანალიზატორი შეცდომით ახდენს დიალექტური ფორმის იდენტიფიცირებას სალიტერატურო ფორმასთან, მაგ.: გეიმსო – გეიმი – N: Sg, Nom, O (Geo Trans) და გეიმსო – იმსება(=ივსება)- V:Sg, Aor, 3, მაგრამ არის შემთხვევები, როცა სალიტერატურო და დიალექტურ ფორმებს შორის ომონიმია სისტემურად ვლინდება. ასეთ შემთხვევაში გამოიყოფა "პროგნოზირებადი" ომონიმიის ჯგუფები და ომონიმია ან ხელით იხსნება, ან მუშავდება ომონიმიის კონტექსტური მოხსნის ავტომატური მექანიზმი. სისტემათაშორისი განსხვავების ამსახველი ომონიმიის კარგი მაგალითია სუბიექტური პირველი და მეორე პირის ზმნურ ფორმათა დამთხვევა, გამოწვეული დიალექტებში უ- პრეფიქსის წინა პოზიციაში სუბიექტური პირველი პირის ვ- ნიშნის დაკარგვით. Geo Trans-ი, ბუნებრივია, მათ სალიტერატურო ვერსიას აანალიზებს:

Geo Trans – უკეთებ/ V:Sg, Prs, 2;

ქდკ – უკეთებ / V: Sg, Prs, 1;

Geo Trans – უკეთებთ / V: Pl, Prs, 2

ქდკ – უკეთებთ / V: Pl, Prs, 1

სისტემათაშორისი ომონიმიის მაგალითია აგრეთვე ფუძის სახით წარმოდგენილი თანხმოვანფუძიანი არსებითები (კაც). ამ ტიპის სახელობითისა (ხევსურულში) და მიცემითის წარმოება საკმაოდ გავრცელებულია დიალექტებში. Geo Trans-ი მათ "აპოზიციურ" – სუბსტანტიურ ატრიბუტულ მსაზღვრელად წარმოადგენს ვითარებითსა და მიცემით ბრუნვებში („კაც ამფიბიას, კაც ამფიბიად" ანალოგიით). მოხსენებაში განხილული იქნება პროგნოზირებადი ომონიმიის სხვა შემთხვევებიც.

არის საკითხები, რომელთა სისტემური შესწავლის გარეშე გრამატიკული ომონიმიის დაძლევის პროცესში პრობლემები წარმოიქმნება. განსაკუთრებით მნიშვნელოვანია ამ მხრივ ტრანსლაციის შედეგად გაჩენილი გრამატიკული ომონიმია, როცა ერთი ფორმა შეიძლება სხვადასხვა მეტყველების ნაწილად კვალიფიცირდეს.

მორფოლოგიური მარკერის შერჩევისას ზოგჯერ წამოიჭრება ნორმატიულ ორთოგრაფიასთან დაკავშირებული სირთულე (მაგალითად: კაი ძალი იარა – „კაი ძალი" ერთად უნდა დავწეროთ თუ ცალკე), იგივე პრობლემა დგება ტოპონიმებთან დაკავშირებით, რომლებიც ჩამწერებთან ხან შერწყმულად არის წარმოდგენილი და ხან გაყრილად, (გვიხდება კორექტივის შეტანა ტექსტებშივე): კაკლიძირი, ხოჯიყანა, მიღმაყანა...

ზოგჯერ ქდკ-ს ლინგვისტური ანალიზის კონცეფცია მოითხოვს Geo Trans-ის მარკირების სისტემაში ცვლილების შეტანას. მაგალითად, Geo Trans-ის ავტორები სალიტერატურო ენისათვის ზმნის ლემად მხოლოდ მყოფადის ფორმას (III პ. მხოლ. რიცხვი) წარმოადგენენ, ხოლო ჩვენ, დიალექტურ დონეზე, ლემად ვიღებთ ხან აწმყოს (უპირატესად უზმნისწინო ზმნური ფორმებისთვის) და ხან მყოფადს (ზმნისწინიანი ზმნური ფორმებისთვის). შესაბამისად შესწორ-

და: ეშველება/ეშველება (Geo Trans-ით „უშველის“), ეჩქარებოდა/ეჩქარება (და არა „აეჩქარა“), ემსახურებოდა/ემსახურება (და არა „მოემსახურება“); ვიპოვი/იპოვის (და არა „პოულობს“)…

მასდარი და **მიმღეობა** Geo Trans-ის სისტემაში პირველი იერარქიის, გრამატიკული ჯგუფის მარკერებია, ჩვენ ისინი გადავიტანეთ მეორე იერარქიის (ფორმაწარმოების/სიტყვაწარმოების დონე) მახასიათებლებად, ხოლო სახელზმნებს გრამატიკულ კლასიფიკატორად მივანიჭეთ: „არსებითი სახელი“ და „ზედსართავი სახელი“.

მოხსენებაში განხილული იქნება კონკრეტული საკითხები, რომლებიც წამოიჭრა იმერული, აჭარული კახური... დიალექტების მასალის ავტომატური ანალიზის ტესტირებისა და ომონიმიის მოხსნის პროცესში.

Analysis of the Results of the Automated Annotation of GDC (Based on the Data of Imeretian, Acharan, Kakhetian Dialects)

Nargiza Surmava, Tsitsino Kvantaliani, Marina Kikonishvili, Marine Beridze
Arn. Chikobava Institute of Linguistics (Georgia)
nargizasurmava@yahoo.com; tsitsino.nino.kvantaliani@gmail.com

The principal automated annotating tool in GDC is the system Geo Trans, designed for morphological analysis of Standard Georgian.

The processes of testing of automated morphological analysis and of the final grammatical annotation will be carried out by means of a special tool, an annotation editor, in which lists of analyzed words (non-ambiguous and ambiguous) are previously uploaded. Currently, the automated testing of data of several dialects (Kartlian, Imeretian, Acharan, Kakhetian) and assignment of markers to contexts are under way. Meanwhile, recommendations are processed for sophisticating the parser and for developing of its dialect module.

In the process of testing and annotation, the activities will proceed in several directions:

- testing of lemmatization results;
- forecasting, identification, elimination of ambiguity between systems (between the standard and a dialect or between various dialects);
- forecasting, identification, elimination of ambiguity, emerged as a result of translation;
- identification of missing words in the standard language parser or identification of mistakes in analysis;
- correction of analysis in accordance with the GDC concept.

The paper will provide a brief description of the typical problems, faced and overcome by the group of researchers while carrying out these activities.

Most of the mistakes are observed at the stage of lemmatization. Mistakes are due to various causes: a corresponding lexical item is missing in the standard morphological dictionary; the lemmatization concept of the standard parser does not correspond to that of the dialect one; the standard language parser erroneously identifies a dialect form with a standard one; e.g. geimso – geimi 'game' – N: Sg, Nom, O (Geo Trans) and geimso – imseba(=ivseba) 'It is filled' – V:Sg, Aor, 3; however, there are cases when ambiguity is systematically attested between a standard and a dialect form. In such cases, groups of "predictable" ambiguity are identified, and ambiguity is either removed manually or an automated mechanism of the contextual elimination of ambiguity is developed. A good example of the ambiguity, reflecting inter-systemic differences, is the coincidence of S1 and S2 verb forms, caused in dialects by the loss of the S1 marker v- in the position preceding the prefix u-. Naturally enough, Geo Trans analyses its standard version:

Geo Trans – uk'eteb / V:Sg, Prs, 2;
DGC – uk'eteb / V: Sg, Prs, 1;
Geo Trans – uk'etebt / V: Pl, Prs, 2
DGC – uk'etebt / V: Pl, Prs, 1

Inter-systemic ambiguity is also well illustrated by means of the consonant-final nouns occurring as stems (k'ac 'man'). The formation of such a nominative (Khevsurian) and a dative is widespread in dialects. Geo Trans presents them as "an appositional" – adnominal attributive modifier in the adverbial and dative cases (analogy of "k'ac ampibias, kac' ampibiad"), the paper will analyze other cases of predictable ambiguity.

There are issues without the systemic study of which the process of overcoming of grammatical ambiguity will be problematic. Particularly significant is grammatical ambiguity, emerged as a result of translation, when one form may be qualified as various parts of speech.

In the selection of a morphological marker, complications with normative spelling sometimes occur (e.g. k'ai ʒali iara – when to spell "k'ai ʒali" solidly or separately). The same problem arises with place-names, being spelt by recorders sometimes solidly and sometimes separately (we have to make corrections in texts); k'akliʒiri, xojiq'ana, miɣmaq'ana…

Sometimes the concept of the linguistic analysis of GDC requires changes in the markup system of Geo Trans. For instance, authors of Geo Trans present a verb lemma for the standard language only as a future form (3rd person, sg.), while we, at a dialect level, take sometimes present (predominantly preverbless verb forms) and sometimes future (for verb forms with preverbs); hence, its was corrected: ešveleba/ešveleba (Geo Trans – "ušvelis"), ečkareboda/ečkareba (and not "aečkara"), emsaxureboda/ emsaxureba (and not "moemsaxureba"), vip'ovi/ip'ovis (and not "p'oulobs")…

In the system of Geo Trans, a masdar and an adjective are grammatical group markers belonging to the first hierarchy; we moved them to the second hierarchy (inflection/word-formation level), and verbal nouns were classified as "Noun" and "Adjective."

The paper discusses specific issues evolving in the process of automated analysis testing of the dialect data of Imeretian, Acharan, Kakhetian, etc. and of elimination of ambiguity.

ქართული ხალხური ცხოველთა ზღაპრების ელექტრონული კატალოგი

მარინე ტურაშვილი
თსუ შოთა რუსთაველის ქართული ლიტერატურის ინსტიტუტი (საქართველო)
mariturashvili@yahoo.com

საარქივო ჩანაწერების ანალიზი ცხადყოფს, რომ ქართული ხალხური ზღაპრები ტექსტების საერთო რაოდენობის 10%-ს შეადგენს, რაც ამ ჟანრის მნიშვნელობის უტყუარი დასტურია. ცხოველთა ზღაპრები კი საზღაპრო ეპოსის ერთ-ერთი მნიშვნელოვანი ჯგუფია, რომელსაც თავისებური ფორმა და შინაარსი აქვს. მათი ფიქსაცია, ისევე როგორც ზეპირსიტყვიერების სხვა ნიმუშებისა, აქტიურად XX საუკუნის მეორე ნახევრიდან იწყება, თუმცა ჯერ კიდევ XIX საუკუნეში საკმაოდ ღირებული ტექსტებია ჩაწერილი.

საქართველოში ზღაპრების კატალოგიზაციის რამდენიმე ნაშრომი არსებობს. ცხოველთა ზღაპრის შესწავლის საქმეში ამ მხრივ განსაკუთრებით უნდა აღინიშნოს **ელ. ვირსალაძის** დამსახურება (1960 წ.). 1970 წლიდან ქართული ზღაპრის სიუჟეტთა საძიებელზე მუშაობდა **თ. ქურდოვანიძე,** ხოლო ცხოველთა ეპოსის კლასიფიკაციასა და სიუჟეტურ საძიებელზე იმუშავა **რ. ჩოლოყაშვილმა** (2004 წ.).

ზღაპრების კატალოგიზაციის ყველაზე სრულყოფილ საერთაშორისო ნაშრომად მიიჩნევა 2004 წელს პროფ. **ჰანს იორგ-უთერის** მიერ გამოქვეყნებული ნაშრომი („The Types of International Folktales, A Classification and Bibliography Based on the System of Antti Aarne and Stith Thompson, by Hans-Iorg Uther, Helsinki), რომელიც ტიპოლოგიური და კომპარატიული ანალიზის საშუალებას არ იძლევა. აღნიშნული ნაშრომი ეყრდნობა არათანაბარ რეპრეზენტატულ მასალებს, შესაბამისად, შეუძლებელია ეროვნული რეპერტუარისთვის დამახასიათებელი ნიშნების დადგენა. ამასთან, ტიპები აღწერილია ორი–სამი წინადადებით, რომელშიც ეროვნულთან ერთად გაუთვალისწინებელია რელევანტურობის განმსაზღვრელი ნიშნებიც.

2010-2012 წლებში შოთა რუსთაველის ეროვნული სამეცნიერო ფონდის ხელშეწყობით განხორციელდა პროექტი **„ქართული ხალხური პროზის კომპარატივისტული ანალიზის ელექტრონული პლატფორმა“,** რომლის ფარგლებში შეიქმნა ქართული ხალხური კომპარატიული ანალიზის ელექტრონული პლატფორმა (იხ. www.folktreasury.ge/Folklore/), რომელშიც დამუშავდა შოთა რუსთაველის ქართული ლიტერატურის ინსტიტუტის ფოლკლორის არქივში დაცული გამოუქვეყნებელი ზღაპრები. აღნიშნული პლატფორმის ფარგლებში შექმნილი სიუჟეტური ტიპის რეზიუმირების ფლექსიური სქემა შესაძლებელს ხდის მოცემული ტიპის ინვარიანტის ცალკეული ნიმუშის მიხედვით შეკუმშვას ან განვრცობას, ითვალისწინებს, ეროვნული რეპერტუარის როგორც ტიპოლოგიური ნიშნების დადგენას, ისე მისგან გადახვევის დაფიქსირებას.

როგორც მოსალონელი იყო, ქართულ ზღაპრულ რეპერტუარში საერთაშორისო კატალოგში გამოყოფილი ცხოველთა ყველა ბლოკის ტიპები ფიქსირდება, ესენია: **გარეული ცხოველები**

(1-99); გარეული და შინაური ცხოველები (100-149); გარეული ცხოველები და ადამიანები (150-199); შინაური ცხოველები (100-219) და **სხვა ცხოველები და საგნები (220-299).**

საარქივო ტექსტების ანალიზის საფუძველზე დადგინდა, რომ ქართულ ფოლკლორულ რეპერტუარში რაოდენობრივად უფრო მეტია გარეულ ცხოველთა ჯგუფის შიგნით განთავსებული ზღაპრები, რომლებიც ზემოთ დასახელებული ზღაპრის საერთაშირისო კატალოგში მოიცავს **1-99 ინდექსს.** ეს ჯგუფი, თავის მხრივ, ორ ქვეჯგუფად იყოფა: **ჭკვიანი მელია (1-69)** და **სხვა გარეული ცხოველები (70-99).**

ქართული ხალხური პროზის კომპარატიული ანალიზის ელექტრონულმა პლატფორმამ საშუალება მოგვცა ასევე დაგვედგინა, რომ ქართული ზღაპრული რეპერტუარისათვის ყველაზე რელევანტური ტიპია **15 („ნათლიის“ მიერ საჭმლის ქურდობა“)**.

საარქივო ტექსტებზე მუშაობის პროცესში განსაკუთრებულ ყურადღებას ტიპთა კომბინაციები იპყრობს. ანალიზმა ცხადყო, რომ ქართული ხალხური ცხოველთა ზღაპრები კომბინაციაშია მხოლოდ თავისივე ბლოკის სხვადასხვა ან ანეკდოტთა ჯგუფის ტიპებთან (მაგ., **2015 – “თხას არ უნდა წავიდეს სახლში”**), თუმცა ერთადერთ შემთხვევაში ფიქსირდება ცხოველთა ზღაპრის ტიპების ჯადოსნური ზღაპრის ტიპებთან კომბინაცია: **248+9+219E**+513+516.**

ამრიგად, ქართული ხალხური პროზის კომპარატიული ანალიზის ელექტრონულ პლატფორმაში დამუშავებული ცხოველთა ზღაპრების საარქივო ტექსტების ანალიზი საინტერესო და მრავალფეროვან სურათს გვიჩვენებს. დადგინდა, რომ ქართულ ფოლკლორულ რეპერტუარში, ისევე როგორც საერთაშორისოში, უფრო გავრცელებულია ერთი ტიპის ზღაპრები, ვიდრე – კომბინირებული, რაც ცხოველთა ზღაპრების მარტივი კომპოზიციით აიხსნება. ასევე საინტერესო სურათს იძლევა ქართულ ფოლკლორულ რეპერტუარში დადგენილი ცხოველთა ზღაპრების ტიპთა კომბინაციები და რელევანტუროზა. აქვე აღვნიშნავთ იმასაც, რომ ქართულ ხალხურ ცხოველთა ზღაპრებს ეროვნული რეპერტუარისთვის დამახასიათებელი თავისებურებებიც მრავლად აქვთ შეძენილი, რისი დადგენის საშუალებასაც სწორედ აღნიშნული ელექტრონული პროდუქტი იძლევა.

Electronic Catalogue of the Georgian Animal Folktales

Marine Turashvili
Shota Rustaveli Institute of Georgian Literature, TSU (Georgia)
mariturashvili@yahoo.com

A folktale is an important genre of folklore. Analyses of the archive recordings shows that animal folktales represent 10% of the texts of the folktales. Animal folktale is an important type of the genre, that has its characteristic forms and content. Recording of animal folktales, as well as of other types and genres of folklore, has intensively started from the second half of the past century, though rather valuable texts were recorded in the 19th century.

There are some works on cataloguing of folktales in Georgia. Professor **E. Virsaladze** composed the catalogue of animal folktales in 1960. Since 1970, **T. Kurdovanidze** has worked on the index of Georgian folktales**. R. Cholokashvili** also worked on the classification and the type index of animal folktales (2004).

The book on cataloguing of folktales by prof. **Hans-Jorg Uther,** published in 2004, is considered to be perfect international work, though it does not provide an opportunity to make comparative analysis of folktales. The mentioned book is not based on the equal representative materials; hence, it is impossible to establish national characteristics, types are described by two, three sentences, not taking into consideration relevance features together with the national characteristics.

In 2010-2012, by the financial support of Rustaveli National Scientific Foundation, we developed the project of **"The Electronic Platform of the Comparative Analysis of the Georgian Folk Prose"**, within the framework of which the electronic platform of the comparative analysis was created (see www.folktreasury.ge/Folklore/), by means of which we processed unpublished folktales preserved in the Folklore Archive of Shota Rustaveli Institute of Georgian Literature. The above mentioned platform makes it possible to extend or to contract the samples of invariants of the given type, to establish national characteristics and to detect deviant types.

As we expected, the Georgian repertoire contains all types of the animal tales of the International Catalogue, these are: **Wild Animals (1-99), Wild Animals and Domestic Animals (100-149), Wild Animals and Humans (150-199), Domestic Animals (200-219), other Animals and Objects (220-299).**

Analyses of the texts revealed that there are much more animal tales of the index 1-99 (Wild Animals). This group is divided into two subgroups: the Clever Fox (1-69) and Other Wild Animals (70-99).

The Electronic platform revealed the most relevant type of the Georgian animal tale – type **15 ("The Theft of Food by Playing "Godfather").**

During working on the archive, our attention was attracted by the combinations of types. The Georgian animal tales are in combination with the tales of different type of their blocs, or with the anecdote type ones (for example: **2015 – "Goat Doesn't Want to Go Home"**) though we have an example with tales of magic: **248+9+219E**+513+516.**

This electronic platform of the comparative analysis of the Georgian folk prose reveals interesting and diverse examples. It was established that in the Georgian folk repertoire as well as internationally there are widespread one-type animal tales rather than combined ones which is explained by their simple composition. There must be noted combinations and relevance of the types of animal tales, as well as national characteristics that were revealed owing to the mentioned electronic product – the electronic platform of the comparative analysis of the Georgian folk prose.

პროექტის – „კიდევ ერთი ნაბიჯი მოსაუბრე ქართული თვითგანვითარებადი ინტელექტუალური კორპუსისაკენ“ – მიზნები და პირველი შედეგები

კონსტანტინე ფხაკაძე, მერაბ ჩიქვინიძე, გიორგი ჩიჩუა,
ინეზა ბერიაშვილი, დავით კურცხალია
საქართველოს ტექნიკური უნივერსიტეტი (საქართველო)
gllc.ge@gmail.com

2015 წლის 27 აპრილიდან, საქართველოს ტექნიკური უნივერსიტეტის ქართული ენის ტექნოლოგიზების ცენტრში, კ. ფხაკაძის ხელმძღვანელობითა და შოთა რუსთაველის ეროვნული სამეცნიერო ფონდის საგრანტო მხარდაჭერით ამოქმედდა პროექტი „კიდევ ერთი ნაბიჯი მოსაუბრე ქართული თვითგანვითარებადი ინტელექტუალური კორპუსისაკენ“, რომელიც, თავის მხრივ, კ.ფხაკაძის ხელმძღვანელობით 2012 წლიდან ცენტრში მოქმედი გრძელვადიანი პროექტის „ქართული ენის ტექნოლოგიური ანბანი“ [1] ერთ-ერთი მეტად მნიშვნელოვანი ქვეპროექტია[1].

დასრულებული სახით მოსაუბრე ქართული თვითგანვითარებადი ინტელექტუალური კორპუსის შექმნა [2] გულისხმობს მუდმივ ავტომატურ განვითარებაში მყოფი ქართული ვებ-კორპუსის აგებას, რომელშიც ჩადგმული სახით იფუნქციონირებს:

1. ქართული სააზროვნო და საკომუნიკაციო სისტემების ლოგიკაზე დაყრდნობით აგებული ქართული ენის ტექნოლოგიური ანბანი;

2. ქართულიდან სხვა ენებზე ორმხრივ მთარგმნელი სისტემები.

ამ მიზნების მიღწევას მოსაუბრე ქართული თვითგანვითარებადი ინტელექტუალური კორპუსი ანუ, მოკლედ, ქართული-ქსელი, შეძლებს იმის ხარჯზე, რომ იგი აღჭურვილი იქნება:

1. ქართული ინტელექტუალური ენობრივი ტექნოლოგიური სისტემებით ანუ ქართული ენისა და ქართული ენის სიღრმისეულ საფეხურზე მდგარი მათემატიკური ენის თავისებურებების გათვალისწინებით აგებული ქართულ-მათემატიკური, მოკლედ, ქართული კომპიუტერული „ტვინით“ და ამ „ქართულ კომპიუტერულ ტვინზე“ დაყრდნობით აგებული ქართული კომპიუტერული „ყურით“, „ყელითა“ და „თვალით“, რაც, მთლიანობაში, კორპუსში ქართული ინტელექტუალური სისტემისა და ქართული ხმოვანი და ვიზუალური ბრძანებებით მართვის სისტემების ჩადგმას ნიშნავს;

2. მასში უკვე ჩადგმულ ქართულ ინტელექტუალურ სისტემაზე დაყრდნობით აგებული ქართული ავტომატური მრავალენოვანი მთარგმნელით, რაც, მთლიანობაში, მანამდე მხოლოდ

[1] ქართული ენის ტექნოლოგიური ანბანის ქვეშ ჩვენ გვესმის ქართული ინტელექტუალური სისტემა ანუ, სხვა სიტყვებით რომ ვთქვათ, ქართული ენის სხვადასხვა სახის ტექსტებით მოცემული სხვადასხვა სახის პრობლემების მაანალიზებელი და ანალიზის შედეგად გამოკვეთილი დასკვნების მაგენერირებელი სისტემა.

ქართული ენით მოსაუბრე, მოაზროვნე და მართვად კორპუსს სხვა ენებით საუბრის, აზროვნებისა და მართვის უნარებითაც აღჭურავს.

ხაზგასასმელია, რომ ამ მიზნების მიღწევა, დღეს უკვე, ქართული ენის ციფრული კვდომის საფრთხისგან დაცვის ეროვნული მიზნისა და პასუხისმგებლობის გათვალისწინებით, სრულიად ცხად აუცილებლობას წარმოადგენს.

ჩვენი ეს კატეგორიული პოზიცია, ამავე საკითხზე წინა წლებში ჩვენ მიერვე გამოკვეთილ ხედვებთან ერთად [3], ძირეულად ეყრდნობა და ითვალისწინებს აგრეთვე მეტაქსელის ანუ მრავალენოვანი ევროპული ტექნოლოგიური ალიანსის ქსელის 200-ზე მეტი ექსპერტის მონაწილეობით განხორციელებული კვლევის „ევროპული ენები ციფრულ ეპოქაში“ შედეგებს.

კერძოდ, ჩვენ ვითვალისწინებთ ამ კვლევის საფუძველზე 2012 წლის 26 სექტემბერს – ენების ევროპულ დღეს მეტაქსელის მიერ გამოქვეყნებული პრესრელიზის „სულ ცოტა 21 ევროპული ენაა ციფრული კვდომის საფრთხის წინაშე – კარგი და ცუდი სიახლეები ენების ევროპულ დღეს“ საგანგაშო შინაარსს.

ასევე, ჩვენ ვითვალისწინებთ მეტა-ქსელის ტექნოლოგიური საბჭოს მიერ 2012 წლის 1 დეკემბერს გამოქვეყნებული ნაშრომით „სტრატეგიული კვლევითი გეგმა 2020 წლის მრავალენოვანი ევროპისათვის“ დაგეგმილი მრავალენოვანი ევროპის ტექნოლოგიური დაფუძნების მეტად მაღალ პერსპექტივებს, რომლის თანახმადაც მეტაქსელის არცთუ ისე შორეული „მიზანია ისეთი მრავალენოვანი ევროპული საზოგადოების ჩამოყალიბება, რომელშიც ყველა მოქალაქეს შეეძლება გამოიყენოს ნებისმიერი მომსახურება, ხელი მიუწვდებოდეს ნებისმიერ ცოდნაზე, ისიამოვნოს ნებისმიერი მედია საშუალებით, გააკონტროლოს ნებისმიერი ტექნოლოგია *თავისივე* მშობლიური ენით“ [4].

გარდა ამისა, ჩვენ აქ ვითვალისწინებთ იმასაც, რომ ქართული ენა ბევრად ჩამორჩება ციფრული კვდომის საფრთხის ქვეშ მყოფი 21 ევროპული ენიდან თითქმის ყველას როგორც ენობრივი რესურსებით უზრუნველყოფის, ისე ტექნოლოგიური მხარდაჭერისა და ქართული ენის ტექნოლოგიური დამუშავების მიზნებზე მიმართული ადამიანური და ფინანსური რესურსების თვალსაზრისით.

ამგვარად, პროექტის საბოლოო მიზანია უკვე არსებული მოსაუბრე ქართული თვითგანვითარებადი ინტელექტუალური კორპუსის საცდელი ვერსიის შემდგომი გასრულებით ზემოაღწერილი სახის ქართული-ქსელის რაც შეიძლება ფართოდ გამოყენებადი ვერსიის აგება და 2020 წლამდე დაგეგმილი მრავალენოვანი ევროპის ტექნოლოგიური დაფუძნების პროცესში ქართული ენის ჩართულობის უზრუნველყოფა. – ეს ჩვენ ქართული ენის ციფრული კვდომის საფრთხისგან დაცვის ერთადერთ გზად გვესახება [5].

მოხსენებისას პროექტის პირველი შედეგების სახით წარმოვადგენთ უკვე არსებული მოსაუბრე საცდელი კორპუსისა და მასში უკვე ჩადგმული ქართული ტექსტების მაანალიზებელი, მეტყველების დამამუშავებელი, ავტომატურად მთარგმნელი და ხმოვანი მართვის სისტემების ახალ – გაუმჯობესებულ ვერსიებს.[1] ამასთან, ხაზს ვუსვამთ, რომ ამ სისტემების ძველი ვერსიები

[1] ხაზგასასმელია, რომ როგორც თავად ეს მოსაუბრე საცდელი ვებ-კორპუსი, ასევე საპრეზენტაციოდ წარმოდგენილი სისტემებიდან უმეტესობა უნიკალურია იმ გაგებით, რომ არცერთ მათგანს სხვა ქართული ანალოგი არ გააჩნია (http://geoanbani.com/).

აგებულია სტუ ქართული ენის ტექნოლოგიზების ცენტრში მოქმედი გრძელვადიანი პროექტის – „ქართული ენის ტექნოლოგიური ანბანი“ ისეთი მნიშვნელოვანი ქვეპროექტების ფარგლებში წარმოებული კვლევებით [6 – 8], როგორებიცაა:

1. „ქართული ენის ლოგიკური გრამატიკის საფუძვლები და მისი გამოყენება საინფორმაციო ტექნოლოგიებში“ (პროექტი დააფინანსა შოთა რუსთაველის ეროვნულმა სამეცნიერო ფონდმა);
2. „ქართული ენის ტექნოლოგიური ანბანის ასაგებად აუცილებელი რიგი სისტემების გაფართოებადი (სწავლებადი) საინტერნეტო ვერსიების შემუშავება“ (პროექტი დააფინანსა საქართველოს ტექნიკურმა უნივერსიტეტმა);
3. „ევროკავშირში ქართული ენით ანუ სადოქტორო თემა – ქართული მეტყველების სინთეზი და ამოცნობა“ (პროექტს აფინანსებს შოთა რუსთაველის ეროვნული სამეცნიერო ფონდი);
4. „ევროკავშირში ქართული ენით ანუ სადოქტორო თემა – ქართული გრამატიკული მართლმწერი (ანალიზატორი)“ (პროექტს აფინანსებს შოთა რუსთაველის ეროვნული სამეცნიერო ფონდი).

ნაშრომი მომზადდა შოთა რუსთაველის ეროვნული სამეცნიერო ფონდის მიერ AR/122/4-105/14 პროექტზე „კიდევ ერთი ნაბიჯი მოსაუბრე ქართული თვითგანვითარებადი ინტელექტუალური კორპუსისაკენ“ გაღებული საგრანტო მხარდაჭერით.

ლიტერატურა:

Pkhakadze K., Chikvinidze M., Chichua G., Maskharashvili A., THE TECHNOLOGICAL ALPHABET OF THE GEORGIAN LANGUAGE – AIMS, METHODS, RESULTS, Reports of Enlarged Session of the Seminar of I.Vekua Institute of Applied Mathematics, Volume 27, pp. 46-49, 2013.

Pkhakadze K., Chikvinidze M., Chichua G., Maskharashvili A., Beriashvili I., AN OVERVIEW OF THE TRIAL VERSION OF THE GEORGIAN SELF-DEVELOPING INTELLECTUAL CORPUS NECESSARY FOR CREATING GEORGIAN TEXT ANALYZER, SPEECH PROCESSING, AND AUTOMATIC TRANSLATION SYSTEMS, Reports of Enlarged Session of the Seminar of I. Vekua Institute of Applied Mathematics, Volume 28, pp. 67-75, 2014.

კ. ფხაკაძე, ქართული ენის უფლებების დაცვისათვის, სამეცნიერო–საგანმანათლებლო ჟურნალი „ქართული ენა და ლოგიკა“, N3-N6, „უნივერსალი“, გვ. 83-111, 2007.

Presented by the META Technology Council, STRATEGIC RESEARCH AGENDA FOR MULTILINGUAL EUROPE 2020, Springer, 2012, pp. 1-87.

კ. ფხაკაძე, ღია წერილი საქართველოს მეცნიერებათა ეროვნულ აკადემიას ანუ ის, რომ ევროპული ენები საფრთხის წინაშეა, სრულიად ცხადს ხდის იმ საფრთხის განსაკუთრებით მაღალ ხარისხს, რომლის წინაშეცაა ქართული! – ანუ, კვლავ ქართული ენის უფლებების დასაცავად!! – ანუ, დროა მივხედოთ ქართულ ენას!!!- მოკლე ვარიანტი, ჟურნალი „ქართული ენა და ლოგიკა“, N7-N8, გვ. 1-20, 2014.

კ. ფხაკაძე, მ. ჩიქვინიძე, გ. ჩიჩუა, ა. მასხარაშვილი, ქართული ენის ლოგიკური გრამატიკის საფუძვლები და მისი გამოყენებანი, (იბეჭდება), 2015.

Pkhakadze K., Chichua G., Chikvinidze M., Maskharashvili A., THE SHORT OVERVIEW OF THE AIMS, METHODS, AND RESULTS OF THE LOGICAL GRAMMAR OF THE GEORGIAN LANGUAGE, Reports of Enlarged Session of the Seminar of I.Vekua Institute of Applied Mathematics, Volume 26, pp. 58-64, 2012.

კ. ფხაკაძე, მ. ჩიქვინიძე, გ. ჩიჩუა, პროექტი „ქართული ენის ლოგიკური გრამატიკის საფუძვლები და მისი გამოყენება საინფორმაციო ტექნოლოგიებში“ და სადოქტორო თემები „ქართული გრამატიკული მართლმწერი (ანალიზატორი)“ და „ქართული მეტყველების სინთეზი და ამოცნობა“, ჟურნალი „ქართული ენა და ლოგიკა“, გვ. 21-36, 2014.

The Aims and First Results of the Project “One More Step Towards Georgian Talking Self-Developing Intellectual Corpus”

Kostantine Pkhakadze, Merab Chikvinidze, Giorgi Chichua,
Ineza Beriashvili, David Kurtskhalia
Georgian Technical University (Georgia)
gllc.ge@gmail.com

Since 27 April, 2015, in the Center for Georgian Language Technology of the Georgian Technical University, under K. Pkhakadze’s leadership, the project “One More Step Towards Georgian Talking Self-Developing Intellectual Corpus” was launched. The project, which is supported by Shota Rustaveli National Science Foundation grant, is a very important sub-project of the long-term project “Technological Alphabet of the Georgian Language" [1], which, in turn, is developed at the center under K. Pkhakadze’s leadership from 2012.[1]

To build completely the Georgian talking self-developing intellectual corpus [2] means to create an automatically developing complete Georgian web-corpus which will be equipped:

1. With the technological alphabet of the Georgian language constructed on the basis of the logic of the Georgian thinking and communication systems;

2. With the two-way translator systems from Georgian into foreign languages.

All these aims can be done by the Georgian talking self-developing intellectual corpus, briefly, Georgian-net, because of it will be equipped:

1. With the Georgian intellectual language technology systems i.e. with the Georgian-mathematical, briefly, Georgian computer “brain”, which will be constructed on the basis of

[1] The Georgian technological alphabet, in other words, the Georgian intellectual system is understood as a system with ability to analyze the different problems of the different types given with the different Georgian texts and to generate the results of the analyzing.

the mathematical theory of the Georgian lingual-mathematical thinking system and, also, it will be equipped with Georgian computer "ear", "throat" and "eye" which, in turn, will be constructed on the basis of the "Georgian computer brain" (To provide the corpus with all these systems, it means to inbuilt the Georgian intellectual system and Georgian voice and visual manager systems in the corpus);

2. With the Georgian multilingual translator system, which will be constructed on the basis of the Georgian intellectual system (To provide the corpus with all these systems, it means to equip before only in Georgian language thinking, speaking, hearing and managing corpus with thinking, speaking, hearing and managing abilities in technologically elaborated foreign languages too).

It should be emphasized, that, today, achieving these goals is obviously necessary in the context of the Georgian national aim and responsibility of saving the Georgian language from the danger of the digital extinction.

Our categorical position is based on and takes into account, together with our views shaped in the previous years [3], the views from META-NET, which is dedicated to building the technological foundations of a multilingual European information society and which on the basis of the research "Europe's Languages in the Digital Age" has published the very alarming press-release "At Least 21 European Languages in Danger of Digital Extinction – Good News and Bad News on the European Day of Languages" on September 26, 2012.

Together with this, we take into account the content of the guide-like paper "Strategic Research Agenda For Multilingual Europe 2020" published on December 1, 2012, where the following is declared to be META-NET's not so far aim: "The goal is a multilingual European society, in which all citizens can use any service, access all knowledge, enjoy all media and control any technology *in their mother tongues*." [4].

Besides this, here we take into account the fact that almost any language from the above identified 21 European languages is much ahead than the Georgian from language resources and technology support points, as well from points of human and financial resources directed toward the technological foundation.

Thus, the final goal of the project is to construct the most widely usable version of Georgian-net of the above described type as it is possible through the extension of the trial version of the Georgian talking intellectual corpus and, also, to provide engagement of the Georgian language in already planned process of technological foundation of the 2020 multilingual Europe. This, we believe, is the only way to save the Georgian language in the digital age [5].

As it was already mentioned, the project is aimed at constructing as widely as possible usable version of the trial version of the Georgian talking intellectual corpus. Thus, at the conference, as first results of the project, we will present the new – improving version of the already existed trial version of the corpus and in it already inbuilt Georgian text analyzer, speech processor, automatic translator and

voice manager systems.[1] It must be emphasized, that older versions of these systems are built as results of the following important sub-projects of the long-term project "Technological Alphabet of the Georgian Language" led by K. Pkhakadze [6 – 8]**.** They are:

1. Project "Foundations of Logical Grammar of Georgian Language and Its Application in Information Technology", which was supported by the Shota Rustaveli National Science Foundation grant;
2. Project "Internet Versions of a Number of Developable (Learnable) Systems Necessary for Creating The Technological Alphabet of the Georgian Language", which was supported by the Georgian Technical University grant;
3. Project "In the European Union with the Georgian Language, i.e., the Doctoral Thesis – Georgian Speech Synthesis and Recognition" which is supported by the Shota Rustaveli National Science Foundation grant.
4. Project "In the European Union with the Georgian Language, i.e., the Doctoral Thesis – Georgian Grammar Checker (Analyzer)", which is supported by the Shota Rustaveli National Science Foundation grant;

We gratefully acknowledge the fact that the paper is published with the Shota Rustaveli National Science Foundation grant for the AR/122/4-105/14 project "One More Step Towards Georgian Talking Self-Developing Intellectual Corpus".

დეტერმინატორები და მოდიფიკატორები ძველ და ახალ ქართულში

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

სახელური ფრაზა, როგორც წესი, ბირთვისა (საზღვრული) და მისი თანმხლები სიტყვის (მსაზღვრელი) ან სიტყვებისაგან შედგება, შესაბამისად, საზღვრულს შესაძლოა რამდენიმე მსაზღვრელი ახლდეს ერთდროულად. თანამედროვე ლინგვისტურ თეორიებში სახელური ფრაზის ბირთვს დართული მსაზღვრელები ორ ჯგუფად იყოფა: მოდიფიკატორებად და დეტერმინატორებად. აღნიშნული კლასიფიკაციის პრინციპს განაპირობებს მსაზღვრელთა ფუნქციური სემანტიკა: მოდიფიკატორი ლექსიკური შინაარსის მატარებელი სიტყვაა და ახდენს სახელური ფრა-

[1] It must be emphasized, that the trial version of this talking web-corpus, as well as the systems, which are to be presented, are unique in the sense that none of them have any other Georgian analogues. (http://geoanbani.com/).

ზის ბირთვის სემანტიკურ მოდიფიკაციას – მიუთითებს გარკვეულ ნიშან-თვისებებზე (მაგ., ლამაზი სახლი). მოდიფიკატორებს მიეკუთვნებიან ზედსართავი სახელები, მიმღეობები და მიმღეობური კონსტრუქციები, ისევე როგორც ნათესაობით ბრუნვაში მდგარი არსებითი სახელები. მოდიფიკატორისაგან განსხვავებით, დეტერმინატორი ახდენს ბირთვის დეტერმინაციას და ორ ქვეჯგუფად იყოფა: პირველად და მეორეულ დეტერმინატორებად. პირველად დეტერმინატორებად ითვლება მაგ. განსაზღვრული და განუსაზღვრელი არტიკლი, მეორეულად კი მაგ. ჩვენებითი, კითხვითი და კუთვნილებით ნაცვალსახელები და კვანტიფიკატორები. ზოგადად, დეტერმინატორი ფუნქციური ელემენტია, რომელიც რეფერენციულობის გამოსახატავად გამოიყენება – ახდენს საზღვრულის დეტერმინაციას განსაზღვრულობა – განუსაზღვრელობის ან დეიქსისის თვალსაზრისით და არსებითად განსხვავდება მოდიფიკატორისაგან.

წინამდებარე კვლევაში, რომელიც ემპირიულ მონაცემებს ეფუძნება (კვლევა ჩატარებულია ქართული ენის ეროვნული კორპუსის ბაზაზე), ვეცადე შემესწავლა დეტერმინატორებისა და მოდიფიკატორების საკითხი ძველსა და ახალ ქართულში და გამერკვია, თანმიმდევრულობის რომელ წესებსა თუ წყობას ემორჩილება მთლიანად სახელური ფრაზა მოდიფიკატორებისა და დეტერმინატორების თვალსაზრისით და რატომ; ეთანხმებიან თუ არა სინტაგმის ბირთვს მოდიფიკატორები და დეტერმინატორები ბრუნვასა და რიცხვში, და, თუ ეთანხმებიან, როგორ? რა არის დეტერმინატორებისა და მოდიფიკატორების მაქსიმალური რაოდენობა სახელის ფრაზაში? რომელ იერარქიას ექვემდებარებიან დეტერმინატორები და მოდიფიკატორები და რომელი მოდიფიკატორები ან დეტერმინატორები შეიძლება იდგეს ყველაზე მოშორებით საზღვრული არსებითი სახელისაგან მსაზღვრელთა ჯაჭვში?

Determiners and Modifiers in Old and Modern Georgian

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

A typical noun phrase is built of a head noun and a modifier or determiner. Of course, if the language allows, several modifiers and/or determiners can accompany the head noun in a noun phrase.

In the ordinary noun phrase (e.g. *a beautiful house*), we have to distinguish between determiners and modifiers. There are two kinds of determiners: the primary and the secondary ones. As primary determiners we consider indefinite articles; as secondary demonstratives, interrogative and possessive pronouns and also quantifiers. Determiners, in general, express reference. Specifically, whether the head noun is referring to something definite or indefinite, something close or distant depends on determiners. In contrast to that, modifiers, like the name itself says, modify the head noun and indicate a certain

quality. Among modifiers we count e. g. adjectives, participles or participle constructions, and genitives (nouns in the genitive, which serve similar purposes like adjectives).

For my research, I tried to figure out, which rules or word order the whole noun phrase follows:

(1.1)

Es	čem-i	ķargad	ašenebul-i	x-is	or-i	lamaz-i	saxl-i
DET	PP	ADV	PART	GEN	NUM	ADJ	HEAD
This$_{Nom}$	my$_{Nom}$	well$_{ADV}$	built$_{Nom}$	wooden$_{Gen}$	two$_{Nom}$	beautiful$_{Nom}$	house$_{Nom}$

' These my well-built two beautiful wooden houses.'

Modern Georgian follows the prepositional word order for determiners and modifiers; that means that the head noun stands at the end, in the rightmost place. The head noun stands in the nominative, so because of the strong case agreement, most modifiers do the same (the only exceptions here is the genitive noun). The grammatical number is lexically expressed by the numeral; although the head noun as well as the modifiers and determiners are singular, the noun phrase is in plural because of the numeral. The determiners and modifiers will never express or, rather, mark plurality, but when the quantity of the head noun is indefinite, the plural is used. Compare (2) with (3):

(2)

or-i	saxl-i
NUM	HEAD
Two$_{Nom}$	house$_{Nom}$

'Two houses.'

(3)

čemi /	es /	lamaz-i	saxl-eb-i
PP	DET	ADJ	HEAD-Pl
My$_{Nom}$	this$_{Nom}$	beautiful$_{Nom}$	houses$_{Nom}$

'These my beautiful houses.'

In contrast to Modern Georgian, Old Georgian preferred other strategies for building a noun phrase. The noun phrase in Old Georgian is branded with postpositional word order; all determiners and modifiers were usually postpositioned.

(4)

Saxl-i	ig-i
HEAD	DET
House$_{Nom}$	that$_{Nom}$

'That/the house.'

As shown in (4), the postpositioned determiner *igi* – which usually is a demonstrative pronoun – serves as a definite article.

Regarding case and grammatical number agreement, head nouns and modifiers in Old Georgian agreed in both, case and grammatical number. It is important to note that genitive modifiers were marked with the genitive case marker to express the affiliation to the prepositioned head noun, but the last modifier additionally received the same case marking as the head noun, see (5):

(5)

saxl=man	mam-isa	čem-isa=man
HEAD	GEN	PP
House=Erg	fatherGenmyGen=Erg	

'My father's house' (ERG)

The modifier could also be in the plural, if the head noun was plural, regardless of whether the modifier could be singular:

(6)

ʒe=n-i	israel-isa=n-i
HEAD	GEN
Son=Pl-Nom	IsraelGen=Pl-Nom

'The sons of Israel.'

Nonetheless, there are examples in Old Georgian where a prepositional and a postpositional noun phrase occur in coordination in the same sentence:

(8.1) OG:

[…]	mis-n-i	ʒe-n-i	da	msaxur-n-i	mis-n-i
[…]	PP	HEAD	CONJ	HEAD	PP
[…]	his/herPl-Nom	sonPl-Nom	and	servant Pl-Nom	his/herPl-Nom

'[…] his/her sons and his/her servants.'

In case of several determiners, the head can stand between them:

(8.2) OG:

[…]momiṭevn-es	me	mraval-n-i	šecodeba-n-i	čem-n-i
[…]V	Pers	P	QUANTHEAD	PP
[…]ForgiveAOR-S3-Pl	meDat	manyNOM-Pl	sinNOM-Pl	myNOM-Pl

'[…] they forgave me my many sins.'

As mentioned before, Modern Georgian prefers the prepositional word order in a noun phrase, but also allows postpositional word order. In case of discontinuity, it can occur with both prepositional and postpositional word order. Notice the dative-marking of the adjective in (9), an example taken from poetry:

(9) MG:

ċrel-s	čavicomdi	ḳaba-sa
ADJ	V	HEAD
Coloured$_{Dat}$	put on$_{PRS-1Sg}$	dress$_{Dat}$

'I would put on a coloured dress.'

The word order shown in the beginning in (1) is one of the possible word orders for this sentence (the other one would be:(9) *es čemi ori lamazi ḳargad ašenebuli xis saxli* – 'These my two beautiful well-built wooden houses.'). Nonetheless, a chance of reading the noun phrase ambiguous (whether with the first word order or the second one) is possible: *ḳargad ašenebuli* (*well-built*) stands in both examples before the genitive noun *xis* (*wooden*; actually *tree* in the genitive), so it could refer to tree (*ḳargad ašenebuli xis saxli* – a house made of a well-built tree). In the present example, this reading is only excluded semantically, because trees are not built.

In my presentation, I will try to show which word order Modern Georgian follows and why. Is the word order in the Modern Georgian noun phrase *ad libitum* or does it follow certain rules? Which determiners or modifiers can stand farthest from the head noun? Do the modifiers and determiners agree (in case or grammatical number) among themselves, and if so, how? What is the maximal amount of determiners and modifiers in a noun phrase? Which hierarchy are the determiners and modifiers liable to? With the help of the GNC (Georgian National Corpus), I will also try to show which word order is used the most and evaluate the results.

ვარირების შესწავლა დიდ ინტერნეტკორპუსებში ლექსიკოგრაფიული აღწერის ამოცანის გადასაჭრელად

სერგეი შაროვი
ლიდსის უნივერსიტეტი (დიდი ბრიტანეთი)
s.sharoff@leeds.ac.uk

ლექსიკოგრაფიულ წყაროდ აუცილებელია დიდი ზომის კორპუსების გამოყენება, რადგანაც ლექსიკონთა შექმნას სჭირდება საიმედო ინფორმაცია სიტყვათა და მათი კონსტრუქციების სიხშირის შესახებ. პირველი დიდი ზომის კორპუსები, როგორებიცაა „ინგლისური ენის ბანკი“ და ბრიტანული ეროვნული კორპუსი, ლექსიკოგრაფიული მიზნებისათვის შეიქმნა. 2000-იანი წლების შემდეგ კორპუსების შემქმნელთა ყურადღება გადავიდა დიდი რაოდენობით არსებული

ისეთი მზა რესურსების გამოყენებაზე, როგორებიცაა ახალი ამბების გამოშვებები (Cieri & Liberman, 2002), ვიკიპედია,[1] ან ინტერნეტი (Baroni et al., 2009, Kilgarriff, 2001, Sharoff, 2006).

და მაინც, ასეთი სახის დიდი ზომის კორპუსები საგრძნობლად განსხვავდება ლექსიკური ერთეულებისა და მათი კონსტრუქციების სიხშირეების რაოდენობების დათვლის თვალსაზრისით. ეს შეეხება (1) სიხშირული ამოფრქვევებით გამოწვეულ ცვლილებებს, როდესაც სიტყვა დიდი სიხშირით გამოიყენება მცირე რაოდენობის დოკუმენტებში, (2) ამ კორპუსების თემატიკასა და ჟანრებს შორის არსებულ განსხვავებებს, (3) დოკუმენტებისა და კორპუსების მოცულობების განსხვავებებს. ასეთი ტიპის სამუშაო შესრულებულია მცირე ზომის ტრადიციულ კორპუსებთან მიმართებით, როგორიცაა ბრიტანეთის ეროვნული კორპუსი (Kilgarriff, 2005) ან ICE-GB (Gries, 2006). და მაინც, დიდი ზომის ინტერნეტკორპუსებს ნაკლებად აქვთ ინფორმაცია ტექსტთა კლასების შესახებ. ამასთანავე, ეს მიდგომები ვერ სწვდება მილიონობით მოკლე დოკუმენტს, რომლებიც ჩვეულებრივ გვხვდება ინტერნეტკორპუსებში. არსებულ გამოკვლევებთან დაკავშირებულ მეორე საკითხს წარმოადგენს სტატისტიკური საიმედოობის ნაკლებობა, რადგანაც ე.წ. „ნარჩენებს“ საკმარისი გავლენა შეიძლება ჰქონდეთ ფართოდ გამოყენებული სტატისტიკური ფუნქციების წარმოებაზე.

წინამდებარე მოხსენებაში წარმოვადგენ ჩემ მიერ განხორციელებულ სამუშაოს, რომელიც გულისხმობს დიდი ზომის ინტერნეტკორპუსებში ვარირებასთან დაკავშირებული საიმედო პროგნოზების გაკეთებას ისევე, როგორც ამ კორპუსების სტრუქტურის გამოვლენას მათს თემატიკასთან და სფეროებთან მიმართებით. ეს სამუშაო მოიცავს საიმედო სტატისტიკურ მონაცემებს მოსალოდნელი სიდიდეების (ადგილმდებარეობა) და მათი დიაპაზონების (განფენილობა) შესახებ, სტატისტიკურ მონაცემებს ვარირების შესახებ იმ ცალკეული დოკუმენტების თავდაპირველ და ვინსორიზებულ აღწერაში, რომლებისგანაც შედგება კორპუსი, ასევე მახასიათებელთა ჯგუფების შედარებას. ვარირების ამგვარი კვლევა მოიცავს ისეთ ერთეულებს, როგორებიცაა სიტყვათა და შესიტყვებათა სიხშირეები, ასევე ზედაპირული ენობრივი მახასიათებლები, როგორებიცაა არსებითი სახელების, ზმნების, ემფატიკური ნაწილაკების ან სემანტიკური კლასების სიხშირეები. წარმოვადგენთ მაგალითებს ukWac-იდან (Baroni et al., 2009), ruWac-იდან (Sharoff et al., 2013) და kaWac-იდან (Daraselia et al., 2014), ასევე ინგლისური, ქართული და რუსული ვიკიპედიებიდან.

[1] http://www.wikipedia.org/

Studying Variation in Large Web-corpora for the Task of Lexicographic Description

Serge Sharoff
University of Leeds (United Kingdom)
s.sharoff@leeds.ac.uk

Lexicographic evidence necessitates the use of large corpora, because development of dictionaries requires reliable information about the frequencies of words and their constructions. The very first large corpora, such as the Bank of English and the BNC, have been collected for lexicographic purposes. Since the 2000s, the attention of corpus developers shifted from manual development of corpora to using large amounts of readily available resources, such as newswires (Cieri and Liberman, 2002), Wikipedia,[1] or the Web (Baroni et al., 2009, Kilgarriff, 2001, Sharoff, 2006).

However, large corpora of this kind differ widely in their estimates of the frequencies of lexical items and their constructions. This is related to (1) differences caused by frequency bursts, when a word is used with high frequency in a small number of documents, (2) differences in the topics and genres of these corpora, and (3) differences in document and corpus sizes. Some related work has been done for small traditional corpora, such as the BNC (Kilgarriff, 2005) or ICE-GB (Gries, 2006). However, large webcorpora lack information about the text classes. Also these approaches do not scale to millions of relatively short documents, as commonly found in Web corpora. Another issue with existing studies concerns the lack of statistical robustness, since outliers can have a very considerable effect on the output of commonly used statistical functions.

In this talk I will present my work on making reliable predictions of variation in large web-corpora, as well as on detection of their structure in terms of their topics and domains. This involves robust statistical estimates of expected values (location) and their ranges (scatter), estimates of variation in original and Winsorised descriptions of individual files constituting a corpus, as well as comparison between clusters of features. The items for studying such variations are the frequencies of words, collocations, as well as surface-level linguistic features, such as the frequencies of nouns, verbs, emphatic particles or semantic classes. I will present examples of ukWac (Baroni et al., 2009), ruWac (Sharoff et al., 2013) and kaWac (Daraselia et al., 2014), as well as from the Wikipedias for English, Georgian and Russian.

References:

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

1 http://www.wikipedia.org/

Cieri, C. and Liberman, M. (2002). Language resources creation and distribution at the Linguistic Data Consortium. In *Proc Conference on Language Resources and Evaluation (LREC'02)*, pages 1327–1333. Las Palmas, Spain.

Daraselia, S., Sharoff, S., and Lortkipanidze, L. (2014). Towards creating a large corpus for Georgian. In *Biennial IVACS Conference*, Newcastle.

Gries, S. T. (2006). Exploring variability within and between corpora: some methodological considerations. *Corpora*, 1(2):109–151.

Kilgarriff, A. (2001). The web as corpus. In *Proc. of Corpus Linguistics 2001*, Lancaster.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus linguistics and linguistic theory*, 1(2):263–276.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Baroni, M. and Bernardini, S., editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna.

Sharoff, S., Umanskaya, E., and Wilson, J. (2013). *A frequency dictionary of Russian: core vocabulary for learners*. Routledge, London.

„ქართლის ცხოვრების“ პარალელური (ქართულ-სომხური) კორპუსი

ცირა ხახვიაშვილი, ნატო ბილანიშვილი
ილიას სახელმწიფო უნივერსიტეტი (საქართველო)
tsira.khakhviashvili@iliauni.edu.ge, nato.bilanishvili.1@iliauni.edu.ge

საპრეზენტაციოდ წარმოდგენილი პორტალი – **„ქართლის ცხოვრების“ პარალელური (ქართულ-სომხური) კორპუსი** 2012-2014 წლებში შეიქმნა ილიას სახელმწიფო უნივერსიტეტის ლინგვისტურ კვლევათა ინსტიტუტში. ეს კორპუსი 2010-2012 წლებში ინსტიტუტის მიერ განხორციელებულ „ვეფხისტყაოსნის“ პარალელურ (ქართულ-ინგლისურ) კორპუსთან ერთად, ქართული ენის უმნიშვნელოვანესი ძეგლების კვლევის სრულიად ახალი ეტაპია. მით უმეტეს, რომ ამ ეტაპზე ჩვენ არ გვაქვს არც ერთი მათგანის კრიტიკული გამოცემა თანამედროვე სტანდარტების გათვალისწინებით. ვგულისხმობთ კრიტიკული გამოცემის მომზადებას ტექსტის სტემის დადგენის საფუძველზე.

პროექტზე მუშაობის პროცესში მომზადდა საქართველოში დაცული ყველა ხელნაწერის (60) აღწერილობა შუა საუკუნეების ხელნაწერი და ბეჭდური ტექსტების კორპუსული გამოცემების სტანდარტის გათვალისწინებით. მასში შევიდა შემდეგი მონაცემები: ხელნაწერის შენახვის ადგილი, შექმნის თარიღი, შექმნის ადგილმდებარეობა, ობიექტის მასალისა და დაზიანებების აღწერა, გვერდების რაოდენობა, ავტორი, გადამწერი, მომგებელი, მინაწერები, ილუსტრაციებისა და გადამწერის ხელის აღწერა, გვერდების ზომები და ა.შ.

მონაცემთა ელექტრონული დოკუმენტირებისას თავიდან გადაისინჯა, გასწორდა და აღინუსხა რიგი უზუსტობანი, რომლებიც ცალკეულ გამოცემებსა და აღწერილობებში იყო. „ქართლის ცხოვრების“ ნუსხების გარდა, პორტალზე განთავსდა ვახუშტი ბატონიშვილის „აღწერა სამეფოსა საქართველოჲსა" და ანტონ კათალიკოსის „საქართველოს ისტორიის“ ტექსტები, რომელთაგან ზოგიერთი აქამდე გამოუცემელი იყო.

კორპუსის პარალელურ ნაწილში წარმოდგენილია „ქართლის ცხოვრების“ უძველესი უცხოენოვანი, ძველი სომხური თარგმანი და ამ სომხური ვერსიის ილია აბულაძისეული ახალი ქართული თარგმანი. ძველი ქართული და ძველი სომხური ტექსტების გასათანაბრებელ ერთეულად გამოყენებულია აბზაცი. პარალელური კორპუსის ნაწილში სამომავლოდ იგეგმება რობერტ ტომსონისეული ინგლისურენოვანი თარგმანის, ასევე „ქართლის ცხოვრების“ გერმანული და ფრანგული თარგმანების განთავსება. პორტალის ვიზუალურ ბლოკში განთავსებულია კორპუსში წარმოდგენილი ყველა ხელნაწერის თითო გვერდის ფოტოასლი.

„ქართლის ცხოვრების“ პროექტი განხორციელდა ნინო დობორჯგინიძის ხელმძღვანელობით. ტექსტური მემკვიდრეობის დოკუმენტირების პროცესს კოორდინატორებას უწევდა სვეტლანა ბერეკაშვილი და ირინა ლობჟანიძე. საძიებო სისტემა და ვებპორტალი მოამზადა პროგრამისტმა გიორგი მირიანიშვილმა. ტექსტების დოკუმენტირებაზე მუშობდნენ: ცირა ხახვიაშვილი, ნატო ბილანიშვილი, გვანცა შუბითიძე, შალვა თუმანიშვილი და ნინო დათაშვილი.

Parallel (Georgian-Armenian) Corpus of *Kartlis Tskhovreba* (Life of Kartli)

Tsira Khakhviashvili, Nato Bilanishvili
Ilia State University (Georgia)
tsira.khakhviashvili@iliauni.edu.ge, nato.bilanishvili.1@iliauni.edu.ge

The portal represented for **Parallel (Georgian-Armenian) Corpus of Kartlis Tskhovreba (Life of Kartli**) was set up at Linguistic Research Institute of Ilia State University in 2012-2014. This corpus is completely new stage of the most important monuments of the Georgian language together with parallel (Georgian-English) corpus of the poem *Vepkhistkhaosani* (“The Knight in the Panther's Skin”) implemented by the Institute in 2010-2012. Moreover, at the moment we do not have any critical editions of them taking into account the modern standards. We mean preparation of critical edition on the basis of determination text stem.

In the process of the project, description of all manuscripts (60) preserved in Georgia was prepared taking into account corpus edition standards of manuscripts and printed texts of the Middle Ages. The following data are included into it: location of manuscript storage, date of creation, location of creation,

description of facilities and material damage, number of pages, author, scribe, customer, notes, illustrations, and description of a scribe's caligraphy, page sizes, etc.

During electronic documenting of data, a number of inaccuracies, occurring in individual publications and descriptions, were revised, corrected and described. In addition to Kartlis Tskhovreba (Life of Kartli) lists, "Description of Kingdom of Georgia" written by Vakhushti Batonishvili and Anton Catholicos texts about "the history of Georgia" were published on the portal, some of which were previously unedited.

The oldest foreign language translation of Kartlis Tskhovreba (Life of Kartli), old Armenian translation, and Ilia Abuladze's new Georgian translation of this Armenian version are represented in the parallel part of corpus.

A paragraph is used as an equation unit for old Georgian and old Armenian texts. In the part of parallel of Corpus, Robert Tomson's English translation, as well as German and French translations of Kartlis Tskhovreba (Life of Kartli) are planned to be publish on the portal in the future. Photocopies of each page of all manuscripts represented in Corpus are placed in the visual block of the portal.

Kartlis Tskhovreba (Life of Kartli) project was implemented under the leadership of Nino Doborjginidze. The documenting process of text heritage was coordinated by Svetlana Berekashvili and Irina Lobjanidze. The query system and web portal were prepared by programmer George Mirianashvili. Text documentation: Tsira Khakhviashvili, Nato Bilanishvili, Gvantsa Shubitidze, Shalva Tumanishvili and Nino Datashvili.

მარკირება მეტყველების ნაწილთა მიხედვით სხვადასხვა ტიპის ენაში: რამდენიმე თეორიული საფუძველი მორფოსინტაქსური ანოტირების სქემებისათვის

ენდრიუ ჰარდი
ლანკასტერის უნივერსიტეტი (დიდი ბრიტანეთი)
a.hardie@lancaster.ac.uk

როდესაც გარკვეულ ენაზე მომუშავე ლინგვისტური თემი ჩაებმება ამ ენისათვის კორპუსის შექმნის საქმეში, გარდა საკორპუსე მასალების თავდაპირველი მობილიზებისა, ერთ-ერთ უპირველეს პრიორიტეტად ჩვეულებრივ გამოდის ზედაპირული ანოტირების მექანიზმის შემუშავება. ზედაპირული ანოტირება გულისხმობს მარკერთა ადეკვატური სიმრავლის არსებობას ანუ ამომწურავ და შინაგანად თანმიმდევრულ სქემას მოცემული ენის მორფოსინტაქსური ანალიზისათვის. და მაინც, ამგვარი სქემის შემუშავება ნაკლებად იქნება კონცეპტუალურად სწორხაზოვანი, რადგანაც მარკერთა სიმრავლის კონსტრუქცია მრავალმხრივადაა დაკავშირებული

მოცემული ენის ტიპოლოგიურ პროფილთან. ამ საკითზე არსებული ლიტერატურის დიდი ნაწილი (ა) ძირითადად ეხება ინგლისურს ან ტიპოლოგიურად მის მსგავს ენებს -- ზედაპირული ანოტირება დასავლეთ ევროპის გარეთ არსებული ენებისათვის სათავებს იღებს მე-20 საუკუნის 90-იანი წლების ბოლოდან; (ბ) უმეტესწილად ფოკუსირებულია ანოტირების პროგრამული უზრუნველყოფის შემუშავებაზე, ვიდრე მარკერთა სიმრავლეებზე, რომლებიც ხშირად „ნაგულისხმევია“.

წინამდებარე მოხსენებაში წარმოვადგენთ ძირითად თეორიულ საკითხებს, რომლებიც გულისხმობს მარკერთა სიმრავლის შექმნას ისეთი ენებისათვის, რომლებიც მსოფლიოს მასშტაბით უფრო მეტ მრავალფეროვნებას გვიჩვენებენ, ვიდრე ეს დასავლეთ ევროპაში გვაქვს. პირველ განსახილველ საკითხს წარმოადგენს საკუთრივ მარკერთა სიმრავლის კონცეპტუალური სტრუქტურა, რომლისათვის რამდენიმე ალტერნატივა არსებობს:

- მარკერთა უნიტარული სიმრავლე
- მახასიათებელთა იერარქია (რასაც სხვაგან ვუწოდე „მარკერთა იერარქიულ-დაშლადი სიმრავლე“)
- მახასიათებელთა მატრიცა
- მახასიათებელთა უწესრიგო სიმრავლე (რასაც არაოფიციალურად „მარკერთა ჩანთას“ უწოდებენ)

ამასთანავე, არსებობს (სულ ცოტა) ორი შემდგომი სტრატეგია, რომლებიც შეიძლება დაინერგოს მორფოსინტაქსურ სქემებში მორფოლოგიურად რთული ენებისათვის: ხელახალი ტოკენიზაციის სტრატეგია და მორფოლოგიური ანალიზის უკუქცევის სტრატეგია.

მარკერთა სიმრავლეების ამ განსხვავებული ტიპების განხილვისას მე წარმოვადგენ, როგორ ურთიერთქმედებს ენის მორფოლოგიური ტიპოლოგია მარკერთა სისტემის კონსტრუქციასთან, მაშინ როცა არ არსებობს მკაცრად განსაზღვრული ურთიერთობა – ასე, მაგალითად, თანამედროვე ინგლისურის შემთხვევაში მოიპოვება სამუშაო მაგალითები მარკერთა სიმრავლის რამდენიმე სხვადასხვა სახისათვის, დაწყებული უნიტარულიდან მახასიათებელთა იერარქიამდე და მახასიათებელთა უწესრიგო სიმრავლით დამთავრებული – გვაქვს ზოგიერთი აშკარა ტენდენცია. მაშინ როცა იზოლირებული ენები (მანდარინული ჩინური, ინგლისური) უპირატესობას ანიჭებენ ან მარკერთა უნიტარულ სიმრავლეებს ან ძალიან ზედაპირულ იერარქიებს, ფლექსიური ენები მოკრძალებულად უპირატესობას ანიჭებენ უფრო ღრმა იერარქიებს, ხოლო ჭარბი და რთული ფლექსიური და/ან აგლუტინაციური მიმართებების მქონე ენები ამჯობინებენ ან მახასიათებელთა უწესრიგო სიმრავლეს ან მახასიათებელთა მატრიცებს, ან მოითხოვენ ხელახალი ტოკენიზაციის სტრატეგიის ან მორფოლოგიური ანალიზის უკუქცევის სტრატეგიის გამოყენებას იერარქიულ სქემასთან ერთად, რათა თავიდან აიცილონ ისეთი იერარქია, რომელიც კონტროლს არ ექვემდებარება.

მოხსენებაში წარმოდგენილი იქნება მარკერთა სისტემის შემუშავების დეტალური მაგალითები რამდენიმე განსხვავებული ენისათვის – ზოგიერთი ლიტერატურიდან იქნება მოყვანილი, ზოგი კი ჩვენი საკუთარი გამოცდილებიდან – მათ შორის, რუსულის, არაბულის, ურდუს, ტიბეტურისა და თურქულისათვის. განსაკუთრებულ ყურადღებას მივაქცევთ ნეპალურსა და ასამურს; ორივე მათგანი ინდოევროპული ენაა, რომლებზეც დიდი გავლენა მოახდინეს მათმა ტიბეტელ-ბირმელმა მეზობლებმა; ასევე გავარკვევთ როგორ განიცადა წარუმატებლობა მახასი-

თებელთა იერარქიის მათდამი მიყენების თავდაპირველმა მცდელობებმა და როგორ გახდა საჭირო მათი გადამუშავება სხვადასხვა სტრატეგიის გამოყენებით. ეს წანამძღვრები წარმოგვიდგენს იმ ფაქტის თვალსაჩინო მაგალითებს, რომ ანოტირების სქემა, რომელიც შემუშავებულია ენის ფუნქციონირების შესახებ არსებული უმართებულო თეორიის საფუძველზე, უბრალოდ ქარწყლდება კორპუსულ მონაცემებთან კონტაქტის კვალად.

დასკვნის სახით შემოგთავაზებთ საწყის კომენტარებს ნეპალურის, ასამურისა სხვა ენებთან დაკავშირებით უკვე გაწეული სამუშაოს შესახებ, რომელიც მიგვითითებს იმ სათანადო გზაზე, რომლის მიხედვითაც შეიძლება შეიქმნას ზედაპირული ანოტირების მარკერთა სისტემა ქართულის მსგავსი ენებისათვის.

Part-of-speech Tagging in Different Kinds of Language: Some Theoretical Bases for Morphosyntactic Annotation Schemata

Andrew Hardie
Lancaster University (United Kingdom)
a.hardie@lancaster.ac.uk

When the linguistic community working on a particular language embarks upon the corpus linguistic enterprise for that language, it is very typical for the first priority – other than the initial development of corpus data resources – to be the creation of a part-of-speech (POS) tagger. POS tagging assumes the existence of an adequate tagset, that is, an exhaustive and internally consistent schema for morphosyntactic analysis of the language in question. Yet the development of such a schema is far from conceptually straightforward, since the design of a tagset interacts extensively with the typological profile of the language in question. Much of the literature on the issue (a) centres either on English, or on typologically similar languages – POS tagging for languages outside Western Europe stems from the mid-to-late 1990s; (b) focuses more on the development of tagger software than on the tagsets, which are often *assumed*.

In this talk, I will lay out the basic theoretical issues involved in thinking about tagset design for languages which exhibit more of the world's variability than is found in Western Europe alone. The first issue to consider is the conceptual structure of the tagset itself, for which a number of options exist:

- A unitary tagset
- A feature hierarchy (what I have elsewhere dubbed a "hierarchical-decomposable tagset")
- A feature matrix
- An unordered feature set (informally referred to as a "bag of tags")

In addition, there exist (at least) two further strategies which can deployed in morphosyntactic schemata for morphologically complex languages: the retokenisation strategy, and the morphological-analysis fallback strategy.

In discussing these different types of tagset, I will outline how the morphological typology of the language interacts with the tagset design. While there is no strongly determinative relationship – so, for instance, in the case of modern English, we can find working examples of several different kinds of tagset, from unitary to feature hierarchy to unordered feature set – there are some clear tendencies. While isolating languages (Mandarin, English) favour either unitary tagsets or very shallow hierarchies, and moderately inflecting languages favour deeper hierarchies, languages with a large amount of complex inflectional and/or agglutinative behaviour favour either unordered feature sets or feature matrices, or else require the use of a retokenisation strategy or morphological-analysis fallback strategy alongside a hierarchical schema, to avoid a unmanageably complex hierarchy.

The presentation will include detailed examples of tagset development for a number of different languages – some drawn from the literature and some from my own experience – including Russian, Arabic, Urdu, Tibetan and Turkish. In particular, I will look at Nepali and Assamese, both Indo-European languages which have been strongly influenced by the agglutinative nature of their Tibeto-Burman neighbours, and explore how initial attempts to treat them in terms of a feature hierarchy failed and had to be reworked using a different strategy. These experiences furnish an informative exemplification of the fact that a schema of annotation driven by mistaken theory about how a language works simply cannot survive contact with the corpus data.

I conclude by offering some initial comments on earlier work with Nepali, Assamese, and other languages can tell us about the appropriate way to approach POS tagset design for a language such as Georgian.

სარჩევი – Content

THE INTERNATIONAL CONFERENCE

„LANGUAGE AND MODERN TECHNOLOGIES – 2015“

საერთაშორისო კონფერენცია

„ენა და თანამედროვე ტექნოლოგიები – 2015“

THE INTERNATIONAL CONFERENCE

„LANGUAGE AND MODERN TECHNOLOGIES – 2015“

საერთაშორისო კონფერენცია

„ენა და თანამედროვე ტექნოლოგიები – 2015“

THE INTERNATIONAL CONFERENCE

„LANGUAGE AND MODERN TECHNOLOGIES – 2015“