სამეტყველო კოდთა მიმართებები ფორუმულ მეტყველებაში

ნათია ამაღლობელი
ილიას სახელმწიფო უნივერსიტეტი (საქართველო)
natiama2000@yahoo.fr

შესავალი

ახალი ტექნოლოგიები, რომლებიც ჩვენი ცხოვრების განუყოფელი ნაწილი გახდა, პირველ რიგში ხელს უწყობს სოციალურ კომუნიკაციასა და ინტერპერსონალურ ურთიერთობებს. განსხვავებული წარმოშობისა და მენტალიტეტის მქონე სხვადასხვა ენის მატარებელ ინდივიდებსა და სოციალურ ჯგუფებს შორის ამგვარი ინტერაქცია საფუძვლად ედება ენათა კონტაქტებს. ჩვენი მიზანია, ქართული ინტერნეტფორუმებიდან მიღებული მონაცემების საფუძველზე აღვწეროთ სამეტყველო კოდთა მიმართებების ის ტიპები, რომლებიც ელექტრონულ კომუნიკაციაში გვხვდება.

ამგვარი მიმართებები პირობითად შეიძლება სამ კატეგორად დავაჯგუფოთ: ბუნებრივ სამეტყველო კოდთა; წერით და ზეპირ სამეტყველო კოდთა; ბუნებრივ და ელექტრონულ სამეტყველო კოდთა მიმართებები.

ბუნებრივ სამეტყველო კოდთა მიმართებები

ბუნებრივ სამეტყველო კოდთა მიმართებებში განვიხილავთ ოთხ ტიპს. ესენია: ნასესხები სიტყვები, ფონეტიკური მიმიკრია, კოდური მონაცვლეობა, სტილთა მონაცვლეობა.

საკვლევი კორპუსის მიხედვით **სესხების** სამ ძირითად ტიპს გამოვყოფთ:

- ტრანსკრიფცია და/ან ტრანსლიტერაცია:
  ოფტოპიკი ← 'off topic',
  დრაივერი ← 'driver'.
  მაზერბორდი ← 'motherboard'.
  არის შემთხვევები, როდესაც არასწორად წარმოთქმული ინგლისური სიტყვა პოპულარული ხდება მომხმარებლებს შორის და ადგილსაც იმკვიდრებს, როგორც მაგ., *computer – კომპუტერი.*
- თარგმანი, რომელიც შეიძლება იყოს სრული (*ფანჯარა* ← 'windows', *თაგვი* ← 'mouse') ან არასრული (*მყარი დრაივი* ← 'hard drive')
- ტრანსფერი, ანუ სიტყვაწარმოებითი მორფოლოგიური მოდელის გადატანა ერთი ენიდან მეორეში. ტრანსფერს მიეკუთვნება:

ა) ინგლისურენოვანი ფუძისა და ქართული აფიქსების შედეგად მიღებული ენობრივი ერთეულები: Linking ← *და-ლინკ-ვა*; Shutting down ← *და-შათ+დაუნ-ებ-ა*; Sending a message← *და-მესიჯ-ებ-ა;* Search ← და-სერჩ-ვა.

ბ) ქართული ფუძისა და ინგლისური აფიქსების შედეგად მიღებული ერთეულები (ეს უკანასკნელი შემთხვევა ძალზე იშვიათია): *შიმშილ-ინგ-ი* ← 'hanger'

საკვლევი მასალის (6800 სიტყვა, რომელიც აღებულია 2009 წლის ივნისის მონაცემებით სამი ქართული საჯარო ინტერნეტფორუმიდან) ანალიზმა აჩვენა, რომ ქართულ ფორუ-

მულ მეტყველებაში სიტყვათა საერთო რაოდენობის 2,18% ნასესხებია. აქედან ნათარგმნია მხოლოდ 0,1%. მომხმარებელს ურჩევნია გამოიყენოს ტრანსლიტერაცია ან ტრანსფერი.

ფონეტიკური მიმიკრიის შემთხვევაში ინგლისური ტერმინები იცვლება იდენტური წარმოთქმის მქონე ქართული სიტყვებით. შედეგად ეს უკანასკნელნი იძენენ ახალ მნიშვნელობას და გადადიან შედარებით დაბალ სტილისტურ რეგისტრში: 'google' ← *გუგული*.

კოდური მონაცვლეობა ჩვენს შემთხვევაში გამოიხატება მონაცვლეობით:

- ქართულსა და უცხო ენას შორის, რაც უმეტესწილად გამოიყენება ციტირების ან აზრის გამოთქმისათვის: *„ალარ ღირს ამაზე საუბარი, delete and forget... I'm OUT.... Peace-Out y'all. მე, მაგალითად, რუსი ბებია მყავდა... Ethnocentrism is the cause of war people…"*
- სტანდარტულ მეტყველებასა და დიალექტების შორის, რაც ძირითადად იუმორის გადმოცემას ემსახურება: *„ხალხო, მომილოცეთ, 20–დან 19 ქულა ავიღე! აბა, რა გეიფიქრეთ თქვენ აქანე? მარტო თქვენ ხართ ჭკვიანები? :))))*

სტილთა მონაცვლეობა ქართულ ფორუმებში გამოიხატება შემდეგი ფორმებით:

- ფორმალურიდან არაფორმალურ მეტყველებაზე გადასვლა და, პირიქით:

  პატივცემულო მოდერატორო, გთხოვთ გაითვალისწინოთ ფართო საზოგადოების ნება და არ გააუქმოთ ზემოთ აღნიშნული თემა, თორე იცოდე, ჩვენც შეგვიძლია რაღაცეები, ბიჭო!!)

  ფორმალური სტილი ფორუმულ მეტყველებაში ძირითადად გამოიყენება ირონიის, მუქარის ან დაცინვის გამოსახატავად. პირისპირი თუ წერითი კომუნიკაციისაგან განსხვავებით, ფორმალური სტილი აქ არ იხმარება უცხო ან მაღალი სოციალური სტატუსის მქონე პირთან ურთიერთობისას, რადგან ონლაინფორუმის გარემო ყველა მომხმარებელს ერთ დონეზე აყენებს.
- სტანდარტული მეტყველებიდან ჟარგონსა თუ სლენგზე გადასვლა. ამ შემთხვევაში სლენგად განვიხილავთ ასევე ელექტრონულ მეტყველებაში გამოყენებულ არასტანდარტულ ორთოგრაფიულ ფორმებს (*2განიზებული, გკ, მგრ).*
- ერთი ენობრივი რეგისტრიდან მეორეზე გადასვლა, მაგ.: *„ლაო, პატალა, გელაფჟლი გაიგე? :-) მოკლედ გეტყვი, აფექტი მსხვერპლის პროვოცირებული უნდა იყოს, ეს სავალდებულო ნიშანია, რეალურად კი, თუ მოინდომეს, მიუყენებ ამ მუხლს, რა პრობლემაა? :-)"*

წერითი და ზეპირ სამეტყველო კოდთა მიმართებები

წერით და ზეპირ სამეტყველო კოდთა მიმართებები ქართულ ინტერნეტფორუმში გამოიხატება ზეპირმეტყველებისთვის დამახასიათებელი ლექსიკური თუ სინტაქსური ხერხების, რიტმისა თუ ემფაზის გამომხატველი საშუალებების (კაპიტალიზაცია, ექსპრესიული პუნქტუაცია, ასოთა გამეორება) გამოყენებით.

ბუნებრივ და ელექტრონულ სამეტყველო კოდთა მიმართებები

ბუნებრივ და ელექტრონულ სამეტყველო კოდთა მიმართებებში განვიხილავთ ფორუმული დისკურსის იმ ასპექტებს, რომლებიც შესაძლო გახდა მედიუმის ტექნიკური მახასიათებლებიდან გამომდინარე:

- **ხატულები**, ანუ ასოების, პუნქტუაციური თუ სხვა ნიშნებისგან შექმნილი ფიგურები, რომლებიც ელექტრონულ მეტყველებაში ემოციისა და მიმო-ჟესტუალობის გამოხატვის საშუალებას იძლევა. ქართულ ფორუმულ მეტყველებაში ხატულების

სამი ტიპის გამოყოფა შეიძლება: ტიპოგრაფიული (:-)); გრაფიკული, ხშირად ანიმაციური (☺) და ვერბალური ("გახარებული სმაილიკი") ხატულები.

- **ჰიპერტექსტი**, რაც ელექტრონულ მეტყველებას უფრო დინამიკურს ხდის.
- **ლათინიზაცია**, რაც ქართული ელექტრონული მეტყველების ერთ-ერთ ყველაზე თვალშისაცემ თავისებურებას წარმოადგენს. ტექნოლოგიური შეზღუდვებიდან გამომდინარე, ზოგჯერ მომხმარებელს ქართული შრიფტის გამოყენების საშუალება არ აქვს და იძულებულია ქართული ასოები ლათინურით ჩაანაცვლოს. ეს ყოველივე დამატებით პრობლემებს ქმნის, რადგან ზოგიერთი ქართული ბგერის ჩასაწერად რამდენიმე ლათინური ნიშნის გამოყენებაა საჭირო. ამასთან, რამდენადაც ლათინიზაციას არაფორმალური ხასიათი აქვს, ქართული ასოების ლათინური „ორთოგრაფია" ხშირად იცვლება. საბედნიეროდ, ლათინიზაციის შემთხვევები მომხმარებელთა გაღიზიანებას იწვევს და მისი გამოყენება ინტერნეტფორუმებში სულ უფრო მცირდება.

საბოლოოდ შეგვიძლია დავასკვნათ, რომ ქართული ფორუმული მეტყველება წარმოადგენს წერითი, ზეპირი და ელექტრონული მახასიათებლებისგან შექმნილ ჰიბრიდულ რეგისტრს, რომელზეც გავლენას ახდენს როგორც თავად მედიუმის ტექნიკური მონაცემები, ასევე საკომუნიკაციო სიტუაცია, ქართული ენის ლინგვისტური მახასიათებლები და მომხმარებლის პირადი მონაცემები.

Speech Code Interrelations in the Georgian Forum Language

Natia Amaghlobeli
Ilia State University (Georgia)
natiama2000@yahoo.fr

New technologies which have become an inseparable part of our life, primarily encourage social communication and interpersonal relationships. Such interactions between social groups or individuals of different backgrounds initiate numerous speech code contacts. The purpose of this paper is to analyse speech code relations taking the example of Georgian online forums.

Such relationships may be grouped into three categories: relationships between natural speech codes, between oral/written speech codes, and those between natural/electronic linguistic codes.

In the category of **natural code relations** we will consider four types of language contact: borrowing, phonetic mimicry, code-switching and style-shifting.

We have singled out three main types of **borrowing** in our research corpus:

1. Transcription or transliteration:
   optoṗiḳi ← 'off topic',
   draiveri ← 'driver'.

mazerbordi ← 'motherboard'.

Sometimes a mispronounced English word becomes popular among forum users.

Ex. *computer* – 'ḳompuṭeri' (ST. 'ḳompiuṭeri').

2. Translation that can be exact (panjara ← 'windows') or approximate (mqari draivi ← hard drive').

3. Transfer of morphological word formation model from one language into the other:

a) Transferred stem and indigenous affix: 'Linking'→ da-*linḳ-va* ; 'Shutting down'→ *da-šat+daun-eb-a*; 'Sending a message'→ *da-mesiǯ-eb-a*.

b) Indigenous stem and transferred affix: š*imšil-ing-i* ← 'hunger'.

The analysis of the whole corpus data (6800 words) has shown that 18% of the total number of words in Georgian online forum are lexical borrowings.

The percentage of the translated loan words in the corpus is equals to only 0, 1 %. Thus, the forum users prefer to use imported words or loan blends.

Phonetic mimicry is the resemblance of English computer terms to semantically different Georgian words. Consequently, the Georgian words assume new meanings and move to a lower stylistic register.

Ex. 'google' → *guguli* (eng. 'cuckoo')

We will be using the term **code-switching** to cover the phenomena of alternating between:

- two languages, which is generally used to insert an opinion or citation : *„aɣar ɣirs amaze saubari* ["Let's not speak more about it"], *delete and forget... I'm OUT.... Peace-Out y'all. me magalitad rusi bebia mq̣avda*…["My grand-mother was Russian"] *Ethnocentrism is the cause of war people…"*
- standard Georgian and its dialect, that is generally used to make fun, mock or imitate: *"xalxo, momilocet, 20dan 19 kula aviɣe! (*standard Georgian) *aba ra geipikret tkven akane? marṭo tkven xart ċḳvianebi?:))))"* (the Gurian dialect)

We can outline the following types of **style–shifting** in Georgian internet forums:

- From informal or casual to formal and vice-versa, ex. *„p̣aṭivcemulo moderat'oro, gtxovt gaitvalisċinot parto sazogadoebis neba da ar gaaukmot zemotaɣnišnuli tema, tore icode, čvenc šegviǯlia raɣaceebi, biċo!!;-)*"

The formal style in the Georgian forum is generally used to express offence, irony, grievance, or threat. In contrast to the face-to-face or written communication, in forum language, the formal style is not used to address strangers or persons of higher social status, because the specificity of the forum environment gives to interaction a peer-to-peer register.

- Shifts to slang or jargon.
- Shifts from one speech register to another: ex. *„lao, paṭala, velapeli gaige? :-) moḳled geṭqvi, apekṭi msxverṗlis ṗrovocirebuli unda iqos, es savaldebulo nišania, realurad ḳi, tu moindomes, miuqeneb am muxls, ra ṗroblemaa? :-)“*

Spoken/written code relations in Georgian forum are realised by syntax and word choice evoking conversational informality, emphasis, rhythm. Syntactic informality often acquires the form of incomplete, elliptical sentences. Another device used to imitate characteristics of speech is the textual indication of emphasis on words or phrases. In order to create the sense of oral emphasis participants often use capital (in case of Latin script) or bold letters (*me šen MI¶VARXAR!* („I LOVE you“), expressive punctuation (*ra magaria*!!!! ('Cool!!!'); *raṭom???* ('why???') or repetition of letters (*HIIIII; magariaaa* ('coooool').

In the category of **natural/electronic code relations** we will consider the features of forum language that are made technically possible by the computer keyboard: emoticons, hypertext and latinization.

- Emoticon (the loan word from *emotion* and *icon*) is a facial expression represented by a combination of punctuation mark, letters or other characters, that viewed from the side resembles a facial expression or, more rarely, gestures. In Georgian forums we can differentiate three types of emoticons: punctuate ("**:-)**"), graphical (☺) and verbal ("*gaxarebuli smailiḳi*" 'happy smiley') emoticons.
- Hypertext that makes the on-line discussion more dynamic.
- One of the most "salient" limitations of Georgian electronic writing is that the Georgian language has its own script and the use of Latin letters for Georgian words is not common in most applications. However, due to technological constraints (keyboard not supporting Georgian scripts or requiring additional functions) users are forced or prefer to use Latin characters that create additional problems. As there are some Georgian letters that have no analogues in the Latin alphabet, users are forced to use two or more Latin letters to express one Georgian letter. As well as Latinization, Georgian has informal character, Latin "orthography" of Georgian words may vary among users or within an individual writer's usage. Fortunately in the internet communication, Latinized Georgian is becoming increasingly uncommon and provokes user irritation.

Finally, we can conclude that the Georgian forum language is a hybrid register with written, spoken and electronic patterns that draws originality from its technological substrate, the situation of communication, linguistic features of Georgian and the user's personal identity.

მრავლობითის წარმოების ზოგიერთი საკითხი ქართული ენის მორფოლოგიურ პროცესორში

ნინო ამირეზაშვილი, ლიანა ლორთქიფანიძე, ლიანა სამსონაძე, ნინო ჯავაშვილი
საქართველოს ტექნიკური უნივერსიტეტი,
არჩილ ელიაშვილის მართვის სისტემების ინსტიტუტი (საქართველო)
l_lordkipanidze@yahoo.com

რუსთაველის ეროვნული ფონდის მხარდაჭერით იქმნება ქართულად წერის ხელშემწყობი კომპიუტერული სისტემა – **ქართული ენის კომპიუტერული სუფლიორი**[1]. მისი პროგრამული უზრუნველყოფისათვის აუცილებელი გახდა ქართული ენის მორფოლოგიური პროცესორის ავტომატური სისტემის რეალიზაცია. აღნიშნული პრობლემის გადასაწყვეტად გამოიყენება ცოდნის დაგროვების ექსპერტული სისტემა – **MESLM** (**M**ultilingual **E**xpert **S**ystem of **L**anguage **M**odelling)[2], რომელშიც ცოდნის შეძენა შესაძლებელია ენის ამოსავალი სიტყვების სტერეოტიპული პარადიგმების საფუძველზე.

ენობრივი მოვლენები გარკვეულ კანონზომიერებას ექვემდებარება, მაგრამ ზოგადი წესების გვერდით არსებობს გამონაკლისებიც. ამიტომ სირთულის მოსახსნელად ლექსიკონის შემდგენელი-ექსპერტი ხშირად იყენებს სინტაქსურ, სემანტიკურ ცოდნას, ხალხში ხშირად გამოყენებულ (შეიძლება არასწორად წარმოებულ, მაგრამ უკვე დამკვიდრებულ) ფორმებს. ამჯერად ყურადღებას გავამახვილებთ ზოგიერთი სახელის მრავლობითი რიცხვის წარმოების საკითხზე. ბევრია ისეთი არათვლადი სახელი, რომელთა მრავლობით რიცხვში გამოყენება უკვე დამკვიდრებულია და არცთუ უმართებულოდ.

ვიცით, რომ ნივთიერებათა სახელებს მრავლობითი რიცხვი არა აქვს, თუმცა ენაში ისინი მაინც გვხვდება მრავლობითი რიცხვის ფორმით – მინერალური წყლები, ქართული ღვინოები, ნახშირწყლები, ოქროები, სითხეები და ა.შ. მართალია, იგულისხმება სხვადასხვა მინერალური წყალი, სხვადასხვა მარკის ღვინო და ა.შ., მაგრამ, როდესაც გვინდა ყველა არსებული ფორმის მორფოლოგიური პარადიგმის აგება, მაშინ წესი, რომ ნივთიერებათა სახელებს მრავლობითი რიცხვი არა აქვს, არ გამოგვადგება. მაგალითად, არსებითი სახელი "**მარილი**": ხომ არ ვიტყვით – საჭმელს მარილები აკლია, მაგრამ ვამბობთ – "**მარილები მაწუხებს**", "**მარილების დაგროვება**". ამის გამო, იძულებული ვართ **მარილი** და, ზოგადად, ასეთი ტიპის სახელები ვაწარმოოთ ებ-იანი მრავლობითის ფორმით.

გადახრა გვხვდება კრებითი სახელების მრავლობითი რიცხვის წარმოების დროსაც: გუნდი, ხალხი, ერი, ჯოგი და სხვ. წესით, მათ მრავლობითი არ უნდა ეწარმოებოდეთ, მაგრამ ვამბობთ – "ფეხბურთის გუნდები", "მსოფლიოს ხალხები" და სხვ.

როგორც აღვნიშნეთ, პარადიგმას ვაგებთ ორივე ტიპის მრავლობითში, თუმცა ზოგიერთ სახელს მხოლოდ ერთი მრავლობითის ფორმა ეწარმოება. იგივე „**ღვინო**", „**წვენი**" გვხვდება ებ-იან მრავლობითში – **ღვინოები, წვენები,** მაგრამ ნარ-თანიანში არავინ იტყვის **ღვინონი** ან წოდებით ბრუნვაში – **ღვინონო, წვენნო**. ადამიანის ორგანო **წელი** ნარ-თანიან

[1] ლორთქიფანიძე ლ., ქართული ენის „კომპიუტერული სუფლიორის" ხელშემწყობი ლექსიკონების პროგრამული უზრუნველყოფა. სსიპ არჩილ ელიაშვილის მართვის სისტემების ინსტიტუტის შრომათა კრებული №13, 2009წ.

2. ლორთქიფანიძე ლ., ენის მორფოლოგიის წარმოდგენა ექსპერტულ სისტემაში. სსიპ არჩილ ელიაშვილის მართვის სისტემების ინსტიტუტის შრომათა კრებული, თბილისი, 2007წ.

მრავლობითში ლამაზად ჟღერს – "**წელთა რხევანი**", მაგრამ ებ-იან მრავლობითში "წელებს" არ ვიტყვით (გამონათქვამი '**წელებზე ფეხს ვიდგამ**' იდიომატურია და წელებში ნაწლავებს გულისხმობენ). ასევეა სიტყვა „ცა", მიუხედავად იმისა, რომ დამკვიდრებულია გამონათქვამები – "მეშვიდე ცა", „მეცხრე ცა", ებ-იან მრავლობითში არ ვიტყვით ცები, თუმცა ვამბობთ – „ცათა სასუფეველი", დერივატი – „ცათამბჯენი" და ა.შ.

აბსტრაქტულ სახელებს, წესით, მრავლობითი რიცხვი არ აქვს, მაგრამ გამონაკლისი აქაც ბევრია და მათი მრავლობითის ფორმები ხშირად გვხვდება მხატვრულ ტექსტებსა თუ ზეპირმეტყველებაში. ესეც გავითვალისწინეთ პარადიგმების აგებისას და ამიტომ ისინი ცალკე ჯგუფად გამოვყავით და "აბსტრაქტული თვლადი სახელების" წესებში მოვათავსეთ. ჩამოვთვლით რამდენიმე მათგანს: ამინდები, ეფექტები, ექსცესები, ეჭვები, იდეები, იმედები, ინსტინქტები, კავშირები, კონკურსები, კონტრასტები, მაქსიმუმები, მინიმუმები, ცდომილებები, მომენტები, მოტივები, ფიქრები, ოცნებები, რეიდები, რიტუალები, თამაშები, ტყუილები, ვიზიტები, ვალდებულებები, ჩვეულებები, შეხედულებები, წვეულებები და სხვ; დერივატებიდან გამოვყოფთ: დღეობები, მკვლელობები, ნათლობები, ამხანაგობები, ნიშნობები ...

მრავლობითის წარმოებასთან დაკავშირებით უცხოური სიტყვების პარადიგმების აგებისას ერთი საყურადღებო შემთხვევაც გვხვდება: მართალია, სიტყვებს – მარკიზა, სენიორა (ფრ.), სინიორა (იტალ.) ქართულშიც გამოხატული აქვთ სქესი, მაგრამ ებ-იან მრავლობითში, ნარ-თანიანისგან განსხვავებით, ეს მნიშვნელობა იკარგება.

როგორც ვიცით, ქართულში, ისე როგორც მრავალ სხვა ენაში, არის კატეგორია სიტყვებისა, რომელთაც მხოლოდ მრავლობითის ფორმა აქვს (Pluralia Tantum). ქართულში ასეთი სიტყვების საკმაოდ მცირეა – ბატონები, ყვავილბატონები, არდადეგები ...

სიტყვა „არჩევნები" მხოლოობით „არჩევანთან" ერთად უკვე დამკვიდრდა ისეთ გამოთქმებში, როგორებიცაა საპარლამენტო არჩევნები, საპრეზიდენტო არჩევნები და სხვ.

სიტყვებს, რომლებიც უმეტეს ლექსიკონებში მრავლობითის ფორმითაა შეტანილი, ჩვენც შევუნარჩუნეთ მრავლობითის ფორმა და გავუკეთეთ ებ-იანი მრავლობითის პარადიგმა. ასეთი სიტყვებიდან ზოგი აღნიშნავს რაიმე გაერთიანებას, კავშირს და მივაკუთვნეთ კრებითებს (ბითლები [The beatles], პრიმატები [primatus] – ძუძუმწოვართა ჯგუფი – მაიმუნები, ნახევრადმაიმუნები, ადამიანები, და ა.შ).

გეოგრაფიულ სახელებთან დაკავშირებით გამოსაცალკევებელი იყო მრავლობითის ფორმით არსებული დასახელებები, როგორებიცაა: კორდილიერები, ანდები, ალპები, იდერლანდები და სხვ. საქართველოში გვაქვს ბაგები, შავშები, ნაფეტვრები და სხვ.

ხშირია ისეთი მითოლოგიური შინაარსის სახელები, რომლებიც აღნიშნავენ ისეთ სიმრავლეს, რომელთაგან ერთეულის გამოცალკევება შეუძლებელია (ალოადები – პოსეიდონის შვილიშვილების საერთო სახელი, დანაიდები – დანაოსის 50 ქალიშვილის საერთო სახელწოდება, დიოსკურები – ზევსის ტყუპების ზედმეტი სახელი და ა.შ.).

ზოგჯერ გაარსებითებულია ზედსართავი სახელი მრავლობითი რიცხვის ფორმით (წითლები, მწვანეები, მდიდრები, ღარიბები და ა.შ.)

მართალია, საწყისი მრავლობით რიცხვს არ აწარმოებს, მაგრამ ხშირად გვხვდება გამონათქვამები: შეძახილები გაისმა; შელოცვებმა უშველა და სხვ.

გამოიყო მიმღეობების ერთი ტიპი, რომელსაც არ ეწარმოება მრავლობითის ფორმა. ესენია -ებ თემისნიშნიანი მიმღეობები (დანატოვები, დანაბარები, განაცხელები, დასაღონები, დასამონები, ნაწამები და ა.შ.).

მრავლობითის ფორმები არა აქვს წინა ვითარების ზედსართავ სახელებსაც, რომლებიც ნაწარმოებია ნა-ებ აფიქსებით (ნახილები, ნაყანები, ნააბანოები, ნავაჟკაცები და ა.შ.).

Some Questions of the Formation of the Plural in the Georgian Morphological Processor

Nino Amirezashvili, Nino Javashvili, Liana Lortkipanidze, Liana Samsonadze
Archil Eliashvili Institute of Control Systems, Georgian Technical University (Georgia)
l_lordkipanidze@yahoo.com

The computer system – Georgian Computer Prompter for the disabled which supports writing in the Georgian language, is being created with the support of the Rustaveli National Scientific Foundation[1]. For the programm realisation of the system, it became necessary to develop the Georgian morphological processor. To solve this problem, **MESLM (M**ultilingual **E**xpert **S**ystem of **L**anguage **M**odelling) was used[2]. This system makes it possible to acquire certain knowledge on the basis of stereotypical word paradigms.

The language phenomena are subject to particular laws, but beside the general rules, there are some exceptions. That's why the dictionary experts, in order to remove such complexities, often use syntactic, semantic knowledge, forms used in everyday language (which may sometimes be incorrect, but which have already become the part of the system of the language). Currently, we are focusing on the problems which emerge while producing plural noun forms. There are number of uncountable nouns which are still used in the plural.

As is known, material nouns are uncountable and thus they are not usually used in the plural. However, we still use: *'mineraluri ცქl-eb-i' – mineral waters, 'kartuli ɣvino-eb-i' – Georgian wines, 'naxširცქl-eb-i' –carbohydrates, 'okro-eb-i' – gold, 'sitxe-eb-i' – liquids,* etc. Although different kinds of mineral waters or different sorts of wine are meant here, when we want to enumerate all the existing forms of the morphological paradigm, we cannot apply the rule according to which material nouns have no plural forms. For example: the noun "salt". We don't say - '*sadils marilebi aklia' - there are not enough salts in the food. However,* we can still say – *'marilebi maცuxebs' – I suffer from salt (disease is meant here). 'marilebis dagroveba' – adjournment of salt*. That's why we are forced to form the plural forms of nouns similar to "salt" in this way.

Deviation from this rule occurs while forming plural forms of collective nouns: team, people, nation, troop, etc. As a rule, they do not form plural forms, but we say *'pexburtis gundebi' – football teams, 'msoplios xalxebi' - peoples of the world, 'sxvadasxva erebi'- various nations, etc.*

As we have mentioned, we build both kinds of plural paradigms, but some nouns form only one

[1] L. Lortkipanidze, The computer realization of applied dictionaries of "computer prompter" for Georgian language. Proceedings of the LEPL Archil Eliashvili Institute of Control Systems №13 2009.
[2] Lortkipanidze L., Presentation of Language Morphology in the Expert System. Proceedings of the LEPL Archil Eliashvili Institute of Control Systems, Tbilisi, 2007.

type of the plural. The words *‘ɣvino’ - wine, ’ɕveni’ - juice* – produce plural forms with -eb suffix, but not with -n, -ta affixes (old ones) – *ɣvino-eb-i, ɕven-eb-i* but nobody says *ɣvino-n-i* or *ɕven-n-i*. The part of body – *‘ɕeli’ - waist* sounds very nice with -n, -ta suffixes – *‘ɕel-ta rxevani’ –a certain way of walking is implied*, but in the expression *‘ɕelebze fexs vidgam’ – ɕelebze* is a dialect form of *‘ɕeli’* and means “guts”. The word *‘ca’ – sky* can be considered. We can say *‘mešvide ca’ – the seventh sky, ‘mecxre ca’ – the ninth sky*, but we can’t say *‘cebi’ – skies*. However, we say *‘ca-ta sasupeveli’ – the kingdom of heaven/God*, the derivate *‘catambƺeni’ – skyscraper etc*.

As a rule, abstract nouns are uncountable and thus they do not from plural forms, but there are a lot of exceptions and their plural forms occur both in literature and in ordinary speech. We took this fact into consideration and while building paradigms we marked them out as a separate group and placed them into the rules of the “uncountable abstract nouns”. There are some examples: *‘amind-eb-i’* – weathers, *‘efeqṭ-eb-i’* – effects, *‘eqsces-eb-i’* – excesses, *‘ide-eb-i’* – ideas, *‘imed-eb-i’* – hopes, *‘insʃinkṭ-eb-i’* – instincts, *‘kavšir-eb-i’* – connections, *‘ḳonḳurs-eb-i’* – competitions, *‘ḳonṭrasṭ-eb-i’* – contrasts, *‘maksimum-eb-i’* – maximums, *‘minimum-eb-i’* – minimums, *‘momenṭ-eb-i’* – moments, *‘moṭiv-eb-i’* – motives, *‘pikr-eb-i’* – thoughts, *‘ocneb-eb-i’* – dreams, *‘reid-eb-i’* – roadsteads, *‘riṭual-eb-i’* – rituals, *‘tamaš-eb-i’* – games, *‘viziṭ-eb-i’* – visits, *‘ɕes-eb-i’* – rules, *‘valdebuleb-eb-i’* – obligations, *‘čveuleb-eb-i’* – customs, *‘šexeduleb-eb-i’* – opinions, *‘ɕveuleb-eb-i’* – parties, ...; From the derivates we emphasise: *‘dɣeob-eb-i’* – birthdays, *‘mḳvlelob-eb-i’* – homicides, *‘natlob-eb-i’* – christenings, *‘amxanagob-eb-i’* – comradeships, *‘nišnob-eb-i’* – engagements, ...

While forming the paradigms of several words of foreign origin, one interesting moment emerged: Although in the Georgian equivalents of the following words – *‘markiza’* – marchioness, *‘seniora’* – seignior (Fr.), *‘siniora’* – seignior (Ital. express gender is expressed though in the plural forms with -eb suffix, as opposed to the forms with -na, -ta affixes, this meaning has been lost.

As we know, in Georgian, as in other languages, there are some categories of words, which have only singular forms (Pluralia Tantum). In Georgian language such words are very few – *‘baton-eb-i’ – infectious disease, ‘ardadeg-eb-i’ – holidays, etc*.

The word ‘arčevn-*eb-i’ – elections* with the singular form *‘arčevani’* settled down in the expressions like: *parliamentarian elections, presidential elections, etc.*

The words which are fixed in the dictionaries with their plural forms, often mean , union of some kind are considered as collective nouns – *‘bitl-eb-i’ [The Beatles], ‘primaṭ-eb-i’ [primatus] – mammal groups, ‘maimun-eb-’ – monkeys, ‘naxevradmaimun-eb-i’ –half monkey, ‘adamian-eb-i’ – people and so on).*

As to the geographical names, many of them are in plural forms, like: *‘and-eb-i’ – Andes, ‘alp-eb-i’ – the Alps, ‘niderlandebi’ – the Netherlands*, etc. In Georgia we have *‘bag-eb-i’ – Bagebi, ‘šavš-eb-i’ – Shavshebi, ‘ napeṭvr-eb-i’ – Napetvrebi, etc.*

There are some mythological nouns used only by plural forms: *'aload-eb-i' – Poseidon grandchildren's common name, 'danaid-eb-i' – Danaos 50 daughter's common name, 'diosḳur-eb-i' – Zeus twins' nickname and so on.*

Sometimes the adjectives with plural forms become substantivised: *'ṭitl-eb-i' – the reds, 'mṭvane-eb-i' – the greens, 'mdidr-eb-i' – the rich, 'γarib-eb-i – the poor, etc.*

The Infinitive does not form the plural, but the following expressions are becoming common: *'darbev-eb-i daiṭqo' – the routs have begun; 'šeʒaxil-eb-i gaisma' – the shouts have been heard, 'šelocv-eb-ma ušvela' – the blessings have helped, etc.*

Adjectives, expressing previous states formed by adding affixes -na-eb, do not form the plural forms either: *'na-xil-eb-i' – used to be fruit, 'na-kan-eb-i' – used to be field of wheat, 'na-abano-eb-i' – used to be a bath, 'na-važḳac-eb-i' – used to be a brave man and so on.*

There are some gerunds, which do not form plural forms at all. These are the gerunds with -eb suffix: *'danaṭov-eb-i' – that is left, 'danabar-eb-i' – it is made a testament, 'ganacxel-eb-i' – that is heated, 'dasaγon-eb-i' – that makes somebody sad, dasamon-eb-i – that enslaves somebody, etc.*

ბიბლიოგრაფია / References

იმედაშვილი ი., უცხო სიტყვათა ლექსიკონი, თბილისი, 1928წ.
ლორთქიფანიძე ლ., ენის მორფოლოგიის წარმოდგენა ექსპერტულ სისტემაში, სსიპ არჩილ ელიაშვილის მართვის სისტემების ინსტიტუტის შრომათა კრებული, თბილისი, 2007წ.
ლორთქიფანიძე ლ., ქართული ენის „კომპიუტერული სუფლიორის" ხელშემწყობი ლექსიკონების პროგრამული უზრუნველყოფა, სსიპ არჩილ ელიაშვილის მართვის სისტემების ინსტიტუტის შრომათა კრებული №13, 2009წ.
მარგველანი ლ., ქართული ენის კომპიუტერული მოდელები, თბილისი, „ინტელექტი", 2008.
მითოლოგიური ლექსიკონი, თბილისი, "განათლება", 1972წ.
ქართული ენის განმარტებითი ლექსიკონი, "მეცნიერება", 1950-68წ.
შანიძე ა., ქართული გრამატიკის საფუძვლები, თბილისი, თბილისის სახელმწიფო უნივერსიტეტი, 1953წ.
თოფურია ვ., გიგინეიშვილი ივ., ქართული ენის ორთოგრაფიული ლექსიკონი, თბილისი, „განათლება" , 1998წ.
უცხო სიტყვათა ლექსიკონი, თბილისი, "განათლება", 1973წ.

წესებზე დაფუძნებული დაპროგრამების გამოყენება ქართული ტექსტების კომპიუტერული მორფოლოგიური, სინტაქსური და სემანტიკური ანალიზისათვის

ჯემალ ანთიძე, ნანა გულუა, დამანა მელიქიშვილი
სოხუმის სახელმწიფო უნივერსიტეტი,
ივ. ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)
jeantidze@yahoo.com, damanamel@yahoo.com

ქართული ტექსტების კომპიუტერული მორფოლოგიური, სინტაქსური და სემანტიკური ანალიზი არის ერთ-ერთი მთავარი კომპონენტი ისეთი პრობლემის გადასაწყვეტად, როგორიცაა მანქანური თარგმანი ქართული ენიდან სხვა ენებზე; ასევე ქართული ტექსტების ორთოგრაფიის ავტომატური შემოწმება და ხელოვნური ინტელექტის სხვა პრობლემები, რომლებიც მოითხოვენ ქართული ტექსტების კომპიუტერულ დამუშავებას. ქართული ტექსტების სრული კომპიუტერული მორფოლოგიური, სინტაქსური და სემანტიკური ანალიზის სისტემა ჯერჯერობით არ არსებობს. ზემოთ აღნიშნული პრობლემის გადაწყვეტა საშური საქმეა, თუ გვსურს გამოვიყენოთ ქართული ენა კომპიუტერთან კომუნიკაციისათვის.

სასრული ავტომატის გამოყენება სრული მორფოლოგიური ანალიზის პრობლემის გადასაწყვეტად, რაც ფართოდაა გამოყენებული დასავლეთ ევროპის ენებისათვის, შეუძლებელია ([1-2]). მეორე მხრივ, სრული ძებნის ალგორითმის გამოყენება ანელებს მორფოლოგიური ანალიზის პროცესს. ამიტომ ჩვენ გამოვიყენეთ მეთოდი, რომელიც აჩქარებს ანალიზის პროცესს სრული ძებნის ალგორითმთან შედარებით([1]). ეს მეთოდი იყენებს შეზღუდვებს იმისათვის, რომ დავადგინოთ მორფემების სწორი შერჩევა. მორფოლოგიური ანალიზატორი ამოწმებს სიტყვიდან უკვე გამოყოფილ სავარაუდო მორფემებს, რამდენად აკმაყოფილებს ის შეზღუდვებს. თუ შეზღუდვა დაკმაყოფილებულია, ანალიზატორი აგრძელებს სხვა მორფემების გამოყოფას. წინააღმდეგ შემთხვევაში, იგი ასრულებს უკუსვლას ახალი ალტერნატივის მოსაძებნად და უკუაგდებს ბოლოს გამოყოფილ მორფემას. ასეთი გზით არასწორი ალტერნატივების უკუგდება ხდება წინასწარ, რაც აჩქარებს ძებნის პროცესს. შეზღუდვები არის ლოგიკური გამოსახულებები, რომლებიც შეგვიძლია შევადგინოთ მორფემათა თვისებებისაგან. ანალიზატორი ამოწმებს გამოყოფილი მორფემის თვისებას, აქვს თუ არა შეზღუდვაში მოცემული მნიშვნელობა, რომელიც განსაზღვრავს გამოყოფის სისწორეს. მორფემის თვისებების მნიშვნელობებს ვადგენთ ქართული ენის მორფოლოგიის მიხედვით.

სრულ კომპიუტერულ მორფოლოგიურ ანალიზში ვგულისხმობთ სიტყვაფორმის ყველა სწორ დაშლას მორფემებად და ყოველი დაშლისათვის მორფოლოგიური კატეგორიების დადგენას. ჩვენ ქართული სიტყვების სრული მორფოლოგიური ანალიზი ამ ანალიზატორით განვახორციელეთ ([3-5]).

ამ სისტემის შესადგენად გამოვიყენეთ შემდეგი ძირითადი მეთოდები და ალგორითმები: თვისებათა სტრუქტურებისათვის განსაზღვრული ოპერაციები; ძებნის ალგორითმი (მორფოლოგიური ანალიზატორისათვის); ზოგადი სინტაქსური გარჩევის ალგორითმი კონტექსტისაგან თავისუფალი გრამატიკისათვის და თვისებებით შეზღუდვების შედგენის მეთოდი. თვისებათა სტრუქტურები ფართოდ გამოიყენება ანალიზის ყველა დონეზე. ჩვენ ვიყენებთ მათ სხვადასხვა ინფორმაციის შესანახავად ლექსიკონურ ერთეულებში და ანალი-

ზის შედეგად მიღებული ინფორმაციის შესანახავად. მორფოლოგიური, სინტაქსური ან სემანტიკური წესის ყოველ სიმბოლოს შეიძლება ჰქონდეს მასთან დაკავშირებულ თვისებათა სტრუქტურა, რომელსაც დასაწყისში ვავსებთ ლექსიკონიდან, ან შეიძლება სისტემა შეივსოს ანალიზის წინა დონეებიდან. თვისებათა სტრუქტურებს და მათზე განსაზღვრულ ოპერაციებს ვიყენებთ თვისებათა შეზღუდვების შესადგენად. ზოგადი გარჩევის ალგორითმით შესაძლებელია მივიღოთ კონტექსტისაგან თავისუფალი გრამატიკით განსაზღვრული წინადადების სინტაქსური და სემანტიკური ანალიზი და ერთდროულად შევამოწმოთ თვისებათა შეზღუდვები, რომლებიც შეიძლება უკავშირდებოდეს გრამატიკის წესებს. თუ შეზღუდვა არ კმაყოფილდება ანალიზის დროს, მაშინ სისტემა უკუაგდებს განსახილველ წესს და აგრძელებს ძებნის პროცესს. შეგვიძლია აგრეთვე გამოვიყენოთ თვისებათა შეზღუდვები მორფოლოგიურ წესებში, ოღონდ, სინტაქსური ან სემანტიკური წესებისაგან განსხვავებით, შეზღუდვები შეგვიძლია მოვათავსოთ წესის ნებისმიერ ადგილას, და არა მხოლოდ ბოლოში. ეს აჩქარებს მორფოლოგიურ ანალიზს, რადგან სისტემა ამოწმებს შეზღუდვებს ადრეულ ეტაპზე და უკუაგდებს სიტყაფორმის არასწორ დაშლას მორფემებად დროულად ([6-7]).

მორფოლოგიურ წესებს ვსაზღვრავთ შემდეგი სახით:

word -> M_1 { C_1 } M_2 { C_2 } . . . M_N { C_N }

სადაც M_i მორფემათა კლასებია და C_i (i = 1, … , N) არასავალდებულო შეზღუდვებია.

სინტაქსურ წესებს აქვთ შემდეგი სახე:

S -> A_1 { C_1 } A_2 { C_2 } . . . A_N { C_N } ;
S -> A_1 A_2 . . . A_N : R { C } ;

სადაც, S არის LHS არატერმინალური სიმბოლო, A_i (i = 1, … , N) არის RHS არატერმინალური ან ტერმინალური სიმბოლოები, C და C_i (i=1,...,N) შეზღუდვებია და R არის სიმბოლოს პოზიციის რეგულატორების სიმრავლე. პოზიციის რეგულატორები განსაზღვრავენ RHS სიმბოლოების რიგს წესში, შესაბამისად, ქმნიან სიმბოლოთა არაფიქსირებულ რიგს. არსებობს პოზიციის რეგულატორების ორი ტიპი:

$A_i < A_j$ ნიშნავს, რომ A_i სიმბოლო უნდა იყოს მოთავსებული სადმე A_j–ს წინ;

$A_i – A_j$ ნიშნავს, რომ A_i სიმბოლო უნდა იყოს მოთავსებული უშუალოდ A_j–ს წინ([1]).

ფორმალიზმი, რომელიც ჩვენ გამოვიყენეთ მორფოლოგიური, სინტაქსური და სემანტიკური ანალიზისათვის, ძალზე მოხერხებულია. მას აქვს მრავალი კონსტრუქცია, რომლებიც აადვილებენ გრამატიკის ფაილის დაწერას. მორფოლოგიურ ანალიზატორს აქვს პრეპროცესორი. ანალიზატორი იყენებს STL სტანდარტულ ბიბლიოთეკას. ეს სისტემა მუშაობს UNIX და WINDOWS ოპერაციულ სისტემებში. ჩვენ შეგვიძლია მისი კომპილაცია და გამოყენება ნებისმიერ სხვა პლატფორმაზე, რომელსაც აქვს თანამედროვე c++ კომპილატორი.

ქართული ტექსტების ჩვენ მიერ განხორციელებულ ექსპერიმენტულ მორფოსინტაქსურ და სემანტიკურ ანალიზს ეს სისტემა უდევს საფუძვლად.

Towards the Use of Rule-Based Programming for Computer Morphological, Syntactic and Semantic Analyses of the Georgian Texts

Jemal Antidze, Nana Gulua, Damana Melikishvili
Sokhumi State University,
Iv. Javakhishvili Tbilisi State University (Georgia)
jeantidze@yahoo.com, damanamel@yahoo.com

The computer morphological, syntactic and semantic analysis of Georgian texts is one of the main components for solving such problems as machine translation from the Georgian language into the other languages, as well as the automated checking of orthography of Georgian texts, and some problems of artificial intelligence, which require computer processing of Georgian texts. The system for complete computer morphological, syntactic and semantic analysis of Georgian texts does not exist yet. If we wish to use the Georgian language to communicate with the computer, the solution of the problem mentioned above is very urgent.

To solve the problem of morphological analysis by using finite automation, which is widely used for the languages from Western Europe is impossible ([1-2]). On the other hand, use of full search algorithm slows the process of morphological analysis. For this reason, we used the method, which is making the analysis process faster compared to the full search algorithm ([1]). This method uses constraints to establish the correct morpheme's selection. Having separated, presumable morphemes from word, the morphological analyser checks them for the satisfaction of their constraints. If the constraint is satisfied, the analyser continues separation of other morphemes. In the opposite case, it performs backtracking to search the new alternative and rejects the last separated morpheme. In this way, the removal of incorrect alternatives happens in advance and thus speeds up the searching process. The constraints are logical expressions, which we can compose from the features of morphemes. The analyser checks if the separated morpheme's feature has the value (given in constraint), which defines the correctness of the separation. We compose the values of morphemes' features according to morphology of the Georgian language.

Under complete computer morphological analysis, we understand all valid splitting of a word-form in morphemes and establishment of morphological categories for each splitting. We realized complete morphological analysis of Georgian words by the analyser ([3-5]).

Basic methods and algorithms, which we used to develop the system, are as follows: operations defined on features' structures; search algorithm (for the morphological analyser); general syntactic parsing algorithm for context free grammar and constraints construction method with the features' structures. The features' structures are widely used on all of the levels of analysis. We use them to hold various information about dictionary entries and information obtained during analysis. Each symbol defined in a morphological, syntactic or semantic rule can have an associated features' structure, which we initially fill from the dictionary, or the system itself fills them from the previous levels of analysis. we use the features' structures and operations defined on them to build up features' constraints. By a general

parsing algorithm, it is possible to get a syntactic and semantic analysis of any sentence defined by a context free grammar and simultaneously check features' constraints, which may be associated with grammatical rules. If the constraint is not satisfied during the analysis, then the system will reject the current rule and the search process will go on. The features' constraints can also be applied to morphological rules. However, unlike the syntactic rules, we can attach constraints at any place within a morphological rule, but at the end. This speeds up the process of morphological analysis, because the system checks constraints early and it rejects incorrect word-form's division into morphemes in a timely manner ([6-7]).

We define morphological rules in the following way:
word -> M_1 { C_1 } M_2 { C_2 } . . . M_N { C_N };
Where M_i are morpheme classes, and C_i (i = 1, … , N) are constraints (optional).

Syntactic and semantic rules have the form:
S -> A_1 { C_1 } A_2 { C_2 } . . . A_N { C_N };
S -> A_1 A_2 . . . A_N : R { C };
Where S is an LHS non-terminal symbol, Ai (i = 1, … , N) are RHS terminal or non-terminal symbols, C and Ci (i =1, …N) are constraints, and R is a set of symbol position regulators. Position regulators define order of RHS symbols in the rule, consequently making symbols non-fixed ordering. There are two types of position regulators:
$A_i < A_j$ means that symbol A_i must be placed somewhere before the symbol A_j;
$A_i – A_j$ means that symbol A_i must be placed exactly before the symbol A_j([1]).

Formalism, which we used for the morphological, syntactic and semantic analysis is highly comfortable for the human use. It has many constructions that make it easier to write a grammar file. The morphological analyser has a built-in preprocessor. The analyser utilizes the STL standard library. The system operates in UNIX and the Windows operating system. We can compile it and use it in any other platforms, which contain a modern C++ compiler. With this system, we have realised experimental morphological, syntactic and semantic analysis of Georgian texts.

ბიბლიოგრაფია / References

Antidze J., Gulua N., On Complete Computer Morphological and Syntactic Analysis of Georgian Texts, Proceedings of the Seventh Inernational Conference "Internet – Education – Science", vol. 1(11), pages 214-217, Vynnitsia, Ukraine, 2010. http://fpv.science.tsu.ge/vinitsa1.mht

Antidze J., Theory of formal languages and grammars, natural languages computer modeling, Nekeri, Tbilisi, 2009, 254 pages.

Antidze J., Gulua N., Mishelashvili D., Nukradze L., On Complete Computer Morphological and Syntactic Analysis of Georgian Texts, Report of International Symposium – Natural Language Processing, Georgian Language and Comuter Technologies, Institute of Linguistics of Georgian Academy of Sciences, Tbilisi, 2009. http://www.ice.ge/conferenciebi/Conf_Fs.html

Melikishvili D., System of Georgian verbs conjugation. Logos Press. Tbilisi, 2001. 310 pages.

Melikishvili D., Hamfris D., Kufunia M., The Georgian Verb: A Morphosyntactic Analysis, Dunwoody Press, USA, 2008, 723 pages.
Antidze J. Gulua N., Software for Processing of Natural Language Texts, Proceedings of Third International Conference "Problems of Cybernetics and Informatics", Volume 1, pages 114-117, Baku, 2010. http://fpv.science.tsu.ge/baku.mht
Antidze J., Gulua N., Software Tools for Computer Realization of Morphological and Syntactic Models of Georgian Texts, International Scientific-Technical Journal - Optoelectronic Information-Power Technologies, #2(20), 2010, pages 98-105.

ქართული ენის თესაურუსი

ავთანდილ არაბული, რუსუდან ასათიანი, მარინე ივანიშვილი,
ეთერ სოსელია, გია შერვაშიძე
საქართველოს მეცნიერებათა ეროვნული აკადემია (საქართველო)
rus_asatiani@hotmail.com

ქართული კულტურისათვის ერთ-ერთ უმნიშვნელოვანეს ამოცანას წარმოადგენს ქართული ენის ისტორიული განვითარების შესწავლა თხუთმეტსაუკუნოვანი მწერლობის ძეგლების საფუძველზე, რისთვისაც შემუშავდა პროექტი „ქართული ენის თესაურუსი“.

პროექტის მიზანია ქართული ენის წერილობით ფიქსირებული პერიოდის კომპიუტერიზებული ლექსიკური არქივის შექმნა და პარალელურად მისი მომზადება გამოსაქვეყნებლად ელექტრონული ვერსიის სახით.

აღნიშნული მიზნის მისაღწევად შესრულდა შემდეგი სამუშაო:

- შედგა ქართული მწერლობის სრული ბიბლიოგრაფია;
- შეიქმნა თესაურუსის მონაცემთა ბაზა;
- შემუშავდა სპეციალური კომპიუტერული პროგრამა თავისი პორტალით, საძიებო სისტემით; ასევე, კონტექსტებად ტექსტების დაშლისა და ლემების ავტომატური გამოყოფის სისტემით (იხ. www.saunje.nekeri.net).
- სალექსიკონო ერთეულები (ლემები) დალაგდა ანბანური წესით, სათანადო ლექსიკური ერთეულის ყველა წერილობით დადასტურებული სიტყვაფორმის საილუსტრაციო კონტექსტის მოტანით;
- ეს სიტყვაფორმები სათანადო ლექსიკური ერთეულის „ბუდეში“ ასევე ანბანური რიგით განლაგდა;
- დათარიღებული ხელნაწერებიდან და გამოცემული ტექსტებიდან ამოღებული შესაბამისი საილუსტრაციო მასალა განაწილდა ქრონოლოგიურად;.
- სიტყვაფორმა „ლექსიკონში“ დაფიქსირდა იმ ფორმით, როგორითაც ის სათანადო ხელნაწერებშია დადასტურებული – შემოკლებებისა და ქარაგმების გათვალისწინებით.

მიუხედავად იმისა, რომ პროექტი მეტად შრომატევადია და დიდ დროს მოითხოვს, მისი განხორციელება საკმაოდ რეალურად გვესახება. ამგვარ „ლექსიკონში“ წარმოდგენილი იქნება ქართული ენის უმდიდრესი მრავალსაუკუნოვანი სიტყვიერი მარაგი „ქართული ენის თესაურუსი“, ანუ „ქართული ენის სრული ლექსიკური საუნჯე“. ის იქნება ანალოგი საქვეყნოდ ცნობილი ბერძნული, ლათინური, ინგლისური, ფრანგული, გერმანული, იტალიური, ებრაული და ზოგი სხვა ენის „თესაურუსებისა“, რომლებზედაც გასულ საუკუნეში დაწყებული მუშაობა დღემდე გრძელდება.

„ქართული ენის თესაურუსის“ საფუძველზე შესაძლებელი გახდება შემდგომში ქართული ენის სხვადასხვა ხასიათის სპეციალური ლექსიკონებისა თუ ენობრივი (კომპიუტერული) მოდელების შექმნა; მაგალითად:

- ქართული ენის ისტორიული და ისტორიულ-ეტიმოლოგიური ლექსიკონები;
- ქართული ენის ონტოლოგიური და განმარტებითი (კომპიუტერული) ლექსიკონები;
- ქართული მართლწერის (ორთოგრაფიის) კომპიუტერული შემოწმება;

- ქართული ენის მორფოლოგიური (კომპიუტერული) მოდელი;
- ქართული ენის სინტაქსური (კომპიუტერული) მოდელი და სხვ.

წარმოდგენილი პროექტი გზას გაუხსნის და საფუძველს ჩაუყრის მრავალ ახალ აქტუალურ პროექტსა თუ წამოწყებას, დააინტერესებს ფართო საზოგადოებას და მეცნიერთა წრეს არა მხოლოდ საქართველოში, არამედ საზღვარგარეთაც.

Thesaurus of the Georgian Language

Avtandil Arabuli, Rusudan Asatiani, Marine Ivanishvili, Eter Soselia, Gia Shervashidze
The Georgian National Academy of Sciences (Georgia)
rus_asatiani@hotmail.com

The Georgian language is one of the South-Caucasian languages with the oldest literary traditions. The Georgian script was devised around 400 A.D. in order to facilitate the dissemination of Christian literature and one of the most important current tasks of the preservation of the Georgian cultural heritage is the study of the historical development of the Georgian language based on the analysis of the Georgian sixteenth-century-old literary monuments. To this end, it seems essential to publish *Thesauri,* and the project *Thesaurus of the Georgian Language* serves this purpose. The goal of the project is a creation of a Thesaurus database – the computer processing of a computerised lexical archive.

Due to the software in which the program *Thesaurus of the Georgian Language* (www.saunje.nekeri.net) is ultimately created, lexical items are arranged in an alphabetic order, each of them being provided with its all attested word-forms together with the illustrative contexts; the word-forms themselves are arranged according to the alphabet in the "nest" for each lexical item; as for illustrative material from the dated manuscripts and published texts, they are arranged chronologically. In this way, it is possible to define more exactly the appearance of a certain word in Georgian and to establish the way of its historical development. Besides, a word-form is given in the form it occurs in respective original manuscripts, preserving the *abbreviations* and *spelling specificities*.

The *Thesaurus of the Georgian Language* will be a significant acquisition not only for Georgian philology, Georgian linguistics and Kartvelology, but also for Georgian culture in general, and it will pave the way to many other projects of the utmost importance: ontological, concise, historical, etymological dictionaries, various kinds of morphological and syntax models for computer processing, etc. It attracts a widespread public and scientific interest of scholars in Georgia as well as abroad. The software product of the *Thesaurus of the Georgian Language* has been produced, and on the grounds of its database a large number of other electronic programs can also be created.

ქართული დიალექტური კორპუსი (ქდკ) კორპუსული ლინგვისტიკის გამოცდილების ფონზე

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

1. შესავალი:

კომპიუტერული ლინგვისტიკა XXI საუკუნის დიდ კულტურულ და სამეცნიერო გამოწვევად იქცა. მან ახალ ტექნოლოგიურ ერაში ახალი კუთხით წარმოაჩინა ჰუმანიტარული და ტექნიკური აზროვნების სინთეზის გარდაუვალობა. კომპიუტერული ლინგვისტიკის ნაწილია კორპუსის ლინგვისტიკა, რომელიც ამუშავებს ლინგვისტური ტექსტური კორპუსების აგებისა და გამოყენების საერთო პრინციპებს კომპიუტერული ტექნოლოგიების საშუალებით.

არსებობის ნახევარსაუკუნოვანი ისტორიის მანძილზე გამოიკვეთა კორპუსის ლინგვისტიკის "სოლიდარულობის" მაღალი ხარისხი (გამოცდილების ურთიერთგაცვლა და გაზიარება; კორპორაციული მუშაობის აუცილებლობა...) და პროგრესის სწრაფი ტემპი. ამიტომ, შეიძლება ითქვას, რომ არ არსებობენ "დამწყები" კორპუსისტები: მათ, ვინც ახლადა ერთვება ენის კორპუსული დამუშავების პროცესში (კორპუსების შექმნა და კვლევა), აქვთ ფუფუნება, ხელთ ჰქონდეთ წინამორბედების უმდიდრესი გამოცდილება და მათზე ბევრად უკეთესი ტექნოლოგიური ბაზა.

2. ისტორიიდან

როგორც ცნობილია, პირველი კომპიუტერული კორპუსი იყო ბრაუნის კორპუსი (აშშ, ბრაუნის უნივერსიტეტი). მისი შემქმნელების, კორპუსის ლინგვისტიკის "პიონერების", ნელსონ ფრენსისისა და ჰენრი კუჩერას, ნაშრომებმა მისცეს მაგალითი სხვა კორპუსის შემქმნელებს.[1]"

ბრაუნის კორპუსის უშუალო მემკვიდრეა ***LOB*** *(ლანკასტერ-ოსლო-ბერგენი)* **კორპუსი** – ბრაუნის კორპუსის "ბრიტანული ასლი", რომლის შექმნის ინიციატივა ეკუთვნის ჯეფრი ლიჩს. ამას მოჰყვა ***BNC*** – ბრიტანული ნაციონალური კორპუსი, ბრიტანული კორპუსის მაგალითზე შეიქმნა ამერიკული ნაციონალური კორპუსი (ANC), რომელიც ეყრდნობა ლინგვისტური მონაცემების კონსორციუმის (LDC) მასალებს და ა.შ. ინგლისურენოვანი კორპუსების პარალელურად ევროპისა და აზიის მრავალ ენაზე იქმნება სხვადასხვა მოცულობისა და სპეციფიკის კორპუსები.

კორპუსის ლინგვისტიკის განვითარების საწყის ეტაპზე წარმოქმნილი პრობლემები უკავშირდებოდა არა მარტო უშუალოდ სამუშაო პროცესს, არამედ ორგანიზაციულ და იურიდიულ საკითხებსაც. ერთ-ერთი ყველაზე რთული და საპასუხიმგებლო იყო საავტორო უფლების პრობლემა. ასევე რთული იყო კორპუსის, როგორც კულტურული იდეის დამკვიდრება და პოპულარიზაცია. აღსანიშნავია, რომ პიონერ კორპუსისტთა "რიტორიკა" დასაწყის ეტაპზე თითქმის ერთნაირი იყო ყველა ქვეყანაში.

1977 წლის თებერვალში ოსლოში შეიკრიბა მკვლევართა მცირე ჯგუფი ამ საკითხებს განსახილველად. აქვე შეიქმნა ცენტრი – ჩანასახი თანამედროვე ინგლისური ენის

[1] Johansson S., From Brown to LOB, ICAME Journal No. 20,1996, 100–104.

კომპიუტერული არქივისა (ICAME– International Computer Archive of Modern English). მისი უმთავრესი მიზნები იყო: მასალის შეკრება, ხელმისაწვდომობა კომპიუტერული დამუშავებისათვის, ინფორმაციის გავრცელება ლინგვისტური კვლევის შესახებ, ტექსტების არქივის თავმოყრა ბერგენის უნივერსიტეტში, ძალების კოორდინაციის ორგანიზება, კვლევების დუბლირების აცილება და სხვ.

ICAME–ის დაარსების მაცნე დოკუმენტი გავრცელდა და გადაეცა ამ დარგში მოღვაწე მეცნიერებს იმისთვის, რომ მხარი დაეჭირათ და ოფიციალურად "მიერთებოდნენ" ICAME პროექტს. ამ პროექტის შედეგიანობაზე მეტყველებს ის ფაქტი, რომ დღეისათვის ICAME-ს ეყრდნობა სამწერლო ინგლისურის სულ ცოტა 9 კორპუსი.

მოვლენები მსგავსად ვითარდებოდა პოსტსაბჭოთა სივრცეშიც. ასეთივე აუცილებლობად იქნა აღქმული რუსული ენის მანქანური ფონდის შექმნა თავის დროზე[1], მოგვიანებით კი შეიქმნა რუსული ნაციონალური კორპუსი (НКРЯ)[2].

ეს გამოცდილება დასაყრდენად იქცა ყველა იმ ქვეყნისთვის, რომელმაც "კორპუსული" გამოწვევის სირთულეები მიიღო და წარმატებით გაუმკლავდა მათ.

დღეისათვის უკვე არსებობს მრავალი ენის ნაციონალური კორპუსი, არსებობს პერიოდული გამოცემები (Internacional Journal of Corpus Linguistics; Corpora; Corpus Linguistics and Linguistic Theory; ICAME Journal; ACL...). დარგის განვითარებაში მნიშვნელოვანი წვლილი შეიტანა კომპიუტერული ლინგვისტიკის ასოციაციამ (ACL).

პოსტსაბჭოთა სივრცეში კორპუსული ლინგვისტიკის თეორიული და პრაქტიკული პრობლემები განიხილება სემინარებსა და სამეცნიერო კონფერენციებზე გამოყენებით კომპიუტერულ ლინგვისტიკაში: „დიალოგი", „მეგალინგი", „კომპიუტერული ლინგვისტიკა" და ა. შ. საქართველოში 1998 წლიდან იმართება საერთაშორისო კონფერენცია: „ქართული ენა და კომპიუტერული ტექნოლოგიები".

3. ქდკ – პირველი ქართული კორპუსული პროექტი

ქდკ პირველი ქართული კორპუსული პროექტია. ის იმთავითვე ჩაფიქრებული იყო როგორც დიდი ტექსტების კორპუსის ერთი სუბკორპუსი. ჩვენ ვქმნით კორპუსის სტრუქტურას, კორპუსში ანოტირების პრინციპებსა და ტექსტურ კოლექციას ერთდროულად. ამასთან, გვიწევს ყველა ეტაპის სამუშაოს შესრულება: მასალის მოპოვება, გაშიფვრა, ძველი ტექსტების რედაქტირება, დიგიტალიზაცია, უნიფიკაცია, კორპუსში ინტეგრირება და სხვ.

ამ ეტაპზე კორპუსი "ვითარდება" ორი მიმართულებით: ივსება ტექსტური ბაზა (ქართული ენის 20-მდე დიალექტის მონაცემები) და მუშავდება პირველადი მორფოლოგიური ანოტირების სისტემა. გარდა ტექსტური მასალისა, ქართული დიალექტების კორპუსში გათვალისწინებულია ლექსიკოგრაფიული და ენციკლოპედიური მასალის (დიალექტური ლექსიკონები, დარგობლივი ტერმინოლოგიის მასალები) ჩართვა.

გარდა ამისა, ვმუშაობთ დიალექტური კორპუსის "სასწავლო სუბკორპუსზე", რომელიც ითვალისწინებს საზღვარგარეთ (ირანი, თურქეთი, აზერბაიჯანი) მცხოვრები ქართველებისთვის ქართული ენის საკუთარი დიალექტის დახმარებით გაცნობასა და სწავლებას.

სამუშაო უპრეცედენტო მასშტაბისაა ქართული სინამდვილისთვის. სწორედ კორპუსის ლინგვისტიკის "სოლიდარულობა" და ამ დარგის მიღწევების ხელმისაწვდომობა გვაბედვინებს მივიღოთ ეს მნიშვნელოვანი გამოწვევა.

[1] Ершов А, Машинный фонд русского языка _ внешняя постановка. Текст машинописный _http://ershov.iis/archive/1984.

[2] Сичинава Д., Национальный корпус русского языка: 2003-2005. М.: Индрик, 2005, 21-30.

ჩვენთვის სახელმძღვანელოდ იქცა პიონერი კორპუსისტების გამოცდილება:

”როცა მუშაობას იწყებდნენ, ფრენსისი და კუჩერა არაპოპულარულები იყვნენ. ახლა, როცა ”კორპუსი გაბატონებულ ტენდენციად იქცა”, უნდა გვახსოვდეს, რომ არ შეიძლება მხოლოდ დინების მიმართულებით სვლა... რაც ჩვენი პიონერებისგან უნდა ვისწავლოთ – ეს არის საკუთარი დამოუკიდებელი აზრის პატივისცემა და სიმამაცე, რომ ნაკადის საწინააღმდეგოდ წავიდეთ[1].“

Georgian Dialect Corpus (GDC) against the Backdrop of the experience of Corpus Linguistics

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

1. Introduction

1. Introduction
Computational linguistics has become a major cultural and scientific challenge of the XXI century, which has revealed the inevitability of joint development of humanitarian and technical thought in the new Technological Era. Corpus linguistics as a part of computational linguistics elaborates the general principles of the composition and use of linguistic text corpora based on computational technologies.

50 years of the existence of corpus linguistics have demonstrated a high degree of “solidarity” among the scholars within the sphere (sharing experience, close cooperation between different institutions, etc.) and equally high rate of progress. Thus, it can be stated that nowadays there are no “beginners” in the field of corpus linguistics, as far as those who engage in this area of research (composition and study of corpora) can rely on the ample experience of their predecessors as well as much better technological basis.

2. Background

As is known, the first computational corpus was the Brown Corpus (The Brown University Standard Corpus of Present-Day American English, USA). The works of its compilers, “pioneers of Corpus Linguistics W.N. Francis and H. Kucera gave a good example to other scholars working in the field[2]“

[1] Johansson S., From Brown to LOB, ICAME Journal No. 20,1996, 100–104.
[2] ibid.

The direct heir of the Brown Corpus is LOB (Lancaster-Oslo- Bergen) corpus which is considered a "British copy" of the Brown Corpus, the initiator of which is J. Leech. This corpus was followed by BNC (British National Corpus) which, in its turn, served as a model for the American National Corpus (ANC). ANC is based on the data of Linguistic Data Consortium (LDC). As well as English data, corpora of various volumes and specifics were complied in quite a few European and Asian languages.

The issues which emerged at the initial stage of the development of Corpus Linguistics were connected not only to the process of work itself, but to organisational and legal issues as well. The most complicated and dangerous responsibilities also included issues of the copyright infringement, cultural aspects of the concept of the corpus, its establishment and dissemination. It is worth noting that the "rhetoric" of the first generation of scholars (the "Pioneers") was similar in all countries.

On February, 1977 a small group of scholars gathered in Oslo to discuss the problems mentioned above. At the same time, the centre which was considered to be an 'embryo' of the International Computer Archive of Modern English (ICAME) aimed at collecting the data, making it accessible for computational processing, localising the archives in Bergen University, coordinating, avoiding double research, etc.

The heralding document of the establishment of ICAME was disseminated through the scholars working in the field in order to support and join the project officially. The fact that ICAME is successful can be proved by the fact that currently, at least nine corpora of literary English rely on it.

Similar occurrences took place in the Post-Soviet countries. The compilation of the computational fund of Russian was considered to be of crucial importance[1]. Later, the National Corpus of the Russian Language (НКРЯ[2]) was created. This experience served as a starting point for all the countries which accepted the challenges of Corpus Linguistics and successfully overcame them. Nowadays, National Corpora of many languages have been designed and periodical journals have been published (International Journal of Corpus Linguistics; Corpora; Corpus Linguistics and Linguistic Theory; ICAME Journal; ACL). In the development of the field Association of Computational Linguistics (ACL) also played a big role. In the Post-Soviet countries theoretical and practical issues of Corpus Linguistics are discussed at seminars and academic conferences in Applied Computational Linguistics: The Dialogue, The Megaling, Computational Linguistics, etc. International Conference "The Georgian language and Computational Technologies" has been held every year, since 1998.

3. Corpus of Georgian Dialects (GDC) - The First Georgian Corpus Project

GDC is the first Georgian corpus project which was initiated as one of the sub-corpora of a large Corpus of texts. Currently we are making a structure of the Corpus, principles of annotation and collecting texts simultaneously. We have to do all stages of the work: collecting materials,

[1] Ершов А. П., Машинный фонд русского языка _ внешняя постановка. Текст машинописный _http://ershov.iis/archive/1984.

[2] Сичинава Д., Национальный корпус русского языка: 2003-2005. М.: Индрик, 2005, 21-30.

decoding, editing of old texts, digitalising, unifying and incorporating them in the Corpus. At this stage the Corpus is developing in two directions: a) the text base is being compiled (with the data taken from up to 20 dialects of Georgian) and b) the preliminary system of morphological annotation is being elaborated.

Besides the textual data, the GDC will include lexicographical and reference material (such as dictionaries of dialects, field terminology, etc.). In addition to this, we are working on the 'learner's sub-corpus' which aims at teaching the standard Georgian language to Georgians living abroad (in Iran, Turkey, Azerbaijan) by means of their own dialects.

The work is of unprecedented size and importance. While working on such a challenging project we rely on the above discussed "solidarity" of Corpus linguistics and availability of the latest achievements of this field of study. Thus, we are following the experience of the "Pioneer" scholars of Corpus Linguistics.

"When starting their work, Francis and Kucera were unpopular. Now, when "Corpus has become a leading tendency", we should remember that it is not wise to follow the stream. What we should learn from our pioneers is to respect our independent opinion and have the courage to dare "go against the flow[1]".

ბიბლიოგრაფია / References

ბერიძე მ., "მეტყველების მეოთხე ფაქტურა და საქართველოს ლინგვისტური პორტრეტი", წახნაგი, ფილოლოგიურ კვლევათა წელიწდეული, 2, მემკვიდრეობა, 2010. გვ.105-121.

Beridze M., Nadaraia D., The Corpus of Georgian Dialects// Fifth International Conference: NLP, Corpus Linguistics, Corpus Based Grammar Research, Svolko-2009.

Johansson S., From Brown to LOB, ICAME Journal No. 20,1996, 100–104.

Сичинава Д., Национальный корпус русского языка: 2003-2005. М.: Индрик, 2005, 21-30.

Ершов А., Машинный фонд русского языка _ внешняя постановка. Текст машинописный _http://er-shov.iis/archive/1984.

McEnery A., Wilson A., Corpus Linguistics. Edinburg, 1996.

Developing Linguistic Corpora:a Guide to Good Practice
http://ahds.ac.uk/guides/linguistic-corpora/preface.htm

http://icame.uib.no/

http://www.ruscorpora.ru

http://americannationalcorpus.org

http://sara.natcorp.ox.ac.uk/lookup.html

[1] Johansson S., From Brown to LOB, ICAME Journal No. 20,1996, 100–104.

კორპუსის ლინგვისტიკა: სინქრონია – დიაქრონია – კონტრასტი (გერმანული და ქართული ენები)

დალი ბახტაძე
ილიას სახელმწიფო უნივერსიტეტი (საქართველო)
d_bachtadze@hotmail.com

კორპუსის ლინგვისტიკამ გერმანულენოვან სივრცეში დიდი გავრცელება პოვა გასული საუკუნის 1990-იანი წლების მეორე ნახევრიდან. ჯერ კიდევ სადავოა, კორპუსის ლინგვისტიკა ზოგადი და გამოყენებითი ლინგვისტიკის ერთ-ერთი მეთოდია თუ დამოუკიდებელი ენათმეცნიერული დისციპლინა. როგორც სამეცნიერო დარგმა, მან ადგილი დაიმკვიდრა კვლევითსა და აკადემიურ სასწავლო დაწესებულებებში (ბერლინის ჰუმბოლდტისა და ბირმინგემის უნივერსიტეტებში შექმნილია კათედრები, როგორც საგანი, შეტანილია გერმანიის მრავალი უნივერსიტეტის პროგრამაში). შესწავლის ობიექტია ენა თავისი არსებობის სხვადასხვა ფორმით, ავთენტური ენობრივი მონაცემების გამოყენება, რომელიც ელექტრონულად დიდი კორპუსების სახითაა დოკუმენტირებული.

კორპუსის ლინგვისტიკამ გადაწყვიტა ოდინდელი დავა თეორიასა და ემპირიას შორის, მათ მიანიჭა კონსტრუქციული თანაარსებობის უფლება, დაადასტურა, რომ ლინგვისტიკის შემდგომ განვითარებასა და ევოლუციას ერთობლივი რესურსები, მეთოდები და სტანდარტები ესაჭიროება. აქედან გამოიკვეთა კორპუსის ლინგვისტიკის მიზანი – გადაისინჯოს არსებული ლინგვისტური ჰიპოთეზები (აღიარონ ან უარყონ), საკვლევი მონაცემების ანალიზით შექმნან ახალი ჰიპოთეზები და თეორიები. ისტორიისა და ლიტერატურის თეორიისაგან განსხვავებით, ლინგვისტიკამ შედარებით გვიან იპოვა საკუთარი „წყაროთმცოდნეობა“ (ნორბერტ რიჰარდ ვოლფი). მას სათავე დაუდო ტექსტების აუდიოჩანაწერების შენახვამ, თანამედროვე ელექტრონული მედიების საშუალებით უდიდესი რაოდენობის ტექსტების თავმოყრამ და მანქანურმა წაკითხვამ. ორი ლინგვისტური დარგი: კორპუსის ლინგვისტიკა და კომპიუტერული ლინგვისტიკა ამჟამად ყველაზე მეტადაა დაინტერესებული ტექსტების კორპუსების კვლევით (იხ. IDS - Mannheimer Korpus; DeReKo - Mannheimer Korpus; DWDS - Kernkorpus “Digitales Wörterbuch“; Korpus „Deutscher Wortschatz“ Leipzig; Schweizer Textkorpus Basel).

სათაურში მოცემული სინქრონია, დიაქრონია და კონსტრასტულობა მოიცავს კორპუსის აქტუალურ სამუშაო არეალს ისტორიული, თანამედროვე თუ ენობრივ-შეპირისპირებით ასპექტში, საშუალებას იძლევა გამოკვლეული და გაანალიზებული იქნეს ტექსტის ერთობლივი კორპუსი (მაგ.: ისეთი ტექსტისა, როგორიც ბიბლიაა) და არა ტექსტის ტიპების მიხედვით ცალკეული ფრაგმენტები, ამონარიდები. კონტრასტულობა ბი- და მულტილინგვური პარალელური კორპუსების შექმნას გულისხმობს (მაგ., რუსმა მკვლევრებმა შექმნეს თანამედროვე გერმანელ და რუს ავტორთა ტექსტების პარალელური კორპუსი, ამ ტექსტებზე მუშაობის მოდელები). გერმანელმა მეცნიერებმა დაამუშავეს გერმანულ-იტალიური, გერმანულ-ჩეხური, გერმანულ-ინგლისური, გერმანულ-ნიდერლანდური კორპუსები. ლინგვისტიკის სხვადასხვა სფეროში ხელთარსებული კონკრეტული მასალები ნათელყოფენ, როგორ ფართოდ ინერგება ლინგვისტურ კვლევებში კორპუსის ლინგვისტიკის მეთოდები.

Corpus Linguistics: Synchrony - Diachrony – Contrast (The German and the Georgian Languages)

Dali Bakhtadze
Ilia State University (Georgia)
d_bachtadze@hotmail.com

Corpus linguistics has widely spread in the German-speaking world since the second half of the 1990s. It is still being argued whether corpus linguistics is one of the methods of general and applied linguistics or an independent linguistic science. As a scientific field, it was introduced into research and academic educational institutions (when the departments were established in Berlin, Humboldt and Birmingham Universities). In addition to this, it's included in the curricula of many German universities as one of the courses. The object of this study is the language in its various forms, the application of authentic lingual data documented electronically in large corpora. Corpus linguistics settled the old dispute between theory and empiry, granted them the right of constructive co-existence, proved that the further development and evolution of linguistics requires joint resources, methods and standards. It made the purpose of corpus linguistics clear – to revise the existing linguistic hypotheses (either recognise or reject them), create new hypotheses and theories based on the analysis of investigated data. Unlike the history and theory of literature, linguistics gained its own "source study" relatively late (Norbert Rihard Wolf, 2004). It started from the storage of audio-records, accumulation of largest quantity of texts by means of modern electronic media and machine reading. Two linguistic fields- corpus linguistics and computer linguistics are now most interested in the investigation of text corpora (see IDS - Mannheimer Korpus; DeReKo - Mannheimer Korpus; DWDS - Kernkorpus "Digitales Wörterbuch"; Korpus „Deutscher Wortschatz" Leipzig; Schweizer Textkorpus Basel).

Synchrony, diachrony and contrast, mentioned in the title, cover all of the working areas of corpus linguistics in its historical, modern or lingual-comparative aspects, allowing the investigation and analysis of the entire corpus (e.g. of texts like the Bible) and not only that of separate fragments and extracts of texts. Contrast implies creation of parallel bi- and multi-lingual corpora (e.g. Russian researchers created parallel corpora of texts of contemporary German and Russian authors, models of working on these texts). German linguists processed German-Italian, German-Czech, -English, -Dutch corpora. Specific materials, available in different spheres of linguistics show, how widely the methods of corpus linguistics are being introduced into linguistic researches.

ინფორმაციის მოპოვების პრინციპები ფონოგრამების ავტომატური ანალიზისთვის

ანსის ბერზინი
რიგის ტექნიკური უნივერსიტეტი (ლატვია)
ansis.berzins@rtu.lv

შესავალი

2007 წელს ჩვენ წინაშე დადგა ენათა ავტომატიზებული შედარების ამოცანა[1]. ამ მიმართულებით მუშაობა დავიწყეთ იმის გარკვევით, რამდენად გამოსადეგი იყო ლატვიაში მანამდე არსებული დიალექტური ფონოგრამები ჩვენი ექსპერიმენტებისათვის.

სამწუხაროდ, დასკვნა არასახარბიელო აღმოჩნდა: არსებული აუდიომასალა უსარგებლო იყო ავტომატიზებული ანალიზისთვის და ფონოგრამების შეკრება ხელახლა გახდა საჭირო. 50-60-იანი წლების მაგნიტურ ჩანაწერებზე დიდ იმედს არც ვამყარებდით მათი სიძველისა და ჩაწერის არაეფექტური ტექნიკური საშუალებების გამო, მაგრამ ძალიან გაგვაოცა ჩვენი მიზნებისთვის ბოლო ორი ათეული წლის მანძილზე მოპოვებული ჩანაწერების სრულმა გამოუსადეგარობამ.

როგორც ჩანს, თანამედროვე ფოლკლორული, დიალექტოლოგიური თუ ზეპირი ისტორიების მასალის მომპოვებელი ექსპედიციების მონაწილეები ყურადღებას არ აქცევენ ფონოგრამის ჩაწერის ხარისხს – მათთვის არსებითია მხოლოდ მოსმენისა და მანუალური გაშიფვრის შესაძლებლობა.

ჩაწერა ხდება სხვადასხვა ჩამწერი საშუალებით, სხვადასხვა ტიპის მიკროფონით, როგორც წესი – ხმოვანი ფაილების შეკუმშული ფორმატებით, გარეხმაურის თანხლებით, მიკროფონსა და ინფორმანტს შორის მანძილის გაუთვალისწინებლად და ა. შ. ყველაფერი ეს კი იწვევს იმას, რომ ამ სახით ჩაწერილი მასალა გამოუსადეგარია ფონოგრამების ავტომატიზებული ანალიზისთვის ჩაწერის არათანაბარი პირობების, გარეშე ხმაურის და შეკუმშვისას ინფორმაციული დანაკარგის გამო.

მოხსენებაში აღვწერთ ხარისხიანი დიალექტოლოგიური ფონომასალის მოპოვების ჩვენ მიერ დამუშავებულ მეთოდიკას.

ძირითადი პრინციპები

პირველი და ძირითადი პრინციპია მასალის ერთგვაროვნება. ჩაწერის ტექნოლოგია არ უნდა დაირღვეს, რაც უზრუნველყოფს იმას, რომ ფონოგრამები ერთმანეთისგან მხოლოდ შინაარსობლივად იქნება განსხვავებული და არა ტექნიკური თვალსაზრისით.

სარწმუნო შედეგების მისაღებად მნიშვნელოვანია, რომ მეტყველება იყოს სპონტანური. ამისთვის საჭიროა ინფორმანტის პროვოცირება, რომ თავად წარმართოს თხრობა. ასევე მნიშვნელოვანია თემის არჩევანი: უკეთესია ერთსა და იმავე თემაზე არსებული ჩანაწერების შედარება. ამიტომ ინფორმანტებს ვესაუბრებით ზოგად და ყველასათვის ახლობელ თემებზე: მშობლებზე, ბებია-ბაბუებზე, მშობლიურ ადგილებზე, სკოლაზე, არმიაზე, ქორწინებაზე, ბავშვებზე, სამსახურზე, მეურნეობაზე, სახლზე, დღესასწაულებზე, მეზობლებზე, შემოგარენზე და ა. შ. ცხოვრების გზის დოკუმენტირების პარალელურად ვაგროვებთ

[1] ენებად ამ შემთხვევაში მიჩნეულია ზეპირი ენობრივი სისტემები, მათ შორის, დიალექტები.

ფოლკლორულ მასალასაც: სიმღერებს, ცეკვებს, წეს-ჩვეულებებს, ზღაპრებს, ანეკდოტებს. ძალიან მნიშვნელოვანია ინფორმატორთან საუბრის საკუთარ დიალექტზე წარმართვა (ამ დიალექტის არასრულფასოვანი ფლობის შემთხვევაშიც კი), რადგან გამოცდილება გვიჩვენებს, რომ სალიტერატურო ენაზე საუბრის დროს ინფორმატორების დიდი ნაწილი ამა თუ იმ ხარისხით გადაერთვება სალიტერატურო კოდზე.

ჩამწერი მოწყობილობის არჩევა

პირველ რიგში, უნდა გადაწყდეს ანალოგური მოწყობილობით ვისარგებლოთ თუ ციფრულით. ანალოგური ჩანაწერი თავისი უწყვეტი ხასიათის გამო უფრო ზუსტად გადმოსცემს ცოცხალ ბგერას, მაშინ, როცა ციფრული ჩანაწერი წარმოადგენს გარკვეული სიხშირის მონაკვეთებს – რაც მაღალია სიხშირე, მით უფრო ზუსტია ბგერა. მიუხედავად ამ საბაზისო ნაკლისა, ციფრულ ჩანაწერს მრავალი უპირატესობა აქვს: მოწყობილობების შედარებითი სიიაფე, ასლის გადაღებისას ხარისხის შენარჩუნების თვისება და სხვ. [1, 2, 3], მაგრამ ყველაზე მთავარი: მასალის ანალიზს ჩვენ ვაპირებთ პერსონალური ელექტროგამომთვლელი მანქანის (პეგმ) მეშვეობით და, შესაბამისად, მუშაობა მოგვიწევს მაინც დისკრეტულ და არა უწყვეტ მონაცემებთან, ამიტომ უწყვეტი ჩანაწერის აუცილებლობა არ გვაქვს.

როგორც უკვე აღვნიშნეთ, ციფრული აუდიოჩანაწერი დროით ღერძზე დისკრეტიზაციის სიხშირით ხასიათდება. ხოლო დროის ფიქსირებული მნიშვნელობის დროს ბგერის მახსიათებელია ანათვალის თანრიგობა, რომელიც აღწერს ბგერის ამპლიტუდის წარმოდგენის სიზუსტეს. მაგალითად, 16 ბიტის თანრიგობის შემთხვევაში ჩვენ გვექნება ამპლიტუდის 2^{16} = 65536 ფიქსირებული დონე. ცხადია, რაც მეტია ასეთი დონეების რიცხვი, მით ნაკლებია ბგერის დამახინჯება და თანმდევი ხმაური. ფართო მოხმარების სტანდარტად (მაგალითად, სტანდარტები, რომლებიც გამოიყენება კომპაქტურ ფირფიტებში) ჩვენს დროში ითვლება 44,1 კჰერცი / 16 ბიტი, ხოლო მუსიკის ჩაწერისას უფრო ხშირად სარგებლობენ 96 კჰერცი / 24 ბიტით [4].

ერთ-ერთი მთავარი მახასიათებელია ჩანაწერის ფორმატი. ციფრული აუდიოჩანაწერი შეიძლება იყოს შეკუმშული ან შეუკუმშავი. შეკუმშვა შეიძლება სხვადასხვა ფორმატისა იყოს, იმის მიხედვით, როგორია შეკუმშვის საშუალება – დანაკარგიანი შეკუმშვა (მონაცემები გაფართოების შემდეგ აღდგება ერთი ბიტის სიზუსტით) თუ უდანაკარგო შეკუმშვა (მონაცემები, რომლებიც მიჩნეულია უმნიშვნელოდ, იკარგება) [5, 6, 7]. მართალია, აუდიომონაცემების სამეცნიერო მიზნებით გამოყენებისას შეიძლება მნიშვნელოვანი იყოს ისეთი მონაცემებიც, რომლებიც აუდიოფორმატების შემქნელებს უმნიშვნელოდ მიაჩნიათ. ამიტომ ჩვენი მიზნებისთვის უფრო მიზანშეწონილია ფორმატები შეკუმშვის გარეშე ან შეკუმშვა დანაკარგის გარეშე. მართალია, უდანაკარგო შეკუმშვის დროს ფაილები დაიკავებს დაახლოებით ორჯერ ნაკლებ ადგილს, მაგრამ მნიშვნელოვნად გაიზრდება მათი დამუშავების დრო (ყოველთვის მათი მოსმენისას ეგმ-ს მოუწევს მასალის გაფართოება და შეკუმშვა), ამიტომაც უფრო რაციონალურია შეუკუმშავი ფაილებით სარგებლობა. ამას უპირატესობა აქვს მატარებლის მწყობრიდან გამოსვლის შემთხვევაშიც, რადგანაც შეუკუმშავი ფაილები უფრო ადვილად და უკეთ ექვემდებარება აღდგენას, მათ შორის ნაწილ-ნაწილ აღდგენასაც.

მეტყველების ამოცნობის სფეროში წარმატებით გამოიყენება ფონოგრამულ მონაცემთა „სატელეფონო ხარისხის“ ბაზებიც კი. მაგ., SpeechDat (8 კჰერცი, A-law[1]), TIMIT (16 битов/16 კჰერცი) [9], PER (8 კჰერცი, A-law და 16 ბიტი/16 კჰერცი) [10], ამიტომ და, აგრეთ-

[1] შეკუმშული ფორმატი, რომელიც 16 ბიტიან ბგერას გარდაქმნის 8 ბიტად.

ვე, ფართო გავრცელებისა და შეთავსებადობის გამო ჩვენ საკმარისად და მიზანშეწონილად ჩავთვალეთ მონაცემების 44,1 კჰერცის ხარისხით ჩაწერა.

სამწუხაროდ, მეცნიერები მონაცემების მოპოვებისას ჩაწერის ფორმატს უმეტესად ყურადღებას არ აქცევენ, სარგებლობენ არაპროფესიული ან ნახევრადპროფესიული აპარატურით, რომლებიც დანაკარგებით შეკუმშვის ფორმატებს იყენებენ.

საექსპედიციო მუშაობისას, რასაკვირველია, დიდი მნიშვნელობა აქვს აპარატურის ზომას. თანამედროვე ჩამწერი აპარატურის დიდი ნაწილი საკმაოდ მცირე ზომისაა, ამიტომ არჩევანის გაკეთება არ არის რთული.

მიკროფონის არჩევა

ხარისხიანი ჩაწერისათვის უნდა გამოვიყენოთ აუცილებლად გარე და არა აპარატურაში ჩამონტაჟებული შიდა მიკროფონი.

ჯერ უნდა ავირჩიოთ მიკროფონის ტიპი მოქმედების პრინციპის მიხედვით [11]. როგორც ჩანს, არჩევანის გაკეთება მოგვიწევს კონდენსატორულ ელექტრეტულს [17] და ელექტროდინამიკურ კოჭა მიკროფონს შორის მათი ტექნიკური თვისებებისა და მისაღები ფასის გათვალისწინებით[1] [12, 13, 14]. ზოგადად, კონდენსატორული მიკროფონები უკეთესია მგრძნობიარეობისა და სიხშირული მაჩვენებლების მიხედვით, მაგრამ ისინი ზედმეტად მგრძნობიარეა გარეშე პირობების მიმართ. თუ გავითვალისწინებთ, რომ მიკროფონი საველე პირობებში უნდა გამოვიყენოთ, ალბათ უმჯობესი იქნება, თუ დინამიკური მიკროფონით ვისარგებლებთ[2]. რასაკვირველია, მას უნდა ჰქონდეს სიხშირის ფართო დიაპაზონი, სასურველია, 15 000 ჰერცი (ანუ უმაღლესი კლასის დინამიკური მიკროფონები), რადგანაც, მაგალითად, ადამიანის, განსაკუთრებით კი, მამაკაცის ხმის სიხშირე შეიძლება იყოს ფართომოხმარების 100 ჰერცზე დაბალი[3].

ყურადღება უნდა მიექცეს მიკროფონის სიხშირული მგრძნობელობის მრუდს: რაც უფრო პირდაპირი და სწორხაზოვანია ის (ანუ სიხშირული დიაპაზონები ჩაიწერება ნატურალური ხმოვანი მიმართებით), მით უკეთესია.

კონსტრუქციულად მიკროფონი უნდა იყოს თავზე დასამაგრებული და რაც შეიძლება ახლოს უნდა განთავსდეს ინფორმანტის პირთან.

რაღა თქმა უნდა, მიკროფონი უნდა იყოს ცალმხრივი მიმართულებისა (ე.წ. კარდიოიდული), რადგანაც ბგერის წყარო – ინფორმანტის პირი – ერთ წერტილში მდებარეობს. მიმართულების კარდიოიდული დიაგრამის დროს ინფორმანტის მეტყველება დაფიქსირდება კარგად, ხოლო გარეშე ხმები – ცუდად.

ჩაწერის ადგილის არჩევა და მომზადება

ხარისხიანი ჩანაწერის მისაღებად აუცილებელია არა მხოლოდ საუკეთესო ტექნიკა, არამედ შესაფერისი პირობები, რაც გულისხმობს გარე ხმაურის შემცირებას.

ჩაწერა უნდა მიმდინარეობდეს დახურულ სივრცეში. თუ სახლი გზის პირასაა, ვირჩევთ საპირისპირო მხარეს მდებარე ოთახს. უნდა ჩამოვხსნათ კედლის საათი, მთელ შენო-

[1] ჩვეულებრივი კონდენსატორული მიკროფონი საჭიროებს გარე კვებას და მგრძნობიარეა დარტყმებისა და კლიმატური პირობებისადმი [16], ნათურიანი (კონდენსატორული ნათურიანი გამაძლიერებლით ნახევარგამტარიანის ნაცვლად) – კარგია, მაგრამ დიდია და ძვირადღირებული; დინამიკური ლენტური – ძვირია და ძალიან მგრძნობიარე [15]; კუთხოვანი – მოძველებულია და თან ცუდი სიხშირული მახასიათებელი აქვს [18].

[2] დინამიკური მიკროფონის უპირატესობაა ისიც, რომ ის კარგად უძლებს გადატვირთვას.

[3] ადამიანის სმენისა და ხმის სიხშირეთა სრული დიაპაზონია 20-20 000 ჰერცი, მაგრამ ინტერვალის საზღვრებთან მიახლოებული მნიშვნელობები უმნიშვნელო რაოდენობით გვხვდება და უმეტესად არ ისმის.

ბაში უნდა გამოვრთოთ ტელეფონები[1], ტელევიზორი და რადიო და მოვითხოვოთ მობინადრეთაგან სიჩუმე.

ПРИНЦИПЫ СБОРА ИНФОРМАЦИИ ДЛЯ АВТОМАТИЗИРОВАННОГО АНАЛИЗА ФОНОГРАММ

А.У. Берзинь
Рижский технический университет (Латвия)
ansis.berzins@rtu.lv

Введение

В 2007 году мы поставили перед собой задачу автоматизированного сравнения языков[2]. Работу в данном направлении мы начали с анализа имеющихся в Латвии диалектологических экспедиционных фонограмм на предмет их пригодности для наших экспериментов. К сожалению, выводы оказались отрицательными: ничто из собранного непригодно и для автоматизированного анализа фонограммы надо собирать сызнова. Если по записям на магнитофонных лентах 50-60 гг. мы особых надежд не чаяли, ввиду их возраста и возможностей техники той поры, то непригодность записей последних двух десятилетий нас удивила.

Оказывается, участники современных экспедиций, будь то фольклорные, диалектологические или по сбору жизнеописаний, не уделяют внимания качеству записи фонограммы – им важна только возможность последующего мануального прослушивания и расшифровки. Запись ведётся разными записывающими устройствами, разными микрофонами, в основном – в сжатых форматах звукового файла, при наличии внешних шумов, не уделяется внимание и фиксированию расстояния от рта информанта до микрофона. Всё это приводит к тому, что звукозаписи, записанные подобным образом, не пригодны для автоматизированного анализа фонограмм, ввиду разных условий записи, слишком большого количества внешних шумов и потерь информации при сжатии.

В данном докладе мы опишем разработанную нами методику собирания качественных диалектологических фономатериалов.

Основные принципы

Первым и основополагающим принципом является однородность материала, т.е., технология записи должна сохраняться, таким образом обеспечивая то, что фонограммы будут различаться только содержательно, а не по техническим причинам.

Для достижения наиболее достоверных результатов важно, чтобы речь была спонтанной, т.е. требуется только провоцировать информанта вопросами, дабы он сам вёл

[1] უნდა გამოვრთოთ მთლიანად, მხოლოდ ხმის გამორთვა საკმარისი არ არის.
[2] Языками мы в данном случае считаем устные языки в самом широком понимании этого слова, в том числе и наречия.

рассказ. Важен также выбор темы: лучше сравнивать рассказы на одну и ту же тему, поэтому мы разговариваем с людьми на темы, пережитые всеми: родители, деды и бабки, родное место, школа, армия, свадьба, женитьба/замужество, дети, работа, хозяйство, дом, праздники, вечеринки, соседи, окрестности и т.п., т.е. документируем жизненный путь, а заодно и собираем фольклорный материал: песни, танцы, описания обрядов, сказки, анекдоты. Исключительно важно с информантом разговаривать на его наречии (пусть даже на плохом), так как опыт показывает, что при разговоре с информантами на литературном языке, даже при просьбе говорить на наречии большинство информантов всё равно в большей или меньшей степени подстраивают свой язык под литературный.

Выбор записывающего устройства

Во-первых, надо решить, пользоваться аналоговыми устройствами или цифровыми. Аналоговая запись более подобно отражает живой звук, так как он сам, по сути, является непрерывным. Цифровая же запись – это разрезы, взятые с определённой частотой – чем частота больше, тем правдоподобнее звук. Несмотря на этот – базовый – недостаток, у цифровой записи много преимуществ: устройства дешевле, при копировании качество не ухудшается и др. [1, 2, 3] Но – самое главное: мы наши данные собираемся анализировать и сравнивать при помощи ПЭВМ, т.е., мы всё равно будем работать с дискретными, а не непрерывными данными, поэтому у нас нет потребности в непрерывности звукозаписи.

Как мы уже указали, цифровая аудиозапись на оси времени характеризуется частотой дискретизации. А при фиксированном значении времени звук характеризуется разрядностью отсчёта, описывающей точность представления амплитуды звука. Например, при разрядности в 16 битов мы будем иметь 2^{16} = 65536 фиксированных уровней амплитуды. Очевидно, что чем таких уровней больше, тем меньше искажений и шума. Стандартом широкого потребления (используется, например, в компактпластинках) в наше время является 44,1 кГц / 16 битов, а при записи музыки чаще пользуются 96 кГц / 24 битами. [4]

Одна из главных характеристик – формат записи. Цифровая аудиозапись может быть несжатой или сжатой. Сжатие может быть в разных форматах, которые подразделяются на способы сжатия с потерями (данные при разжатии восстанавливаются с точностью до бита) и без потерь (данные, которые разработчиками признаны несущественными, теряются). [5, 6, 7] Так как при использовании аудиоданных в научных целях существенными могут оказаться и такие данные, которые разработчикам аудиоформатов кажутся несущественными, то для наших целей подходящими являются только форматы несжатые и сжатые без потерь. Хоть и при сжатии без потерь файлы будут занимать приблизительно в два раза меньше места, но существенно возрастёт время их обработки (каждый раз при их прослушивании, ЭВМ будет их разжимать и сжимать), посему более рационально пользоваться несжатыми данными. Это также даст преимущество при порче или поломке носителя, так как несжатые данные легче и лучше восстанавливаются, в том числе и по частям.

В области распознавания речи успешно применяются базы данных фонограмм даже

так называемого „телефонного“ качества, например, SpeechDat (8 кГц, A-law[1]), TIMIT (16 битов/16 кГц) [9], PER (8 кГц, A-law и 16 битов/16 кГц) [10], поэтому, а также учитывая более широкую распространённость и совместимость, мы посчитали достаточным и целесообразным данные записывать в качестве 44,1 кГц / 16 битов.

К сожалению, учёные, собирающие данные, на формат записи зачастую вообще не обращают внимания, а устройствами ползуються непрофессиональными или полупрофессиональными, в которых по умолчанию используются сжатые с потерями форматы.

Для экспедиционной работы, безусловно, важен размер устройства. Большая часть современных записывающих устройств по размерам невелика, поэтому выбор не составит проблемы.

Выбор микрофона

Для качественной записи надо пользоваться внешним, а не встроенным в записывающее устройство, микрофоном.

Во-первых, надо выбрать тип микрофона по принципу действия [11]. Очевидно, нам придёться выбирать между конденсаторным электретным [17] и электродинамическим катушечным, ввиду их технических свойств и доступной цены.[2] [12, 13, 14] В среднем, конденсаторные лучше по чувствительности и частотным показателям, но они также более чувствительны к внешним условиям. Учитывая то, что микрофон будет использоваться в экспедиционных условиях, наверное лучше воспользоваться динамическим микрофоном[3]. Конечно, он должен иметь широкий диапазон воспроизводимых частот, желательно 30-15 000 гц (т.н., динамический микрофон высшего класса), так как, например, частоты человеческого, в особенности мужского, голоса бывают и ниже ширпотребных 100 гц[4].

Следует обратить внимание и на кривую частотной чувстительности микрофона: чем она прямее и более пологая (т.е. частотные диапазоны записываются в натуральном громкостном соотношении), тем лучше.

По конструкции необходимо пользоваться микрофоном, крепящимся на голове и находящимся у рта информанта: это позволяет избежать колебаний громкости речи при передвижении головы.

Разумеется, микрофон должен быть одностороннего направления (т.н. кардиоидный), так как источник звука – рот информанта – находится в одной точке. При кардиоидной диаграмме направленности речь информанта будет улавливаться хорошо, а окрестные шумы – плохо.

[1] Сжатый формат, кодирующий 16 битовый звук 8 битами. [8]

[2] Обыкновенный конденсаторный нуждается во внешнем питании и чувствителен к ударам и климатическим воздействиям [16], ламповый (конденсаторный с ламповым предусилителем вместо полупроводникового) – более громоздкий и дорогой, хоть и хороший, динамический ленточный – дорогой и чувствителен к дуновениям [15], угольный – устаревший и с плохой частотной характеристикой [18].

[3] Преимуществом динамического микрофона является также его устойчивость к перегрузкам.

[4] Полный диапазон частот человеческого слуха и голоса 20-20000 гц. Но значения, близкие к границам интервала, встречаются в несущественном количестве и большинством не различаются на слух.

Выбор и подготовка места записи

Для получения качественной записи необходимо не только пользоваться качественной техникой, но и проводить её в соответствующих условиях, в которых внешние шумы минимизированы.

Запись надо проводить в закрытом помещении, при закрытых окнах и дверях. Если дом стоит возле дороги, то лучше выбрать комнату с противоположной стороны. Надо снять или заглушить настенные часы, выключить холодильник и другие произвольно включающиеся электроприборы, выключить все телефоны[1], выключить телевизоры и радио, даже если они находятся в других помещениях, попросить домочадцев не шуметь и не входить.

Библиография / References

От аналоговой записи – к цифре,
http://its-journalist.ru/Articles/ ot_analogovoj_zapisi_k_cifre.html

Звукозапись, цифровая или аналоговая?
http://www.midi.ru/forumd.php?id=181648

Дубровский Д.Ю. Чем цифровая запись лучше аналоговой?
http://demorecord.ru/analogsound.html

Музыченко Е.В. Принципы цифрового звука. 1998-1999,
http://www.websound.ru/articles/theory/digsnd.htm

Audio file format,
http://en.wikipedia.org/wiki/Audio_file_format

Сжатие без потерь,
http://ru.wikipedia.org/wiki/Сжатие_без_потерь

Сжатие данных с потерями,
http://ru.wikipedia.org/wiki/Сжатие_данных_с_потерями

A-Law Compressed Sound Format,
http://www.digitalpreservation.gov/formats/fdd/fdd000038.shtml

Salvi, G. Mining Speech Sounds. KTH: Stockholm, 2006. pp. 18-19.

Melin, H. Automatic speaker verification on site and by telephone: methods, applications and assessment. KTH: Stockholm, 2006. pp. 103-104.

Микрофон,
http://ru.wikipedia.org/wiki/Микрофон

Характеристики микрофонов,
http://ingibit.rigalink.lv/info/c2/mikro01.html

Сравнение конденсаторных и динамических микрофонов.
http://www.microphone.ru/articles/paragraph_1.html

Динамические, конденсаторные микрофоны и фантомное питание,
http://midi.ucoz.ru/publ/1-1-0-16

[1] Именно выключить – отключения звука недостаточно.

Костоломов В. Ленточные микрофоны. 2000,
http://www.oktava-mics.net/shop/a-2/lentochnye_mikrophony.html
Конденсаторный микрофон,
http://ru.wikipedia.org/wiki/Конденсаторный_микрофон
Электретный микрофон,
http://ru.wikipedia.org/wiki/Электретный_микрофон
Угольный микрофон,
http://ru.wikipedia.org/wiki/Угольный_микрофон

დიალექტური ლექსიკონები კორპუსში (ქდკ) და ნახევრადავტომატური ლემატიზაციის საკითხები

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

შესავალი: ქართული დიალექტების კორპუსი იქმნება რუსთაველის ეროვნული სამეცნიერო ფონდის მხარდაჭერით და გულისხმობს ქართული ენის ქვესისტემების ვრცელი, ლინგვისტური და მეტალინგვისტური ანოტირების აპარატით აღჭურვილი კორპუსის შექმნას[1]. ქდკ-ს კონცეფციაში გამორჩეული ადგილი უჭირავს ერთ მნიშვნელოვან მომენტს, რომელიც ჯერჯერობით არ გვხვდება საერთაშორისო კორპუსულ გამოცდილებაში. ეს არის ლექსიკონების, როგორც ტექსტური კომპონენტის ინტეგრირება კორპუსში.

1. ლექსიკონი, როგორც ტექსტური კომპონენტი კორპუსში:

ლექსიკონი რთული ტექსტური ორგანიზაციის მქონე ლინგვისტური და კულტურული პროდუქტია. დიალექტური ლექსიკონი ერთენოვანი განმარტებითი ლექსიკონის ერთგვარი ნაირსახეობაა.

განმარტებითი ლექსიკონის სალექსიკონო სტატია განსაკუთრებული სახეობის ტექსტია, რომლის ცენტრშიც დგას მოკლედ და ნათლად ჩამოყალიბებული დეფინიცია (განმარტება); მის ტექსტურ ღირებულებას კიდევ უფრო აფართოებს საილუსტრაციო მასალა და ლექსიკოგრაფიული პარამეტრიზაციის სხვადასხვა "ბლოკი", რომლებიც მოიცავენ სალექსიკონო ერთეულის შესახებ მდიდარ დამატებით ინფორმაციას (არაძირითადი მნიშვნელობები, ფრაზეოლოგიზმები, სინონიმები (მითითებებები სხვა სიტყვაზე ან ლექსიკონზე), გადატანითი მნიშვნელობები, სიტყვის უცხოური წარმომავლობა და სხვ.). ასე რომ, ქდკ-ში ტექსტურ კომპონენტად შეგვაქვს ლექსიკონის "მარჯვენა" მხარე, როგორც სპეციფიკური სტრუქტურის მქონე (ლექსიკოგრაფიული პარამეტრებით ფრაგმენტირებული) ტექსტი.

კორპუსში ინტეგრირდება სალექსიკონო მთავარი ფორმაც – ლექსიკონის "მარცხენა" მხარე და ის ისევე მიემართება სალექსიკონო სტატიას, როგორც სიტყვაფორმა მიემართება კონტექსტს ჩვეულებრივ კონკორდანსში. გარდა ამისა, ცალკე ტექსტურ ფრაგმენტად "განიხილება" საილუსტრაციო ფრაზა და მისი შემადგენელი ყველა სიტყვაც ადგილს იმკვიდრებს საერთო კონკორდანსში.

კორპუსის არქიტექტურაში ტექსტების შეტანის რედაქტორთან ერთად გათვალისწინებულია ლექსიკონების დამატების რედაქტორიც, რომელიც შესაძლებლობას იძლევა აისახოს ქართული დიალექტური ლექსიკონების ყველა აღნიშნული ლექსიკოგრაფიული თავისებურება.

ლექსიკონის (ან სხვადასხვა სახის სალექსიკონო მასალის) შეტანა კორპუსში მისი რეპრეზენტატულობის ხარისხის გაზრდის ეფექტური საშუალებაა. განსაკუთრებით ეს ით-

[1] Beridze M., Nadaraia D., Corpus of Georgian Dialects, NLP, Corpus Linguistics, Corpus Bazed Grammar Research, Fifth Internacional Conference "slovko -2009", 25-35.

ქმის დიალექტურ კორპუსზე, რადგან დიალექტური ტექსტების თემატური, ჟანრობრივი ან სტილური ”დაბალანსება” გაცილებით პრობლემურია, ვიდრე სალიტერატურო ენის ტექსტებისა.

3. დიალექტური ლექსიკონები ქდკ-ში, როგორც ნაწილობრივი მორფოლოგიური ანოტირების საშუალება:

ლექსიკონების კორპუსში ინტეგრირებით ჩვენ შევეცადეთ არა მარტო ახალი რაკურსით წარმოგვეჩინა ის მჭიდრო კავშირი, რომელიც ლექსიკონსა და კორპუსს შორის არსებობს, არამედ პრაქტიკულად გამოგვეყენებინა ტრადიციული ლექსიკოგრაფიის პრინციპებით შემუშავებული ლექსიკონები კორპუსულ ”მშენებლობაში”.

როგორც ცნობილია, დღეს არც ერთი ავტორიტეტული ლექსიკონი და გრამატიკა კორპუსთან კავშირის გარეშე არ იქმნება (მათ ასეც ეწოდებათ: კორპუსზე დაყრდნობილი), მაგრამ ვერც კორპუსები იქმნება ლექსიკოგრაფიული და გრამატიკული ცოდნის გამოყენების გარეშე: სწორედ ამ ცოდნაზე დაყრდნობით ხორციელდება მრავალდონიანი ლინგვისტური ანოტირება კორპუსში, რომლის გარეშეც ის მხოლოდ ტექსტების კოლექცია ან ელექტრონული ბიბლიოთეკა იქნებოდა. ეს ”შეკრული წრე” – ურთიერთდაყრდნობილობა და ურთიერთგანსაზღვრულობა არის კორპუსული და ტრადიციული (თეორიული და კომპიუტერული) ლინგვისტიკის ურთიერთმიმართების მთავარი გამოხატულება[1].

ლექსიკონის სიტყვანის (ლექსიკონის მარცხენა მხარის) გამოყენების შესაძლებლობა კორპუსის მორფოლოგიური ანოტირების ელემენტად ამ ”წრიული” კავშირის ერთი კერძო გამოვლინებაა.

4. ლემატიზაცია და მეტყველების ნაწილების მიხედვით ანოტირება ქდკ-ში:

ამჟამად ჩვენ ვამუშავებთ მორფოლოგიური ანოტირების სისტემას კორპუსში (ქდკ). ამ გზაზე პირველი ნაბიჯია ლემატიზაცია. თუკი ამომწურავი მორფოლოგიური აღწერილობის მქონე სალიტერატურო ენათა კორპუსებში ავტომატური ლემატიზაცია იოლად გადასაჭრელი და ტრივიალური პრობლემაა, დიალექტურ კორპუსში, რომელშიც 20-მდე ქვესისტემის ვრცელი ტექსტური კოლექციაა ინტეგრირებული, ეს საკმაოდ რთულ ამოცანას წარმოადგენს. დიალექტურ კორპუსთა უმეტესობაში ეს პროცესი ხელით ხორციელდება.

ლემატიზაციის პროცესი ჩვენს კორპუსში სალიტერატურო ფორმაზეა ორიენტირებული – ხდება დიალექტური და სალიტერატურო ლემის ”გათანაბრება, გაიგივება”. ამ საფეხურამდე, ბუნებრივია, გასავლელია თვით დიალექტური ტექსტის ლემატიზაციის საფეხური. პარალელურად ვაწარმოებთ მეტყველების ნაწილის მიხედვით ანოტირებას:

I – ლემატიზაცია და სალიტერატურო ფორმასთან ”გათანაბრება” ასეთი სქემით ხდება:

- დიალექტური სიტყვაფორმა / / დილექტური მთავარი ფორმა
- სიტყვაფორმის სალიტერატურო შესატყვისი // მთავარი ფორმის სალიტერატურო შესატყვისი

[1] *Sinclair J.*, Corpus and Text: Basic Principles *(Tuscan Word Centre)* Developing Linguistic Corpora: a Guide to Good Practice / Edited by Martin Wynne, http://ahds.ac.uk/guides/linguistic-corpora/preface.htm; Filmore Ch., Corpus Linguistics or Computer-aided armchair Linguistics, 1992.

II – მეტყველების ნაწილების მიხედვით ანოტირება:

ჩვენ გადავწყვიტეთ, გამოვიყენოთ დიალექტური ლექსიკონის "მარცხენა მხარე" ნაწილობრივი ლემატიზაციისა და მეტყველების ნაწილების მიხედვით ანოტირებისათვის შემდეგი სახით:

ა) სალექსიკონო ერთეული (მთავარი ფორმა) ბაზას მიეწოდება, როგორც სიტყვიერი მასალა და ამავე დროს, როგორც ტექსტებში სალექსიკონო ფორმით დადასტურებული ლექსემების იდენტიფიკატორი: ტექსტში დადასტურებული ყველა მთავარი ფორმა, რომელიც დაემთხვევა ლექსიკონში მოცემულს, ავტომატურად იქნება მარკირებული, როგორც ლემა.

ბ) ზმნის ლემატიზაციისას ჩვენ ლემად (მეთაურ ფორმად) ვუთითებთ, როგორც საწყისს, ისე მესამე პირის ფორმას. იგივე პრინციპია გატარებული ქართული ენის განმარტებით და ქართულ დიალექტურ ლექსიკონთა უმრავლესობაშიც. ამიტომ, ლექსიკონების მასალის ამ (კორპუსის ლემატიზაციისა და მორფოლოგიური ანოტირების) ფუნქციით გააქტიურება ზმნის მესამე პირის ფორმათა იდენტიფიკაციის საშუალებასაც მოგვცემს;

გ) ლექსიკონის დამატების რედაქტორი გარდა ამა თუ იმ ლექსიკონის ლექსიკოგრაფიული თავისებურებების ამსახველი ინფორმაციული ბლოკებისა, შეიცავს პირველადი მორფოლოგიური ანოტირების მარკერებს: სახელი (არსებითი, ზედსართავი, რიცხვითი, ნაცვალსახელი), ზმნა, ფორმაუცვლელი სიტყვა (ზმნიზედა, თანდებული, შორისდებული, ნაწილაკი, კავშირი). ამგვარად მარკირებული სალექსიკონო ერთეული არა მარტო ახდენს ტექსტურ მასალაში დამოწმებულ ფორმათა იდენტიფიკაციას, არამედ იძლევა მათი პირველადი მორფოლოგიური მონიშვნის საშუალებასაც.

დ) თუ ლექსიკონში დიალექტური ფორმის განმარტება მარტივი სალიტერატურო შესატყვისითაა გადმოცემული და სიტყვა სალექსიკონო ფორმით ტექსტშიც დასტურდება, მაშინ შესაძლებელი იქნება როგორც ლემის, ისე სალიტერატურო შესატყვისის მიწერაც.

ე) ამ გზით შესაძლებელი იქნება მხოლოდ იმ მასალის ლემატიზაცია და ნაწილობრივი მორფოლოგიური ანოტირება, რომელიც ტექსტურ მასივში მეთაური ფორმით არის დამოწმებული, თუმცა დიალექტური კორპუსის ლემატიზაციის ავტომატური პროცესი აქ არ დამთავრდება. დარჩენილი (არამეთაური ფორმით წარმოდგენილი) სიტყვახმარებების ანოტირება კორპუსში ხელით უნდა განხორციელდეს.

თუმცა ეს ეტაპიც შეიძლება გახდეს „ნახევრადავტომატური", თუკი ლექსიკონების სიტყვანის ზემოთ აღნიშნულ ფუნქციას გავააქტიურებთ. ხელით ანოტირების სხვადასხვა ამოსავალი „მიმართულება" ამის სხვადასხვა შესაძლებლობას იძლევა:

- მიმართულება: **სიტყვაფორმა** > **ლემა** (თუ ეს სიტყვა ლექსიკონში შესულია) მეტყველების ნაწილის მარკერის ავტომატურად მიწერის საშუალებას იძლევა, რადგან მეთაურ ფორმას (ლემას) ეს მარკერი წინასწარ აქვს მინიჭებული.
- მიმართულება: **სიტყვაფორმა** > **სიტყვაფორმის სალიტერატურო შესატყვისი** სალიტერატურო ენის მორფოლოგიური ანალიზატორის საშუალებით სალიტერატურო ლემის ავტომატურად მიწერის საშუალებას იძლევა. ამ შემთხვევაშიც, ბუნებრივია, მიეწერება მეტყველების ნაწილის მარკერი.
- დიალექტური კორპუსის ლემატიზაციისას ყველა იმ შემთხვევაში, რომელიც **დიალექტურ მეთაურ ფორმას** სალიტერატურო ენაში **მარტივი მეთაური** ფორმა შეესაბამება, შესაძლებელი იქნება სრული მორფოლოგიური ანოტირება – სალიტერატურო ენის ანალიზატორის საშუალებით ყველა გრამატიკული მა-

ხასიათებელი მიენიჭება დიალექტურ ლემას სალიტერატურო ლემასთან მისი „დაკავშირების“ საშუალებით.

ვფიქრობთ, ლექსიკონის, როგორც ერთგვარი ”ინსტრუმენტის”, ლემატიზაციისა და პირველადი ანოტირების პროცესში ჩართვის იდეა შესაძლოა ეფექტურად იქნეს გამოყენებული იმ ენათა ლინგვისტური კორპუსების შექმნისას, რომელთაც არ აქვთ ამომწურავი მორფოლოგიური აღწერილობა (რომელთა კომპიუტერული დამუშავება ჩამორჩება თანამედროვე მოთხოვნებს).

Dialect Dictionaries in the Corpus (GDC) and the Issues of Semi-automatic Lemmatisation

M. Beridze, L. Lortkipanidze and D. Nadaraia
Arn.Chikobava Institute of Linguistics (Georgia)
marine.beridze@gmail.com, l_lordkipanidze@yahoo.com, dnad@itex.ge

Introduction

The Corpus of Georgian Dialects is compiled with the assistance of Shota Rustaveli National Science Foundation and implies the creation of a vast Corpora of subsystems of the Georgian language equipped with the apparatus of linguistic and metalinguistic annotation[1].
In the concept of the Corpus of Georgian Dialects **(GDC)** one significant issue which is new to the international experience of Corpus linguistics is the integration of dictionaries into the corpus as a textual component.

1. The Dictionary as a textual component in the Corpus

A dictionary is a linguistic and cultural product with a complex textual organisation. A Dialect dictionary of a language is a type of monolingual explanatory dictionary. A lexical article is a specific text with a brief and clearly organised definition. Its textual value is further enhanced by illustrations and various ‘blocks’ of lexicographic parameterisation which include additional information about the dictionary item (polysemantics, phraseological units, synonyms (references to other words or dictionaries), metaphoric meanings, borrowed words, etc. Thus, we include the “right” side of the dictionary as a text of a specific structure (that is fragmented by means of lexicographic parameters) in the GDC.

[1] Beridze M., Nadaraia D., Corpus of Georgian Dialects, NLP, Corpus Linguistics, Corpus Based Grammar Research, Fifth International Conference “slovko -2009”, 25-35.

The Corpus also incorporates the major, “left” side and it refers to the lexical article in the same way as a word-form to the context in the concordance. In addition to this, the illustration phrase is considered to be a separate textual fragment and each word constituent has its own place in the concordance.

In the corpus, along with the editor of the text insertion, the editor of the dictionary addition is envisaged, which enables us to reveal all of the above discussed lexicographic characteristics of the GDC.

Obviously, inclusion of the dictionary (or dictionary materials of various kinds) into the Corpus is an effective way of increasing its representativeness. This is particularly true for the Corpus of Dialects, as thematic, genre or stylistic “ balancing” of dialect texts is far more problematic than those of literary texts.

2. Dialect Dictionaries in GDC as a means of partial morphological annotation

By integrating dictionaries into the Corpus, we have tried to reveal the close connection between them as well as to practically integrate the dictionaries, compiled by applying principles of traditional lexicography, into the Corpus Building.

As is known, nowadays, not a single authoritative dictionary and grammar can be created without being connected to the Corpus (what is more, they are referred to as Corpus-based). But, also corpus can not be created without applying lexicographic and grammatical knowledge to the material. Moreover, multilevel annotation in the Corpus is based on this type of knowledge. Without this, it would be only a collection of texts or an e-library.
This circle, interrelation and interdependence is the main expression of mutual relations of the traditional (Theoretical and Computational) and Corpus linguistics.[1] The possibility of application of the “left” hand side of the dictionary as an element of morphological annotation of the Corpus is, arguably, one of the manifestations of this “circle” relationship.

3. Lemmatisation and annotation according to the part of speech categories in GDC

Currently we are in the process of elaborating the system of morphological annotation in the Corpus (GDC). In this respect, the first step is lemmatisation. If lemmatisation is an easy, even a trivial issue in the Corpora of literary languages possessing exshaustive morphological description, in the dialectal corpora, as a rule, lemmatisation is done manually. The process of lemmatisation in our Corpus is oriented on the literary form- as, in this process, “equalisation” of dialectal and

[1] Sinclair J., Corpus and Text: Basic Principles *(Tuscan Word Centre)* Developing Linguistic Corpora: a Guide to Good Practice / Edited by Martin Wynne, http://ahds.ac.uk/guides/linguistic-corpora/preface.htm; Ch. Filmore, Corpus Linguistics or Computer-aided armchair Linguistics, 1992.

literary lemmas takes place. Naturally, before this stage is reached, the process of lemmatisation of the dialectal text itself should be finished. Simultaneously, the material should also be annotated according to the part of speech category.

I _Lemmatisation and „equalisation“ with the literary form is done according to this scheme:

- Dialectal word-form / / dialectal head form
- Literary equivalent of the word-form / / Literary equivalent of the head form
-

II _Annotation according to the category of parts of speech

We decided to use the “left”- hand side of the dialectal dictionary for both processes, the partial lemmatisation and annotation according to the part-of-speech category:

a) the lexical unit (head form) is transmitted to the database as a lexical material as well as the identifier of the lexemes attested in the database. All the main forms in the text which will coincide with the forms in the dictionary will automatically be marked as lemmas.

b) While conducting lemmatisation of the verb, we provide two forms of the paradigm as lemmas (head forms): the infinitive and the third person form. The same principle is used in the Explanatory Dictionary Georgian as well as in the majority of dictionaries of Georgian dialects. Thus, the actualisation of these functions (lemmatisation and morphological annotation) will make it possible to easily identify the third person forms of the verb in the database.

c) Together with the information blocks which reflect the lexicographic peculiarities of the dictionary, the tool for editions includes the markers of the primary morphological annotations: nominal parts (noun, preposition, numeral, pronoun), verb, the unchangeable parts of speech (adverbs, postpositions, interjections, particles, conjunctions). Dictionary items marked in such a way will not only identify the forms presented in the textual material but also enable us to conduct primary morphological tagging.

d) If a dialectal form is defined by its simple literary equivalent, and the form by which this word occurs in the text coincides with that of the dictionary, it will be possible to include both the lemma and its literary equivalent.

e) This will make it possible to do lemmatisation and partial morphological annotation of those words which are presented in the text by their main forms. However, the process of lemmatisation of the dialectal corpus will not finish here. The left part of the word forms should be annotated manually.

This stage can be “semi-automatic” if the above mentioned functions of dictionaries are activated. The various ‘directions’ of manual annotation give us the following possibilities:

- Direction **word-form** > **Lemma** (if the word is already part of the dictionary) makes it possible to ascribe the part of speech marker automatically, as the lemma has already obtained it.
- Direction **word-form> literary equivalent of the word-form** makes it possible to ascribe automatically the literary lemma with the help of the morphological analyser of the literary language. Naturally, in this case, the part of speech marker will also be ascribed.
- In all of the cases when the lemmatisation of the Dialect Corpus takes place and when the main form of the dialect corpus is corresponded by a simple main form of the literary language, it will be possible to conduct complete morphological annotation, all the grammatical characteristics will be ascribed to the dialect lemma with the help of its 'linking' to the literary lemma.

Arguably the idea of including the dictionary as one of the tools in the process of lemmatisation and primary annotation can be used effectively while building the linguistic corpora of the languages the computational processing of which have not been completed.

პროექტები კორპუსის ლინგვისტიკის მიმართულებით ილიას სახელმწიფო უნივერსიტეტში

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

მოხსენებაში წარმოდგენილი იქნება ილიას სახელმწიფო უნივერსიტეტის ლინგვისტურ კვლევათა ცენტრის ორი პროექტი:

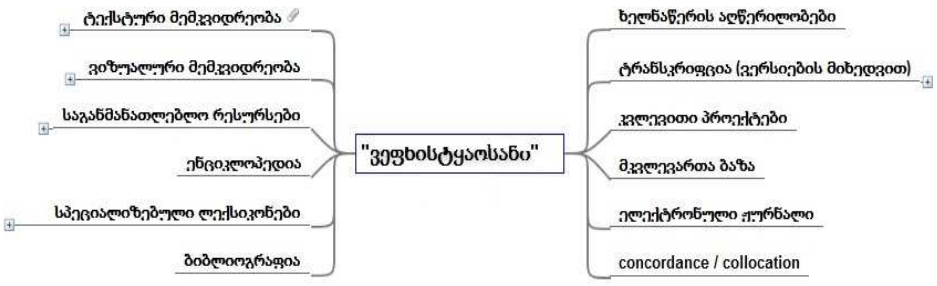
- „ვეფხისტყაოსნის“ ქართულ-ინგლისური კორპუსი;
- თანამედროვე ქართული ენის კორპუსი (1991–2011 წლები).

1. „ვეფხისტყაოსნის“ ქართულ-ინგლისური კორპუსი არის პირველი ეტაპი პროექტისა "ვეფხისტყაოსნის" ელექტრონული კორპუსი, რომლის მიზანია ხელი შეუწყოს ამ პოემასთან დაკავშირებული მონაცემების შეგროვებას, სისტემატიზაციას, გამოქვეყნებას და კვლევას კორპუსული ლინგვისტიკის მეთოდოლოგიით. კვლევის შედეგები, რომლებიც განთავსდება პორტალის ბიბლიოგრაფიულ, ტექსტოლოგიურ, ვიზუალურ და სხვა ბლოკებში (იხ. ვებ-პორტალის სქემა), შესაძლებლობას მისცემს როგორც ქართველ, ისე უცხოელ მეცნიერებს, იკვლიონ პოემის ტექსტის საკმაოდ ბუნდოვანი ისტორია, უმდიდრესი ტექსტური (ხელნაწერები, გამოცემები) და ვიზუალური (ილუსტრაციები, ყდა) მემკვიდრეობა, სხვადასხვა პერიოდში, განსაკუთრებით ვახტანგ მეექვსის ეპოქაში, პოემის ძირითადი ტექსტის ირგვლივ სხვადასხვა მიზნით შექმნილი მეტატექსტები.

ამ ეტაპზე პროექტის პრიორიტეტულ ამოცანებს წარმოადგენს:

1. ინტერნეტ-პლატფორმის // ვებ-პორტალის განვითარება;
2. ”ვეფხისტყაოსნის” ქართული (ყველა გამოცემა) და ინგლისური პარალელური ტექსტების კორპუსის შექმნა (სამომავლოდ მულტილინგვური კორპუსის შექმნის შესაძლებლობით);
3. ბიბლიოგრაფიის მოდულის შექმნა;
4. სხვადასხვა ტიპის ელექტრონული ვებ-მოდულების (ენციკლოპედია, ლექსიკონები და ა.შ) განვითარება.

ქვემოთ წარმოდგენილია „ვეფხისტყაოსნის“ ვებ–პორტალის სქემა.



2. თანამედროვე ქართული ენის კორპუსი

პროექტი 2011 წლიდან ხორციელდება ილიას უნივერსიტეტის ლინგვისტურ კვლევათა ცენტრში. იგი ემსახურება თანამედროვე ქართული ენის კორპუსის შექმნას საერთაშორისო პრაქტიკაში აღიარებული მეთოდოლოგიისა და მიღებული სტანდარტების გათვალისწინებით, მათ შორის:

XML (extensible Markup Language)[1]

ლაიპციგის გლოსირების წესები (Leipzig Glossing Rules)[2]

ძირითადი სტანდარტი: სეგმენტაცია (ISO 24614), ანოტაცია (ISO 24611, 246121 და 24615), მახასიათებლები (ISO 24610) და ლექსიკონები (ISO 24613)

დამატებითი სტანდარტი: მონაცემთა კატეგორიები (ISO 12620), ენობრივი კოდები (ISO 639 ან IETF BCP-47), სკრიპტ-კოდები (ISO 15924), ქვეყნის კოდები (ISO 3166), თარიღი (ISO 8601) და უნიკოდი (ISO 10646) [ISO-TC37].

თანამედროვე ქართული ენის კორპუსი იქმნება ლინგვისტიკასა და ტექნოლოგიურ მეცნიერებებში აღიარებული თანამედროვე მეთოდოლოგიით; ამავე პრინციპებითა და მეთოდოლოგიით არის შექმნილი ამერიკული ნაციონალური კორპუსი (American National Corpus ANC; იგულისხმება დეკოდირება XML ფორმატში კორპუსის დეკოდირების სტანდარტის შესაბამისად, Corpus Encoding Standard, XML ფორმატისათვის).

Projects on Corpus Linguistics at Ilia State University

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

The paper will present two projects of the Center of Linguistic Research at Ilia State University:

- *Georgian-English Corpus of the Epic Poem, "The Knight in the Panther's Skin";*
- *Corpus of Contemporary Georgian (1991–2011).*

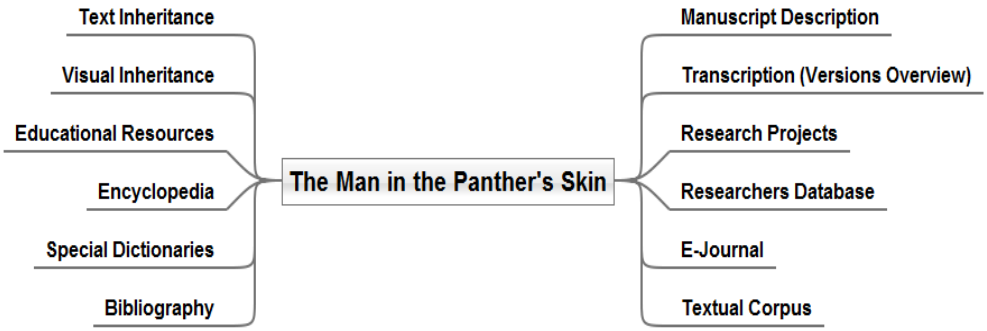
Brief Description of the Projects:

1. Georgian-English Corpus of the Epic Poem, *"The Knight in the Panther's Skin"* is the first part of the "Online Corpus of "The Knight in the Panther's Skin", a project aimed at supporting collection, systematisation and publishing of all of the data related to the poem as well as conducting research through the Corpus Linguistic Methodology. The research results which will be incorporated in bibliographical, textual, visual and other blocks (see the scheme of the Web Portal), will offer an opportunity to both Georgian and foreign scholars to study the misty history

[1] www.xml-ces.org
[2] http://www.eva.mpg.de/lingua/resources/glossing-rules.php

of the poem's text, rich textual (manuscripts, publications) and visual (illustrations, covers) heritages as well as meta-texts created in different periods of time (especially during the reign of the King Vakhtang VI).

Currently, the high priority targets of the project include the following:
a. To develop the platform for Internet // web-portal;
b. To create the Corpus of Georgian (including all the existing publications) and English parallel texts of the poem "The Knight in the Panther's Skin" (with an opportunity for developing further Multilingual Corpus);
c. To create the Bibliography Module;
d. To develop online web-modules of different kinds (encyclopedia, dictionaries etc.).



2. Corpus of Contemporary Georgian Language (1991-2011). The project aims at planning and creating the analysis and annotation software for contemporary Georgian Language. It will be implemented by Ilia State University Centre of Linguistic Research. The proposal presented includes project goals and objectives, project implementation plan with its stages and timeline designed for the next two years.
The main goal of the project is to create the analysis and annotation software for the Corpus of Contemporary Georgian Language (1991-2011). The Corpus is considered to be the most important part of Georgian National Corpus, the megaproject of Humanities. In addition to the project planned, i.e. the Corpus of Contemporary Georgian, the sub-corpora included in the Georgian National Corpus should be those of Old, Medial, New and the 20th Century Georgian language. The proposal contains linguistic and technological standards for the Corpus of Contemporary Georgian, timelines for developing the design of the Corpus, storing data, creating databases, as well as the stages of data elaboration and data annotation. It establishes procedures to create search systems, computer programs and a morphological analyser. The proposal focuses on project development perspectives displaying the importance of the project product, i.e. the Corpus of Contemporary Georgian (1991-2011), for the development of Georgian as a state language facing the current challenges of globalisation.

Methodology

The project aims at creating the analysis and annotation software using international methodology and widely accepted standards, including:
XML (eXtensible Markup Language)[1]
Leipzig Glossing Rules[2]
Key standards: word segmentation of written texts (ISO 24614), lexical markup framework (ISO 24613), feature structures (ISO 24610) and syntactic annotation framework (ISO 24611 and 24615).

Additional standards: specification of data categories and management of a data category registry for language resources (ISO 12620), codes for the representation of names of languages (ISO 639 or IETF BCP-47), script-codes (ISO 15924), country codes (ISO 3166), data elements and interchange formats (ISO 8601) and universal coded character set (ISO 10646).

[1] www.xml-ces.org
[2] http://www.eva.mpg.de/lingua/resources/glossing-rules.php

ზმნის სრული წარმომქმნელი კომპიუტერული მოდელი და კონცეპტუალური საკითხები

ლალი ეზუგბაია, თედო უთურგაიძე
არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველო)
ezugbaia@ice.ge

ზმნის სრული წარმომქმნელი მოდელის კომპიუტერული დამუშავებისათვის გადამწყვეტია უღლების პარადიგმების აგება, რაც ზმნის უღლების ტიპებად კლასიფიკაციასაც გულისხმობს. უკანასკნელ ათწლეულში არაერთი საინტერესო თვალსაზრისი გამოითქვა უღლების ტიპებთან დაკავშირებით, მაგრამ კონცეპტუალური სიახლით გამოირჩეოდა დამანა მელიქიშვილისა და თედო უთურგაიძის მოსაზრებები. აქვე აღვნიშნავთ, რომ კომპიუტერული პროგრამისათვის მთავარია, შეძლოს კონკრეტული ზმნური ფუძის უღლების პარადიგმის აგება და ზმნათა დაჯგუფება პარადიგმაში გამოყენებული მოდელის მიხედვით. თუ სრულად გამოვრიცხავთ ზმნისწინს და მისგან დამოუკიდებლად შევეცდებით უღლების პარადიგმის აგებას, ვაწყდებით რამდენიმე არსებითი ხასიათის პრობლემას:

1. ნებისმიერი ზმნური ფორმა წარმოადგენს მორფემათა კანონზომიერ თანმიმდევრობას. რანგების თეორიის თანახმად, ცენტრში დგას ზმნური ფუძე /ზოგჯერ ემთხვევა ძირს/ და, შესაბამისად, მარცხნივაა პრეფიქსთა რიგი ანუ სათანადო კატეგორიის მარკერები (გვარი, ვერსია, ობიექტის ნიშანი, სუბიექტის ნიშანი), ხოლო მარჯვნივ – სუფიქსთა რიგი, ასევე სათანადო კატეგორიების მარკერები (თემის ნიშანი, რომელიც ჩვენთვის მწკრივის ზეკატეგორიის მარკერია, სავრცობი, კონკრეტული დრო-კილოს მაწარმოებელი, პირის ნიშანი, რიცხვის ნიშანი). ცხადია, და ეს კარგად გამოჩნდა კომპიუტერული მოდელირების პროცესში, ქართული ენის ძირითადი ზმნური ფონდი ფონოტაქტიკის ძლიერ ზეწოლას განიცდის განურჩევლად პარადიგმის სახეობისა. აღმოჩნდება, რომ ზმნები ერთნაირად მოირგებენ უღლების მოდელებს, მაგრამ სხვაობა დაჩნდება ან რომელსამე მწკრივში, ან რომელსამე კომბინაციაში. ამდენად უღლების პარადიგმატული ტიპები მორფონოლოგიური კვალიფიკაციით გამოიყოფა (შდრ. ბრუნების მორფონოლოგიური ტიპები).

2. უღლების ტიპების კომპიუტერული მოდელირების პროცესში თანაბრად მნიშვნელოვანია ყველა გრამატიკული კატეგორია. ტრადიციული დაყოფა უღლებისა და წარმოქმნის კატეგორიებად არაფერს ნიშნავს: თუ კატეგორიის შემოყვანით ფორმის სტრუქტურაში ცვლილება ხდება, მაშასადამე, ის უღლების კატეგორიაა. მაგალითად, **წერ-ს** ზმნის პარადიგმები სამსავე ვერსიაში რომ აიგოს, ერთი ცვლილება უნდა განხორციელდეს: კომპიუტერული პროგრამის თანახმად, ქცევის კატეგორიის უჯრაში უნდა ჩაისვას **Ø**, ან **ი**, ან **უ**, შესაბამისი წესების დაცვით, ხოლო საზედაო სიტუაციის **ა** მაქცევრის დართვისას ასევე ახალი პარადიგმა უნდა აიგოს. იგივე ითქმის ნებისმიერ სხვა კატეგორიაზე.

3. უღლების ტიპების ბაზაზე სრული წარმომქმნელი მოდელი უნდა შეიქმნას, ანუ ნებისმიერი სახელური ფუძისგან ზმნის გენერაციის შესაძლებლობა უნდა ჰქონდეს კომპიუტერულ პროგრამას ასევე ყველა გრამატიკული კატეგორიის მიხედვით. ეს, თავის მხრივ, მოითხოვს **სახელის ბრუნების** ტიპების გათვალისწინებას.

4. ყველაზე რთულია ყველა შესაძლო ზმნისწინიან ფორმათა კომპიუტერული მოდელის შექმნა. როგორი აზრთა სხვადასხვაობაც არ უნდა არსებობდეს ზმნისწინის ფუნქციების შესახებ, ცხადია, ყველა დაჩნდება ფორმალურ დონეზე, რომლის მოდელირება ორ პრობლემას აწყდება:

ა. არც ერთი ზმნა არ დაირთავს იმავე ზმნისწინებს ისეთივე დანიშნულებით, როგორც ნებისმიერი სხვა ფუძე;

ბ. ზმნისწინიან ფორმათა სემანტიკური მოდელირება შეუძლებელი ხდება ზმნურ ფუძეთა სემანტიკური ველების გამოყოფის გარეშე, რადგან ხშირად ზმნისწინის სემანტიკის ცვლას ზმნური ფუძის ლექსიკური მნიშვნელობაც იწვევს;

5. თავისთავად უღლების მორფონოლოგიური ტიპების ბაზაზე უნდა მოხდეს სემანტიკური მოდელირება, რაც გზაა ქართული ზმნის სრული წარმომქმნელი მოდელის შექმნისკენ.

Complete Computational Model of the Verb Generation and Conceptual Issues

Lali Ezugbaia, Tedo Uturgaidze
Arn.Chikobava Institute of Linguistics (Georgia)
ezugbaia@ice.ge

In order to elaborate the complete derivational model of the verb in a computer mode, it is of crucial importance to build paradigms of conjugation which, in its turn, implies classification of verbs according to the types of conjugation. In the course of the last ten years quite a few interesting points of view have been made regarding the typology of verb conjugation, although those by D. Melikishvili and T.Uturgaidze were distinguished by their conceptual novelty. It is worth noting that it is essential for a computer program to be able to build the conjugation paradigm for a specific verb-form and to group the verbs according to the model used in the paradigm. If we exclude the preverb and make an attempt to build a paradigm without it, we will face several important issues:

1. Any verb-form is a regular succession of morphemes. According to the theory of rankings, the centre is occupied by the verb-stem (sometimes it coincides with the root). Consequently, on the left the set of prefixes or the markers of corresponding categories are arranged (such as voice, version, object and subject markers), on the right side - the set of suffixes with the markers of corresponding categories (for instance, the theme marker, which we consider to be the marker of a supra-category of the screeve as well as tense and mood formants, markers of the person and number). Clearly, this emerged in the process of computer modelling. The core verb fund of Georgian is under pressure of phonotactics notwithstanding the type of paradigm. It will be shown that the verbs will follow the similar models of conjugation though the difference will be revealed either in some screeve or in some combination. Thus, the paradigmatic types of conjugation are singled out by means of morpho-phonological qualification (compare morpho-phonological types of declination)

2. In the process of computer modelling of the types of conjugation, each grammatical category is equally important. Traditional classification in the categories of conjugation and derivation does not prove anything: if the introduction of the category causes changes in the form of the structure, then we deal with the category of conjugation. For instance, in order to build the paradigms of the verb **çers (he/she/it writes)** in all of the three versions, one change should be made according to the computer program. More specifically **Ø**, **i** or **u** should be inserted in the box of the category of version without violating the rules. A new paradigm should be built by adding the category of the upper location of the version . This is true for any other category as well.

3. A complete derivative model should be created on the basis of the types of conjugation, or the computer program should be able to generate the verb from any nominal stem, according to all of the grammatical categories. This means that the types of the nominal **declination** should be taken into account.

4. The most difficult part of this research is creation of the computer model of all the possible pre-verbal forms. In spite of the difference in views about the functions of the preverb, all of the latter will be revealed on the formal level. The modelling of the formal level will have to overcome two problems:

a) None of the verbs add the same pre-verb in the similar function to any other verbal stem

b) Semantic modelling of pre-verbal forms becomes impossible without singling out semantic fields of verbal forms as the change in the semantics of the pre-verb, among other things, is frequently caused by the lexical meaning of the verb stem

5. Semantic modelling should be based on the morpho-phonological types of conjugation which is the way towards the creation of the complete derivative model of the Georgian verb.

ქართველ ებრაელთა ლექსიკონის კომპიუტერული ბაზის შექმნა

რეუვენ ენოხი
არიელის უნივერსიტეტი (ისრაელი)
msar@mscc.huji.ac.il, reuvene @ariel.ac.il

მოხსენებაში განხილული იქნება ის პრინციპები და მეთოდიკა, რომელთა გამოყენებითაც შეიქმნება კომპიუტერული ბაზა ქართველ ებრაელთა ლექსიკონზე სამუშაოდ.

იგულისხმება ინდექსაციის სისტემის დამუშავება სახელთა ბრუნებისა და ზმნის უღვლილების სისტემისათვის, რაც შესაძლებლობას მოგვცემს გაირჩეს ქართული ენის "რეგულარული" ფორმები, და საკუთრივ, ქართველ ებრაელთა მეტყველების თავისებურებები.

სპეციალური მეთოდიკის მეშვეობით გამოცალკევდება აგრეთვე ფონეტიკურ მოვლენათა საფუძველზე აღმოცენებული განსხვავებული ლექსიკური ერთეულები, ივრითული სიტყვები, რომლებიც ქართველ ებრაელთა მეტყველებაშია გადმოსული (შესაბამად – სახეცვლელად თუ გარკვეული ცვლილებებით) და ადგილობრივ ქართულ დიალექტებში ხმარებული სიტყვები, რომლებიც ასევეა შესული ებრაელთა სიტყვათხმარებაში (უცვლელად თუ სახეცვლილი სახით).

Creation of the Computer Database of the Dictionary of Georgian Jews

Reuvene Enoch
Ariel University (Israel)
reuvene@ariel.ac.il, msar@mscc.huji.ac.il

The paper deals with the principles and methodology which will underlie the computer database of the dictionary of Georgian Jews. First of all, there will be elaboration of the indexing system for the systems of conjugation and declination which will enable us to distinguish 'regular' forms of Georgian from the specific features of the Georgian Jews' speech. Applying specific methodology, it will be possible to single out lexical items created on the basis of phonetic processes, words from Ivrit which have already entered the vocabulary of Georgian Jews (unchanged or with certain changes) and words used in local, Georgian dialects which are also attested in the speech of Georgian Jews either in modified or original forms.

ენის სინტაქსური მოდელირების პრინციპებისათვის

თამარ ვაშაკიძე, ნინო ჯორბენაძე, თეა ბურჭულაძე
არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველო)
tamarvashakidze@yahoo.com

ქართული ენის სინტაქსური ანალიზატორის შექმნა ბევრ სირთულეს უკავშირდება: მანქანამ უნდა განსაზღვროს მეტყველების ნაწილის სინტაქსური ფუნქცია ისეთ შემთხვევებშიც, როცა ეს უკანასკნელი ერთი და იმავე ფორმით არის წარმოდგენილი სხვადასხვა პოზიციაში და, შესაბამისად, სინტაქსური სტატუსიც განსხვავებული აქვს.

ენის სინტაქსური მოდელირებისათვის, უპირველეს ყოვლისა, საჭიროა ამა თუ იმ ლექსიკური ერთეულის ყველა შესაძლო ფორმის გათვალისწინება შესიტყვება-სიტყვათშეხამების დონეზე.

მოხსენებაში წარმოდგენილი იქნება, თუ რა წესების გამოყენებით დააფიქსირებს მანქანა ფორმობრივად ერთი და იმავე სიტყვის სხვადასხვა სინტაქსურ ფუნქციას, მაგალითად, ლექსიკური ერთეული **კარგი** ამ ფორმით (ზედსართავი სახელის სახ. ბრ.) წინადადებაში შეიძლება იყოს:

- განსაზღვრება (**კარგი** მეგობარი / წიგნი / მეტყველება...)
- შედგენილი შემასმენლის სახელადი ნაწილი (ის **კარგი** არის / იყო / იქნება...)
- გასუბსტანტივების შემთხვევებსა თუ საზღვრულჩავარდნილ პოზიციებში: ა) სუბიექტი (ძირითადად ზღაპრებში: **კარგი** ამბობს, გათენდებაო, ცუდი – არაო...), ბ) პირდაპირი ობიექტი (მგონი, რაღაც **კარგი** დაწერა მეგობრის შესახებ...)
- თანხმობის სემანტიკის გამომხატველი ნაწილაკი (ამ საღამოს გელოდები! – **კარგი**, მოვალ...).

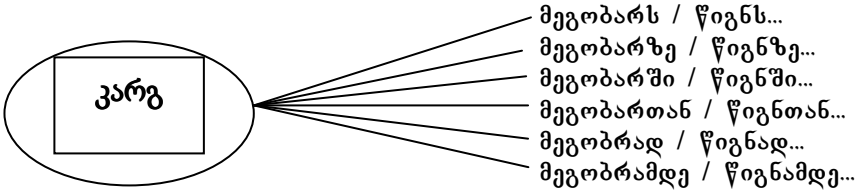
I. მანქანა **განსაზღვრების** სინტაქსური სტატუსით ამოიცნობს **კარგი** სიტყვას, თუ მას მოსდევს სახელობითი, ნათესაობითი ან მოქმედებითი ბრუნვის ფორმით წარმოდგენილი არსებითი სახელი, მასდარი ან მიმდეობა (**კარგი** მეგობარი / მეტყველება / ნაწერი; **კარგი** მეგობრის / მეტყველების / ნაწერის; **კარგი** მეგობრით / მეტყველებით / ნაწერით).

II. **შედგენილი შემასმენლის სახელადი ნაწილის** ფუნქციით დააფიქსირებს მანქანა **კარგი** ფორმას, თუ მას მოსდევს: ა) **არის** ზმნის შეკვეცილი ვარიანტი **ა** გარკვეულ პოზიციებში (ამინდი დღეს **კარგია**...), ბ) მეშველი ზმნები (ამინდი **კარგი** არის / იყო / იქნება...), გ) მეშველი ზმნის ფუნქციით გამოყენებული ზმნური ფორმები (კომპიუტერი **კარგი** გამოდგა...).

III. გასუბსტანტივების შემთხვევებსა თუ საზღვრულჩავარდნილ პოზიციაში მანქანა **სუბიექტის** სტატუსს მიანიჭებს **კარგი** ფორმას, თუ მარტივი წინადადება შეიცავს გარდამავალ ზმნა-შემასმენელს, რომელიც წარმოდგენილია I სერიის ნებისმიერი მწკრივის ფორმით (**კარგი** აშენებს / აშენებდა / აშენებდეს / ააშენებს / ააშენებდა / ააშენებდეს) ან – გარდაუვალს (სამივე სერიის გათვალისწინებით: **კარგი** იმალება / დაიმალა / დამალულა...), ხოლო – **პირდაპირი ობიექტის** ფუნქციით ამოიცნობს, თუ მარტივ წინადადებაში წარმოდგენილია გარდამავალი ზმნა-შემასმენელი II ან III სერიის ნებისმიერი მწკრივის ფორმით (**კარგი** ააშენა / ააშენოს / აუშენებია / აეშენებინა / აეშენებინოს).

IV. თანხმობის სემანტიკის გამომხატველ **ნაწილაკად** დააფიქსირებს მანქანა **კარგი** ფორმას, თუ მას მოსდევს სასვენი ნიშნები (**კარგი**, მოვალ / გეთანხმები / შეგხვდები... **კარგი**! – შევწყვიტოთ კამათი; ახლა ზღაპრები წავიკითხოთ. – **კარგი**) ან ფორმები – **რა**, **ერთი** (**კარგი რა**, ნუ გამიბრაზდები; **კარგი ერთი**, თავი დამანებე).

მოხსენებაში გათვალისწინებული იქნება აგრეთვე ის სინტაქსური პოზიციები, რომელთა საშუალებითაც მანქანა ამოიცნობს ნებისმიერი ბრუნვის ფორმით (აგრეთვე ფუძის სახით) დადასტურებული **კარგი** სიტყვის სინტაქსურ სტატუსს, მაგალითად, ერთ-ერთი სამანქანო წესი ასეთი იქნება: წინადადებაში წარმოდგენილი ყველა **კარგ** ფორმა განსაზღვრებაა:



აღსანიშნავია, რომ მანქანისათვის მიწოდებული **ძირითადი წესები** უნივერსალურია და, მაშასადამე, ისინი ვრცელდება სხვა მსაზღვრელ სიტყვებსა თუ შედგენილ შემასმენელზე. საჭიროების შემთხვევაში კი გამოიყენება **დამატებითი წესები** (რომლებიც ძირითადად შეეხება ელიფსურ წინადადებებს).

Towards the Principles of Syntactic Modelling of a Language

Tamar Vashakidze, Nino Jorbenadze,Tea Burchuladze
Arn. Chikobava Institute of Linguistics (Georgia)
tamarvashakidze@yahoo.com

Creating a syntactical analyser of the Georgian language is connected with many difficulties: a machine has to define a syntactic function of a word even in those cases when the latter, in the same forms, occurs in different positions with different syntactic statuses.

First of all, for a successful syntactic modelling of a language it is essential to take into account all of the possible forms of any lexical entry on the level of word-combinations and collocations.

The paper will represent the rules by means of which a machine can fix different syntactical functions of formally identical words; e. g.: a lexical entry **ḳargi** "good" (nom. case of an adjective) can occur in a sentence with this form: 1) determinative (**ḳargi megobari** / **çigni** / **meṭkqveleba**... "a good friend / book / speech...); 2) a nominal part of a compound predicate (**ḳargi aris** / **iqo** / **ikneba**... "It is / was / will be good"...); 3) in cases of substantisation and in the position without a modified noun: a) a subject (mainly in fairy-tales: **ḳargi ambobs gatendebao**, **cudi – arao**... "A good (man) says, dawn will break, a bad (man) – it will not"...); b) a direct object

(**mgoni, raγac ḳargi daçera megobris šesaxeb**... “I think he/she has written something good about his/her friend”...); 4) a particle expressing agreement (**am saγamos gelodebi! – ḳargi, moval**... “This evening I will be waiting for you! – well, I’ll come”).

I. A machine will recognize a word **ḳargi** according to a syntactic status of an **attribute**, if it is followed by a noun, a verbal noun (masdar) or a participle in nominative, genitive or instrumental cases: (**ḳargi megobari** / **meṭqveleba** / **naçeri** – “a good friend / speech / a piece of writing”; **ḳargi megobris** / **meṭqvelebis** / **naçeris** – “ of a good friend / speech / a piece of writing”; **ḳargi megobrist** / **meṭqvelebit** / **naçerit** – “with a good friend / speech / a piece of writing”).

II. A machine will recognise a **ḳargi** form as having the function of a **nominal part of a compound predicate** if it is followed by: a) a elided variant of a verb **aris** (“is”) - **a** in certain position (amindi dγes **ḳargia** – “ The weather is fine today”); b) **auxiliary verbs** (amindi **ḳargi** aris / iqo / iknеba... – “Weather is /was / will be fine...”); c) **verbal forms used with the function of an auxiliary verb** (ḳompiuṭeri ḳargi gamodga... – “A computer appeared to be good...”).

III. In cases of substantisation and in the position without a modified noun a machine will give a status of a **subject** to a **ḳargi** form, if a simple sentence contains: a transitive verb-predicate that is represented in any screeve of I series (**ḳargi** ašenebs / ašenebda / ašenebdes / aašenebs / aašenebda / aašenebdes – “A good (man) builds / built / would be building / will build / would have built / would have be building”) – or an intransitive verb-predicate (in three series: **ḳargi** imaleba / daimala / damalula – “A good (man) hides / hid / has hidden itself’...), or – a machine will recognize it with the function of a **direct object**, if a transitive verb-predicate is represented in a simple sentence in any screeve of I or III series (**ḳargi** aašena / aašenos / aušenebia / aešenebina / aešenebinos – “He built / would build / has built / had built / would have built a good (building...)”...)/

IV. A machine will recognise a **ḳargi** form as a **particle** expressing agreement, if it is followed by punctuation marks (**ḳargi**, moval / getanxmebi / šegxvdebi... – “Good, I’ll come / I agree with you / I’ll meet you...”; **ḳargi**! – ševçqviṭot ḳamati – “Good! – let’s stop disputing”; axla zγaṗrebi çaviḳitxot – ḳargi. – “Now let’s read fairy-tales. – Good”) or the forms – **ra**, **erti** (ḳ**argi ra**, nu gamibrazdebi – “Come on, don’t get angry with me”; ḳ**argi erti**, tavi damanebe – “Enough, leave me”).

The paper will take into account those syntactical positions, through which a machine will recognize a syntactical status of a **ḳargi** form evidenced in any case form (in the form of a stem, as well); e. g. : one of the machine rules will be like this: all **ḳarg** forms represented in a sentence are attribute:

megobars / çigns... to a good friend / book...
ḳarg megobarze / çignze... **about a good friend / book...**
megobarši / çignši... in a good friend / book...

megobartan /çigntan... with a good friend / book...
megobrad / çignad... as a good friend / book...
megobramde / çignamde... as far as a good friend / book...

It should be underlined, that the basic rules supplied to a machine are universal and thus, they can be used for other determinative words and compound predicates. If needed, additional rules will also be used (which basically are formulated for elliptical sentences).

ლიტვური ენის ინსტიტუტი და საინფორმაციო საზოგადოება: ა ამოცანები და გამოწვევები

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

თანამედროვე ტექნოლოგიების გავრცელებამ და საზოგადოების ცოდნის დონის ამაღლებამ ახალი გამოწვევების წინაშე დააყენა ლიტვური ენა, რომელსაც დღეს საინფორმაციო საზოგადოების გაზრდილი მოთხოვნების დაკმაყოფილება უწევს. ინტეგრაციული და ინტერდისციპლინარული მეცნიერების განვითარება მოითხოვს ენობრივი კვლევის ახალ, თანამედროვე მეთოდებს. ენის გამოყენების არეალის გაფართოება თავისთავად სვამს ახალი ენობრივი რესურსების კომპილაციისა და ადვილად ხელმისაწვდომი ელექტრონული ბაზების – ლინგვისტური არქივების შექმნის, დამუშავებისა და გამოქვეყნების საჭიროებას.

ევროკავშირის მულტიკულტურულ და მრავალენობრივ სივრცეში აქტუალური გახდა ეროვნული და ენობრივი იდენტობის შენარჩუნების საკითხი. ამ მიზნისათვის აუცილებელია ლიტვური ენის აქტიური კვლევა და საზოგადოებისათვის მისი შედეგების გაცნობა. ყველა ამ პროცესში აქტიურად მონაწილეობს ლიტვური ენის ინსტიტუტი, რომელიც წარმოადგენს ენების კვლევის კომპეტენტურ ცენტრს.

ამჟამად ინსტიტუტი დაკავებულია ფუნდამენტური თეორიული და გამოყენებითი კვლევებით, რომლებიც ეფუძნება უახლეს საინფორმაციო ტექნოლოგიებსა და რომელთა შედეგადაც იქმნება კომპიუტერული (CD-ROM da on-lain) პროდუქტები. ინსტიტუტი საზოგადოებას ასევე სთავაზობს ინტერნეტ პროდუქტებს www.kalbosnamai.lt, www.lkz.lt, www.likit.lt).

შევეცდებით, უფრო დაწვრილებით წარმოვადგინოთ ზოგიერთი პროექტი, რომლებიც ამჟამად მუშავდება ინსტიტუტში. ერთ-ერთი მათგანია სამუშაო ლიტვური ენის ციფრულ ლექსიკონზე (Lietuvių kalbos žodynas) _ მილიონზე მეტი სალექსიკონო ბარათის გაციფრება-ზე.

ეს იქნება თეზაურუსის ტიპის ლექსიკონი და მოიცავს ლიტვური ენის ყველა ქრონოლოგიურ და ადგილობრივ, ლოკალურ ვარიანტს. ლექსიკონი ეყრდნობა 1941-2002 წლებში გამოქვეყნებულ ტექსტებს, რომლებიც მოიცავს 11 500 000 სიტყვაფორმას და მათ 500,000 ამოსავალ ფორმას.

მეორე წყარო, რომელიც უმნიშვნელოვანესია საერთაშორისო ინდოევროპული ფუნდამენტური კვლევის თვალსაზრისით, არის ძველი ლიტვური ტექსტები, რომელთა კომპიუტერიზაცია ინსტიტუტში დაიწყო 1995 წელს.

ეს მონაცემთა ბაზა შედგება ოთხი ნაწილისაგან: კორპუსი, კონკორდანსი, ინდექსები და ციფრული გრაფიკა. სპეციალური შრიფტი პალემონასი (**palemonas)** სწორედ ამ წყაროების გრაფემებისათვის შეიქმნა. ვერც ერთი ისტორიული გამოკვლევა ვერ აუვლის გვერდს ძველ წერილობით წყაროებსა და დიალექტებს. ინსტიტუტის ციფრული დიალექტური არქივი ლიტვაში ყველაზე დიდ დიალუქტურ კოლექციას წარმოადგენს.

ინსტიტუტს აგრეთვე აქვს თანამედროვე ლიტვური ენის ელექტრონული ბაზები. გამორჩეული ყურადღება ექცევა ტერმინოლოგიას. ტერმინოლოგიური წყაროები ინახება მონაცემთა ბაზაში **ლიტვური ტერმინოლოგიის სინონიმია,** რომელშიც აღნუსხულია საერთაშორისო, ლიტვური და ჰიბრიდული ტერმინების სინონიმური რიგები.

ბოლოს, საჭიროა ვისაუბროთ ენობრივი კვლევის პოლიტიკის პრობლემებზეც. ევროკავშირის ენობრივმა უმცირესობებმა უნდა ითანამშრომლონ ენობრივი კვლევების თვასაზრისით. ეს მნიშვნელოვანია ფუნქციური ტენდენციების ჩამოყალიბებისათვის. ამჟამად რამდენიმე მონაცემთა ბაზა (ენის კონტექსტუალური გამოყენების ბაზა, ენის გამოყენების გზამკვლევი, საგნობრივი ბიბლიოგრაფია) მუშავდება კვლევითი მიზნებისათვის, თუმცა ზოგიერთი მათგანი შეიძლება გამოვიყენოთ ენის შესწავლის მიზნებისათვისაც.

დღესდღეობით, ერთ-ერთი უმთავრესი ამოცანაა სხვადასხვა მონაცემთა ბაზის ორმხრივი თავსებადობის შესაძლებლობების გამოძებნა და შესწავლა. ლიტვური ენის ინსტიტუტის კიდევ ერთი საპასუხისმგებლო მოვალეობაა ლიტვური მონაცემთა ბაზების ინტერნეტში განლაგება და მათი გავრცელება აკადემიურ სივრცეში.

Lithuanian Language Institute and Information Society: Tasks and Challenges

Dr. Jolanta Zabarskaitė
Lithuanian Language Institute (Lithuania)
jzabarskaite@yahoo.com

The spread of modern technologies and the rise of knowledge of present new challenges to the Lithuanian language is meeting the needs of information for society. The development of integrative and interdisciplinary fields of knowledge requires new up-to-date techniques of language research. The widening of the spheres of language usage leads to the compilation of new language resources and the creation of easily accessible databases as linguistic archives, their processing and publication. Belonging to the multi-cultural and multi-linguistic area of the European Union, the need arises to foster the national and linguistic identity by activating the research of the Lithuanian language and acquainting society with the results of this activity. In all of these processes the Institute of the Lithuanian Language is to play a vital role as a competent centre of language studies.

Currently the Institute is engaged both in fundamental theoretical and applied research, in which state-of-the-art informational technologies are used and computerised (CD-ROM and online products are created). The Institute offers Internet products, meant for the general public www.kalbosnamai.lt, www.lkz.lt, www.likit.lt).

Some projects, conducted at the Institute, can be presented in greater detail. One of them is the digitisation of *Lietuvių kalbos žodynas* (Dictionary of the Lithuanian Language) and millions of paper slips of the Dictionary files. This dictionary is of the thesaurus type and encompasses all chronological and local varieties of Lithuanian. The publication of the Dictionary spanned the

period between 1941 and 2002. The text of the Dictionary comprises 11,500,000 words, distributed among 500,000 entries.

Another source, fundamental to international Indo-European scholarship, is old Lithuanian writings. Their computerisation began at the Institute in 1995. This database consists of four parts: the corpus, concordances, indexes and digitised graphics. Special font *Palemonas* had to be created for the graphemes of the sources.

Not a single historical linguistic study can dispense with the data of old writings and dialects. The digital Dialect Archive of the Institute is the largest accumulation of the dialectal heritage in Lithuania.

The Institute also holds an accumulation of modern language usage, terminology being its particular concern. Terminological resources are kept in the database Synonymy of Lithuanian Terminology which stores lines of synonyms – various combinations of international, Lithuanian and hybrid terms.

Finally some problems of language policy should be dealt with. The minor languages of the European Union should cooperate in their linguistic research endeavours. This is important in establishing the functioning tendencies, as well as in the creation of the programs of machine translation and multilingual language resources. At present several databases (of language use, of usage advice, of subject bibliography) are being produced for scholarly purposes; some of them can also be used for Lithuanian language learning purposes.

At present one of the main tasks is to seek the reciprocal compatibility of different databases. The other great responsibility of the Institute of the Lithuanian Language is the placement the Lithuanian databases on the Internet and their spread in the academic sphere.

ქართულ-გერმანული ელექტრონული ლექსიკონის შედგენის ძირითადი პრინციპების შესახებ

რუსუდან ზექალაშვილი
ივ. ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)
rusudanz@gmail.com

ელექტრონულ სივრცეში ქართული ენის დამკვიდრებისთვის განსაკუთრებით მნიშვნელოვანია თარგმნითი ლექსიკონების შექმნა. ამ მხრივ ჯერ კიდევ ბევრი რამაა გასაკეთებელი. ქართულენოვან კომპიუტერულ ლექსიკოგრაფიაში წინგადადგმული ნაბიჯია ინგლისურ-ქართული დიდი ონლაინლექსიკონი, გერმანულში კი ამ თვალსაზრისით საქმე არც ისე კარგადაა. ინტერნეტში შეიძლება მხოლოდ მიხაელ იელდენის მოკლე ლექსიკონის მოძიება, რომელიც გარკვეულ დახმარებას უწევს მომხმარებელს, მაგრამ არ შეიცავს ვრცელ ინფორმაციას (სიტყვათა რაოდენობის სიმცირისა და თვით მონაცემთა სიმწირის გამო).

თარგმნითი ელექტრონული ლექსიკონის შექმნისთვის აუცილებელი წინაპირობაა უცხოური და ქართული ონლაინლექსიკოგრაფიის გამოცდილების შესწავლა და გაზიარება (მაგალითად, „ლინგვოსა“ და „მულტიტრანსის“ ტიპის ლექსიკონები, ინგლისურ-ქართული ლექსიკონი და სხვ.).

დაწყებულია მუშაობა ქართულ-გერმანული ონლაინლექსიკონის შესაქმნელად, რომელიც დაეფუძნება ჩვენ მიერ გამოცემულ თარგმნით ბეჭდურ ლექსიკონებს, მაგრამ, ცხადია, არ იქნება მათი სტერეოტიპული ანალოგი.

ლექსიკონი მომხმარებელს მისცემს დამატებითი ფუნქციების გამოყენების შესაძლებლობას: ორივე ენის სიტყვების სრულ გრამატიკულ, სემანტიკურ და სტილისტიკურ დახასიათებას, ძიების მაქსიმალურ შესაძლებლობას სხვადასხვა პარამეტრის მიხედვით (მეტყველების ნაწილთა ფორმაცვალება, სიტყვაწარმოება, კომპოზიცია, შესიტყვებები, ფრაზეოლოგია და სხვ.), რასაც უზრუნველყოფს სიტყვაწარმოებითი მოდელები, სახელური და ზმნური პარადიგმები, მეთაურ სიტყვებთან ნაწარმოები ლექსიკური ერთეულებისა და კომპოზიტების მითითება, ზედსართავ სახელებთან – ხარისხის ფორმების, ზმნებთან – გარდამავლობის, გვარისა და მართვის, მწკრივთა ფორმების, შესაბამისი საწყისისა და მიმღეობების ჩვენება, სანიმუშო მაგალითების მოყვანა კონოტაციების გათვალისწინებით, ფრაზეოლოგიური ერთეულებისა და ანდაზების ადეკვატური ერთეულების შერჩევა (ან აღწერილობითი ახსნა-განმარტება).

ცხადია, ლექსიკონში აისახება ლექსიკის სისტემური მიმართებები: პოლისემია და ომონიმია, სინონიმია და ანტონიმია, რაც შესაძლებლობას მოგვცემს, რომ ძიება მოხერხდეს დასახელებული კომპონენტების მიხედვითაც. სიტყვების თემატური კლასიფიკაცია გაზრდის ლექსიკონის გამოყენების შესაძლებლობებს.

უცხო ენის შემსწავლელთა მოთხოვნების გათვალისწინებამ განაპირობა გერმანული სიტყვების ფონეტიკური და გრამატიკული დახასიათების აუცილებლობა: მითითებული იქნება სიტყვის მახვილი, ნასესხებ სიტყვათა ტრანსკრიფცია, მორფოლოგიურ-სინტაქსური ინფორმაცია, გამოყენების სფერო და სიტყვის ემოციური შეფერილობა.

კლავიატურის მოდელის გადართვით სასურველ ენაზე მომხმარებელს ექნება სიტყვების ამ ენაზე აკრეფის საშუალება, ხოლო ლექსიკონის მოხერხებული ინტერფეისი საშუა-

ლებას მისცემს მათ, მოძებნონ არა მარტო ცალკეული სიტყვის ან შესიტყვების, არამედ ფრაზეოლოგიური ერთეულისა და ანდაზის ეკვივალენტი მეორე ენაზე.

მოხმარების გასაიოლებლად ლექსიკონში გამოვიყენებთ შემოკლებებსა და სპეციალურ სიმბოლოებს. შესაძლებელი იქნება ახალი სიტყვების დამატება, შესწორებების შეტანა, რაც თანდათან შეავსებს და დახვეწს ლექსიკონს.

დღეისათვის მზად გვაქვს ლექსიკონის ციფრული ვერსია. მისი მრავალდონიანი ლექსიკოგრაფიული პარამეტრიზაციის გათვალისწინებით ვამუშავებთ ლექსიკონის ბაზის სტრუქტურას. გვინდა, რომ ეს ლექსიკონი იყოს პარალელური ტექსტური კორპუსების შექმნისკენ გადადგმული ერთი ნაბიჯი (ქართულ-გერმანული ან მრავალენოვანი კორპუსებისთვის), რაც დააჩქარებს ჩვენი ენის ინტეგრაციას მსოფლიო ელექტრონულ სივრცეში.

Towards of the Basic Principles of Producing the Georgian-German Electronic Dictionary

Rusudan Zekalashvili
Iv. Javakhishvili Tbilisi State University (Georgia)
rusudanz@gmail.com

In order to secure a place for the Georgian language in electronic space, it is very important to produce inter-lingual dictionaries. Much still needs to be done in this sphere though the Comprehensive English-Georgian Online Dictionary is a great step forward in Georgian computational lexicography. In German, there exists only a brief dictionary by Michel Yelden via the Internet that helps the customer but only to some extent. This dictionary does not include detailed information due to the limited number of words and scarcity of the data.

In order to produce a translational electronic dictionary it is crucial to study and share the experience of foreign and Georgian online lexicography (for instance, “Lingvo”, a Georgian-English on-line dictionary, etc).
The work has started on compilation of a Georgian-German online dictionary which will be based on the translational published dictionary though will not be their stereotypical analogue.

The Dictionary will enable the reader to use additional functions, such as complete characteristics of grammatical, semantic or stylistic parameters; maximum opportunities of the search according to various parameters (such as transformation of parts of speech; word composition; composition; collocation; phraseology) which is provided by the word formation models; nominal and verbal paradigms, indication of lexical items and composites derived from headwords, degree forms (in the case of adjectives), indication of transitivity, voice and agreement, forms of screeves, infinite and participle forms, samples and illustration with the connotations, selection of adequate idioms and proverbs (or written explanation) .

It is obvious that the dictionary should include systemic lexical interrelations: polysemy and homonymy, synonymy and antonymy that will give us an opportunity to search according to the listed components. Thematic classification of words will increase the possibilities of the dictionary usage.

The necessity for a phonetic and grammatical description of German words arose due to the demands of learners of German: the dictionary will indicate the stress of the word, transcriptions of loan-words, morphological and syntactic information, emotional colour and the domain of usage of words.

By switching a keyboard to the language desired, users will type any word and will be given a chance to look up not only words and collocations, but, also, the equivalents of phrasal units and of proverbs in the other language.

We will use abbreviations and special symbols to facilitate the use of the dictionary. New words can be added and corrections can be made, so that the dictionary will expand and be revised.

We have already completed the digital version of the dictionary. The basic structure of the dictionary is being processed considering its multilevel lexicographic parameters. We would like this dictionary to be a step towards to creating a parallel text corpora (Georgian-German or multilingual corpora), that will make the process of the integration of our mother language into the world electronic space easier.

ქართული სიტყვათშეხამებითი მოდელები პროგრამაში „Sketch engine“

დევიდ თაგუელი
საინფორმაციო ტექნოლოგიების კვლევითი ინსტიტუტი (შოტლანდია)
dtugwell@gmail.com

შესავალი

ელექტრონული კორპუსები სულ უფრო და უფრო ხელმისაწვდომი ხდება დაინტერესებული საზოგადოებისათვის. ეს კი ხელს უწყობს ლექსიკოგრაფიული სამუშაოების წინსვლასა და განვითარებას მასალის მომცველობითობისა და ობიექტური კვლევის თვალსაზრისით; იგივე ვითარებაა ქართულშიც. მაგალითისათვის საკმარისია თუნდაც რეიფილდის შრომა (2006), რომელშიც ფართოდ არის გამოყენებული ქართული კორპუსული მონაცემები. თავდაპირველად ვრცელ ტექსტურ კორპუსებზე სამუშაოდ გამოიყენებოდა კონკრეტული საკვანძო სიტყვის ძიების მექანიზმი (KWIC), თითოეული სიტყვა წარმოდგენილი იყო კონტექსტისეული ფორმით. ტექსტური მასივების მუდმივი ზრდა სულ უფრო ართულებს ამ მეთოდით მუშაობას. ადვილია კორპუსში მოიძიო და შეისწავლო კონკრეტული სიტყვის 40 განსხვავებული კონტექსტური გამოვლინება, მაგრამ წარმოუდგენლად რთულია ამ მეთოდით მუშაობა 400, 4000 ან 40.000 სიტყვაფორმის შემთხვევაში. სწორედ უსასრულოდ ვრცელ ლექსიკოგრაფიულ მონაცემებში ავტომატური ძიების გასამარტივებლად შეიქმნა სიტყვათშეხამებითი მოდელების საძიებო სისტემა (კილგარიფი, თაგუელი, 2002) რომელსაც უკვე იყენებენ წამყვანი ლექსიკონების გამომცემლები როგორც გაერთიანებულ სამეფოში, ისე მის საზღვრებს გარეთ. მოხსენებაში ვისაუბრებთ, თუ რა არის აუცილებელი მსგავსი ქართული პროგრამის შესაქმნელად; განვიხილავთ, კიდევ რა შეუძლია შემოგვთავაზოს სიტყვათშეხამებითი მოდელების საძიებო სისტემამ (Sketch engine) ქართული ენის რესურსების დამუშავებისთვის.

1. სიტყვათშეხამებითი მოდელები

სიტყვათშეხამებითი მოდელი არის კონკრეტულ კორპუსში დამოწმებულ სიტყვათშეხამებათა (ლექსემის ლექსიკურ-გრამატიკულ მიმართებათა) ავტომატურად წარმოქმნადი მოკლე აღწერილობა ანუ რეზიუმე. ეს არის სემანტიკური სიტყვათშეხამებები (კოლოკაციები), რომლებიც ჯგუფდება გრამატიკულ-სემანტიკური მიმართებების მიხედვითაც. მაგ., კონკრეტული არსებითი სახელის სიტყვათშეხამებითი მოდელი გვიჩვენებს ყველა იმ ზმნის ჩამონათვალს, რომელიც სუბიექტად შეიწყობს ამ სახელს და აგრეთვე იმ ზმნების ჩამონათვალსაც, რომლისთვისაც ის ობიექტია. აქვე იქნება მოცემული იმ ზედსართავ სახელთა სიხშირული ნუსხაც, რომელიც შეიძლება შეგვხვდეს ამ არსებითი სახელის მსაზღვრელად; გვექნება წინდებულთა ჩამონათვალიც ამ სახელთან მიმართებით. სიხშირული მონაცემების მიხედვით სიტყვაფორმათა გამოვლინებების ნახვა საშუალებას აძლევს ლექსიკოგრაფს, სწრაფად გამოყოს ძირითადი მნიშვნელობა ლექსიკონში შესატანად. იქვე, კოლოკაციის ღილაკის გამოყენებით, შესაძლებელია კორპუსში დამოწმებული ყველა კოლოკაციის ნახვა.

სიტყვათშეხამებითი მოდელის გენერირებისათვის, უპირველეს ყოვლისა, აუცილებელია ენის ანოტირებული კორპუსი; ე.ი. თითოეული სიტყვა უნდა იყოს ინდექსირებული მეტყველების ნაწილთა მიხედვით, სხვა მახასიათებლების მიხედვით, ლემების (მეთაური

ფორმების) მიხედვით, ეს ნიშნავს იმას, რომ თითოეული სიტყვის სალექსიკონო ფორმა მარკირებულია. მეორე მოთხოვნა, რაც აუცილებლად უნდა დაკმაყოფილდეს, არის „სიტყვათშეხამებითი მოდელების“ გრამატიკის შექმნა, რომელიც გვაძლევს საძიებო სიტყვათშეხამების შესახებ მოკლე გრამატიკულ ინფორმაციას და აჯგუფებს ერთეულებს აღნიშნული ლინგვისტური მახასიათებლების მიხედვით.

დღეისათვის სიტყვათშეხამებითი მოდელების საძიებო სისტემა მოიცავს 42 ენის კორპუსს სიტყვათშეხამებითი მოდელებითურთ, რომელშიც ათასობით ასეთი ერთეულია წარმოდგენილი.

2. ქართული სიტყვათშეხამებითი მოდელების შექმნა

იმისათვის, რომ შეიქმნას ქართული სიტყვათშეხამებითი მოდელები, უპირველეს ყოვლისა, უნდა გვქონდეს ქართული ტექსტების ვრცელი კორპუსი. ამის შესრულება იოლია, რადგან ვებგვერდებზე უამრავი ქართული ტექსტია განთავსებული. იმისათვის, რომ უფრო გამოსადეგი იყოს კორპუსის ტექსტები, რომლებიც, როგორც წესი, მოიცავს დაბეჭდილ და გამოუქვეყნებელ ბელეტრისტიკას, ჟურნალებსა და აკადემიურ გამოცემებს, ყოველ ტექსტს უნდა მიეთითოს ჟანრი, წყარო, სათაური და ავტორი (თუ ცნობილია). ამგვარი მონიშვნის საშუალებით ავტომატურად იქმნება სუბკორპუსი; დღეისათვის, ქართულ კორპუსში დაახლოებით 50 მილიონი სიტყვაა და ეს რაოდენობა საკმარისია ყველაზე გავრცელებული სიტყვების საინტერესო სიტყვათშეხამებითი მოდელების შესაქმნელად.

ჩვენი სამუშაოს შემდეგი ეტაპია კორპუსის ინდექსაცია და ლემატიზაცია. ეს საქმე კი უნდა დავიწყოთ რეიფილდის კორპუსის ქართული ნაწილის მანქანური ვერსიის გამოყენებით (2006). დასაწყისისათვის შეიძლება ნაწილობრივი ანოტირებაც, გამომდინარე იქიდან, რომ გარკვეული საკითხები ჯერ კიდევ საკამათოა ქართული ზმნის მორფოლოგიაში; თუმცა, ამგვარი ანოტირებაც ალბათ საინტერესო იქნება კონკრეტულ გრამატიკულ მიმართებებზე დასაკვირვებლად. ამის შემდეგ ამ სიტყვათა ინდექსების მიხედვით უკვე უნდა შეიქმნას ქართული სიტყვათშეხამებითი მოდელები, რომლებიც ასახავენ ძირითად გრამატიკულ მიმართებებს სახელებს, ზმნებს, ზედსართავებსა და ზმნიზედებს შორის.

3. სიტყვათშეხამებითი მოდელების საძიებო სისტემის სხვა შესაძლებლობები

სიტყვათშეხამებითი მოდელების საძიებო სისტემა კორპუსის ანალიზატორია, რომელიც, როგორც წესი, ქსელშივეა ხელმისაწვდომი. შესაძლებელია მისი ჩამოტვირთვაც. სისტემა არა მხოლოდ ავტომატურად აგებს სიტყვათშეხამებითი მოდელებს, არამედ მასშტაბური კორპუსის ანალიზის მრავალფეროვან საშუალებებსაც მოიცავს.

ესენია:

- სიტყვათა ჩამონათვალი, ძირითადი საძიებო (KWIC) ფუნქციები, კორპუსისა და ქვეკორპუსის შედარებები;
- ორი სიტყვის განსხვავებულ მახასიათებლებზე დაკვირვება შესაძლებელია მოდელის განსხვავებების აღნიშვნით. ამგვარად შეიძლება წარმოჩინდეს ის შემთხვევები, როცა სიტყვებს აქვთ საზიარო კოლოკაციები, ასევე შემთხვევები, როდესაც კოლოკაცია სიტყვათა წყვილიდან მხოლოდ ერთ-ერთისთვის არის რელევანტური. ეს მომენტი უაღრესად არსებითია არასრულ სინონიმებს შორის სხვაობათა გამოსავლენად (შდრ.: „strong“ da „powerful“ ინგლისურში).
- თითოეული სიტყვისათვის ცალკეულად გამოითვლება დისტრიბუციული თეზაურუსი. ეს გვიჩვენებს კორპუსში სიტყვათა ჩამონათვალს, რომელიც გაწყობილია კონკრეტული სიტყვის რეალიზების ყველაზე ხშირი კონტექსტების მიხედვით.

● ფუნქცია WebBootCaT-ის საშუალებით ავტომატურად აიგება ვებგვერდიდან აღებულ ძირეულ სიტყვათა კორპუსი. ეს ფუნქცია განსაკუთრებით მნიშვნელოვანია სპეციფიკური კორპუსებს ასაგებად.

ჩამოთვლილი ფუნქციები ზრდის კორპუსის აპლიკაციათა მნიშვნელობას არა მხოლოდ ლექსიკოგრაფიული, არამედ სხვა სახის ლინგვისტური კვლევებისთვისაც; იგი გამოდგება აგრეთვე ენის შესასწავლად და ენის ავტომატური ანალიზისათვის.

4. დასკვნა

ამ მოხსენებაში ნაჩვენებია, რა უპირატესობა აქვს ქართული ენისთვის, რომლის ლინგვისტური რესურსები ნაკლებადაა განვითარებული, ისეთი კომერციული პროდუქტის გამოყენებას, როგორიცაა სიტყვათშეხამებითი მოდელების საძიებო სისტემა (Sketch engine) და, შესაბამისად, დიდ კორპუსებში ინფორმაციის დამუშავებისა და მიღების სფეროში ათეული წლის განმავლობაში დაგროვებული გამოცდილების გამოყენებას.

ეს პლატფორმა მუდმივ განვითარებადია. ის საშუალებას აძლევს ლინგვისტს მოახდინოს სპეციფიკური ლინგვისტური საკითხების არასაკმარისი რესურსების კონცენტრირება ტექსტური კორპუსის კოლექციაში და მისი საფუძვლიანი ანალიზიც. აგრეთვე, ის მოიცავს ენის არსში წვდომის მდიდარ საშუალებებს და გვთავაზობს კვლევის ახალ გზებს.

Georgian Word Sketches in the Sketch Engine

David Tugwell
Information Technology Research Institute (Scotland)
dtugwell@gmail.com

0. Introduction

The increasing availability of electronic text corpora has resulted in tremendous improvements for lexicography in terms of coverage and objectivity. This has also been the case for Georgian with, for instance, the extensive use of corpus data in Rayfield (2006). Previously, the standard tool for viewing these large text corpora was a display of Keyword in Context (KWIC), where every instance of the word being looked at is shown in the context in which it occurs. However, as the amount of available text grows, this method becomes increasingly problematic – it is one thing to examine forty occurrences of a word in a corpus one-by-one, but this soon becomes unfeasible when there are 400, 4000 or even 40,000 occurrences. The Word Sketch (Kilgarriff & Tugwell, 2002) was devised in order to deal with such large volumes of data by automatically finding lexicographically-interesting patterns and is now used by many leading dictionary publishers in the UK and elsewhere. This paper examines what is needed to produce similar Word Sketches for Georgian and how these, and other utilities provided by the Sketch Engine[1] system, might be of use for the development of Georgian language resources.

[1] Available at sketchengine.co.uk

1. Word Sketches

A Word Sketch is an automatically-generated summary of a word's significant patterns of behaviour in a particular corpus. These patterns are essentially collocational in nature, but sorted according to grammatical relations. For example, a Word Sketch for a given noun will typically display a list of the most salient[1] verbs having the noun as subject and another list of verbs taking the noun as object. Similarly, there will be a saliency-ordered list of those adjectives that most typically modify the noun and possibly more complex relations involving prepositions. Viewing this list of most typical usages makes it possible for the lexicographer to quickly draw up the senses needed for a dictionary entry. Actual examples from the corpus can be accessed by clicking on the collocation.

The calculation of Word Sketches requires in the first place an electronic corpus of the language that has been tagged, i.e. each word has been assigned a part of speech according to some tagset, and lemmatised, i.e. the dictionary form has been marked for each word. The second requirement is a "sketch grammar" that specifies what patterns are to be shown in the Word Sketch and how they are to be found in the corpus using pattern-matching over the part-of-speech tags. At present, the Sketch Engine contains corpora for 42 languages with Word Sketches being available for around a dozen of these.

2. Producing Georgian Word Sketches

To produce Word Sketches for Georgian, the first requirement is to collect a large corpus of Georgian texts. This is quite straightforward given the large amount of text available on the Web. To increase the usefulness of the corpus texts where mainly collected from large depositories, covering published and unpublished fiction, journals, and academic writing, where each text could be assigned attributes such as genre, source, year, title and author (where known). Such mark-up allows the automatic creation and comparison of sub-corpora. At present, there are around 50 million words in this Georgian corpus, which is sufficient to produce interesting Word Sketches for most common words.

The next step is to tag and lemmatise the corpus and a start has been made on this task by using a machine-readable version of the Georgian side of Rayfield (2006). Although there are many intricacies of verbal morphology in particular that are still to be tackled, even a partial tagging allows many interesting relations to be found. A sketch grammar was then written on the basis of these word tags to cover the basic grammatical relations between nouns, verbs, adjectives and adverbs.

3. Other facilities in the Sketch Engine

The Sketch Engine is a corpus analysis tool, usually accessed online but also available to download, that not only automatically constructs Word Sketches, but also makes available a wide range of other utilities for large corpora. These include:

[1] The Sketch Engine offers various alternative statistical measures for calculating saliency.

- Word lists, basic KWIC functions, corpus and sub-corpus comparisons.
- The contrast in the behaviour of two words can be displayed with a “Sketch Difference”, highlighting areas where the words share collocations and also where a collocation is significant only for one of the pair. This has been shown to be very useful to tease apart the differences in usage between words that are inexact synonyms (such as “strong” and “powerful” in English).
- A distributional thesaurus is also automatically calculated for each word. This shows an ordered list of the words in the corpus that occur in the most similar contexts to the word in question.
- The WebBootCaT utility allows for the automatic construction of a corpus from the web given a set of seed words. This is particularly useful for quickly constructing corpora in specialist domains.

This extends the application of the Sketch Engine to cover not only lexicographic research, but also other linguistic research, language learning, and tasks in the automatic analysis of language.

4. Conclusion

This paper has attempted to demonstrate the advantages for a language such as Georgian, where linguistic resources are less developed, of using a commercial resource such as the Sketch Engine, and thus being able to make use of more than a decade of research into the processing and display of information in large corpora. The platform is in constant development and allows the linguist to concentrate scarce resources on the language specific issues of collecting a corpus of texts and carrying out a basic linguistic analysis of this, while allowing a wealth of insights and new ways of viewing the language.

ბიბლიოგრაფია / References

Kilgarriff A., Tugwell D., Sketching words. In Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins. Marie-Hélène Corréard (Ed.) EURALEX: 125-137, 2002.

Rayfield D. (ed.), A Comprehensive Georgian-English Dictionary. Garnett: London, 2006.

ლიტვური ენის უახლესი ელექტრონული ლექსიკონები და ლიტვური დიალექტური კორპუსები

ვიდას კავალიაუსკასი
ვილნიუსის პედაგოგიური უნივერსიტეტი, ლიტვური ფილოლოგიის განყოფილება (ლიტვა)
vk1119@gmail.com

საინფორმაციო ტექნოლოგიების განვითარება ენათმეცნიერების წინსვლას უწყობს ხელს და ლინგვისტური მიღწევების პრაქტიკული გამოყენების მრავალ ეფექტურ შესაძლებლობას ქმნის. ელექტრონული ლექსიკონები და სხვადასხვა მიზნით მეცნიერულ საფუძველზე შექმნილი ენის კორპუსები უფრო და უფრო პოპულარული ხდება ლიტვაში.

მოხსენებაში განხილულია სტანდარტული ლიტვური ენის სამი უნივერსალური ლექსიკონი, რომლებიც სხვადასხვა სამეცნიერო და პრაქტიკული მიზნებისათვის შეიძლება გამოიყენონ როგორც ფონეტიკოსებმა, მორფოლოგებმა, ლექსიკოლოგებმა, სინტაქსისა და სხვა ფილოლოგიური დარგების სპეციალისტებმა, ასევე სტუდენტებმა და, ზოგადად, ლიტვურ ენაზე მოლაპარაკე საზოგადოებამ. საგულისხმოა, რომ დიალექტური ტექსტები და კორპუსები გადატანილია CD ფორმატში.

1. ლექსიკონი

1.1 **მე-20 საუკუნის პრესის ენაში ხშირად გამოყენებული სიტყვები: სიხშირეთა ელექტრონული ლექსიკონი (CD)** (პროექტის მენეჯერი პროფ. ა. პაკერისი; დაიბეჭდა 2004 წელს). ეს პუბლიკაცია მოიცავს ლექსიკონის წინასიტყვაობასა და კომპაქტურ დისკს (CD), რომელზეც ჩაწერილია ლექსიკონი MS Access and MS Word ფორმატში. ლექსიკონი ეყრდნობა ლიტვური წიგნებისა და მეოცე საუკუნის ლიტვური პერიოდიკის ტექსტურ მასალას (ტექსტები შეირჩა გარკვეულ მეთოდოლოგიაზე დაყრდნობით: 1925, 1935, 1945, 1955, 1965, 1975, 1985, 1995 წლები).

ლექსიკონი საშუალებას იძლევა გამოვიყენოთ იგი როგორც ელექტრონული ლექსიკონის ნებისმიერი სახე, მაგალითად, ძიება შეიძლება დაეყრდნოს მთავარ სიტყვებს (ანბანის, საშუალო სიხშირის ინდიკატორებზე დაყრდნობით) ან სიტყვების გარკვეულ გრამატიკულ ფორმებს (ანბანის, ზოგადი საშუალო სიხშირის ინდიკატორებზე დაყრდნობით).

ლექსიკონი საშუალებას იძლევა მოვძებნოთ კონკრეტული მთავარი სიტყვა, მისი რაიმე ფორმა ან ნაწილი (მაგალითად, სიტყვა სპეციფიკური სუფიქსითა თუ პრეფიქსით, ან სიტყვა ამა თუ იმ გრამატიკული ფორმით – მაგალითად, მხოლობითი რიცხვის აკუზატივით). ლექსიკონი იძლევა არა მხოლოდ მეთაური სიტყვის ან კონკრეტული სიტყვაფორმის ძიების საშუალებას, არამედ წარმოგვიდგენს მთელ პასაჟს იმ ტექსტიდან, რომელშიც იგი აქტუალიზდება.

ლექსიკონი შეიცავს დაახლოებით 30 000 დამოუკიდებელ ლექსიკურ ერთეულს, დაახლოებით 100 000 სიტყვაფორმასა და მათ 550 000 ამოსავალ ფორმას.

ამ ლექსიკონს კიდევ ერთი პრაქტიკული დანიშნულება აქვს: მომხმარებელს შეუძლია მისთვის საინტერესო სიტყვებისა და მათი ფორმების „კალათაში“ ჩაყრა, ახალი მასალის ფაილების შექმნა, მათი რედაქტირება და გამოყენება საჭიროებისამებრ.

ლექსიკონში აღრიცხული სიტყვების გამოყენების სიხშირის მაჩვენებელი არის ინფორმაციის ბრწყინვალე წყარო კოდირებისა და ლიტერატურულ ენაში პოტენციური ვა-

რიანტების სტანდარტიზაციის თვალსაზრისით. სტანდარტული სიტყვების ლექსიკონი შეიქმნა მკვლევრების, მასწავლებლებისა და სტუდენტებისათვის, რომლებსაც აინტერესებთ სიტყვათა გამოყენება, მათი ვალენტობა, სიხშირე, ლექსიკისა და გრამატიკის საკითხები და სხვ.

1.2 **საკუთარი სახელები მეოცე საუკუნის პრესაში: სიტყვათა სიხშირის ელექტრონული ლექსიკონი (CD)** (პროექტის მენეჯერი – პროფ. ა. პაკერისი, გამოქვეყნდა 2005წ). ეს ელექტრონული ლექსიკონი არის მომდევნო ნაწილი **მე-20 საუკუნის პრესის ენაში ხშირად გამოყენებული სიტყვების ლექსიკონისა: სიტყვათა სიხშირის ელექტრონული ლექსიკონი**, რომელიც 2004 წელს გამოქვეყნდა. ეს ლექსიკონი განკუთვნილია მკვლევრების, მასწავლებლებისა და სტუდენტებისათვის, რომლებიც დაინტერესებულნი არიან საკუთარი სახელების, ტოპონიმებისა და სხვა საკუთარი სახელების აღმნიშვნელი სიტყვების (თეონიმების, ზოონიმების, კოსმონიმების) გამოყენების თავისებურებებით, ვალენტობითა და სიხშირით, მართლწერის წესების თავისებურებებითა და მათი განვითარებით.

1.3 დ. მიკულენიენეს, ა. პაკერისისა და ბ. სტუჯიას მიერ შედგენილი წიგნი **სალიტერატურო ლიტვური ენის მახვილის ლექსიკონი** (2007, 2008) შეიცავს 2 ელექტრონულ ლექსიკონს (CD). **ხშირად გამოყენებულ სიტყვათა ლექსიკონი** შეიცავს თანამედროვე ლიტვური ენის ლექსიკონიდან ამოღებულ დაახლოებით 45 000 სიტყვას (ეს ლექსიკონიც ელექტროფორმატით გამოქვეყნდა); სიტყვების გრამატიკული და ფონეტიკური ვარიანტები დამოუკიდებელი სიტყვების ფუნქციითაა წარმოდგენილი ამ ლექსიკონში, რაც მათ ავტომატურ ძიებას აადვილებს. თითოეულ სიტყვას ერთვის ინფორმაცია მისი გრამატიკული კატეგორიის შესახებ; ლექსიკური მნიშვნელობები აიხსნება იმ შემთხვევაში, როდესაც ჰომოგრაფებიც განსხვავდება მახვილით. ლექსიკონს თან ერთვის ახლად შემოსული საერთაშორისო სიტყვები და ხშირად გამოყენებული ტერმინები, რომლებიც ადრინდელ თანამედროვე ლიტვური ენის ლექსიკონში არ შედიოდა. ლექსიკონში არ შედის ვერნაკულარიზმები, მოძველებული, არასტანდარტული სიტყვები და იშვიათი მცენარეებისა და ცხოველების აღმნიშვნელი ლექსიკა. **საკუთარი სახელების ლექსიკონი** შედგენილია სხვადასხვა წყაროზე დაყრდნობით. იგი შეიცავს ქვეყნებისა და მათი დედაქალაქების, ოლქების, დიდი პროვინციების აღმნიშვნელ სიტყვებს, გავრცელებულ ლიტვურ სახელებსა და გვარებს, ლიტვური ქალაქების, დიდი ტბების, მდინარეების დასახელებებს; ლიტვის ისტორიისა და კულტურისათვის მნიშვნელოვან მიკროტოპონიმებს; კულტურის, ხელოვნების, მეცნიერებისა და პოლიტიკის სფეროებისათვის დამახასიათებელ, ლიტვისა და მსოფლიოსათვის მნიშვნელოვან სახელებს. ყველა ერთეული მოწონებულია ლიტვური ენის სახელმწიფო კომისიის მიერ. ლექსიკონი შეიცავს 16 000 საკუთარი სახელის დასახელებას.

საკუთარი და საზოგადო სახელების აღმნიშვნელი სიტყვების ლექსიკონები თავდაპირველად შეიქმნა სწავლისა და სწავლების პროცესში გამოსაყენებლად, თუმცა მათი გამოყენება სხვა მიზნებითაც შეიძლება. ისინი გამოდის MS Excel ფორმატითა და პირველი ლიტვური შრიფტით **Palemonas**, რაც მათ გამოყენებასა და ახალი სიტყვების დამატებას აადვილებს. საჭიროებისამებრ, შეიძლება კერძო კომენტარების დამატებაც.

2. ლიტვური დიალექტების კორპუსები

ლიტვური ენის სახელმწიფო კომისია ახორციელებს გრძელვადიან პროგრამას (2001 წლიდან) სახელწოდებით: **დიალექტებისა და ეთნიკური ტოპონიმების შენარჩუნება**, რომელშიც აქტიურად მინაწილეობენ ლიტვური ენის ინსტიტუტი და უმაღლესი განათლების დაწესებულებები. პროგრამის მიზანია ცალკეული ლიტვური დიალექტების კორპუსების შედგენა (ოფიციალურად ეს არის ორი დიალექტი, რომლებიც მოიცავს მთელ რიგ ქცევებსა და თქმებს). ამ პროექტის შედეგად შეიქმნება ორი ნაშრომი: **წიგნი**, რომელიც აღ-

წერს დიალექტის თავისებურებებს და მოიცავს სპეციალური ტრანსკრიფციის გამოყენებით ჩაწერილ დიალექტურ ტექსტებს, კომენტარებს ტექსტის განსაკუთრებული ადგილების შესახებ (ენობრივ არქაიზმებს, ფონეტიკურ, პროსოდიულ, ლექსიკურ და გრამატიკულ თავისებურებებს). მას ახლავს **კომპაქტური დისკი** (CD) (ელექტრონული კორპუსის ვარიანტით), რომლის საშუალებითაც მომხმარებელს შეუძლია მოუსმინოს წიგნში დაბეჭდილ ამა თუ იმ დიალექტზე შექმნილი ტექსტების აუდიოჩანაწერებს. უკვე გამოქვეყნდა ლიტვური ენის - დიალექტების ამსახველი 20-ზე მეტი კრებული.

განხილული ელექტრონული საშუალებები (ლექსიკონები, დიალექტური კორპუსები) შეიქმნა ლიტვური ენის სახელმწიფო კომისიის მიერ განხორციელებული სხვადასხვა პროგრამის ფარგლებში. ამ ბოლო დროს, უმაღლესი საგანმანათლებლო დაწესებულებების მკვლევრები აქტიურად მონაწილეობენ ახალ პროექტში სახელწოდებით **სალიტერატურო ლიტვურ ენაში, მის დიალექტებსა და სახესხვაობებში მიმდინარე ცვლილებების კვლევის პროგრამა (2011-2020)**. ამ სტრიქონების ავტორმა, ხუთი უნივერსიტეტის ლინგვისტებთან ერთად მოამზადა პროექტი: **მახვილი ახალგაზრდების მეტყველებაში: სტანდარტული ენის ნორმები და სალაპარაკო ენაში მახვილის ხმარების ტენდენციები**, რომელიც განხორციელდება 2011-2013 წლებში და რომელიც დააფინანსა ლიტვური ენის სახელმწიფო კომისიამ.

The most Up-to-date Lithuanian Electronic Dictionaries and Lithuanian Dialect Corpora

Vidas Kavaliauskas
Vilnius Pedagogical University (Lithuania)
vk1119@gmail.com

The progress of information technology stimulates the development of the science of linguistics and opens up a number of effective possibilities for the practical application of linguistic achievements. Electronic dictionaries and corpora of the language of different purposes compiled on scientific grounds have recently seen their growing popularity in Lithuania. The present report will cover 3 universal dictionaries of standard language, which may be used by phonologists, morphologists, lexicologists, specialists of syntax and other fields, students and the entire speaking society for various scientific and practical purposes, as well as electronic bases of dialectal texts, certain corpora, published in CD format.

Dictionaries.

1.1. **"Common Press Words of the 20th Century: Electronic Frequency Dictionary (CD)"** (project manager – Prof. A. Pakerys; published in 2004). The publication constitutes the preface to the dictionary and compact disc (CD), which contains the dictionary in *MS Access* and *MS Word* format. It is based on the passages from Lithuanian books and

periodicals published in Lithuania in the 20th century (the texts were selected from each decade by means of a certain methodology: 1925, 1935, 1945, 1955, 1965, 1975, 1985 and1995). This dictionary enables the selection of any form of the electronic dictionary, which either includes headwords (by alphabet, backward alphabet, general average frequency), or individual grammatical forms of words (by alphabet, backward alphabet, general average frequency). With this dictionary, you may look for a specific headword, its form or a certain part (for example, words with a certain suffix or prefix only or certain forms – e.g. accusative singular, etc). The dictionary does not only display the required word or its form but shows the entire passage from the text, where it was used. The dictionary contains around 30,000 independent lexical units, nearly 100,000 of their different forms and approximately 550,000 cases of usage of those forms. The obvious practical value of the publication: the user of the electronic dictionary may put the words of interest and their forms into a "basket", create new data files, edit them and use them for different purposes. Word frequency recorded in the dictionary is an excellent source of information for the codifiers of accentuation and other norms in respect of standardisation of potential variants in the standard language. The dictionary of standard words is designed for scientists, teachers and students who are interested in the usage of words and their forms, valency, frequency, development of lexis and grammar, etc.

1.2. **"Proper Press Words of the 20th century): Electronic Frequency Dictionary (CD)"** (project manager – Prof. A. Pakerys; published in 2005). This electronic dictionary is a sequel to the publication *Common Press Words of the 20th Century: Electronic Frequency Dictionary* released in 2004. The dictionary of proper press words of the 20th century is devoted to scientists, teachers and students interested in the usage, frequency and valency of personal names, place names and other proper words (theonyms, zoonyms, cosmonyms, etc) peculiarities of spelling and their development.

1.3. The book compiled by D. Mikulėnienė, A. Pakerys and B. Stundžia **Accentuation Dictionary of Standard Lithuanian** (2007, 2008) contains 2 electronic dictionaries (CD). ***Dictionary of Common Words*** includes around 45,000 words from the *Dictionary of Modern Lithuanian Language* (this dictionary was also published in electronic format). Grammatical, phonetic variants of words function as independent units in the dictionary to enable their easier finding by means of automatic search. Each word is supplemented by the information on the part of speech; lexical meanings are explained in such cases where homographs differ from each other by their accentuation. The dictionary is supplemented by new international words and terms of wide usage, which have not been previously included in the *Dictionary of Modern Lithuanian Language.* The dictionary does not contain vernacularisms, obsolete, non-standard words, as well as names of rare species of plants and animals, etc. ***Dictionary of Proper Words*** is from different sources. The dictionary contains the names of world countries and their capitals, counties, larger provinces, more common Lithuanian names and surnames, names of Lithuanian cities, towns, larger lakes, rivers, as well as names of places of smaller scope, which are important to the history and culture of Lithuania, surnames of most well-known figures of

culture, art, science and politics of the world and Lithuania, which are all approved by the State Commission of the Lithuanian Language. The dictionary includes a total of approximately 16,000 proper words. **The dictionaries of common and proper words** were primarily compiled for the purpose of accentuation learning and teaching but they can also be used for other purposes. They are given in *MS Excel* format (original Lithuanian font *Palemonas*), which enables the users to use them easily, to supplement with relevant words, if required, to add personal comments, etc.

2. Lithuanian Dialect Corpora.

The State Commission of the Lithuanian Language implements a long-term programme (since 2001) entitled *Preservation of Dialects and Ethnic Place Names.* Lithuanian linguists from the Institute of the Lithuanian Language and higher education institutions play an active role in the programme. The goal of the programme is to compile the corpora of individual Lithuanian dialects (officially, there are 2 dialects, which are further subdivided into a number of subdialects and speeches), subdialects and speeches. Usually, 2 publications come to life: **a book,** which describes the linguistic peculiarities of a subdialect and publishes dialectal texts written by means of a special transcription, including the comments on certain places in texts (old linguistic forms, peculiarities of phonetics, accentuation, lexis, grammar), and **compact disc (CD)** (electronic corpus variant), where the user may listen to the audio recording of dialectal texts published in the book. The total of more than 20 individual sets representing individual subdialects of the Lithuanian language has already been published.

The discussed electronic means (dictionaries, dialect corpora) were published in the framework of various programmes implemented by the State Commission of the Lithuanian Language. Lately, the scientists from higher education institutions have actively participated in the new *Programme of Functioning and Research of Changes of Standard Lithuanian, Subdialects and Other Varieties of Language for 2011–2020.* The author of this report, along with the linguists from other 5 universities, has prepared the project **Youth Accentuation Trends: Norms of Standard Language and Accentuation Tendencies of Spoken Language**, which will be implemented in 2011–2013 and which is funded by the State Commission of the Lithuanian Language.

Keywords: Lithuanian electronic dictionaries, electronic dialect corpora (CD), norm of the standard language, State Commission of the Lithuanian Language, linguistic projects

საქართველოს პარლამენტის ეროვნული ბიბლიოთეკის ლექსიკონთა ერთიანი ბაზა: პრობლემები და პერსპექტივები

გიორგი კილაძე
საქართველოს პარლამენტის ეროვნული ბიბლიოთეკა (საქართველო)
gkiladze@nplg.gov.ge

ლექსიკონი არა მხოლოდ პირადი ან ბიბლიოთეკის მკითხველთა (მათ შორის ინტერნეტ) საჭიროებისთვის აუცილებელი წიგნი ან ციფრული რესურსია, არამედ ეროვნული კულტურის შემადგენელი ნაწილიც. ლექსიკონის განმარტებათაგან ერთ-ერთს მოვიშველიებთ. მისი ავტორია ფრანგი მწერალი ანატოლ ფრანსი: „ლექსიკონი – ეს ანბანურად დალაგებული მთელი სამყაროა“.

მოხსენებაში წარმოვადგენთ ჩვენს მოსაზრებას ამ სამყაროს მხოლოდ ერთ – „ელექტრონულ ნაწილზე“. ფაქტია, რომ დღეს ელექტრონული ლექსიკონებისადმი მომხმარებელთა მოთხოვნა გაცილებით მეტია, ვიდრე „ჩვეულებრივ“ – ე. წ. „ქაღალდის ლექსიკონებზე“. განსხვავებულია მომხმარებლის ფსიქოლოგიური განწყობა ლექსიკოგრაფიული პროდუქციის ამ ორი ტიპისადმი. ელექტრონული ლექსიკონების მომხმარებელი ვერ ეგუება სიტყვათა რაოდენობის შეზღუდვას ლექსიკონში, მას აქვს ლექსიკონის მუდმივი განახლებისა და შეცდომების მყისიერად გასწორების მოლოდინი, ხშირად კი – ლექსიკონის ავტორებთან ინტერაქტიური კავშირის საშუალებით საკუთარი მოთხოვნებისა და შენიშვნების გაზიარების შესაძლებლობაც.

საქართველოს პარლამენტის ეროვნულ ბიბლიოთეკაში ლექსიკონების შექმნის ტრადიცია დიდი ხანია არსებობს. დღესდღეობით მის ოფიციალურ ვებგვერდზე 13 ლექსიკონია განთავსებული, მათი ნახვა შესაძლებელია ნებისმიერი ინტერნეტმომხმარებლისთვის და ამიტომ მათი ჩამოთვლისგან თავს შევიკავებთ.

საქართველოს პარლამენტის ეროვნული ბიბლიოთეკის ლექსიკონთა ერთიანი ბაზა აკმაყოფილებს შემდეგ მოთხოვნებს:

1. ქართული უნიკოდის სრული მხარდაჭერა
2. სიტყვის, განმარტების ძიება შესაძლებელია როგორც ერთ, ისე ყველა ლექსიკონში
3. სრულადაა გაქართულებული მართვის დაფა
4. ხელმისაწვდომია ნებისმიერი დაინტერესებული პირისა თუ ორგანიზაციისთვის
5. ლექსიკონთა განლაგება შესაძლებელია თემატურად.
6. დიზაინი შეესაბამება ეროვნული ბიბლიოთეკის ვებგვერდის დიზაინს (brand nplg).
7. შეუზღუდავია ლექსიკონების დამატება

ასეთია არსებული რეალობა, თუმცა ვთვლით, რომ ეს მხოლოდ კარგი დასაწყისია. ელექტრონული ლექსიკონების შექმნა, თუნდაც მისი მხოლოდ გაციფრება (ელექტრონული ვერსიების თუ ელექტრონული ასლების მომზადება) სულ მცირე სამ პრობლემასთანაა დაკავშირებული:

1. არასათანადო ტექნიკური ბაზა
2. შეზღუდული რესურსები (ადამიანური, ფინანსური)
3. ლექსიკონების სისრულე

ლექსიკონების ერთიანი ბაზის ჩამოყალიბება რთული და ხანგრძლივი პროცესია.

ჩამოთვლილი სამი პრობლემა ურთიერთკავშირშია და, სამწუხაროდ, შეუძლებელია მათი „ერთმანეთის მიყოლებით“ გადაწყვეტა. უდავოა, რომ საკითხის მეტ-ნაკლებად გადაჭრა კომპლექსურ მიდგომას მოითხოვს.

იმისათვის, რომ დაკმაყოფილდეს თანამედროვე მომხმარებლის მოთხოვნა, საჭიროა, მას ამა თუ იმ სიტყვის განმარტება მიეწოდოს სწრაფად და კომფორტულად.

არის ერთი „ვერაგი“ კანონზომიერება: „ლექსიკონში მხოლოდ იმ სიტყვას ვერ ვპოულობთ, რომელიც გვჭირდება“. ასე რომ, მომხმარებლისთვის ზემოთ ჩამოთვლილი პრობლემები ერთ ძირითადზე დაიყვანება – ლექსიკონების სისრულეზე. ამასთან, მას უკვე აღარ აკმაყოფილებს ამა თუ იმ სიტყვის მხოლოდ ერთი, „მონოგანმარტება“, შეიძლება ითქვას, იგი ლექსიკონების შემქმნელებისგან პოლიგანმარტებას ითხოვს.

როგორ შეიძლება „ლექსიკონის სისრულის“ უზრუნველყოფა? როგორც წესი, ამის მისაღწევად სამი გზა არსებობს:

1. ადამიანური რესურსის მობილიზება
2. მონაცემთა იმპორტი
3. მომხმარებლისთვის ლექსიკონთა შევსების უფლების მინიჭება

ლექსიკონების შევსების კიდევ ერთი გზაა ე.წ. „ონლაინშევსება“, რომელიც მხოლოდ ერთი წელია ეროვნულ ბიბლიოთეკაში დამკვიდრდა: ვერ პოულობთ სიტყვას? მოგვწერეთ და განმარტება თქვენს ელფოსტაში აღმოჩნდება. ეს გაცილებით სწრაფი გზაა, ვიდრე „ქაღალდის ლექსიკონებში“ ძიება.

საქართველოს პარლამენტის ეროვნული ბიბლიოთეკაში ლექსიკოგრაფიის განყოფილება არ არსებობს. არსებული ლექსიკონების გაციფრებასა და ახლის შექმნასაც **(იხ. საქართველოს ისტორიულ ძეგლთა ბიბლიოგრაფიული ლექსიკონი, CIVI ენციკლოპედიური ლექსიკონი)** ბიბლიოთეკის სხვადასხვა განყოფილების თანამშრომლები ახორციელებენ.

უახლოეს მომავალში შესაძლებელი გახდება მრავალენოვანი ძიება (რუსულ, ინგლისურ და სხვა ენებზე) სხვადასხვა ტიპის ლექსიკონებში **(მაგ. English-Georgian, ფსევდონიმების ლექსიკონი, უცხო სიტყვათა ლექსიკონი** Civil **ენციკლოპედიური ლექსიკონი, სამოქალაქო განათლების ლექსიკონი, რუსულ-ქართული ლექსიკონი, საბიბლიოთეკო ტერმინების განმარტებითი ლექსიკონი** და მრავალი სხვა).

The Integrated Database of the Dictionaries of the National Parliamentary Library of Georgia: Problems and Perspectives

Giorgi Kiladze
The National Parliamentary Library of Georgia (Georgia)
gkiladze@nplg.gov.ge

A dictionary is not only a book necessary for personal use but also an integral part of national culture. Anatole France defines the dictionary as *the whole universe arranged in alphabetical order. The* Topic of this speech will be only one "digital part" of this universe. Demand for electronic dictionaries is now higher than for their paper counterparts. The latter cannot be revised and expanded until the next edition is prepared to be published, whereas the former are continuously and easily updated and edited in response to feedback.

Now, going back to the main topic, the National Parliamentary Library of Georgia has a long-standing tradition of building dictionaries. Currently its official website incorporates 13 dictionaries, the listing of which is not necessary inasmuch as they are easily accessible.

The integrated database of the dictionaries of the National Parliamentary Library of Georgia meets the following demands:

1. complete support of Georgian Unicode;
2. word search and definition search can be done in each of them separately as well as in all of them simultaneously;
3. management panel is available in Georgian;
4. availability for persons and organisations;
5. dictionaries can be arranged thematically;
6. design (theme) corresponds to the National Library website design;
7. dictionaries can be added in unlimited numbers.

The building of web dictionaries, or even digitalisation of paper ones, involves at least three kinds of problem:

1. an inadequate technical basis;
2. limited human and financial resources;
3. incompleteness of the dictionaries.

The formation of an integrated database of dictionaries is a long and complex process. The above listed three problems are interrelated, therefore impossible to be solved one after another. Undoubtedly, complex treatment is necessary for the problems to be solved.

Word definitions should be provided to modern readers quickly and comfortably in order to satisfy their demand. “Unfair” law works in the realm of dictionaries: “we can find in the dictionaries anything but the word definitions that we really need”. Thus all of the above-mentioned problems can be brought down to one: incompleteness of dictionary data. Readers are no longer satisfied with one definition. They want several definitions of the words with multiple meanings. How to ensure completeness of dictionary? There are three different ways for that:

1. to involve as many employees as possible in the building/digitalisation of dictionaries despite the limited human and financial resources;
2. import of data;
3. to enable the readers to participate in the building process.

There is one more method for the completion of dictionaries, which was adopted by the National Library a year ago, provisionally called *online completion* — if you are not able to find the definition you are looking for, email us the word and it will soon appear in your mailbox. This way is much faster than searching a paper dictionary.

Lastly, a department of lexicography does not exist at the National Parliamentary Library of Georgia. Dictionaries (Bibliographical Dictionary of the Historical Monuments of Georgia, Civil Encyclopedic Dictionary) are being built/digitalised by the employees of different departments. A multilingual search will be possible in the near future in general and specialized dictionaries, such as an English-Georgian Dictionary, Dictionary of Pseudonyms, Dictionary of Foreign Words, Civil Encyclopedic Dictionary, Dictionary of Civil Education, Russian-Georgian Dictionary, Dictionary of Library Terms, will be available.

უმეშველზმნო წარმოების ზმნათა ფონემატური აგებულების მიმართება ზმნის სინტაქსურ და მორფოლოგიურ ყალიბებთან

მანანა კობაიძე
მალმოს უნივერსიტეტი (შვედეთი)
manana.kock.kobaidze@mah.se

მეშველზმნიანი და უმეშველზმნო წარმოების ზმნათა ერთმანეთისგან განცალკევება საშუალებას გვაძლევს, რომ უმეშველზმნო წარმოების (პირდაპირი წყობის) ზმნებისთვის ჩამოვაყალიბოთ რამდენიმე მარტივი წესი. ამ წესებიდან ზოგი იმდენად მარტივია, რომ შესაძლებელია მათი ჩამოყალიბება მხოლოდ ზმნის ფუძის ფონემატურ აგებულებაზე დაყრდნობით.

1. თუ აწმყოს პირველი პირის ვინის რიგის ფორმაში ზმნა მთავრდება -VC, -CCი, ან Vი, მიმდევრობით (სადაც V ხმოვანს აღნიშნავს და C თანხმოვანს, ხოლო -CC ერთზე მეტ თანხმოვანს), ამ ზმნის სუბიექტი ბრუნვაცვალებადი სუბიექტია. მაგალითად: ***ვხატავ, ვწერ, ვჭამ, ვათბობ, ვდრეკ, ვჩეჩ, ვცხოვრობ, ვთამაშობ, ვფხან, ვჩქმეტ, ვთლი, ვფხაჭნი, ვტყუი, ვღმუი*** და ა.შ.

ეს წესი მოქმედებს მიუხედავად იმისა, თუ რას წარმოადგენს ზმნის ფუძის ბოლო VC მარცვალი: თემის ნიშანს, ფუძედრეკად ელემენტს, თუ უთემისნიშნო ზმნის ძირის ნაწილს.

2. თუ აწმყოს პირველი პირის ვინის რიგის ფორმაში ზმნა მთავრდება VCი მიმდევრობით, სადაც VC არის ნებისმიერი მარცვალი გარდა **-ებ, -ობ,** ან **-ევ** მარცვლებისა, ამ ზმნის სუბიექტიც ბრუნვაცვალებადი სუბიექტია. მაგალითად: ***ვმღერი, ვყვირი, ვკვნესი, ვიცინი, ვტირი, ვღნავი***, და ა.შ.

3. თუ აწმყოს პირველი პირის ვინის რიგის ფორმაში ზმნა მთავრდება **ებ-ი, -ობ-ი, ან -ევ-ი** მიმდევრობით, მაშინ ამ ზმნის სუბიექტი ბრუნვაუცვლელი სუბიექტია: მაგალითად, ***ვთბები, ვიხრჩობი, ვერევი, ვეცოდები, ვპირდები...***

გარდა ამ სამი წესისა, შეიძლება ჩამოყალიბდეს მარტივი წესები იმის შესახებაც, თუ რომელი სავრცობით იწარმოებს ზმნა უწყვეტლის მწკრივებს, აგრეთვე იმის შესახებ, თუ რა პრეფიქსით იწარმოებს ზმნა მყოფადის წრეს (და, შესაბამისად, II სერიის ფორმებს): ზმნისწინით (ტელიკური ზმნები) თუ ხმოვანი პრეფიქსით (ატელიკური ზმნები).

წესი სავრცობის შესახებ:

1. თუ ზმნა აწმყოს ვინის რიგის პირველი პირის ფორმაში მთავრდება -VC ან -CCი მიმდევრობით, უწყვეტელში მისი სავრცობია -დ (***ვხატავ-დ-ი, ვცხოვრობ-დ-ი, ვთლი-დ-ი, ვფხაჭნი-დ-ი***). გამონაკლისებია -CCი მიმდევრობით დამთავრებული 5 ზმნა: ***ვხტი, ვთრთი, ვკრთი, ვძრწი, ვქრი,*** და ეს ზმნებიც აწმყოს წრის გარდა სხვა მწკრივებში მარცვლოვან ფუძეს იჩენენ. ასევე, გამონაკლისთა რიცხვს მიეკუთვნება არქაული ფორმით შემორჩენილი ***ვიჭირვი, ვიბრძვი, ვილტვი...*** ზმნები.

გასათვალისწინებელია აგრეთვე ომონიმური წყვილი *ისვრის*: ***ისვრის – ისროდა ისარს; ისვრის – ისვრიდა (შარვალს)***

2. თუ ზმნა მთავრდება VC-ი ან V-ი მიმდევრობით, მისი სავრცობია ***-ოდ-ი*** (***ვმღერ-ოდ-ი, ვთბებ-ოდ-ი, ვტყუ-ოდ-ი***). გამონაკლისია 3 ზმნა: ***ვყიდი, ვტენი, ვწონი.***

ასევე მარტივია წესი **მყოფადის წარმოების შესახებ ყველა იმ ზმნისთვის, რომელიც -ი ხმოვნით ბოლოვდება:**

1. **-CCი** დაბოლოების მქონე და **-ებ-ი, -ობ-ი,** ან **-ევ-ი** დაბოლოების მქონე ზმნები მყოფადის წრეს მარტივად – ზმნისწინის დართვით იწარმოებენ (***ვფხაჭნი – დავფხაჭნი, ვთლი – გავთლი, ვთბები – გავთბები***). გამონაკლისები აქაც იგივე ზმნებია, რაც ზემოთ: *ვხტი, ვთრთი, ვკრთი, ვძრწი, ვქრი, ვიჭირვი, ვიბრძვი, ვილტვი.*

2. ზმნები, რომლებიც მთავრდება V-ი ან VC-ი მიმდევრობით, სადაც VC არის ნებისმიერი მარცვალი გარდა -ებ, -ობ, ან -ევ მარცვლებისა, მყოფადის წრეს ხმოვანი პრეფიქსით იწარმოებენ (***ვტყუი – მოვიტყუებ, ვღმუი – ვიღმუვლებ, ვტირი – ვიტირებ)***.

მხოლოდ VC მარცვლით დაბოლოებული ზმნებისთვის საჭიროა რამდენიმე დამატებითი წესის ჩამოყალიბება, რათა შესაძლებელი გახდეს იმის გარჩევა, თუ რომელი ზმნაა მათგან ტელიკური (მყოფადში ზმნისწინიანი) და რომელი ატელიკური (მყოფადში ხმოვანპრეფიქსიანი). ეს წესებია:

1. აწმყოში **უმარცვლო ძირის მქონე** (მათ შორის, შეკუმშული უმარცვლო ძირის მქონე) **ზმნები მყოფადის წრეს, ჩვეულებრივ, ზმნისწინით იწარმოებენ**: *ვკლავ – მოვკლავ, ვათბობ – გავათბობ* (გამონაკლისებია: *ცნობს* და *გრძნობს* ზმნები).

2. **მარცვლოვანი ძირისგან ნაწარმოები ტელიკური და ატელიკური ზმნები ერთმანეთისგან ან აფიქსაციით განსხვავდებიან, ანდა ძირის სტრუქტურით**:

2ა. **-ავ-თემისნიშნიანი ზმნების გარდა** ყველა დანარჩენი ზმნა, რომელიც **მხოლოდ VC მარცვლით მთავრდება,** მყოფადის წრეს ზმნისწინით იწარმოებს, თუკი ამ ზმნას ნეიტრალური ქცევის ფორმა ხმოვანი პრეფიქსით ეწარმოება: აშენებს...

2ბ. **-ავ-თემისნიშნიანი ზმნების გარდა** ყველა დანარჩენი მარცვლოვანი ძირის მქონე **თემისნიშნიანი ზმნა** მყოფადის წრეს ხმოვანი პრეფიქსით იწარმოებს, თუკი ამ ზმნას ნეიტრალური ქცევის ფორმა ხმოვანი პრეფიქსის გარეშე ეწარმოება: ***კანკალებს, ჰპატრონობს, ანგარიშობს***..., გამონაკლისებია: ***სტაცებს, ჰბადებს, ჰკადრებს***... სულ 13 ზმნა.

2გ. **-ავ თემისნიშნიანი მარცვლოვანი ძირის ზმნებიდან -ა-ხმოვნის შემცველი ძირები** (გარდა ***ქანავს*** ზმნისა) მყოფადის წრეს ზმნისწინით იწარმოებენ (***ხატავს, ბარავს, ბლანდავს***...).

2დ. **-ავ თემისნიშნიანი მარცვლოვანი ძირის ზმნებიდან -ე-ხმოვნის შემცველი ძირები** მყოფადის წრეს ზმნისწინით იწარმოებენ (გამონაკლისებია: ***ნებავს, ღელავს, ფეთქავს, ღვენთავს, წვეთავს, ცეკვავს***).

3. **ფუძედრეკადი ძირის მქონე ზმნები** მყოფადს, ჩვეულებრივ, ზმნისწინით იწარმოებენ. გამონაკლისებია: ***ვფრენ, ვქშენ, ვფშვენ, ვღრენ, ვჩქმეტ*** (თუმცა ზოგ მათგანს ზმნისწინიანი მყოფადის ვარიანტიც მოეპოვება: ***დავჩქმეტ***. სხვებისთვის ზმნისწინის დართვა დროს არ ცვლის: ***დავფრენ, შევღრენ***...) დამახასიათებელია, რომ ამ ზმნათაგან სწორედ *ვჩქმეტ* ზმნას შეუძლია –ავ სუფიქსი დაირთოს ე-ხმოვნის ცვლილების გარეშე: ***ვჩქმეტავ,*** რაც ამ ზმნის ტელიკურობის კიდევ ერთი გამოვლენაა).

4. **თემისნიშნის დართვის უნარის უქონელი არაფუძედრეკადია** სულ 18 ზმნა. მათგან მყოფადს ზმისწინით იწარმოებს 13 ზმნა (***წერს, ჰფხანს, ქსოვს*** და მისთ.) და ხმოვანი პრეფიქსით 5 ზმნა (***ჟღერს, ჩქეფს, ყეფს, ძგერს, შხეფს***).

მეორე სერიის მორფოლოგიაც ნათლად აჩვენებს ფონემატური სტრუქტურის მნიშვნელობას უღლების პარადიგმებისათვის. მოხსენებაში მეორე სერიის წარმოების საკითხებიც დაწვრილებით არის წარმოდგენილი.

About Relations between Phonemic and Morphological and Syntactic Patterns of Nonauxiliary formed verbs

Manana Kobaidze
Malmoe University (Sweden)
manana.kock.kobaidze@mah.se

Two large groups of verbs are identifiable in Georgian: verbs forming the present tense by means of auxiliary verbs and verbs forming the present tense without auxiliary verbs. Separation of auxiliary and non auxiliary formed verbs makes it possible to formulate some simple **rules regarding non auxiliary formed verbs**. Some of these rules may be defined merely based on the phonematic structure of verb stems.

Subject marking rules

1. If a verb marked by the *v*-set markers ends in the sequence –VC, -CCi or Vi in the first or second person in the present tense, the subject of the verb is a case alternating subject, e.g.: *vxat'av* "I paint", *vc'er* "I write", *vdrek'* "I bend", *vcxovrob* "I live", *vtli* "I peel", *vt'q'ui* "I lie", etc. (The symbol V indicates a vowel, the symbol C means a consonant, and CC is a sequence of two or more consonants).

This rule holds irrespective whether the last segment VC is a thematic marker, an ablauting syllable or a part of a root.

2. If a verb marked by the *v*-set markers ends in the sequence –VC-i in the first and second persons in the present tense, and the sequence VC is any syllable except for *-eb*, *-ob* and *-ev* syllables, the subject of the verb is a case alternating subject, e.g.: *vmgheri* "I sing", *vq'viri* "I shout", *vicini* "I laugh"….

3. If a verb marked by the *v*-set markers ends in the sequence –VC-i in the first and second persons in the present tense, and the sequence VC is either *-eb*, *-ob* or *-ev* syllables, the subject of the verb is not a case alternating subject, e.g.: *vtbebi* "I warm myself", *vp̣irdebi* "I promise him something", *vecodebi* "he feels sorry for me"…

Augment distribution rules are also simple:

1. If a verb marked by the *v*-set markers ends in the sequence -VC or -CC-*i* in the first or second person in the present tense, its augment is *-d* (*vxaṭav-d-i* "I was painting", *vcxovrob-d-i* "I lived", *vtli-d-i* "I was peeling"…). There are exceptions, 5 verbs ending in CC-*i* sequence take the augment *-od*: *vxṭi* "I am jumping", *vtrti* "I am trembling", *vḳrti*, *vdzrc̣i*, *vkri;* some verbs preserved in archaic form are also exception from this rule as they end in CC-*i* sequence, but take the augment *-od-*, e.g., *vič̣irvi* "I am in trouble", *vibrdzvi* "I struggle", *vilṭvi* "I desire, I try to reach", *visc̣rapvi* "I am in a hurry for something", *visvri* "I am shooting".

2. 1. If a verb marked by the *v*-set markers ends in the sequence -V-*i* or -VC-*i* in the first and second persons in the present tense, it takes the augment *-od* (*vmgherodi* "I was singing", *vtbebodi* "I was warming myself" *vtqʼuodi* "I was lying"…). There are three exceptions: *vqʼididi* "I was selling it", *vṭenidi* "I was pressing it into something, *vc̣onidi* "I was weighing it".

The future formation rules:

1. Verbs ending in the segment -CC-*i* or in *-eb-i*, *-ob-i* or *-ev-i* morphemes form the future tense by adding preverbs (*vtli* "I peel it" – *ga-vtli* "I'll peel it", *vtbebi* "I'm warming myself" – *gavtbebi* "I'll warm myself",). The exceptions are the above mentioned verbs: *vxṭi* "I am jumping", *vtrti* "I am trembling", *vḳrti, vdzrc̣i* and archaic *vic̆arvi* "I am in troubles", *vibrdzvi* "I struggle", *vilṭvi* "I desire, I try to reach", *visc̣rapvi* "I am in a hurry for something", *visvri* "I am shooting".

2. Verbs ending in the sequence V-*i* or VC-*i* where the segment VC is any syllable except for the *-eb*, *-ob* or *-ev* form the future tense by adding vowel prefixes (*vṭqʼui* "I lie" – *moviṭqʼueb* "I'll lie", *vṭiri* "I cry" – *viṭireb* "I'll cry"...)

Regarding **verbs ending in the syllable VC**, it is necessary to form several rules in order to define if they are telic or atelic and, consequently, if they form the future tense by adding pre-verbs or vowel prefixes:

1. If a verb stem ends in a thematic marker VC and the verb root is non syllabic, the verb takes a pre-verb in the future tense.

vḳlav "I kill"– *movḳlav* "I'll kill", *vatbob* "I warm it" – *gavatbob* "I'll warm it" (exceptions: *vcnob* "I recognize" and *vgrdznob* "I feel").

2. If a verb stem ends in a thematic marker VC and the verb root is syllabic, the verb is either telic (and takes a preverb in the future tense) or atelic (and takes a vowel prefix in the future tense). Telic and atelic verbs with syllabic roots differ from one another either in their morphological structure of the stem or in their phonematic structure of the root:

2a. All verbs ending in the syllable VC except the *-av* thematic verbs form the future tense by adding preverbs if these verbs have vowel prefixes in the neutral version: aSenebs...

2b. All verbs with syllabic roots marked by any VC thematic markers except for the *-av* thematic marker are atelic and form the future tense with a vowel prefix if these verbs do not have a vowel prefix in a neutral version: *ḳanḳalebs* "He is trembling", *hp̣aṭronobs* "He takes care of him", *angarishobs* "He is counting",..., with the exception of one group consisting of 13 verbs (*sṭacebs* "He grabs it", *hbadebs* "It gives birth to smth."…)

2c. Verbs that have monosyllabic roots and take a thematic marker *-av* are usually telic and form the future tense by means of preverbs if the verb root contains a vowel *-a xaṭavs* "He is painting", *baravs* "He is spading" ... The only exception is the verb *kanavs* "He is swinging"... It contains the vowel -a, however, it is an atelic verb.

2d. Monosyllabic verb roots taking a thematic marker *-av* are usually telic and form the future tense by means of preverbs if the verb root contains a vowel *-e*: *tesavs* "He is sowing it", *xexavs* "He is scratching it"... The exceptions are: *nebavs* "He would like", *ghelavs* "He is excited",

petkavs “It is beating”, *ghventavs* “It is dribbling”, *c’vetavs* “It is dribbling”, *cekvavs* “He is dancing”.

3. Ablauting verbs form the future tense by adding preverbs. There are few exceptions: *vpren* “I fly”, *vpšven* “I am breathing with noise”, *včkmet’* “I nip” (although some of these exceptions can also have a parallel form of the future tense formed with a preverb, e.g., *davčkmeṭ* “I’ll nip”. It is noteworthy that only this verb among these five exceptions can take the thematic marker *-av* without changing the vocalisation *-e-* into vocalisation *-i-* (*vškmeṭav* “I’m nipping”), which manifests that this verb shows the features of telic verbs in this respect too.

4. Apart from thematic verbs (where I also include ablauting verbs, as well as non thematic verbs having parallel thematic forms) there are 18 verbs which are neither ablauting nor able to take a thematic marker. 13 verbs out of these 18 verbs form the future tense by adding preverbs (*ɕers* “He is writing”, *ksovs* “He is knitting”, etc.) and 5 verbs by adding vowel prefixes (*ǯɣers* “It’s sounding”, *ɢeps* “It’s barking”…).

The impact of phonematic structure of verb stems on the conjugation paradigm is exposed in the second series too. This issue is also touched upon in the paper.

იდიომური გამოთქმების ქართულ-ახალბერძნული ლექსიკონი

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

რეზიუმე

წინამდებარე სტატიაში წარმოდგენილია ენის შემსწავლელებისა და პროფესიონალური მთარგმნელებისათვის გათვალისწინებული ორენოვანი ლექსიკონი - *იდიომური გამოთქმების ქართულ-ახალბერძნული ლექსიკონი*. აღნიშნული ლექსიკონი შექმნილია "იდიომური გამოთქმების ქართულ-ახალბერძნული ლექსიკონის" პროექტის ფარგლებში. პროექტის შედეგთან ერთად განხილულია ძირითადი მეთოდოლოგიური პრინციპები, რომლებსაც ეფუძნება ლექსიკონის სტრუქტურა და მისი ძირითადი მახასიათებლები. კერძოდ, გაანალიზებულია ლექსიკონის მაკრო- (სიტყვათა ჯგუფების ლემებისა და მათი კომპონენტების წარმოქმნის მექანიზმები) და მიკრო-სტრუქტურები. ლექსიკონი მოიცავს იდიომურ გამოთქმათა 11800-მდე ერთეულს.

შესავალი

ნებისმიერი ლექსიკონი წარმოადგენს მონაცემთა ბაზას, რომლის ძირითადი დანიშნულებაა ინფორმაციის შენახვა კონკრეტული ენობრივი ერთეულების, ჩვენს შემთხვევაში, იდიომური გამოთქმების შესახებ. შენახვის პროცესი იმგვარად უნდა იყოს წარმოდგენილი, რომ მომხმარებელს (კომპიუტერული პროგრამის ან ლექსიკონის გამოქვეყნებული ვერსიის შემთხვევაში) შეეძლოს ამ მასალის გამოყენება თავისი მიზნებისათვის. ნებისმიერი ლექსიკონის შექმნის პროცესი გარკვეული თანმიმდევრობით მიმდინარეობს და სხვადასხვა ეტაპის არსებობას გულისხმობს. ლექსიკონის შექმნას, როგორც წესი, წინ უძღვის: ერთეულების შერჩევა, ლემათა (ჩვენს შემთხვევაში, ამოსავალი იდიომების განსაზღვრა) ნუსხის შედგენა, სიტყვა-სტატიის ფორმისა და შიდამითითებების განსაზღვრა, კორპუსის შედგენა, კონკორდანსის საშუალებით იდიომების გამოყენების შესაძლო ვარიანტების მოძიება და ა.შ.

მეთოდოლოგია

თანამედროვე კორპუსული ლექსიკოგრაფიის მიდგომების გათვალისწინებით (ატკინსი 2008, სინკლერი 1996, ოოი 1998), ლექსიკონის შედგენა ეფუძნება ქართული ტექსტების კორპუსს (386 ერთეული, თითოეულ ტექსტში 40000 სიტყვა) და ბერძნულ ნაციონალურ კორპუსს (ΕΘΕΓ)[1]. ასევე, მონაცემთა კატეგორიების გასამართავად გამოვიყენეთ ISO-ს სტანდარტი (ISO 12620).

პროექტის მოკლე აღწერა

სტატიაში წარმოდგენილია ენის შემსწავლელებსა და პროფესიონალურ მთარგმნელებზე ორიენტირებული ბილინგვური ლექსიკონი - *იდიომური გამოთქმების ქართულ-ახალბერძნული ლექსიკონი*. იგი შექმნილია "იდიომური გამოთქმების ქართულ-ახალბერ-

[1] იხ. http://hnc.ilsp.gr/find.asp

ძნული ლექსიკონის" პროექტის ფარგლებში[1] და მოიცავს იდიომურ გამოთქმათა 11801 ერთეულს.

იდიომი წარმოადგენს არასტანდარტულ ლექსიკურ ერთეულს, რომლის მნიშვნელობაც არ უდრის მისი შემადგენელი ნაწილების (კომპონენტების) მნიშვნელობათა ჯამს (თაყაიშვილი 1961). მთარგმნელობითი თვალსაზრისით, იდიომური გამოთქმების გადმოცემა უცხო ენაზე რთულია, მაგრამ შესაძლებელია სხვადასხვა მთარგმნელობითი ტექნიკის საშუალებით (შენაცვლება, ადაპტაცია და ა.შ.)

ბერძნულ ნაციონალურ კორპუსსა და ქართული ტექსტების კორპუსზე დაყრდნობით, ლექსიკონის მონაცემთა ბაზის შექმნას დასჭირდა სამი ძირითადი ეტაპის გავლა. პირველი ეტაპი მოიცავდა ქართულენოვან მონაცემთა ბაზის შექმნას, მეორე ეტაპი - მოძიებული ერთეულების თარგმნას, და მესამე - შექმნილი ფორმების საბოლოო შეჯერებას. სქემის სახით ეს შემდეგნაირად შეიძლება წარმოვადგინოთ:

	ლემა	**გული გაუქვავდება**
	ლემის №	2091
	მნიშვნელობა	ულმობელი გახდება
	თარგმანი	**κάνω την καρδιά μου πέτρα, θα γίνω σκληρόκαρδος**
	სიტყვასიტყვითი თარგმანი	η καρδιά θα γίνει πέτρα
1	ქართულენოვანი მაგალითი	... ბატკნებს ჰკლავენ სააღდგომოდ და სასურველია პაწია ბავშვებს არ აცქერინონ ამ სურათზე, რათა მათ **გული არ გაუქვავდეს**, სისხლს არ შეეჩვიონო (ჭ. ლომთ. „თეთრი ღამე“)
	თარგმანი	... θυσιάζουν αρνάκια την Ανάσταση και καλύτερο να μην το δουν μικρά παιδιά για **να μην γίνει η καρδιά τους πέτρα**, να μην συνηθίζουν το αίμα (Τσ. Λομτ. «Άσπρη νύχτα»)
	ბერძნულენოვანი მაგალითი	«Μόνο ο Ρότσα, ως υπεύθυνος προπονητής, συναισθανόμενος ότι πρέπει, λόγω καθήκοντος, **να κάνει την καρδιά του πέτρα** και να ετοιμάσει την ομάδα όσο το δυνατόν καλύτερα, είπε, από...υποχρέωση περισσότερο, δύο κουβέντες» (Χ.Κ. Τεγόπουλος Εκδόσεις «Με βαριά καρδιά»)

გარდა ამისა, შემუშავდა ლექსიკონის მაკრო- (სიტყვათა ჯგუფების საწყისი ლემებისა და მათი კომპონენტების წარმოქმნის მექანიზმი) და მიკროსტრუქტურა მისი ძირითადი მახასიათებლებით.

ლექსიკონის შექმნა დასრულებულია, მოხსენებაში წარმოდგენილია ლექსიკონის მაკრო- და მიკროსტრუქტურები; განხილულია მისი განვითარების შესაძლებლობები.

[1] აღნიშნული პროექტი შესრულებულია შოთა რუსთაველის ეროვნული სამეცნიერო ფონდის მხარდაჭერით, გრანტი № Y-04-10. წინამდებარე პუბლიკაციაში გამოთქმული ნებისმიერი მოსაზრება ეკუთვნის ავტორს და შესაძლებელია არ ასახავდეს ფონდის შეხედულებებს.

Georgian – Greek (Modern Greek) Idiomatic Dictionary

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

Summary

In this paper I present the *Modern Georgian – Modern Greek Dictionary of Idioms*, a bilingual dictionary targeted at language learners and professional translators which has been compiled in the framework of a project for Young Scientists. After presenting the results of the project, I discuss the main methodological principles that underlie its construction and elaborate the main features of the dictionary. In particular, I report on the macrostructure (the organisation of the headwords and form production mechanisms) and the microstructure of the dictionary. The dictionary consists of 11801 entries of idioms.

Background (Introduction)

All dictionaries can be considered as databases. The purpose of such databases is to store the vocabulary of a language and to provide additional information about specific words (in our case about idiomatic units). Correct keeping should be represented in a way to allow the user (in the case of computer based dictionary software) and the reader (in the case of a published dictionary) to extract the information they need. The process of Dictionary compilation consists of certain sequences and different stages, including: selecting the units; defining a list of lemmas (in our case that of idioms); determining a form of entries; defining cross-references; providing a corpus of written texts; selecting possible variants of idioms by means of concordance etc.

Methodology

Following these approaches of modern corpus based lexicography (Atkins 2008, Sinclair 1996, Ooi 1998 and others), a compilation of a dictionary is based on the corpus of Georgian Texts (386 units, each of 40000 words) and the Hellenic National Corpus (ΕΘΕΓ)[1]. In addition to this, we have used ISO 12620 for the specification of data categories and management of a Data Category Registry for language resources.

Brief Description of the Project

In this paper I present the *Modern Georgian – Modern Greek Dictionary of Idioms*, a bilingual dictionary targeted at language learners and professional translators that has been compiled in the framework of a project for Young Scientists[2].

Any idiomatic expression is a non-standard lexical unit considered as: "a group of words which have a different meaning when used together from the one it would have if the meaning of each word (component) were taken individually" (Takaishvili, 1961). Translation of idioms from one language into another language

[1] see http://hnc.ilsp.gr/find.asp

[2] This project was supported by the Shota Rustaveli National Science Foundation and by Grant No Y-04-10. Any idea in this publication belongs to the author and may not represent the opinion of the Foundation.

can be implemented by using various techniques (transposition, modulation, adaptation etc.) applicable to a particular case.

Based on the Hellenic National Corpus (ΕΘΕΓ) and Corpus of Georgian Texts, we have created a database, which has progressed through the following stages: compilation of Georgian database, translation of selected units and editing of dictionary entries. The translated database can be represented as follows:

	Lemma	**გული გაუქვავდება**
	Number	2091
	Meaning	გახდება ულმობელი
	Translation	**κάνω την καρδιά μου πέτρα, θα γίνω σκληρόκαρδος**
	Word-by-word translation	η καρδιά θα γίνει πέτρα
1	Georgian example	... ბატკნებს ჰკლავენ საღდგომოდ და სასურველია პაწია ბავშვებს არ აცქერინონ ამ სურათზე, რათა მათ **გული არ გაუქვავდეს**, სისხლს არ შეეჩვიონო (ჭ. ლომთ. „თეთრი ღამე“)
	Translation	... θυσιάζουν αρνάκια την Ανάσταση και καλύτερο να μην το δουν μικρά παιδιά για **να μην γίνει η καρδιά τους πέτρα**, να μην συνηθίζουν το αίμα (Τσ. Λομτ. «Άσπρη νύχτα»)
	Greek example	«Μόνο ο Ρότσα, ως υπεύθυνος προπονητής, συναισθανόμενος ότι πρέπει, λόγω καθήκοντος, **να κάνει την καρδιά του πέτρα** και να ετοιμάσει την ομάδα όσο το δυνατόν καλύτερα, είπε, από...υποχρέωση περισσότερο, δύο κουβέντες» (Χ.Κ. Τεγόπουλος Εκδόσεις «Με βαριά καρδιά»)

In particular, I report on the macrostructure (the organisation of headwords in word families and groups of derived words and compounds that highlight word production mechanisms) and the microstructure of the dictionary.

Taking into account that the compilation stage of this dictionary is complete, I present the macrostructure and microstructure of the dictionary and describe the further stages of its development.

ბიბლიოგრაფია/References

B.T. Sue Atkins, *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press, 2008.
J. M. Sinclair, Corpus to Corpus: A Study of Translation Equivalence. *International Journal of Lexicography* , 171-178, 1996.
V.B. Ooi, *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press, 1998.
A. Takaishvili, The issues of the Georgian Phraseology, 1961.

lexicographic Sources
A Comprehensive Georgian-English Dictionary. Garnett Press, 2006.
D.N. Stavropoulos, A. H. *Oxford English-Greek Learner's Dictionary*. Oxford University Press, 1998.

Αθανάσιος, Θ. *Λεξικό ξενόφερτων λέξεων στην ελληνική γλώσσα υπό ΣΚΟΥΡΤΗΣ.* Αθήνα: Πύρινος Κόσμος, 1988
Βλαχόπουλος, Σ. *Λεξικό των ιδιωτισμών της νέας ελληνικής.* Κλειδάριθμος, 2007.
Γιαννουλέλλης, Γ. *Νεοελληνικές ιδιωματικές λέξεις δάνειες από ξένες γλώσσες.* Αθήνα : Κείμενα, 1982.
Δημητρίου, Α. *Λεξικό νεοελληνισμών: Ιδιωτισμοί, στερεότυπες μεταφορές και παρομοιώσεις, λέξεις και φράσεις από την καθαρεύουσα.* . Αθήνα: Γρηγόρη , 1995
Κριαράς, Ε. *Νέο ελληνικό λεξικό της σύγχρονης δημοτικής γλώσσας, γραπτής και προφορικής.* Αθήνα: Εκδοτική Αθηνών, 1995.
Λεξικό ομόηχων λέξεων και τονικών παρωνύμων: απο το Ομηρο μέχρι σήμερα υπό ΠΑΠΑΛΕΞΗΣ. Αθήνα : Δημόκριτος , 1998.
Λεξικό των Συνωνύμων της Νεοελλήνικης. Αθήνα : Αθήνα, 1970.
Μαρκαντωνάτος, Γ. *(Αθήνα). Λεξικό Αρχαίων, βυζαντινών και λογίων φράσεων της Νέας Ελληνικής.* 1992: Αθήνα.
Μπαμπινιώτης, Γ. *Λεξικό της Νέας Ελληνικής γλώσσας, Κέντρο λεξικολογίας.* Αθήνα: Αθήνα, 1998.
Oniani A., Georgian Idioms, Tbilisi, pubishing house: Nakaduli, 1966.
Chikobava, Arn., The explanatory dictionary of Georgian Language, Georgian national academy of sciences, 1950-1964; 2008.

დიდი ინგლისურ-ქართული ონლაინლექსიკონი, როგორც ელექტრონული კორპუსი

თინათინ მარგალიტაძე
ივ.ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი (საქართველო)
tinatin@margaliti.ge

გასული საუკუნის 70-იანი წლებიდან კომპიუტერმა და კომპიუტერიზებული რესურსების შექმნამ გადატრიალება მოახდინა ლექსიკოგრაფიაში.

„ ... უკვე მრავლად არის იმის დამამტკიცებელი საბუთი, რომ არსებობს მნიშვნელოვანი ენობრივი ფაქტები, რომლებიც მრავალი საუკუნის განმავლობაში ენის კვლევის მიუხედავად, არსად არ არის რეგისტრირებული ... მეორე მხრივ, ენობრივი მოვლენები, რომლებსაც რეგულარულად გვთავაზობენ, როგორც ინგლისური ენის ფუნქციონირების ნორმას, ფაქტობრივი მონაცემებით არასაკმარისად დასტურდება“, – წერდა 1985 წელს ჯონ სინკლერი.

იმისმა გაცნობიერებამ, რომ ელექტრონული კორპუსების მეშვეობით ვლინდებოდა ისეთი ენობრივი ფაქტები, რომლებიც ადრე არსად არ იყო დოკუმენტირებული, გამოიწვია თავდაპირველი ელექტრონული კორპუსების არნახული ზრდა. ინგლისური ტექსტის ბირმინჰემის კოლექცია, რომელიც 1980 წლისათვის 20 მილიონ სიტყვას მოიცავდა, 1990 წლისათვის 320 მილიონ სიტყვამდე გაიზარდა.

ბირმინჰემში შექმნილი კომპიუტერული რესურსები ეპოქალური მნიშვნელობის აღმოჩენად იქცა, რამაც განსაზღვრა ელექტრონული კორპუსების სწრაფი განვითარება, მათი ძირითად წყაროებად გამოყენება ლექსიკონების შექმნის პროცესში.

მართლაც, თუკი გავიხსენებთ იმ ფაქტს, რომ ოქსფორდის დიდი, ოცტომიანი ლექსიკონი ათი მილიონი საილუსტრაციო მასალის საფუძველზე შეიქმნა, რომელსაც მთელი ინგლისის მასშტაბით 800 მოხალისე აბარათებდა და რომლის დახარისხებას სარედაქციო გუნდმა თითქმის ათი წელი მოანდომა, ძნელი წარმოსადგენი არ უნდა იყოს ის შესაძლებლობები, რომლებსაც მრავალმილიონიანი ელექტრონული კორპუსები უქმნის ლექსიკოგრაფებს.

აღსანიშნავია, რომ დონალდ რეიფილდი იყო პირველი მეცნიერი ინგლისურ-ქართულ ლექსიკოგრაფიაში, რომელმაც დიდი ქართულ-ინგლისური ლექსიკონის პროექტისათვის ელექტრონული კორპუსი შექმნა. აღნიშნული კორპუსი მოიცავდა ორმოცამდე ქართულ რომანს და ორ მილიონამდე ერთეულის შემცველ საგაზეთო კორპუსს. დონალდ რეიფილდმა ასევე პირველმა გამოიყენა უნივერსიტეტის ლექსიკოგრაფიულ ცენტრში შექმნილი დიდი ინგლისურ-ქართული ლექსიკონი, როგორც ელექტრონული კორპუსი (მას ჰქონდა აღნიშნული ლექსიკონის თერთმეტი ტომის ელექტრონული ვერსია).

ერთენოვანი ელექტრონული კორპუსების კვალდაკვალ ჩნდება ორენოვანი ელექტრონული კორპუსები: ტერმინოლოგიური პარალელური ტექსტების კორპუსები (იურიდიული, ტექნიკური, სამედიცინო და სხვა); მხატვრული ლიტერატურის პარალელური კორპუსები - გერმანულ-ინგლისური, ფრანგულ-გერმანული, იტალიურ-ინგლისური, ჩინურ-ინგლისური, ჩინურ-ფრანგული და ა.შ.

როდესაც ივ. ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტის ინგლისური ფილოლოგიის კათედრამ დიდი ინგლისურ-ქართული ლექსიკონის შედგენაზე დაიწყო მუშაობა გასული საუკუნის 60-იან წლებში, პროექტზე მუშაობის ერთ-ერთ მნიშვნელოვან ეტაპად ორენოვანი სიმფონიის შედგენა იგეგმებოდა. ლექსიკონის სარედაქციო გუნდს, რომელიც მუშაობას ლექსიკონზე 80-იანი წლებიდან შეუდგა, ეს მასალა არ უნა-

ხავს. როგორც ჩანს, ამგვარი სიმფონიის შექმნა უპერსპექტივოდ მიიჩნიეს პროექტის იმჟამინდელმა სამეცნიერო ხელმძღვანელებმა. ორენოვანი სიმფონია თარგმნილ ლიტერატურას უნდა დაჰფუძნებოდა. ყველასათვის ცნობილია, რომ იმ პერიოდში მხატვრული ლიტერატურა დედნიდან იშვიათად ითარგმნებოდა, ქართული თარგმანები ინგლისური ლიტერატურის რუსული თარგმანებიდან სრულდებოდა. ქართველი მთარგმნელების უმრავლესობა დედნის ტექსტს ხშირად საკმაოდ თავისუფლად უდგებოდა და სათანადოდ არ ითვალისწინებდა ენობრივი ეკვივალენტობის პრობლემებს. იმავე მოსაზრებებიდან გამომდინარე, დიდი ინგლისურ-ქართული ლექსიკონის რედაქტორების გადაწყვეტილებით, ლექსიკონის წყაროებად გამოყენებულ იქნა ინგლისური ენის დიდი განმარტებითი ლექსიკონები და მათში თავმოყრილი საილუსტრაციო მასალა, ამოკრებილი მდიდარი ინგლისურენოვანი ლიტერატურიდან, ბიბლიიდან, პრესიდან და ა.შ.

მას შემდეგ, რაც ინგლისურ ლექსიკოგრაფიაში გაჩნდა პირველი, ელექტრონულ კორპუსებზე დაფუძნებული ლექსიკონები სიხშირული პრინციპებით შერჩეული ლექსიკით, ეს ლექსიკონები აქტიურად გამოიყენება ლექსიკონის სარედაქციო სამუშაოებში, რითაც დიდ ინგლისურ-ქართულ ლექსიკონში შესული კლასიკური საილუსტრაციო მასალა უფრო თანამედროვე კონტექსტებით შეივსო. ლექსიკონის სიტყვა-სტატიებზე მუშაობა 25 წელი მიმდინარეობდა და დღესაც გრძელდება. დღეისათვის მასში შესულია 110 000 სიტყვა-სტატია, მაქსიმალურად სრულადაა ასახული ინგლისური სიტყვების პოლისემია, ფრაზეოლოგია, შესიტყვებები, მოცემულია რამდენიმე ასეული ათასი ინგლისური მნიშვნელობის ქართული ეკვივალენტი. ლექსიკონში პრაქტიკულად თავი მოიყარა დიდძალმა ფაქტობრივმა მასალამ როგორც ინგლისურ, ისე ქართულ ენასთან მიმართებით. ლექსიკონი თავად იქცა ერთგვარ ორენოვან კორპუსად, რომელიც შეიძლება სხვა ლექსიკოგრაფიულ პროექტებში იქნეს გამოყენებული. მოხსენებაში წარმოდგენილი იქნება საილუსტრაციო მასალა დიდი ინგლისურ-ქართული ლექსიკონის კორპუსის სხვა ლექსიკოგრაფიულ პროექტებში გამოყენების პერსპექტივების საჩვენებლად.

Comprehensive English-Georgian Online Dictionary as an Electronic Corpus

Tinatin Margalitadze
Iv. Javakhishvili Tbilisi State University (Georgia)
tinatin@margaliti.ge

Since the 1970s the advent of the computer and computerised resources have revolutionised lexicography.

"*… there is now ample evidence of the existence of significant language patterns which have gone largely unrecorded in centuries of study … on the other hand, there is a dearth of support for some phenomena which are regularly put forward as normal patterns of English,*" wrote John Sinclair back in 1985.

The realisation of the fact that electronic corpora can reveal previously undocumented language patterns led to an unprecedented increase in the volume of early electronic corpora. The Birmingham Collection of English Text which comprised 20 million words by 1980 has been enlarged by 1990 to comprise as many as 320 million words. Computerised resources generated in Birmingham became a groundbreaking discovery which largely determined the rapid development of electronic corpora and their use as main sources in process of creation of dictionaries.

In fact, if we bring to mind the fact that the 20-volume Oxford English Dictionary was created on the basis of ten million quotation slips with illustrative material copied by 800 volunteers throughout the whole territory of England, which was later being sorted out by the members of the editorial team for almost ten years, we can easily imagine the wide prospects offered to lexicographers by multimillion-strong electronic corpora.

It is worth noting that Professor Donald Rayfield was the first scholar in English-Georgian lexicography to compose an electronic corpus for the project of the Comprehensive Georgian-English Dictionary (2006). The said corpus comprised up to forty Georgian novels and a newspaper corpus including up to two million items. Professor Rayfield was also the first lexicographer to use the Comprehensive English-Georgian Dictionary, created in the Lexicographic Centre of Tbilisi State University, as an electronic corpus (he had at his disposal the electronic version of eleven fascicles of the said Dictionary).

Following monolingual electronic corpora, bilingual electronic ones began to appear: terminological corpora of parallel texts (legal, technological, medical, etc); parallel corpora of belles-lettres texts: German-English, French-German, Italian-English, Chinese-English, Chinese-French, etc.

When the Chair of English Philology of Tbilisi State University began in the 1960s to work on the compilation of a Comprehensive English-Georgian Dictionary, the compilation of a bilingual concordance was one of the significant stages of the project. The editorial team, which began to work on the Dictionary in 1980s, have not seen the material in question. Apparently, the then academic supervisors of the project must have deemed the creation of such a concordance to have no prospects. Bilingual concordance had to be based on translated pieces of literature. However, as we know, in that period of our history belles-lettres were rarely translated directly from the original, as Georgian translations were habitually made from respective Russian translations of English books. The majority of Georgian translators treated original texts rather freely, without giving due consideration to the problems of linguistic equivalence. Based on the above-mentioned reasons, the editors of the Comprehensive English-Georgian Dictionary decided to rely upon major English explanatory dictionaries comprising ample illustrative material selected from numerous pieces of English-language literature, from the Bible, from newspapers and magazines, and so on, using them as the sources for the Dictionary. Following the appearance in the English lexicography of the first electronic corpus-based dictionaries with the wordlists selected according to frequency

principles, the editorial board of the Comprehensive English-Georgian Dictionary began to actively use the said dictionaries. This practice has contributed to the enrichment of the classic illustrative material included in the Dictionary with more recent contexts. The work on word-entries of the Dictionary has been carried out for 25 years and is continuing at the present time. To date, it includes 110,000 entries, to the maximum extent possible reflects polysemy of English words, English phraseological units, presents Georgian equivalents of several hundred thousand English meanings. As a result, the Dictionary has accumulated ample factual material with respect to both English and Georgian languages. The Dictionary has itself become a kind of bilingual corpus which can be used in other lexicographic projects. The paper will present illustrative material to highlight the prospects for the use of the corpus of the English-Georgian Dictionary in other lexicographic projects.

ქართული ენის შემსწავლელი ელექტრონული კურსი

თამარ მახარობლიძე, მარიამ მანჯგალაძე
არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველო)
mariam@ice.ge

ქართული ენის ელექტრონული კურსი გულისხმობს უცხოელთათვის პრაქტიკული ენის შემსწავლელი პროგრამის შექმნას. იგი იქმნება ინგლისურენოვანი აუდიტორიისთვის და არის გაკვეთილების ციკლი, რომელიც საშუალებას აძლევს კომპიუტერის ნებისმიერ მომხმარებელს ეტაპობრივად, საფუძვლიანად შეისწავლოს ქართული ენა - ლექსიკა, გრამატიკა, ფრაზეოლოგია; გამართოს მეტყველება და, გარკვეულწილად, დახვეწოს წერის კულტურა.

საერთოდ, ენის შემსწავლელი ნებისმიერი კურსი მოიაზრებს კომპლექსური ტიპის ამოცანებს. კერძოდ, აქ გამოიყოფა, ერთი მხრივ, წმინდა მეთოდოლოგიური და პედაგოგიური საკითხები და შესაბამისი ამოცანები, მეორე მხრივ, საუბარია ტექნოლოგიურ პროცესებზე, რაც გულისხმობს კურსის შინაარსის (კონტენტის) პროგრამული და დიზაინერული მხარეების შესაბამისობას საბოლოო პროდუქტის მისაღებად.

ქართული ენის ელექტრონული კურსის შესადგენად უპირველესი ამოცანაა კონკრეტული სასწავლო მიზნის განსაზღვრა. საუბარია ენის ცოდნის შესაბამის დონეზე. ჩვენ ვეყრდნობით განათლებისა და მეცნიერების სამინისტროს მიერ შემუშავებულ პროექტს: „ქართულის, როგორც უცხო ენის ფლობის დონეების აღწერილობა“, რომელიც შეიქმნა ევროპის საბჭოს ენის ფლობის ზოგადევროპული აღწერილობის საფუძველზე (**CEFR**). ქართული ენის ელექტრონულმა კურსმა ეტაპობრივად უნდა დაფაროს ყველა დონე.

ელექტრონული კურსის შესადგენად უმნიშვნელოვანესია შინაარსის (კონტენტის) ამოცანების გადაჭრა. ჩვენს სასწავლო სივრცეში არის ქართველი და უცხოელი სპეციალისტების მიერ გამოცემული რამდენიმე სახელმძღვანელო, რომელთა დადებითი და უარყოფითი მხარეების ანალიზი ძალიან წაადგება ქართული ენის ელექტრონული კურსის შექმნის პროექტს. ასეთი ანალიზი გამოავლენს სხვადასხვა სახის მეთოდური მიდგომების სუსტსა და ძლიერ მხარეებს.

ქართული ენის შემსწავლელი ელექტრონული კურსის შედგენა მოითხოვს ენობრივი ბაზის ლინგვისტურ ანალიზს პროგრამულ ამოცანებთან მიმართებით. ჩვენ ვიყენებთ ენის ელექტრონული სწავლების თვალსაზრისით მსოფლიოს არაერთ ქვეყანაში უკვე აპრობირებულ საავტორო პროგრამას eXelearning-ს (საჭიროებისამებრ შეიძლება გამოვიყენოთ სხვა პროგრამებიც, მაგ., Hot Potatoes), ხოლო ლინგვისტთა და მეთოდისტთა ჯგუფი ენის ფლობის შესაბამისი დონისათვის (1, 2, 1, 2 დონეები) ქმნის გაკვეთილების ციკლს სათანადო სავარჯიშოებით. კომპიუტერის მომხმარებელთა ფართო აუდიტორია სხვადასხვა ენისა და კულტურის მატარებელია, შესაბამისად, ქართული ენის სწავლების კომპიუტერული ვერსიის გაკვეთილები იგება ამ თავისებურებათა გათვალისწინებითაც. პარალელურად, პროგრამისტის, მხატვარ-დიზაინერისა და ანიმაციის სპეციალისტის დახმარებით ხდება პროექტის სათანადო ინსტრუმენტებით აღჭურვა და გაფორმება.

მნიშვნელოვანია კიდევ ერთი გარემოება: ენობრივ მონაცემთა ტიპოლოგიური ანალიზი გვიჩვენებს, რომ ქართული ენის სპეციფიკა წარმოქმნის რიგ სირთულეებს, რაც შემდგომში შესაბამისად აისახება ელექტრონული კურსის ფორმატშიც, მაგალითად, წინადადების თარგმანი სავარჯიშოებისათვის ერთ-ერთი ყველაზე მნიშვნელოვანი ელემენტია. იმ

დროს, როდესაც ინგლისური ენისთვის, გამონაკლისების გარდა, მხოლოდ ერთი ვარიანტია დასაშვები, ქართული ენა იძლევა თანაბარი შესაძლებლობების რამდენიმე ვარიანტს. ეს გარემოება, რა თქმა უნდა, ცვლის სავარჯიშოს ასაგებ პროგრამულ ამოცანებს. **რაც უფრო მდიდარი არჩევანი აქვს ენას სინტაქსურად ან სტილის მიხედვით, მით უფრო რთულდება პროგრამისტის ამოცანა და მით უფრო მარტივდება და იზღუდება ელექტრონული სავარჯიშოების ფორმების არჩევანი**. წინადადების თარგმნის შემთხვევაში ქართულისთვის დასაშვები ხდება მხოლოდ ერთი ტიპის სავარჯიშოს ფორმა – "აირჩიეთ სწორი ვარიანტი" – რაოდენობრივად კი შეიძლება დავუშვათ, რომ სწორი ვარიანტი იყოს ერთი, ორი ან სამი. ეს ტიპოლოგიური სხვაობა, უპირველეს ყოვლისა, განპირობებულია იმით, რომ ინგლისური ენა შედარებით მწირი (ზმნური და სახელური) მორფოლოგიის ფონზე სიტყვათა მკაცრი რიგის სინტაქსით ხასიათდება. სამაგიეროდ, ზმნური მორფოლოგიისა და სახელთა ბრუნების შესწავლისათვის აუცილებელია სხვადასხვა ტიპის სავარჯიშოების გამოყენება.

ქართული ენის ელექტრონული კურსის აგებისას გამოვიყენებთ ყველა ხელმისაწვდომ უფასო ინტერნეტრესურსს. კურსში გაკვეთილებად იქნება წარმოდგენილი თანდათანობითი ლექსიკა და ფრაზეოლოგია, დოზირებული გრამატიკა, შესაბამისი ტექსტები და სავარჯიშოები. კურსს დართული ექნება ზმნის უღლების ნიმუშები და სასარგებლო ბმულები.

დასასრულ, წარმოვადგენთ "**ქართული ენის შემსწავლელი ელექტრონული კურსის**" საპილოტე ვერსიას.

E-Learning course of the Georgian Language

Tamar Makharoblidze, Mariam Manjgaladze
Arn. Chikobava Institute of Linguistics (Georgia)
mariam@ice.ge, ateni777@yahoo.com

The e-Learning course of Georgian intends to teach the Georgian language to an English speaking audience in the electronic format. The course will be a cycle of the lessons providing lexical (including phraseology) and grammatical materials step by step. The course will have the additional files at the end – samples of the verb conjugations and the useful links.
Such e-dearning courses usually have two sides of the coin: A. Content objectives and B. Technical objectives. Concerning the content, it would be very useful to review the existing materials – the hard copies of the manuals of the practical Georgian in order to find out the strong and the weak sides of the proper methodologies and to find out the best model for teaching. The first task is to define the goal - which level of the language knowledge this course will cover. By the MOE group recently the templates for the Georgian language levels were exposed and this was performed according to the same type of document of the European Council - CEFR. The e-Learning course of Georgian will cover these levels step by step.

Technologically the eXelearning program and some other free resources (such as Moodle, Hot Potatos, etc.) could be used and the multimedia files will enrich the visual and audio forms for this product.
It's important to mention that the specifics of the Georgian language in some cases will dictate the specific frames for the exercises. For example the free word order of Georgian in opposition to the English language limits the forms of the exercises for translation of the sentences form English into Georgian. **The richer choice in syntax and style makes more limits for the choice of e-Learning format for the exercises. The typological diversity is exposed as a mirror opposition for such cases.**

Finally the pilot version of the e-Learning course of Georgian will be presented at the conference.

ქართული ენისა და ლიტერატურის სწავლების დისტანციური რეკომენდაციები საზღვარგარეთ არსებული საკვირაო სკოლების პედაგოგთათვის

სალომე ომიაძე
ივ. ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი,
არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველო)
salomiadze@yahoo.com

წარმოვადგენთ საზღვარგარეთ არსებული საკვირაო სკოლებისათვის განკუთვნილ პროგრამას ქართულ ენასა და ლიტერატურაში, რომელიც საქართველოს განათლებისა და მეცნიერების სამინისტროს, დიასპორის საკითხებში საქართველოს სახელმწიფო მინისტრის აპარატისა და მასწავლებელთა პროფესიული განვითარების ეროვნული ცენტრის დაკვეთით შეიქმნა. რადგან საკვირაო სკოლების მასწავლებლებიც დიასპორის წარმომადგენლები არიან, ამიტომ მათთვის საგანგებოდ შემუშავდა დისტანციური მომზადების კურსი: პროგრამის პრეზენტაცია, სამოდელო გაკვეთილი და სწავლების ძირითადი რეკომენდაციები, რომელთა აუდიოვიდეოვერსიები განთავსებულია ქართული დიასპორის ვებგვერდზე (იხ. http://www.diaspora.gov.ge/).

აღნიშნული პროგრამა 6-12 წლის ასაკობრივი ჯგუფისათვის შემუშავდა. მისი ძირითადი ლიტერატურული ნაწილი საქართველოს სასკოლო პროგრამას მიჰყვება – საკითხავი მასალა ამოკრებილია მოცემული ასაკობრივი ჯგუფისათვის განკუთვნილი სხვადასხვა სახელმძღვანელოდან. ეს მიდგომა ნაკარნახევია იმით, რომ დიასპორის წარმომადგენლის საქართველოში ნებისმიერ დროს დაბრუნების შემთხვევაში მან თავისუფლად შეძლოს ქართულ საგანმანათლებლო სივრცეში სრულყოფილი ინტეგრაცია. თუმცა, უნდა აღინიშნოს, რომ ლიტერატურულ ნაწარმოებთა ჩამონათვალი სარეკომენდაციო ლიტერატურის სრულსა და შეუცვლელ სიას არ წარმოადგენს. მასწავლებელს შესაძლებლობა ეძლევა, საკუთარი მოსწავლეების დონისა და შესასწავლი ენობრივი საკითხების გათვალისწინებით, თავად შეარჩიოს ავტორებიც, ტექსტებიცა და მათი მოცულობაც. შეიძლება ტექსტი ყოველთვის ლიტერატურული არც იყოს, ვგულისხმობთ, მაგალითად, სპონტანურად გამართულ დიალოგს მასწავლებელსა და მოსწავლეებს შორის, მოსწავლის ჩანახატს, ზეპირ მონათხრობს, საგაზეთო სტატიასა თუ სხვ., მაგრამ შერჩეულმა მასალამ მოსწავლეებს ენობრივი გრძნობა, ალღო უნდა განუვითაროს და გამოუმუშაოს პრაქტიკული ენობრივი უნარ-ჩვევები. ამ შემთხვევაში ერთადერთი რეკომენდაცია შემდეგია: საკითხავი მასალა შესასწავლ ენობრივ საკითხს უნდა მიესადაგებოდეს. ენისა და ლიტერატურის ამგვარი დაკავშირება ეფექტური გზაა ორივეგან სასურველი შედეგის მისაღწევად.

პროგრამის ენობრივი ნაწილი გრამატიკული ცნებებისა და წესების ტრადიციული სისტემური სწავლების ნაცვლად აგებულია გრამატიკის ელემენტების გამოყენებით შექმნილ იმგვარ სახალისო სავარჯიშოებსა და დავალებებზე, რომლებიც მოსწავლეებს დაანახვებს და შეაგრძნობინებს ქართული ენის ნაირგვარ შესაძლებლობას და საკუთარ დამოკიდებულებას გაუჩენს მათ ქართული ენისა და ლიტერატურის მიმართ, რაც კიდევ უფრო მნიშვნელოვანია იმ მოზარდებისათვის, რომლებიც ქართულენოვან საზოგადოებაში არ იზ-

რდებიან, თუმცა იზრდებიან იმ ადამიანთა გარემოში, რომლებიც საკუთარი ქვეყნის გარეთაც ინარჩუნებენ თვითიდენტიფიკაციას ისტორიულ სამშობლოსთან.

„გავიცნოთ ერთმანეთი“ (სამეტყველო ეტიკეტი); „სამშობლო“; „მშობლიური ენა მეტია, ვიდრე ენა“; „ქართული და სხვა ენები“; „ქართული სალიტერატურო ენა და დიალექტები“; „იაკობ გოგებაშვილი და „დედა ენა“; „ენა და გარე სამყარო“ და სხვ. ის საკითხებია, რომელთაგან პროგრამაში ზოგიერთი საკომუნიკაციო სიტუაციის თემადაა წარმოდგენილი, ზოგიერთი დისკუსიის საგანს ქმნის, ზოგიერთი საკითხავ მასალაშია ჩართული, ზოგიერთის გასაცნობად კი საკომუნიკაციო ამოცანის შესრულებაა საჭირო. სწორედ ამ და სხვა საკითხთა არსსა და სწავლების გზებზე ვრცლად მოხსენებაში ვისაუბრებთ.

Distance Recommendations for the Teaching of the Georgian Language and Literature for Teachers of Sunday Schools Functioning Abroad

Salome Omiadze
Iv. Javakhishvili Tbilisi State University,
Arn. Chikobava Institute of Linguistics (Georgia)
salomiadze@yahoo.com

This presentation deals with the program of the Georgian language and literature for Sunday Schools functioning abroad, which was created by the commission of the Ministry of Education and Science of Georgia, the Office of the State Minister of Georgia for Diaspora Issues and the National Centre for Professional Development of Teachers. As the teachers of Sunday Schools are diaspora representatives themselves, a course of distance training has been designed specifically for them. This included a presentation of the program, a model lesson and basic recommendations for teaching, the audio versions of which are available on the website of the Georgian Diaspora (see: http://www.diaspora.gov.ge/).

This program has been elaborated for the groups of 6-12 year old pupils. Its main literary part follows the Georgian school curriculum – reading material is collected from various textbooks intended for the given age group. This approach is selected considering that any representative of the diaspora living abroad upon their return to Georgia at any time should be able to integrate freely and fully into the Georgian educational system. However, it should also be noted that the list of recommended literary works is neither complete nor unchangeable. The teacher is given a freedom, taking into account the level of their class and the language issues to be explained, to select the authors and texts as well as their number and level. Selected texts may not always be literary and fixed for example a spontaneous dialogue between a teacher and pupils, a pupil's essay, an oral narrative, a newspaper article, etc. can also be included in the curriculum. However, the selected material should develop flair for the language in their students as well as enhance their

practical language skills. In this case the only recommendation which can be given to the teachers is the following: the reading material should be suitable for the language structure under study. Such linking of the language and literature is an effective way of obtaining a desirable result in both of these aspects.

The language section of the program, contrary to the traditional system of teaching grammatical concepts and rules, is based on entertaining exercises and tasks created by the use of grammar elements, which will be demonstrated to the pupils and make them feel the diverse possibilities of the Georgian language. In addition to this, it will cultivate in them their own attitude towards the Georgian language and literature. This factor is even more significant for adolescents who are brought up among the people retaining self-identification with their historical homeland even outside its borders.

Some of the issues and topics which are represented in the program as subjects for discussion or possibly for inclusion in the reading material are shown below. The essence and ways of teaching these and some other questions will be discussed in detail in the presentation:

Introducing Oneself **(conversation etiquette);** ***Homeland; Mother Tongue is More than a Language; Georgian and Other Languages; The Georgian Literary Language and Dialects; Iakob Gogebashvili and Deda Ena ("Mother Tongue"); Language and the Outer World,*** and so on.

The Use of Information Technologies for Interpreting Fiction

Liana Petrosyan, Satenik Arakelyan
V. Brusov Yerevan State Linguistic University (Armenia)
satenika@yahoo.com, ma.programmes@brusov.am

This article deals with some certain problems relating to the interpretation of fiction.

As a matter of fact, the practice of teaching some subjects on analysis and interpretation of texts shows that students, in most cases, “do not recognize" the text as they cannot find proper ways for their correct interpretation. Accordingly, very often the lecturer himself has to interpret the text for the auditorium from the beginning to the end. Consequently a student's work comes to naught, as a student turns from an explorer into a passive consumer of ready-made information.

Mainly such misunderstanding concerns the texts of modernism and postmodernism, which require a reader’s high activity and his competence as an explorer. This is primarily due to the fact that the concept of the text undergoes a change in linguistics, which entails a new approach to its main characteristics, cohesion and integrity. In traditional linguistics of a text, which originally served as the grammar of a text, cohesion was brought to the foreground, and it acted as a syntactic connection between inline elements sufficient for the existence of the text as a coherent and recognizable identity.

Nowadays, the presence of micro-cohesion only is insufficient for adequate comprehending, as even if these relations are evidently correct, it is difficult to ascribe the interpretable unit to the text. Cohesion is a formal and grammatical coherence of a text and it is defined with different types of linguistic relations of sentences within a text. Coherence in its turn covers semantic and pragmatic aspects of a semantic cohesion of a text as well. If the text is "grammatical" (if it does not formally break the language rules of a text formation) it is considered to have cohesion. Coherence is manifested in “the communicative success” of a text, explicating global connections and promoting text perception properly. It is expressed in the repetition of certain "motives" and "themes": the key objects, facts, structures, explicitly or implicitly expressed in the work, and it also creates their intersection within macrostructures, defining the scope of the topic.

Thus, it becomes clear that global cohesion has a more general nature and characterises all the text-formation areas. Macrostructure is needed for adequate description of the semantic area, and therefore it is necessary for the establishment of the global connectedness of the text. In addition, the online text comprehension postulated by van Duck, at the current stage, requires authentic solutions to this problem.

We offer a solution to this problem by developing the students' specific competence of a text analysis through the use of Learning Management System (MOODLE).
We will visually demonstrate one of the modules of the lesson. For the analysis we have chosen an excerpt from the prologue of the novel "Petersburg" by A. Beliy. Since the topic and semantics of a work are mainly set by its title, the first module of this lesson is turned exactly to the topic of this novel. First, students should become familiar with a small prologue, after which the instructor forms a task concerning the first stage of text perception.

Firstly, within the module the lecturer himself gives certain materials, which are designed to familiarise students with the notion of the text and learn its basic characteristics by properly selected Internet links. Thereafter, the student must briefly summarise the material in the learning environment by answering the instructor's questions, and show which of these criteria are met, and which are broken in the studied text. Secondly, there is another option as well, it is when the instructor displays the material chosen by himself in the learning environment in advance. In this case, the student's work is simplified because the instructor selects the definite material which is necessary for the proper analysis. The same procedure is also provided to identify the role of the title.

With the help of "concordance" program the instructor asks a student to calculate how frequently and in which environment the given word is used in the text. Such a task enables a student to pay attention to the semantically meaningful units of a text, which are so obvious, that they are unnoticeable in case of careless reading.

At the next stage there occurs the comprehension of meaningful units of a text. With this purpose it is necessary to attract the students' attention to the establishment of extra-textual cohesions, which, very often, they do not even assume. The instructor forms a task in such a way that with the help of properly selected Internet resources they could get acquainted with the phenomenon of the “Petersburg myth” in order to highlight its main characteristics and arrange it according to the schemes proposed by the instructor. The instructor shapes tasks with Hot-Pot, with the aim to be able to verify students' proper comprehension of the independent reading material.

In the next task the studied work is superimposed on the already created abstract scheme and is compared to it to identify the common traits that are characteristic of the model and our work, and which differ from one another.

When some preparatory work has been carried out at the instructor’s discretion, the students are offered general comprehensive questions for the complex analysis of the text. Thus, as we have seen, the given work in the system of MOODLE is comparable with the right solution to the puzzle displayed in the last task.

ИСПОЛЬЗОВАНИЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ ПРИ ИНТЕРПРЕТАЦИИ ХУДОЖЕСТВЕННОГО ТЕКСТА

Лиана Петросян, Сатеник Аракелян
Ереванский Государственный Лингвистический Университет им. В. Брюсова (Армения)
satenika@yahoo.com, ma.programmes@brusov.am

В данной статье мы хотели бы затронуть некоторые проблемы, связанные с интерпретацией художественных текстов.

Дело в том, что практика преподавания ряда предметов по анализу и интерпретации текстов показывает, что студенты, в большинстве случаев, «не узнают» текста, поскольку не могут найти пути его правильной интерпретации. Поэтому очень часто преподавателю самому приходится от начала до конца толковать текст аудитории, в результате чего работа студента сводится к нулю, так как студент из исследователя превращается в пассивного потребителя уже готовой информации.

Особенно такое непонимание касается текстов модернизма и постмодернизма, которые требуют высокой активности читателя и его компетентности как исследователя. Это, в первую очередь, связано с тем, что понятие текста претерпевает изменение в лингвистике, что влечет за собой и новый подход к его основным характеристикам: связности, целостности, цельности. В традиционной лингвистике текста, которая первоначально выступала как грамматика текста, на первом плане была категория связности (когезия), которая выступала в качестве синтаксической связности между внутритекстовыми элементами, достаточными для существования текста как целостной и узнаваемой единицы. В современную эпоху наличие только микросвязности оказывается недостаточным для адекватного понимания, поскольку даже при очевидности правильности этих связей трудно отнести интерпретируемую единицу к тексту. Когезия – это формально-грамматическая связанность текста и определяется она различными типами языковых отношений между предложениями внутри текста. Когерентность же охватывает и семантико-прагматические аспекты смысловой связности текста. Если текст «грамматичен», то есть формально не нарушает языковых правил создания текста, то считается обладающим когезией. Когеренция проявляется в «коммуникативном успехе» текста, эксплицируя глобальные связи и способствуя правильному пониманию текста. Она выражается в повторении определенных «мотивов» и «тем»: ключевых объектов, фактов, структур, эксплицитно или имплицитно выраженных в произведении, а также создает их пересечение в рамках макроструктуры, определяя референтную область применения данной темы.

Таким образом, становится ясно, что глобальная связь имеет более общую природу и характеризует все текстообразующее пространство. Макроструктура оказывается необходима для адекватного описания семантического пространства, и, соответственно, необходима для установления глобальной связности текста. Кроме того, постулируемое ван Дейком онлайн постижение текста на современном этапе требует аутентичных решений данной проблемы.

Мы предлагаем решение данной проблемы путем выработки у студентов определенных компетенций анализа текста с помощью применения Системы управления обучением (MOODLE).

Наглядно представим один из модулей данного урока. Для анализа мы выбрали отрывок из Пролога романа А. Белого «Петербург». Поскольку тема и семантика произведения, как правило, задаются его заглавием, то первый модуль данного урока обращен именно к теме данного романа. Сначала студенты должны ознакомиться с небольшим прологом, после чего преподаватель формирует задание относительно первого этапа понимания данного текста.

Во-первых, преподаватель сам дает в модуле определенные материалы, которые призваны ознакомить студентов с понятием текста и узнать о его основных характеристиках посредством правильно подобранных интернет-ссылок. После этого студент должен кратко изложить материал в обучающей среде, ответив на вопросы преподавателя, и показать, какие из данных критериев соблюдены, а какие нарушены в исследуемом тексте. Во-вторых, возможен и другой вариант, когда преподаватель заранее помещает выбранный им самим материал в обучающую среду. В таком случае упрощается работа студента, поскольку преподаватель отбирает именно тот материал и в том объеме, который необходим для правильного анализа. Та же самая процедура предусмотрена и для выявления роли заглавия.

С помощью программы «конкорданс» преподаватель просит подсчитать, как часто употребляется данное слово в тексте и в каком окружении. Такое задание позволяет обратить внимание студента на семантически значимые единицы текста, которые настолько очевидны, что незаметны при невнимательном чтении.

На следующем этапе происходит осмысливание значимых единиц текста. Для этого необходимо привлечь внимание учащихся к установлению внетекстовых связей, о которых они зачастую и не подозревают. Преподаватель формирует задание таким образом, чтобы при помощи правильно подобранных интернетресурсов, они смогли бы ознакомиться с явлением «Петербургского мифа» с целью выделить его основные характеристики и распределить по предложенным преподавателем схемам. А для того чтобы преподаватель смог проконтролировать правильность понимания студентами самостоятельно прочитанного материала, преподаватель формирует задания с помощью Hot-Pot.

В следующем задании рассматриваемое произведение накладывается на уже созданную абстрактную схему и сравнивается с ней для выявления тех общих черт, которые характерны для модели и нашего произведения, и которые отличают одно от другого.

После того, как по усмотрению преподавателя будет проведена подготовительная работа, студентам в последнем задании будут предложены вопросы общего характера для комплексного анализа текста. Таким образом, как мы смогли убедиться, данная работа в системе MOODLE оказывается сравнимой с правильным решением пазла в последнем задании.

ABBYY Lingvo Platform: ელექტრონული ლექსიკონების შექმნისა და პუბლიკაციის თანამედროვე ტექნოლოგიები

ანა რუბინშტეინი
კომპანია ABBYY (რუსეთი)
Anna_ru@abbyy.com

მოცემული ნაშრომში მიმოვიხილავთ კომპანია ABBYY-ში კომპიუტერული ლექსიკოგრაფიის დარგში დაახლოებით ოცი წლის მანძილზე მიმდინარე კვლევებსა და სამუშაოს, რომელთა შედეგადაც შეიქმნა ისეთი ლექსიკოგრაფიული ინსტრუმენტები, როგორებიცაა: ABBYY Lingvo Content–ის ლექსიკონების შესაქმნელი პროფესიონალური სისტემა, ელექტრონული ლექსიკონი ABBYY Lingvo, მობილური ლექსიკონი ABBYY Lingvo Mobile, ლექსიკონის სერვერული ვერსია ABBYY Lingvo Server და სალექსიკონო ონლაინ პორტალი ABBYY Lingvo.Pro.

ამ წლების განმავლობაში კომპანიაში მუშავდებოდა არა მხოლოდ პროგრამული პროდუქტები და დანართები, არამედ მიმდინარეობდა ლინგვისტური კვლევითი სამუშაოები სემანტიკის, სინტაქსისა და ლექსიკოგრაფიის დარგებში, მზადდებოდა ლექსიკონები გამოსაცემად – ამ მომენტისათვის უკვე 40–ზე მეტი ლექსიკონია შექმნილი კომპანია ABBYY–ის ლექსიკოგრაფთა მიერ სხვა ავტორებთან თანამშრომლობით.

ABBYY Lingvo Platform

1989 წელს შექმნილი ელექტრონული ლექსიკონი ABBYY Lingvo გამუდმებით ვითარდებოდა. იგი იძენდა ახალ ფუნქციებს, რითაც იზიდავდა ყოველწლიურად სულ უფრო და უფრო მეტ მომხმარებელს.

უკვე 1993 წელს მას გაუჩნდა ინსტრუმენტი, რომელიც მომხმარებელს საკუთარი ლექსიკონების Lingvo-ს ფორმატში გარდასახვის საშუალებას აძლევდა. იმ დროიდანვე გახდა შესაძლებელი Windows-ის პროგრამებთან ინტეგრაცია (დანართებიდან უშუალო გადაყვანა "ცხელი კლავიშების" გამოყენებით), სხვადასხვა – როგორც საერთოლექსიკურ, ისე სპეციალიზებულ – ლექსიკონებს შორის გადართვის შესაძლებლობა ერთი პროგრამის ფარგლებში, გადამკვეთი ბმულების გამოყენება. ხოლო 1997 წელს ლექსიკონ Lingvo-ს მე-5 ვერსიას უკვე ჰქონდა მორფოლოგიური მხარდაჭერა და სრულტექსტოვანი ძიების შესაძლებლობა, რითაც მომხმარებელს საშუალება ეძლეოდა სიტყვა მოეძებნა არა მარტო საწყისი ფორმის მიხედვით, არამედ ადვილად მოეპოვებინა სიტყვათხმარებათა რეალური მაგალითები.

ლექსიკონის დახვეწა არ შეჩერებულა, მას გაუჩნდა მრავალი ახალი ფუნქცია და შესაძლებლობა, როგორიცაა მიზანმიმართული თარგმანი, ერთდროული ძიება ყველა ენის მიხედვით, სიტყვის ყველა გრამატიკული ფორმის ერთდროულად ნახვის, სიტყვის წარმოთქმის მოსმენის შესაძლებლობები. ამჟამად პროგრამა შეიცავს 220 ლექსიკონს 20 ენობრივი მიმართულებით, დანართს ABBYY Lingvo Tutor-ის სიტყვათა შესასწავლად, ილუსტრაციებს, ვიდეოგაკვეთილებს, ენის მატარებელთა მიერ გახმოვანებულ სასაუბროებს. კომპანიის გამოკვლევების თანახმად, ABBYY Lingvo-ს ლექსიკონებს მსოფლიოს 10 მილიონზე მეტი ადამიანი იყენებს.

კომპიუტერული ტექნოლოგიების განვითარებასთან ერთად მნიშვნელოვნად გაფართოვდა ლექსიკონის გამოყენების სფეროც: დღეს ABBYY Lingvo წარმოადგენს ელექტრონულ ლექსიკონს PC და Mac-ზე დასაყენებლად, დანართს მობილური მოწყობილობებისთვის (ტელეფონები, სმარტფონები, iPad-ები, ელექტრონული წიგნები), ლექსიკონის სერვერულ ვერსიას intranet (ABBYY Lingvo Server)-ისათვის, ამავე დროს არსებობს საინტერნეტო პორტალის სახით.

ელექტრონული დანართების – ABBYY Lingvo Desktop, ABBYY Lingvo Mobile, ABBYY Lingvo Server და ABBYY Lingvo Content – ერთობლიობა შეადგენს ABBYY Lingvo Platform–ს. ყველა ეს პროდუქტი უზრუნველყოფს მომხმარებლისათვის სწრაფ და მოხერხებულ მისაწვდომობას კომპანიის სერვერზე, რომელზეც ინახება ლექსიკონები და საცნობარო კონტენტი.

ამგვარად, ABBYY Lingvo Platform-ი, რიგი ლექსიკოგრაფიული ინსტრუმენტების გაერთიანებით, დახმარებას უწევს ლექსიკოგრაფებსა და გამომცემლობებს ახალი ლექსიკონების შექმნასა და თანამედროვე მომხმარებლისათვის ხელმისაწვდომი ლექსიკონების სხვადასხვა ელექტრონულ მატარებელზე გამოქვეყნებაში. ამ ჯაჭვში შეუცვლელ ინსტრუმენტად წარმოგვიდგება ABBYY Lingvo Content-ი.

ABBYY Lingvo Content

20 წლის მანძილზე კომპანია აქტიურად ქმნიდა და ავითარებდა საკუთარ ლექსიკონებს, ახორციელებდა მასშტაბურ ლექსიკოგრაფიულ პროექტებს; პარალელურად ასობით ლექსიკონი იქმნებოდა ABBYY Lingvo-ს მომხმარებელთა მიერ. ამიტომაც აუცილებელი გახდა ლექსიკონთა შესადგენი პროფესიული სისტემის შემუშავება Lingvo-ს ფორმატში მომუშავე გამომცემლობათა, ლექსიკოგრაფთა ჯგუფებისა და ინდივიდუალური ავტორების მოთხოვნათა დასაკმაყოფილებლად.

ABBYY Lingvo Content სისტემა იქცა მოხერხებულ ინსტრუმენტად ახალი ლექსიკონების შესაქმნელად და არსებულ ლექსიკონებზე სამუშაოდ, რომელიც არა მხოლოდ ლექსიკონების შექმნის საშუალებას იძლევა მოხერხებულ სარედაქტორო ინტერფეისში, არამედ მიღებული მონაცემების განსხვავებულ ფორმატებში ექსპორტირების შესაძლებლობასაც მათი, როგორც ბეჭდური, ისე ელექტრონული ან საინტერნეტო (on-line) სახით პუბლიკაციის მიზნით.

ABBYY Lingvo Content სისტემას აქვს კლიენტ-სერვერული არქიტექტურა. მთელი ინფორმაცია შენახულია მონაცემთა ბაზაში SQL სერვერზე, მომხმარებლებს მონაცემები შეჰყავთ მათსავე კომპიუტერზე დაყენებულ დანართებში. ეს საშუალებას იძლევა საიმედოდ იყოს დაცული მიღებული მონაცემები.

სალექსიკონო სტატიის შექმნისას ABBYY Lingvo Content-ში ლექსიკოგრაფი მაშინვე ანაწილებს ინფორმაციას ზონების მიხედვით – მაგალითები, იდიომები, განმარტებები, აღნიშვნები და ა. შ. ლაგდება სალექსიკონო სტატიის საჭირო ნაწილებში, რომლებიც ავტომატურად ინახება მონაცემთა ბაზაში სტრუქტურირებული სახით. ეს საშუალებას იძლევა:

- შევძლოთ ინტელექტუალური ძიება სალექსიკონო სტატიებში და სტატიების გადარჩევა ნებისმიერი პარამეტრით;
- ერთდროულად ვიმუშაოთ რამდენიმე ლექსიკონთან;

- ლექსიკონები და ლექსიკონთა ჯგუფები შევუდაროთ ერთმანეთს, ამასთან სიტყვანის მიხედვით განსხვავებები გამოვიყენოთ ლექსიკონის გასაახლებლად;
- მივიღოთ ნებისმიერი სახის სტატისტიკა: რამდენ სტატიას შეიცავს მოცემული ლექსიკონი, რამდენია მასში იდიომი, მაგალითი, თარგმანი და სხვა ნებისმიერი ელემენტი.
- გაუმჯობესდეს სიტყვიერი მასალის მიღების ხარისხი, რამდენადაც სისტემა იძლევა მასალის სტრუქტურირებისა და უნიფიცირების საშუალებას.

ABBYY Lingvo Content-ი ლექსიკოგრაფებს, რედაქტორებსა და სალექსიკონო პროექტის ყველა მონაწილეს ათავისუფლებს რუტინული შრომისაგან, მისი მეშვეობით ავტომატიზებულია მრავალი მოქმედება და გასაქანი ეძლევა სალექსიკონო სტატიის შექმნისა და რედაქტირებისათვის შემოქმედებით გადაწყვეტილებებს. მაგალითად, ისეთი შრომატევადი ამოცანები, როგორებიცაა სალექსიკონო სტატიებში დამოწმებების გადამოწმება, სახელის გადარქმევა და ნუმერაციის შეცვლა, ხორციელდება ავტომატურად.

ABBYY Lingvo Content-ის ერთ-ერთი მნიშვნელოვანი ფუნქციაა სალექსიკონო პროექტის სამუშაოების ადმინისტრირების შესაძლებლობა. დანართი სალექსიკონო პროექტის მონაწილეებს (ლექსიკოგრაფებს, რედაქტორებს, კორექტორებს) შორის სამუშაოს განაწილებისა და ლექსიკონზე მუშაობის ყველა ეტაპზე თვალის გადევნების საშუალებას იძლევა.

ეს შესაძლებელს ხდის, რომ რამდენიმე ლექსიკოგრაფმა ერთდროულად იმუშაოს ერთ სალექსიკონო პროექტზე, მაშინაც კი, თუ ისინი იმყოფებიან სხვადასხვა ქვეყანაში.

ამგვარად, ABBYY Lingvo Content-ში ჩადებულია როგორც სალექსიკონო პროექტის დაგეგმარებისა და წარმართვისათვის აუცილებელი, ისე ლექსიკოგრაფის სალექსიკონო სტატიის შექმნისას მუშაობის შემამსუბუქებელი საშუალებანი.

ამჟამად სისტემას წარმატებით გამოიყენებენ ლექსიკონთა ავტორები და ავტორთა კოლექტივები რუსეთიდან, ფინეთიდან, პოლონეთიდან, გერმანიიდან, ასევე მსხვილი გამომცემლობები, როგორიცაა ABBYY Press, Larousse და სხვ.

ABBYY Lingvo.pro

ლექსიკონზე, ცნობარზე ან გლოსარიუმზე მუშაობის პროცესში ლექსიკოგრაფი აწყდება სიტყვის მოხმარების, საჭირო მნიშვნელობის ან თარგმანის მაგალითების მოსაძიებლად საიმედო წყაროსა და მოხერხებული ინსტრუმენტის არჩევანის პრობლემას. იმისათვის, რომ დავაკმაყოფილოთ ეს და სხვადასხვა მიზნობრივი აუდიტორიის – როგორც ლექსიკოგრაფებისა, ისე ლექსიკონის სხვა საბოლოო მომხმარებლების – მრავალი სხვა მოთხოვნა, კომპანია ABBYY-მ შეიმუშავა სალექსიკონო და კორპუსული პორტალი Lingvo.pro.

ABBYY Lingvo.pro – ესაა ონლაინ პორტალი, რომელიც შეიცავს კონტენტის სხვადასხვა სახეობას (მათ შორის ისეთი გამომცემლობებისა, როგორიცაა Oxford და Collins), მოხერხებულ საძიებო ინსტრუმენტებს, ხარისხიან მორფოლოგიურ მხარდაჭერას და, იმავდროულად, ახდენს ორენოვანი პარალელური ტექსტების დიდი მასივების აკუმულირებას სიტყვათხმარებისა და გამოთქმების რეალური მაგალითებით.

დასკვნა

ჩვენს დროში მომხმარებლისთვის აუცილებელია უბრალო და მარჯვე ინსტრუმენტი, რომელიც მას საშუალებას მისცემს ერთდროულად გადაწყვიტოს მრავალი ამოცანა: მოძებნოს სიტყვის მნიშვნელობა ლექსიკონში, ცნობარსა ან ენციკლოპედიაში, თარგმნოს სიტყვა და გამოთქმა სხვა ენაზე, ისწავლოს ახალი სიტყვები და ა.შ. ამავე დროს ავტორები და გამომცემლები დაინტერესებული არიან, რომ მათი კონტენტი მივიდეს საბოლოო მომხმარებლამდე მისთვის მაქსიმალურად მოხერხებული ფორმით.

სწორედ ამიტომ ABBYY ამუშავებს სალექსიკონო პროგრამულ პროდუქტებს, რომლებიც ერთიანდება სისტემაში ABBYY Lingvo Platform. ამის წყალობით კონტენტის შემქმნელებს ხელთ აქვთ ინსტრუმენტი ლექსიკონის შესაქმნელად და მისი სხვადასხვა საშუალებით წარდგინების პროგრამული გარსი, საბოლოო მომხმარებლებს კი (ლექსიკოგრაფები, ტერმინოლოგები, მთარგმნელები და უცხო ენის შემსწავლელი პირები) – მარტივი და მოხერხებული მისაწვდომობა ლექსიკონებთან პერსონალური კომპიუტერით, მობილური მოწყობილობებით და ინტერნეტში.

ეს მოდელი წარმატებით რეალიზდება ABBYY კომპანიის მიერ როგორც საკუთარი ლექსიკონების პუბლიკაციისთვის, ისე უმსხვილესი რუსული და საზღვარგარეთული გამომცემლობების კონტენტებისთვის, როგორიცაა Oxford, Collins, Compact Verlag, Русский язык – Медиа, РУССО, Перун და სხვ.

ABBYY Lingvo Platform: современные технологии создания и публикации словарей в электронном виде

Рубинштейн Анна
Компания ABBYY (Россия)
Anna_ru@abbyy.com

Данная работа посвящена обзору более чем двадцатилетних исследований и разработок в области компьютерной лексикографии, ведущихся в компании ABBYY и нашедших свое воплощение в таких лексикографических инструментах, как профессиональная система для создания словарей ABBYY Lingvo Content, электронный словарь ABBYY Lingvo, мобильный словарь ABBYY Lingvo Mobile, серверная версия словаря ABBYY Lingvo Server и словарный онлайн портал ABBYY Lingvo.Pro.

На протяжении этих лет компания не только разрабатывала программные продукты и приложения, но и проводила исследования в области лингвистики, семантики, синтаксиса и лексикографии, занималась созданием словарей – на данный момент уже более 40 словарей создано лексикографами компании ABBYY в сотрудничестве с внешними авторами.

ABBYY Lingvo Platform

Созданный в 1989 году электронный словарь ABBYY Lingvo постоянно развивался, в нем появлялись новые функции, с каждым годом привлекающие к нему все больше пользователей.
Уже в 1993 году в нем появился инструмент, позволяющий пользователям преобразовывать собственные словари в формат Lingvo. Тогда же стали возможными интеграция с программами Windows (перевод прямо из приложений с помощью использования «горячих» клавиш), возможность переключаться между различными словарями (как общелексическими, так и специализированными) в составе одной программы, использование перекрестных ссылок. А в 1997 г. 5-я версия словаря Lingvo уже включала морфологическую поддержку и полнотекстовый поиск, которые позволяли пользователям искать слово не только в его начальной форме, а также с легкостью находить реальные примеры словоупотребления.

Развитие словаря не стояло на месте, в нем появилось множество новых функций и возможностей, таких как перевод по наведению, поиск одновременно по всем языковым направлениям, возможность посмотреть все грамматические формы слова, прослушать произношение слова. Сейчас программа содержит 220 словарей по 20 языковым направлениям, приложение для заучивания слов ABBYY Lingvo Tutor, иллюстрации, видеоуроки, разговорники, озвученные носителями языка. Согласно исследованиям компании, более 10 миллионов человек в мире пользуются словарями ABBYY Lingvo.
С развитием компьютерных технологий значительно расширилась и сфера использования словаря: сегодня ABBYY Lingvo представляет собой электронный словарь для установки на PC и Mac, приложение для мобильных устройств (телефонов, смартфонов, iPad, электронных книг), серверную версию словаря для intranet (ABBYY Lingvo Server), а также существует в виде онлайн портала.

Совокупность электронных приложений ABBYY Lingvo Desktop, ABBYY Lingvo Mobile, ABBYY Lingvo Server и ABBYY Lingvo Content составляют ABBYY Lingvo Platform. Все эти продукты обеспечивают пользователям быстрый и удобный доступ к серверу компании, на котором хранятся словари и другой справочный контент.

Таким образом, ABBYY Lingvo Platform, объединяя ряд лексикографических инструментов, помогает лексикографам и издательствам создавать новые словари и публиковать их на различных электронных носителях, доступных современным пользователям словарей. Незаменимым инструментом в этой цепочке становится система для создания словарей ABBYY Lingvo Content.

ABBYY Lingvo Content

На протяжении 20 лет компания активно создавала и развивала собственные словари, реализовывала крупные лексикографические проекты; параллельно сотни словарей создавались пользователями программы ABBYY Lingvo. Поэтому необходимой стала разработка

профессиональной системы для создания словарей для удовлетворения нужд издательств, групп лексикографов и индивидуальных авторов, работающих в формате Lingvo.

Система ABBYY Lingvo Content стала удобным инструментом для создания новых и работы с существующими словарями, позволяя не только создавать словари в удобном редакторском интерфейсе, но и экспортировать полученные данные в различные форматы для их последующей публикации, как в печатном, так и в электронном или on-line виде.

Система ABBYY Lingvo Content имеет клиент-серверную архитектуру. Вся информация хранится в базах данных на SQL сервере, пользователи вводят данные в приложение, которое установлено на их компьютере. Это позволяет надежно хранить полученные данные.

При создании словарной статьи в ABBYY Lingvo Content лексикограф сразу распределяет информацию по зонам – примеры, идиомы, толкования, пометы и т.д. оказываются в нужных частях словарной статьи, которая автоматически сохраняется в базе данных в структурированном виде. Это позволяет:

- осуществлять интеллектуальный поиск по словарным статьям и делать выборки статей по любым параметрам
- работать с несколькими словарями одновременно
- сравнивать словари и группы словарей между собой, при этом разницу по словнику можно использовать для обновления словаря
- получать любые виды статистики: сколько статей содержит данный словарь, сколько идиом, примеров, переводов и любых других элементов
- улучшить качество получаемого словарного материала, поскольку система позволяет структурировать и унифицировать материал

ABBYY Lingvo Content помогает избавить лексикографов, редакторов и всех участников словарного проекта от рутинной работы, автоматизируя многие действия и оставляя простор для творческих решений при создании и редактировании словарной статьи. Например, такие трудоемкие задачи, как проверка, переименование и перенумерация ссылок между словарными статьями, осуществляются автоматически.

Одной из важных функций ABBYY Lingvo Content является наличие средств для администрирования работ над словарным проектом. Приложение позволяет распределять работы между участниками словарного проекта – лексикографами, редакторами, корректорами, и отслеживать ход работ на любом этапе работы над словарем. Это дает возможность нескольким лексикографам одновременно работать над одним словарным проектом; при этом они могут находиться в разных странах.

Таким образом, в ABBYY Lingvo Content заложены средства как для планирования и ведения словарного проекта, так и для облегчения работы лексикографов при создании словарных статей.
В настоящее время систему успешно используют авторы словарей и авторские коллективы из России, Финляндии, Польши, Германии, а также крупные издательства, такие как ABBYY Press, Larousse и др.

ABBYY Lingvo.pro

В процессе работы над словарем, справочником или глоссарием, лексикограф сталкивается с проблемой выбора надежного источника и удобного инструмента для поиска примеров употребления, нужного значения или перевода слова. Для того чтобы удовлетворить эти и многие другие нужды различной целевой аудитории: как лексикографов, так и конечных пользователей словарей, компания ABBYY разработала словарный и корпусный портал Lingvo.pro.
ABBYY Lingvo.pro – это онлайн портал, который содержит различные виды контента (в том числе от таких издательств, как Oxford и Collins), удобные инструменты для поиска, качественную морфологическую поддержку и в то же время аккумулирует большие объемы двуязычных параллельных текстов с реальными примерами употребления слов и выражений.

Заключение
В наше время пользователю необходим простой и удобный инструмент, который позволит ему решать множество задач одновременно: искать значение слова в словаре, справочнике или энциклопедии, переводить слова и выражения на другие языки, заучивать новые слова и т.д. В то же время авторы и издатели заинтересованы в том, чтобы донести свой контент до конечного пользователя в максимально удобной для него форме.

Именно поэтому ABBYY разрабатывает словарные программные продукты, которые объединяются в систему ABBYY Lingvo Platform. Благодаря этому создатели контента получают инструмент для создания словаря и оболочку для его презентации различными способами, а конечные пользователи (лексикографы, терминологи, переводчики и люди, изучающие иностранный язык) - простой и удобный доступ к словарям с ПК, мобильных устройств и онлайн.
Эта модель успешно реализуется компанией ABBYY как для публикации собственных словарей, так и контента от крупнейших российских и зарубежных издательств, таких как Oxford, Collins, Compact Verlag, Русский язык – Медиа, РУССО, Перун и др.

ბიბლიოგრაფია / Библиография

Atkins S. and Rundell M., The Oxford Guide to Practical Lexicography by Oxford University Press Inc., New York, 2008.

Kuzmina V. and Rylova A., The ABBYY Lingvo Electronic Dictionary and the ABBYY Lingvo Content Dictionary Writing System as Lexicographic Tools. In Proceedings of eLexicography in the 21st century: New challenges, new applications (pp.131-133). Louvain-la-Neuve, 2009.

Рылова А., Использование профессиональных приложений для создания словарей в работе лексикографа и авторского коллектива / Новое в теории и практике лексикографии: синхронный и диахронный подходы: материалы VIII Международной школы-семинара, Иваново, 10-12 сент. 2009 г. – Иваново: Иван. гос. ун-т, 2009. с.382-386.

Rylova, A., Electronic Dictionary and Dictionary Writing System: how this duo works for dictionary user's needs (ABBYY Lingvo and ABBYY Lingvo Content case) InDykstra, A. (Ed.), Schoonheim, T. (Ed.), Proceedings of the XIV Euralex Intenational Congress. Euralex Intenational Congress. Leeuwarden, 6-10 July 2010 (pp. 1151). Leeuwarden: Fryske Akademy, 2010

Piotrowski T., Mobile Dictionaries: Situations of Use. In Proceedings of eLexicography in the 21st century: New challenges, new applications (pp.181-182). Louvain-la-Neuve, 2009.

Heid U., Aspects of Lexical Description for Electronic Dictionaries. In Proceedings of eLexicography in the 21st century: New challenges, new applications (pp.1-3). Louvain-la-Neuve, 2009.

რუსული ენის ნაციონალური კორპუსი და ლინგვისტური კორპუსები პოსტსაბჭოთა სივრცეში: კორპუსის ტექნოლოგიების ერთობლივი პროექტები და სტანდარტები

დიმიტრი სიჭინავა
რუსეთის მეცნიერებათა აკადემიის რუსული ენის ინსტიტუტი (რუსეთი)
mitrius@gmail.com

რუსული ნაციონალური კორპუსი სხვადასხვა კორპუსული პროექტის ფარგლებში განხორციელდა და გამოქვეყნდა 2004 წ. (http://ruscorpora.ru/).

რამდენადაც მოხსენებაში საუბარი შეეხება ყოფილი სსრკ-ს ქვეყნებს, აღვნიშნავთ, რომ მათთან თანამშრომლობა სრულიად სხვადასხვა მიმართულებით შეიძლება წარიმართოს:

1. პარალელური (თარგმნითი) ორენოვანი ან მრავალენოვანი კორპუსები;
2. რუსული ენის კორპუსების შექმნა ამ ქვეყნებში (რუსული ნაციონალური კორპუსის (НКРЯ) სუბკორპუსების სახით;
3. მონაწილეობა თანამედროვე სტანდარტებისა და არარუსული მონოლინგვური კორპუსების განვითარებაში, გამოცდილებისა და ტექნოლოგიების გაცვლა.

რუსული ნაციონალური კორპუსის ფარგლებში განხორციელებული პარალელური კორპუსები ძირითადად ეყრდნობიან ბოლო ხანებში ევროპაში შექმნილი სლავური პარალელური კორპუსების მოდელს (მაგ., ParaSOL, ASPAC, ჩეხური ნაციონალური კორპუსი და სხვ.). ჩვენ ვთანამშრომლობთ კოლეგებთან როგორც სხვადასხვა პოსტსაბჭოთა ქვეყანაში, ისე რუსეთის ფედერალურ რეგიონებში. შექმნილია უკრაინულ-რუსული, რუსულ-უკრაინული და რუსულ-ბელორუსული თარგმნითი კორპუსები (http://ruscorpora.ru/search-para.html). ამას მოჰყვა მუშაობა რუსულ-ბაშკირულ კორპუსზე.

თანამშრომლობის ეს მიმართულება აქტუალურია პოსტსაბჭოთა სივრცის დიდი ნაწილისთვის: დიდძალი ნათარგმნი მასალა ჯერ კიდევ საბჭოთა ერაში შეიქმნა, ახალი პოლიტიკური და კულტურული სიტუაციების შესაბამისად ჩნდება ახლადთარგმნილი ტექსტებიც. ამ თარგმანების შესწავლას ლინგვისტურის გარდა ისტორიული და კულტურული ღირებულებაც აქვს. მათი საშუალებით, შეიძლება ითქვას, გამოვლინდება ლინგვისტური "კოლონიზაციისა" და "ემანსიპაციის" ინტენსიური პროცესი.

რუსული ენა პოსტსაბჭოთა სივრცეში (რუსეთის ტერიტორიის ჩათვლით) წინა პერიოდის ტოტალიტარული უნიფიკაციისგან განსხვავებით ძალიან დიდ მრავალფეროვნებას ავლენს.

ვლინდება ახალი რეგიონული კომუნიკაციური ვარიანტები, მაგ., რუსული ენა დაღესტანში (რომელიც აქ lingua franca-ს სტატუსით ფუნქციონირებს და ნახურ-დაღესტნურ ენათა გრამატიკისა და ფონეტიკის ძლიერ ზეგავლენას განიცდის), ასევე: რუსული ენა შუა აზიის ქვეყნებში.

უკრაინასა და ბელორუსიაში მკვიდრდება და შეისწავლება რუსულ-უკრაინულ/ბელორუსიული კოდების შერევის შედეგად მიღებული ვარიაციული ფორმები. ეს სახესხვაობები თავმოყრილია ტექსტურ კოლექციებში და წარმოდგენილია კორპუსებში. ამ ტიპის მასალა ქმნის ერთ-ერთ ქვემიმართულებას რუსულ ნაციონალურ კორპუსში და აერთიანებს რუსული ენის რუსეთის საზღვრებს გარეთ გავრცელებულ ვარიანტებს (აშშ, ფინეთი...).

რუსული ნაციონალური კორპუსი იძლევა თანამშრომლობის შესაძლებლობას მონოლინგვური არარუსული და არასლავური კორპუსების შესაქმნელად. რუსულ კორპუსზე დაყრდნობით შეიქმნა ახალი კორპუსები:

- აღმოსავლეთ სომხური ნაციონალური კორპუსი (http://eanc.net);
- იდიშის კორპუსი, რომელიც შეიქმნა რეგენსბურგის უნივერსიტეტთან ერთად;
- ბელორუსული კორპუსი;
- რუსეთის ხალხთა ენების კორპუსები: ბაშკირული, ოსური, ლეზგიური, აღულური, კალმიკური, ბურიატული, ვეპსური, ალაუტური, შორის ენა);
- ენობრივ უმცირესობათა გენეტიკური ან რეგიონალური ნიშნის მიხედვით გაერთიანებული კორპუსები: უმცირესობათა თურქული, ვოლგის რეგიონის დიალექტური ტექსტები (თურქული და ფინურ-უნგრული), ციმბირის უმცირესობათა ენები (სამოედური, კეტური, ევენკური) დაღესტნის უმცირესობათა ენები, ახლო აღმოსავლეთის უმცირესობათა ენები (ეს უკანასკნელი კატეგორია შეიცავს საველე სამუშაოების შედეგად მოპოვებულ მასალებს, რომლის ორგანიზატორიც იყო მოსკოვის სახელმწიფო უნივერსიტეტის თეორიული და გამოყენებითი ლინგვისტიკის დეპარტამენტი).

ჩამოთვლილი კორპუსები სხვადასხვა პლატფორმაზეა განთავსებული, თუმცა ყველა მათაგანი ეყრდნობა რუსული ნაციონალური კორპუსის მიერ დანერგილ ტექნოლოგიებსა და გამოცდილებას.

პარადოქსულად ჟღერს, მაგრამ სომხური კორპუსი გარკვეული თვალსაზრისთ რუსულ ნაციონალურ კორპუსზე უკეთესიც კია, რადგან მისი აგებისას გათვალისწინებულ იქნა ის შეცდომები, რომლებიც რუსულ ნაციონალურ კორპუსზე მუშაობისას იქნა დაშვებული.

ამგვარად, რუსული ნაციონალური კორპუსი ბოლო ხუთი წლის განმავლობაში თანამშრომლობის სხვადასხვა ფორმას ავითარებს „მეზობელი“ ენების კორპუსების შესაქმნელად.

The Russian National Corpus and Linguistic Corpora in the Former USSR Countries: Joint Developments in Corpora Technologies.

Dmitri Sitchinava
Institute of the Russian Language (Russia)
mitrius@gmail.com

The Russian National Corpus (the RNC: http://ruscorpora.ru/), inaugurated in 2004, is involved into interaction with many other corpora projects. As far as the languages of the former USSR is concerned, the possible directions of the interaction may be different:

- building of parallel (translation) bilingual (polylingual) corpora;
- collecting corpora of the Russian language in these countries (subcorpora of the RNC)
- participation in the development of common standards and technologies for non-Russian monolingual corpora, sharing experience and technologies.

The parallel corpora built with the RNC participation follow the pattern of the parallel Slavic corpora that develop in recent years in Europe (eg ParaSOL, ASPAC, Czech National Corpus and other parallel projects). We work on these projects together with our colleagues in respective countries (or federal regions of Russia). Currently Ukrainian-Russian/Russian-Ukrainian and Russian-Belorussian/Belorussian-Russian translation corpora are developed (http://ruscorpora.ru/search-para.html), and a Russian-Bashkir project will follow suit. This direction can be interesting for the most part of the Post-Soviet/Eastern Bloc domain: a vast amount of translations dates from the Soviet era, and new translated texts continue to emerge within the new political and cultural situation. The studying of these translations and of their evolution can have not only linguistic, but also historic and cultural value – they exhibit an intense linguistic "colonization" and then a new "emancipation", so to say.

The Russian language in the post-Soviet domain (including some territories within Russia!) exhibits much more diversity as compared to the previous period of the totalitarian unification. New regional communicative varieties emerge, as, for example, the Russian language in Daghestan (where it functions as a *lingua franca* with a considerable grammatical and phonetic influence of the Nakh-Daghestanian languages) or in the Central Asia. In Ukraine and Belarus, various forms of mixed Russian-Ukrainian/Belarussian language are gaining ground and beginning to be studied. All these varieties are to be collected and put into corpora. This material is in line with the other corpora of Russian language outside Russia (in the United States or Finland), that are also developed within the RNC.

The last but not the least – the RNC is active in collaboration with the purposes of building monolingual non-Russian and non-Slavic corpora. The new corpora developed with its participation include:

the East Armenian National Corpus (http://eanc.net)
the Yiddish Corpus developed together with the Regensburg University;
the Belorussian Corpus;
corpora of the languages of Russia: Bashkir, Ossetic, Lezgian, Aghul, Kalmyk, Buryat, Veps, Alutor, Shor;
joint corpora of minor languages of Russia by genetic and/or areal grouping: minor Turkic languages, dialect texts of the Volga region (Turkic and Finno-Ugrian), minor Siberian languages (Samoyed, Ket, Evenki), minor Daghestanian languages, minor Far East languages. This last category includes the material of the fieldwork organized by the Department of Theoretical and Applied Linguistics of the Moscow State University.
These corpora do not share a fully common platform, however they use the technologies and share the experience of the RNC team. Paradoxically, the Armenian Corpus is in some sense "better" than the RNC, as it was built avoiding some mistakes of the first stages of the RNC development. Thus, the RNC in the latest few year develops into different directions of collaboration with the corpora of the linguistic "neighbors" of the Russian language.

დიალექტური კორპუსული გამოცდილება და ქართული დიალექტური კორპუსი

ნარგიზა სურმავა, მარინა ბერიძე
არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველო)
marine.beridze@gmail.com, nargizasurmava@yahoo.com

შესავალი

ზეპირმეტყველებისა და დიალექტური კორპუსების შექმნა თანამედროვე ლინგვისტიკის ნოვატორული მიმდინარეობაა. სამეტყველო ტექსტურ მასივებზე ორიენტირებული კვლევა განვითარების ახალ პერსპექტივებს სახავს როგორც თეორიული, ისე პრაქტიკული ლინგვისტიკის ყველა სფეროში, მათ შორის დიალექტოლოგიაში.

დიალექტური ტექსტური კორპუსების რესურსებზე დაყრდნობილი კვლევა ორივე ენობრივი სტრატის – სალიტერატურო ენისა და დიალექტური მეტყველების – სიმეტრიულად, თანაბარ დონეზე შესწავლის ახალ, უპრეცედენტო შესაძლებლობებს ქმნის. ამასთან, დიალექტური კორპუსების მეშვეობით შესაძლებელი ხდება დიალექტური მეტყველების ფუნქციური, კომუნიკაციური და კოგნიტიური, ლინგვოკულტურული და სხვ. ასპექტებით კვლევა.

დიალექტური მასალის ახალი ტექნოლოგიით კვლევის იდეა თან ახლდა კომპიუტერის დამკვიდრებას ლინგვისტური კვლევის სფეროში, ხოლო კორპუსის ლინგვისტიკის განვითარების კვალდაკვალ ვითარდება "კორპუსის დიალექტოლოგიაც".

დღეისათვის დიალექტური მასალის კორპუსული დოკუმენტირებისა და კვლევის საკმაო გამოცდილება არსებობს. დიალექტური კორპუსების ნაწილი (ბრიტანული კორპუსი, რუსული ენის ტიუბინგენის კორპუსი...) სალიტერატურო ენის მასალას ეყრდნობა (სალიტერატურო ტექსტებში დამოწმებულ დიალექტიზმებს), თუმცა არსებობს მრავალი დიალექტური კორპუსი, რომელიც სპეციალურად მოპოვებულ დიალექტურ მასალას აერთიანებს.

დიალექტური ენობრივი მასალის არამატერიალური კულტურის ერთ-ერთ უმნიშვნელოვანეს კომპონენტად მოაზრება ახალ კორპუსულ გამოწვევებს სვამს თანამედროვე ლინგვისტთა წინაშე. სათანადო ადგილს იმკვიდრებს ეს თემა ენობრივი პოლიტიკის ახალ კონცეფციებშიც.

დიალექტური კორპუსების ზოგადი დახასიათება

არსებული დიალექტოლოგიური კორპუსები განსხვავდება მიზანდასახულობით – სტრატეგიით. მორფოლოგიურ, მორფო-სინტაქსურ და ფონეტიკურ სტრატეგიაზე ორიენტირებული კორპუსები იქმნება შესაბამის სფეროებში საკვლევი მასალების ეფექტურად მოძიებისათვის. სტრატეგია განსაზღვრავს დიალექტური მასალის ტრანსკრიფციისა და ლემატიზაციის მეთაური ფორმის არჩევანს.

დიალექტური კორპუსები ხშირ შემთხვევაში ორიენტირებულია ტრადიციულ საკვლევ პრობლემატიკაზე (დიალექტთა სტრუქტურული დახასიათებები, შეპირისპირებითი ანალიზი კილოებისა სალიტერატურო ენასა და სხვა ქვესისტემებთან და სხვ.). მაგალითად, ამ მიზანს ემსახურება რუსული ნაციონალური კორპუსის (НКРЯ) დიალექტური ქვეკორპუსი.

მასში ფონეტიკური ტრანსკრიფციით წარმოდგენილია მხოლოდ დიალექტური სიტყვაფორმები (სალიტერატურო ენასთან საერთო სიტყვაფორმებისთვის გამოყენებულია ჩვეულებრივი ანბანური ჩაწერა). ამგვარად დიალექტიზმები დიფერენცირებულია საერთო ენობრივი მასალისაგან, ხოლო ლემატიზაცია (სიტყვაფორმების ერთ ლემასთან გაერთიანება) წარმოებს ზოგ შემთხვევაში სალიტერატურო, ხოლო ზოგ შემთხვევაში დიალექტური ლექსემის მიხედვით.

თუ НКРЯ-ს დიალექტური ქვეკორპუსის მთავარი ამოცანაა დიალექტი წარმოადგინოს როგორც საერთო-სახალხო ენის სპეციფიკური ტერიტორიული ნაირსახეობა, სულ სხვა მიზანს ისახავს სარატოვის მულტიმედიური დიალექტოლოგიური ტექსტური კორპუსი СДК, რომელშიც დიალექტი წარმოდგენილია როგორც კონკრეტული სამეტყველო კოლექტივის საურთიერთო კულტურულ-კომუნიკაციური მოდელი, თვითკმარი კულტურულ-კომუნიკაციური წარმონაქმნი.

ქართული დიალექტური კორპუსი (ქდკ)

ქართული დიალექტური კორპუსი (ქდკ) ეყრდნობა მასზე ადრე შექმნილი კორპუსების გამოცდილებას, თუმცა კორპუსის სტრუქტურა, მეტატექსტური და მორფოლოგიური ანოტირების პრინციპები იქმნება ქართული ენობრივი სინამდვილის გათვალისწინებით.

ჩვენს კორპუსში თავიდანვე გამოვრიცხეთ დიფერენცირებული მიდგომა დიალექტური და სალიტერატურო ფორმების მიმართ და დიალექტურ მასალაში ორივე ამ მონაცემს ერთ სისტემის (დიალექტურის) კუთვნილებად განვიხილავთ (შდრ: НКРЯ).

ჩვენს კორპუსს განსაკუთრებით ფასეულს ხდის ინფორმაციული სისრულე – მასში აისახება რამდენადმე ღირებული ყველა დიალექტური ტექსტი და სიტყვაფორმა. ხელით ჩაწერილ ადრინდელ ტექსტებთან ერთად ქდკ-ში განთავსდება დიდი რაოდენობით აუდიო და ვიდეო ტრანსკრიპტები (სამომავლოდ გათვალისწინებულია ტრანსკრიპტებისა და აუდიო-ვიდეო ფაილების პარალელურად წარმოდგენის პროგრამული უზრუნველყოფა).

ქდკ იქმნება პირველ რიგში სხვადასხვა მიმართულების ლინგვისტური კვლევებისათვის, მაგრამ ინფორმაციული მომცველობითობა და მრავალმხრივი მეტაანოტირების სისტემა, ორთოგრაფიულთან მიახლოებული გრაფემული ჩაწერა, კოდირებული ენისათვის დამახასიათებელი პუნქტუაციის სისტემა მას გამოსაყენებლად მოხერხებულს ხდის ეთნოლოგიური და ფოლკლორული კვლევებისათვის.

Dialectological Corpus Experience and the Georgian Dialect Corpus

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

Introduction

The creation of dialectal and oral corpora is considered to be an innovative method in contemporary linguistics. Such research, oriented on oral texts, reveals new prospects of development in all of the spheres of theoretical as well as practical linguistics, including dialectology.

The research based on the material of dialectal text corpus creates unprecedented opportunities for studying both strata of the language - literary language and dialectal speech - on equal levels. As well as this, while studying dialectal corpora, it also becomes possible to explore functional, communicative, cognitive and linguocultural aspects of dialectal speech.

The idea of applying new technologies to the research of dialectological data has become possible only after introducing computer technologies into the domain of linguistic research. "Corpus-based Dialectology" has been developing together with Corpus Linguistics. Nowadays there is a good experience of research and collection of such data. Part of the dialectal corpora (British corpus, Russian corpora in Tubingen, etc.) are based on the data of literary languages (and on dialectal words attested in literary texts). However, there exist quite a few dialectal corpora containing specially collected dialectal data.

The dialectological linguistic material is classed as one of the significant components of non-materialistic culture. This puts new challenges to contemporary linguists. Consequently, this issue is increasingly considered in new theories of language policy.

General Characteristics of Dialectological Corpora

The Dialectal Corpora differ in aims and strategies. The corpora oriented on morphological, morpho-syntactic and phonetic strategies are created for effective collection of the research data in corresponding domains. The strategy defines the choice of the major method of transcription and lemmatisation of the data.
In many cases the dialectal corpora are oriented on traditional research issues (such as, structural characteristics of the dialects, comparative analysis of grammatical moods and literary languages and other subsystems, etc). For instance, the dialectal sub-corpus of the Russian National Corpus (RNC) is designed to serve this purpose. In this corpus the phonetic transcription is used to present only dialectal word forms whereas alphabetical notation is used for the word forms which are also testified in literary language. Thus, the dialectal forms are differentiated from the general language material whereas lemmatisation (which means grouping the word forms with one Lemma) is carried out either according to the literary or dialectal lexeme.
The major goal of the dialectal sub-corpus of the RNC is to present a dialect as a territorial variety of a national language whereas Saratov multimedia dialectal textual Corpus (SDC) considers a dialect as a cultural-communicative model of the speech community, a self-sufficient cultural-communicative entity.

Georgian Dialectological Corpus (GDC)

Georgian Dialectological Corpus (GDC) is based on the previous experience of the creation of the corpora, though the structure of the corpus as well as the principles of metatextual and

morphological annotations is created after taking consideration of the Georgian reality.

The differential approach to the dialectal and literary forms was excluded from the very beginning and thus both of these data are classed as one system in our data (cf. RNC).
The completeness of information makes the corpus created by us especially valuable as it contains all of the valuable types of dialectal texts and word forms. Together with the previous texts written by hand, GDC will contain a great number of audio and video transcripts (the programe for presenting transcripts and audio-visual files simultaneously is being elaborated).

GDC is being created for linguistic research of various types. However, the information capacity and a multifaceted system of meta-annotation, grapheme transcription which is close to orthography, punctuation system typical of coded language makes it valuable for the ethnological and folkloristic research as well.

ბიბლიოგრაფია/References

Wood G., Dialectology by computer, International Conference on Computational Linguistics, 1969.
Крючкова О., Гольдин В.Е., Текстовый диалектологический корпус как модель традиционной сельской коммуникации, Труды международной конференции по компьютерной лингвистике Диалог – 2008.
Крючкова О., Гольдин В.Е., Корпус русской диалектной речи: концепция и параметры оценки, Труды международной конференции по компьютерной лингвистике Диалог _ 2011.
Летучий А., Корпус диалектных текстов: задачи и проблемы// Национальный корпус русского языка: 2003-2005. М. Индрик, 2005.
Szmrecsanyi B., Hernandez N., Manual of Information to accompany the *Freiburg Corpus of English Dialects Sampler* ("FRED-S"). English Department University of Freiburg, 2007.

სახელთა ფუძეების კუმშვა-უკუმშველობისათვის ქართულში (არსებით სახელთა მოდელირების საკითხისათვის) უკუმშველი ფუძეები

თედო უთურგაიძე, მარიამ მანჯგალაძე
არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველო)
mariam@ice.ge

არსებით სახელთა მორფონოლოგიური მოდელის აგებისას იკვეთება რიგი საკითხებისა, რომელთა წესებად ჩამოყალიბების შემთხვევაში ქართული ენის მოდელირების პროცესი უფრო ზუსტი და ტექნიკური აღწერის თვალსაზრისით მოხერხებული ხდება.

მაგალითად, არსებით სახელთა ფუძეები შემადგენელთა მიხედვით არ იცვლება მხოლობითსა და მრავლობით რიცხვში, რადგან **-ებ** და **-თ** სუფიქსები ფუძეს არ განეკუთვნებიან. ფუძეთა ვარიანტები იქმნება მხოლოდ კუმშვის დროს.

1. არსებით სახელთა ფუძე **არ იკუმშება**, თუ ფუძე ბოლოვდება ორი თანხმოვნით, მაგ.:

ფერდი > ფერდის
შუშაბანდი > შუშაბანდის
დუქარდი > დუქარდის...

2. არსებით სახელთა ფუძე **არ იკუმშება**, თუ კუმშვის შედეგად თავს მოიყრის ერთი კლასის სამი თანხმოვანი (**-C_1 C_1 C_1 , -C_2 C_2 C_2, -C_3 C_3 C_3, -C_4 C_4 C_4)** მაგ.:

ჟურნალი > ჟურნალის
მკურნალი > მკურნალის
დასაბამბავი > დასაბამბავის
საწარბავი > საწარბავის...

გამონაკლისი: **ამბავ > *ამბვის > ამბის (ამ-ბის)**, ამ მაგალითში ხდება ვ- ფონემის დაკარგვა, მაგრამ ეს ფონეტიკური ცვლილება ადასტურებს ძირითად წესს.

3. თუ არსებითი სახელის ფუძეში თავს მოიყრის სამი თანხმოვანი, რომლის შემადგენლობაში სონორს მოსდევს ერთი და იმავე კლასის ორი თანხმოვანი, ფუძე არ შეიკუმშება, მაგ.:

მანდატი > მანდატის
კონკორდატი > კონკორდატის
რეზულტატი > რეზულტატის
ტრანსპლანტატი > ტრანსპლანტატის...

ჩვენ მიერ წარმოდგენილი მიმართებები არსებითია ქართულ ენაში დამარცვლის თეორიისათვის და მნიშვნელოვანია არსებით სახელთა მოდელირების თვალსაზრისით.

Towards the peculiarities of the Reduction and Non-reduction of Stems in Georgian (Towards the Issue of Modelling of Noun Stems) Non-reducing stems

Tedo Uturgaidze, Mariam Manjgaladze
Arn. Chikobava Institute of Linguistics (Georgia)
mariam@ice.ge

In the process of building a morphonological model of nouns, a number of issues emerge which, if formulated as rules, will make the process of modelling of the Georgian language more accurate and technically convenient to describe.
For example, the stems of nouns do not change in singular and plural according to their components as *–eb* and *–t* suffixes do not belong to a stem. The variants of the stems are formed only during the process of reduction.

1. If the noun stems end in two consonants they are not reduced. For example:
   dukardi – dukardis
   perdi – perdis
   shushabandi – shushabandis.
2. If three consonants of the same class are clustered together in a noun as a result of the reduction, the stems are not reduced (**-C_1 C_1 C_1 , -C_2 C_2 C_2, -C_3 C_3 C_3, -C_4 C_4 C_4)** . For example:
   žurnali – žurnalis
   mḳurnali – mḳurnalis.
   dasabambavi – dasabambavis
   saṣarbavi – saṣarbavis...

Exceptions: **ambav > *ambvis > ambis (am-bis).** In this example the syllable **v** is lost but this phonetic change follows the main rule.

3. If three consonants are gathered in a stem of a noun in which a sonorous consonant is followed by two consonants of the same class, the stem will not be reduced. For example:
   mandaṭi > mandaṭis
   ḳonḳordaṭi > ḳonḳordaṭis
   rezulṭaṭi > rezulṭaṭis
   ṭransplanṭaṭi > ṭransplanṭaṭis...

These relations are essential for the theory of the syllabification and modelling of nouns.

ტერმინი „რეპრეზენტაციული“ ქართულ კორპუსულ ლინგვისტიკაში

ნათია ფუტკარაძე
არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველო)
putkaradzenatia@gmail.com

”რეპრეზენტაციულობა[1]" (representativness) მასალის შერჩევის თვისებაა. ტექსტები „რეპრეზენტაციული“ შერჩევის მეთოდით განთავსდება კორპუსში, რათა კონკრეტული ენა რაც შეიძლება სრულფასოვნად იქნეს წარმოდგენილი. შედეგად ვიღებთ „რეპრეზენტაციულ“ ლინგვისტურ კორპუსს, რომელიც კონკრეტული ენის მიკრო-მოდელია ენის ზოგადი სტრუქტურის დარღვევისა და სისტემატურ შეცდომათა დაშვების გარეშე[2].

ტერმინი ”რეპრეზენტაციული” (representative) სოციალური მეცნიერებებიდან (ფსიქოლოგია, სოციოლოგია, სტატისტიკა...) გავრცელდა ემპირიული კვლევის ისეთ სფეროში, როგორიცაა კომპიუტერული ლინგვისტიკა და კორპუსის ლინგვისტიკა. სოციალური მეცნიერებების მკვლევართა მიერ ქართულენოვან თუ ქართულად ნათარგმნ ნაშრომებში პარალელურად გამოიყენება ტერმინები: ”რეპრეზენტატული” და ”რეპრეზენტაციული”[3]. ბოლო პერიოდში კი სოციოლოგები, ქართული ენის ნორმების გათვალისწინებით, შეთანხმდნენ ტერმინზე: **"რეპრეზენტაციული" (> „რეპრეზენტაცია“)**.

სიტყვა ”Representative” რაიმეს სიმბოლიზების, წარმოდგენის მნიშვნელობით ინგლისურმა ენამ XIV საუკუნის დასაწყისში ისესხა ფრანგულიდან (ძვ. ფრ.: ”**representatif**”), ფრანგულში კი დამკვიდრებულია შუა საუკუნეების ლათინურიდან: ”**repræsentativuus**”; მისი პირვანდელი ფორმაა ძვ. ლათინური: **repræsentare**, რომელშიც ზნური ძირია esse (“ყოფნა”), præ - პრეფიქსი ”წინ”, re- კი აწმყო დროის მიმღეობაზე დართული ინტენსივობის მეორეული პრეფიქსი.

ამგვარად, ლათინურიდან მომდინარე სიტყვა ფრანგულის გავლით დამკვიდრდა ინგლისურ ენაში, ბოლო პერიოდში კი ერთ-ერთ უმნიშვნელოვანეს ტერმინად იქცა ინგლისურენოვან სამეცნიერო ლიტერატურაში, კერძოდ, **კორპუსის ლინგვისტიკაში**. როგორც კორპუსის ლინგვისტიკის ტერმინი ”representative” ინგლისურიდან აქტიურად ვრცელდება სხვა ენებში (ყველგან, სადაც ხორციელდება სრულფასოვანი ლინგვისტური კორპუს(ებ)ის დამუშავება).

---

[1] აღნიშნული ტერმინი გამოიყენება, როდესაც საუბარია ენობრივი მთლიანობიდან (სტატისტიკაში მიღებული ტერმინით: ”პოპულაციიდან”/population) საანალიზო ნედლი მონაცემების შერჩევის საკითხებზე: ”**შერჩევა უნდა იყოს პოპულაციის რეპრეზენტაციული**” (მაგ., რეპრეზენტატულობის თვალსაზრისით გამოირჩევა ჰელსინკის ინგლისური ტექსტების კორპუსი, რომელიც შეიცავს ინგლისური ენის ყველა სახესხვაობაზე შესრულებულ სხვადასხვა ჟანრის ნაწარმოებებს და ისეთ სოციოლინგვისტურ ცვლადებს, როგორიცაა სქესი, ასაკი, განათლება და სოციალური სტატუსი).

[2] “რეპრეზენტატული შერჩევით” უნდა მივიღოთ საკვლევი ობიექტის ქვესტრუქტურების სრულფასოვანი ერთობლიობა (სოციალურ და პოლიტიკურ ტერმინთა ლექსიკონი-ცნობარი, 2004).

[3] იხ., მაგ., ჰ. კისი, სტატისტიკა სოციალურ მეცნიერებებში, 2007 /აკადემიური თარგმანი/. რ. გერიგი, ფ. ზიმბარდო, ფსიქოლოგია და ცხოვრება /აკადემიური თარგმანი/ 2009...

განსახილველი ტერმინის ინგლისურიდან სხვა ენებზე თარგმნისას ხშირად მიმართავენ კალკირების ხერხს, რაც "გამორიცხავს მთარგმნელის ნებსით თუ უნებლიე ინტერპრეტაციების საფრთხეს და ტერმინის ზუსტ თარგმანს უზრუნველყოფს"[1]. ამის კარგი მაგალითია რუსულ კორპუსულ კვლევებში ამ ტერმინის თარგმნის ცდაც: представительность, თუმცა პარალელურად ვხვდებით რუსულ მორფოლოგიურ ყალიბში მოქცეულ ინგლისურ ფუძესაც: Репрезентативность;

დღეს კორპუსის ლინგვისტიკის საერთაშორისო ტერმინოლოგია ძირითადად ინგლისური ენის გზით არის დამკვიდრებული და ეს ბუნებრივიცაა. ცნობილია, რომ უცხოენოვანი ტერმინების დამკვიდრებისას არსებითია, თუ სად ჩამოყალიბდა და რომელ ენაზე განვითარდა ესა თუ ის სამეცნიერო დარგი. კერძოდ, პირველი ლინგვისტური კორპუსები შეიქმნა ინგლისსა და აშშ-ში, შესაბამისად, ინგლისურ ენაზე ჩამოყალიბდა კორპუსის ლინგვისტიკის **(კლ)** მეტაენაც. როგორც აღვნიშნეთ, კორპუსული კვლევის ტექნოლოგიების გავრცელებასთან ერთად ინგლისურენოვან ტერმინთა სისტემაც შეიჭრა სხვადასხვა ენაში; ამგვარ გავლენას ვერც ქართული ასცდა.

ქართულ სამეცნიერო ნაშრომებში ცნება: "რეპრეზენტაცია" (represenation) გარკვეულ შემთხვევებში ნათარგმნია როგორც "**წარმოდგენა**": "ბუნებრივი ენების ელექტრონულ-ციფრული **წარმოდგენის** აუცილებლობის კონცეფცია.../ Concept of the need for digital representation of languages"[2].

ქართულ ენათმეცნიერულ კვლევებში ერთი და იმავე ავტორების პუბლიკაციებში გვერდიგვერდ შეიძლება შეგვხვდეს ამ ტერმინის ორი ვარიანტი: "რეპრეზენტატიული", "რეპრეზენტატული" (მაგ., "ტექსტის, ლექსიკონისა და სამეცნიერო ლიტერატურის მონაცემებით იქნება შესაძლებელი ნამდვილად **რეპრეზენტატიული** კორპუსის შედგენა"[3]; შდრ.," რეპრეზენტატულობა ნიშნავს, რომ ეროვნული (ნაციონალური) კორპუსი უნდა მოიცავდეს (შესაძლებლობის ფარგლებში), მოცემულ ენაში განვითარების გარკვეულ პერიოდში არსებულ წერილობითი და ზეპირი ტექსტების ყველა ტიპს (სხვადასხვა ჟანრის მხატვრულ, პუბლიცისტურ, სასწავლო, სამეცნიერო, საქმიან ურთიერთობათა ამსახველ, საუბრულ, დიალექტურ და ა.შ. ტექსტებს), შესაბამისი პერიოდის ენობრივ რეალობაში მათთვის განკუთვნილი წილობრივ–პროპორციული თანაფარდობის (ბალანსის) შეძლებისდაგვარად დაცვით"[4]).

მკვლევრებს, მარინე ბერიძესა და ნარგიზა სურმავას, "რეპრეზენტაციულობის" ცნების ანალიზისას კორპუსის ლინგვისტიკაში შემოაქვთ ქართული ფუძისგან ნაწარმოები შუალედური ტერმინი "**მომცველობითი**". წარმოდგენილი ქართულენოვანი ტერმინი რამდენიმე მიზეზის გამოა საინტერესო:

1. ქართველი მკითხველისთვის ტერმინ "მომცველობითის" მნიშვნელობა გასაგებია, ყოფითი ლექსემებისგან ფორმალურ-სემანტიკური თვალსაზრისით გამოირჩევა და ორაზროვნებას არ იწვევს;

2. ტერმინი, როგორც აღმწერი, სემანტიკურად გამჭვირვალეა;

3. ტერმინი ერთცნებიანია, ამდენად იოლი გამოსაყენებელია;

---

[1] მელიქიშვილი დ., ფილოლოგიური ძიებანი, 2009, გვ. 492.

[2] კაპანაძე ო., ენათმეცნიერების საკითხები, I-II, 2010, გვ. 257

[3] ბერიძე მ., "მეტყველების მეოთხე ფაქტურა" და საქართველოს ლინგვისტური პორტრეტი, წახნაგი, II, 2010, გვ. 110

[4] ბერიძე მ., სურმავა ნ., კორპუსის ლინგვისტიკის ძირითადი ცნებანი და მათთვის ქართული შესატყვისების შერჩევის ცდა, ბუნებრივ ენათა დამუშავება, VI, კონფერენციის მასალები, თბ. 2008.

4. რაც ყველაზე მთავარია, ტერმინი ქართული ძირისა და აფიქსებისაგან შედგება.

ჩვენი აზრით, კარგი იქნება, თუ დამკვიდრება აღნიშნული ტერმინი; სამწუხაროა, რომ ამავე ავტორთა ნაშრომებში ტერმინი "მომცველობითი", გამოყენების თვალსაზრისით, ვერ ცვლის „რეპრეზენტაციულს".

კორპუსის ლინგვისტიკის საანალიზო ტერმინთა საკმაოდ ვრცელი მასალა მოვიძიეთ; სამწუხაროდ, ამ დარგის ქართული კანონიკური ტერმინოლოგია და ძირითადი აბრევიატურები ჯერ მხოლოდ ჩანასახში არსებობს; ეს შეიძლება იმით იყოს გამომწვეული, რომ **კლ** ახალგაზრდა სამეცნიერო დარგია (თუმცა, მიუხედავად ამისა, იგი ძალიან სწრაფად ვითარდება სხვადასხვა ქვეყანაში). კორპუსის ლინგვისტიკის განვითარება სასიცოცხლო მნიშვნელობისაა ყველა თანამედროვე ენისთვის, შესაბამისად, საშურია ამ დარგის ქართულენოვანი ტერმინოლოგიის დამკვიდრება. აქედან გამომდინარე, აუცილებელია:

- ტერმინთა წინასწარი ლოგიკურ-ცნებითი ანალიზი;
- ტერმინ-აღმწერებისგან შემდგარ ცნებათა იერარქიული მოდელის შემუშავება;
- ქართულენოვან საანალიზო ერთეულთა - სავარაუდო ტერმინთა - მოძიება სპეციალური ლიტერატურიდან (უპირველეს ყოვლისა, თემატური კონფერენციების მასალებიდან)[1];
- ტერმინოლოგებთან ერთად მოსახერხებელ ტერმინზე შეთანხმება და სამეცნიერო ნაშრომებში დამკვიდრება.

The Term "Representative" in Georgian Corpus Linguistics

Natia Putkaradze
Arn. Chikobava Institute of Linguistics (Georgia)
natia.putkaradze@ice.ge

"Representativeness" is the property of the data selection. The method used for the selection of texts is the method of representative selection. The process aims at presenting a particular language representatively as completely as possible. The representativeness transforms the digital collection of texts into linguistic corpus. As a result , the "representative" linguistic corpus is built, which is a micro model of the particular natural language, without destroying the main structure and tendencies of it.

The term "Representative" was first introduced into the Behavioural Sciences (Psychology, Sociology, Statistics, etc.) and later, into the empirical sciences such as Computational Linguistics and Corpus Linguistics. The term "representative" has two loan translation versions in Georgian: "reprezenṭaciuli" *("reprezentaciuli")* and "reprezenṭaṭuli" ("reprezentatuli"). According to the textbooks and works in Behavioural Sciences (e.g. The Georgian academic translation of the textbook:

[1] В. П. Захаров, Тезаурус по корпусной лингвистике, წყარო: http://textualheritage.org/content/view/361/168/lang,english/

Statistical Concepts for the Behavioural Sciences, 3rd Edition, by Kiess, Harold, published by Pearson Education, etc.) the term “reprezenţaciuli” (“reprezentaciuli”) is preferred.

The word “Representative” itself came into English from French via Latin; although the term in the meaning of "serving to represent" dates from the late 14c. from O.Fr. representatif (early 14c.), from M.L. repræsentativus, from L. repræsentare, Meaning "standing for others" is known from 1620s; in the political sense of "holding the place of the people in the government, having citizens represented by chosen persons" is first recorded in the 1620s. As a noun, it was first recorded in the 1640s; however, first used in the 1690s in the meaning of "member of a legislative body" (Douglas Harper online etymology dictionary).

The developers of the linguistic corpora (Leech 1992, Bibber 1993, Sinclair 1996) insisted on the importance of representative selection for Corpus Linguistics, thus focus has been shifted and the term is now generally used as one of the most important terms in Corpus Linguistics, first of all in Great Britain and the United States. The field was born and developed in the above mentioned countries which presuppose that the corpus terminology is spreading from English to other languages. Generally, International terms of the corpus linguistics in other languages were coined according to the English ones. That is why the meta-language of the Corpus Linguistics (CL) is based on English. The Development of the CL technologies initiated the spread of English/international terminology in other languages and Georgian is not an exception.

Specifically, the term “representative” is borrowed from English by other languages using the method of literal translation; “In the case of calquing involuntary interpretations are avoided” (D. Melikishvili: 2009, 492). Russian term: представительность is a good example of calquing, though the “hybrid” version of this word is commonly used to refer to this concept at the same time - “Репрезентативность”: English word is used as a root of the verb and is formed according to the Russian morphological rules.

In Georgian linguistics, according to some instances, the sense of “Representation" is translated as ”ςarmodgena” [“present” (n.)]; e.g.: "bunebrivi enebis elekţronul-cipruli **ςarmodgenis** aucileblobis ķoncepcia..." Concept of the need for digital representation of Natural languages…”O. Kapanadze, Issues of Linguistics, 2010, p. 257); in other cases, usually, the Georgian “hybrid” terms ”რეპრეზენტაციული” (“**reprezenţaţiuli**”) and ”რეპრეზენტაცული” (**“reprezenţaţuli**”) are used by one and the same authors; The Scholars M. Beridze and N. Surmava attested the Georgian word ”მომცველობითi” [“momcvelobiti” ("including")] as a corresponding term (2007). We are of the opinion, that the presented Georgian term is feasible for several reasons:

1. The term is clear for Georgian speakers and differs from everyday words by its formal and semantic structure.
2. The term, as a descriptor is semantically motivated.
3. The term is mono-semantic, that is why it is easy to decode its meaning.

4. The most important thing is that the term is designed using the Georgian root and affixes.

However, there is no consensus about what word/words could replace the hybrid term “reprezentatuli“ in Georgian. Unfortunately, the term "მომცველობითი" has not been used to replace it even in the works of the above-mentioned scholars.

We have significantly large data of the Corpus terminology which we have collected during our weekly seminars in Corpus linguistics; unfortunately, the canonical terminology and the basic abbreviations of this field do not exist yet. Probably the reason is that the CL is the young field (though developing effectively). To set the Georgian CL terminology depends on further developments in the field Corpus Linguistics and vice *versa*; hence it follows that it is important to:

- analyse and conduct research on underlying concepts of the terms;
- design the hierarchical sample of the term-descriptors;
- collect the Georgian equivalents from the CL works (such as CL conference materials and articles);
- provide equivalents or coin appropriate terms with the terminologists.

ქართული ენის ლოგიკური გრამატიკა და ქართულენოვანი კომპიუტერი

კონსტანტინე ფხაკაძე, ლაშა აბზიანიძე, ალექსანდრე მასხარაშვილი, ნიკოლოზ ფხაკაძე, მერაბ ჩიქვინიძე
საქართველოს ტექნიკური უნივერსიტეტი (საქართველო)
gllc.ge@gmail.com

2010 წელს საქართველოს ტექნიკურ უნივერსიტეტში დაფუძნდა ქართული ენის ტექნოლოგიზების სასწავლო-სამეცნიერო ცენტრი. ცენტრი ამუშავებს გრძელვადიან პროექტს „ქართული ენის ლოგიკური გრამატიკა და ქართულენოვანი კომპიუტერი“. პროექტი ჩამოყალიბდა სახელმწიფო-მიზნობრივი პროგრამის – „კომპიუტერის სრული პროგრამულ-მომსახურეობითი მოქცევა ბუნებრივ ქართულ ენობრივ გარემოში“ – კვლევითი ამოცანების შემდგომი განვითარების საფუძველზე და მიზნად ისახავს ქართული ენის ბუნების სრულად ამსახველი მათემატიკური თეორიის, ანუ ქართული ენის ლოგიკური გრამატიკის შემუშავებას და ქართულენოვანი, ანუ ინტელექტუალური უნარებით აღჭურვილი ქართული კომპიუტერული სისტემის აგებას.

ქართული ინტელექტუალური კომპიუტერული სისტემის აგების ერთადერთი გზაა ქართულ ენაში არსებული ინტელექტუალური გამოთვლების მოდელირება, რაც ქართული ენის სრული მათემატიკური თეორიის შემუშავებას გულისხმობს. აქ ”ქართული ენაში” მოიაზრება ქართული ბუნებრივი ენობრივი სისტემა, რომლის შემადგენლებიცაა ქართული სამწერლობო, სამეტყველო და ”სააზროვნო” ენები. ამგვარად, მოხსენებაში განვიხილავთ ქართული ენის მათემატიკური თეორიის, ანუ ქართული ენის ლოგიკური გრამატიკის ფუნდამენტურ საკითხებს. ესენია:

1. ლოგიკური ბრუნებისა და ლინგვისტური პრედიკატის ცნება, რომელთა საფუძველზე შემუშავებულია ქართული ენის არსებითი და ზედსართავი სახელებისა და ზმნების მორფოლოგიური, სინტაქსურ, ლოგიკური და სემანტიკური ბუნების ახალი მათემატიკური ხედვა. ასევე განხილულ იქნება ქართულ და ინდოევროპულ ენობრივ სისტემებს შორის არსებული პრინციპული განსხვავებები;

2. ქართული ენის I საფეხურის მათემატიკური თეორია, რომელიც სრულად აღწერს ქართული ენის ძირეულ ნაწილს (ქართული ენის თხრობითი წინადადების დონეს). კერძოდ, წარმოდგენილი იქნება ქართული ენის I საფეხურის მათემატიკური თეორიის ფორმალური, სემანტიკური და ლოგიკური აქსიომატიკა;

3. ”ქართული გამაფართოებელი წესები”, რომელთა საფუძველზეც ქართული ენის ძირეული ნაწილის სიტყვებს განვიხილავთ, როგორც ” შ. ფხაკაძისეულ შემამოკლებელ სიმბოლოებს”;

4. ქართული ენის ძირეული ნაწილის წინადადებების ქართული ენის I საფეხურის მათემატიკურ თეორიაზე დამყვანი (მთარგმნელი) ფორმალური წესები.

ქართული ენის ლოგიკური გრამატიკის ფუნდამენტური საკითხების მიმოხილვის შემდეგ მოკლედ შევეხებით ამ კვლევების შედეგად უკვე შემუშავებულ ახალ ენობრივ ხედვას და ჩვენ მიერ ადრე წარმოდგენილ ”ქართული ენის თეზისებს”[1].

[1] ფხაკაძე კ., აბზიანიძე ლ., მასხარაშვილი ს.„ ქართული ენის თეზისები, ი.ვეკუას სახელობის გამოყენებითი მათემატიკის ინსტიტუტის სემინარის მოხსენებები, ტ.34, გვ.108–121, 2008;

Towards the Logical Grammar of the Georgian language and the Georgian Language Computer

Konstantine Pkhakadze, Lasha Abzianidze, Aleksandre Maskharashvili, Nikoloz Pkhakadze, Merab Chikvinidze
Georgian Technical University (Georgia)
gllc.ge@gmail.com

The Scientific-Educational Centre for Georgian Language Technology at the Georgian Technical University was established in 2010. The centre works on the long-term project "Logical Grammar of the Georgian Language and Georgian Intellectual Computer". The project was formed on the basis of the further development of the state-priority programme “Free and Complete Programming Inclusion of a Computer in the Georgian Natural Language System” and it aims to elaborate the Logical Grammar (LG) of the Georgian Language (GL), i.e. the Mathematical Theory (MT) which will fully describe the nature of the GL and, as well as that, to construct the Georgian Computer (GC), i.e. the basic, intelligent Georgian Intellectual Computer System (GICS).

The only possible way of constructing basic, intelligent GICS, i.e. the only way of creating the fullscale intelligent computational model for the GL (GL is decoded as Georgian Natural Language System (NLS), the constituents of which are Georgian Written Language (GWL), Georgian Spoken Language (GSL), Georgian Language of “thought’ (GTL)) is to establish this through the full-scale software modelling of the LG of the GL, which, as was mentioned above, is planned to be the MT, which will fully describe the nature of the GL. Thus, this presentation will discuss the grounding questions of the LG of the GL which are:

1. The concepts of logical declination and linguistic predicates, on the basis of which new

mathematical approaches related to the morphological, syntactic, logical and semantic nature of Georgian nouns, adjectives, verbs are elaborated. In addition to this, the principal differences between the Georgian and Indo-European pronouns, verbs and conjugation systems will be underlined in the paper;

2. The first stage MT of GL, which will fully describe the nature of the Core Part (CP) of the GL. More precisely, we will present formal, semantic and logical axiomatisations of the 1st stage MT of the GL.;
3. The Georgian Extensions Rules (GERs), with the help of which the words of the CP of the GL are understood as the contracting, i.e. abbreviating symbols;

4. The formal rules of reducing (translating) context-free sentences of the CP of the GL into their semantic equivalents in the 1st stage MT of the GL and vice versa.

After discussing the grounding qustions of the LG of the GL, we will briefly overview the innovative approach to the language elaborated following this research.

მრავალდარგოვან ტერმინთა ელექტრონული ლექსიკონი

ლია ქაროსანიძე, რატი სხირტლაძე
არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველი)
karosanidze@yahoo.com

ქართული ენა უცხოენოვანი ტერმინების მოძალებას ყოველთვის განიცდიდა. განსაკუთრებული სიმწვავით ეს საკითხი X-XII საუკუნეებში – ბერძნულთან, XIX-XX საუკუნეებში – რუსულთან, დღეისათვის კი ინგლისურთან მიმართებით გამოიკვეთა.

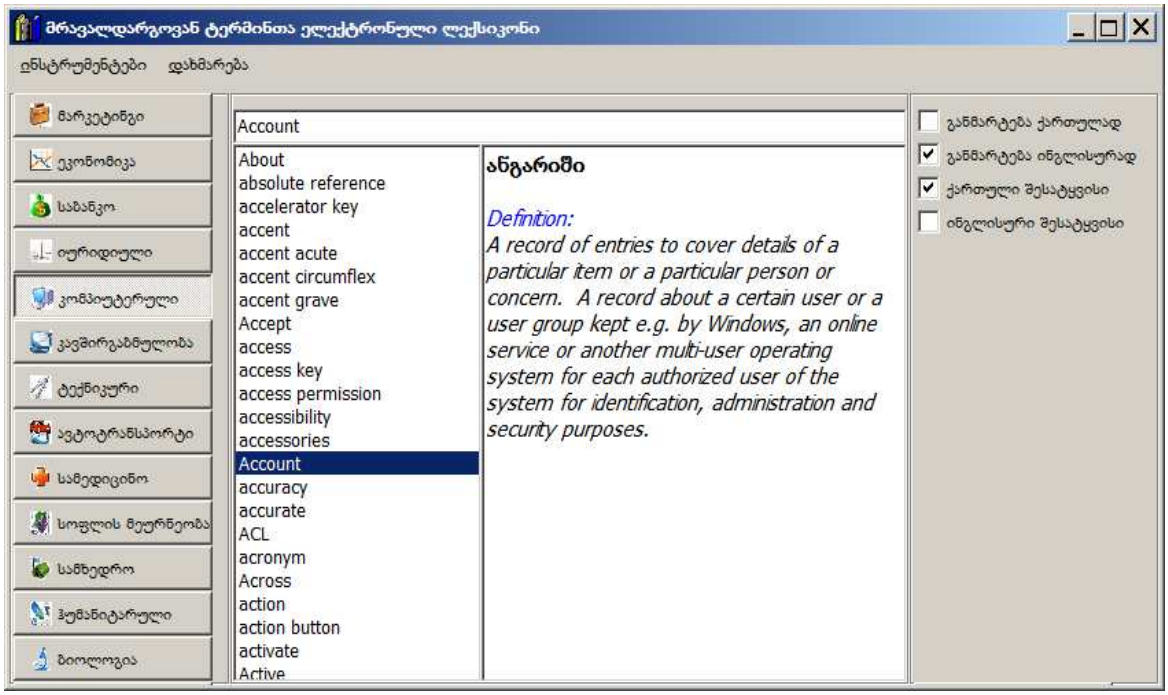
X-XII საუკუნეების ქართულ მწიგნობრებს უნდა ვუმადლოდეთ ისეთი ტერმინების დამკვიდრებას, როგორებიცაა: აგებულება, არსი, არსებითი, გონება, მდგომარეობა, მეცნიერება, მეტყველება, მოძღვრება, სივრცე, ხელოვნება და სხვ. ამ ტერმინთა შემქმნელთათვის უმთავრესი ბერძნული ტერმინების ქართული ენის ბუნებასთან მისადაგება იყო. ასეთივე წარმატებით გაართვეს თავი ქართველმა მოღვაწეებმა ქართულ ენაში რუსულიდან შემოსული ტერმინების გაქართულების პროცესს. დღეს კი ქართული მეტყველება აჭრელებულია ისეთი სიტყვებით, როგორებიცაა: დააშათდაუნე, დაარესტარტე, პორტფოლიო, ივენტი, ჩეტი, ნიკი, ვორნი, იუზერი, მოდერი, ფრეიმი, ოფერი და ა.შ. მიუხედავად იმისა, რომ დახლებზე ცალკეული პირებისა თუ ჯგუფების მიერ გამოცემულ სხვადასხვა დარგის ტერმინოლოგიურ ლექსიკონს წააწყდებით, საზოგადოებისთვის უცნობი რჩება, რომელი ფორმა უნდა მივიჩნიოთ ნორმად - მაუსი თუ თაგუნა, ჰარი თუ პრობელი ....

ნორმირებულ ტერმინთა ლექსიკონის მასაღებად, უპირველესად, საჭიროა მთელი რიგი სამეცნიერო საკითხის კვლევა. მე-12 საუკუნის ქართველ მწიგნობართა ტერმინოლოგიური მუშაობა დღემდე მისაბაძი უნდა იყოს ჩვენთვის. ერთ ასეთ მაგალითს მოვიყვანთ: ბერძნულში სახელწოდება ათენი მრავლობით რიცხვშია, რომლის შესახებ ეფრემ მცირე წერს: ჩემს ენას ეს ტრადიცია არა აქვსო და ვიცით, რომ დამკვიდრდა ათენი და არა ათენები (შდრ. რუსული ფორმა). ეს ერთი მაგალითიც საკმარისია იმის წარმოსადგენად, თუ როგორ არ ახვევდნენ თავს მშობლიურ ენას სხვა ენისთვის დამახასიათებელ კანონებს. ასეთი მიდგომა ტერმინოლოგიისადმი მხოლოდ ორივე ენის „რაბამობის“ ცოდნის შედეგადაა შესაძლებელი. მე-20 საუკუნეში ბევრი ისეთი ფორმა დამკვიდრდა, რომლებიც ქართული ენის ბუნებისთვის მიუღებელია: გაყიდვები, ტკივილები და სხვა მრავალი.

ლინგვისტ-ტერმინოლოგებს ამ მხრივ ძალიან ბევრი სამუშაო აქვთ. უპირველესი და გადაუდებელი საქმე კი ინგლისური და ქართული ენების შესაბამისობის შესწავლაა. მხოლოდ სამეცნიერო თეორიული დაკვირვებები შეიძლება გახდეს ქართული ენის სამეცნიერო ტერმინოლოგიის ელექტრონული ლექსიკონის შექმნის საფუძველი.

ელექტრონული ლექსიკონი, რომლის შექმნაზეც ვმუშაობთ, აღჭურვილი იქნება ტერმინთა ძიების მოქნილი სისტემით. ძიება შესაძლებელი იქნება როგორც ტერმინის ინგლისური და ქართული მეთაური სიტყვის (ან მისი ნაწილის), ისე განმარტებაში მონაწილე სიტყვების მიხედვით. მომხმარებელს საშუალება ექნება ტერმინები იპოვოს როგორც მისთვის საინტერესო დარგის, ისე მთელი ბაზის მიხედვით. თუ ტერმინს სხვადასხვა დარგში განსხვავებული მნიშვნელობა აქვს, ამის შესახებ მომხმარებელი სრულ ინფორმაციას მიიღებს. შეიქმნება ტერმინთა ელექტრონული ლექსიკონის როგორც ინტერნეტვერსია, ისე კომპიუტერული პროგრამა. მისი ჩამოტვირთვა ყველასთვის იქნება ხელმისაწვდომი. ასევე შეიქმნება პროგრამული ინსტრუმენტები, რომლებიც ლინგვისტ-ტერმინოლოგებს მისცემს

ეფექტური მუშაობის საშუალებას. კომპიუტერულ პროგრამას, სავარაუდოდ, შემდეგი სახე ექნება:

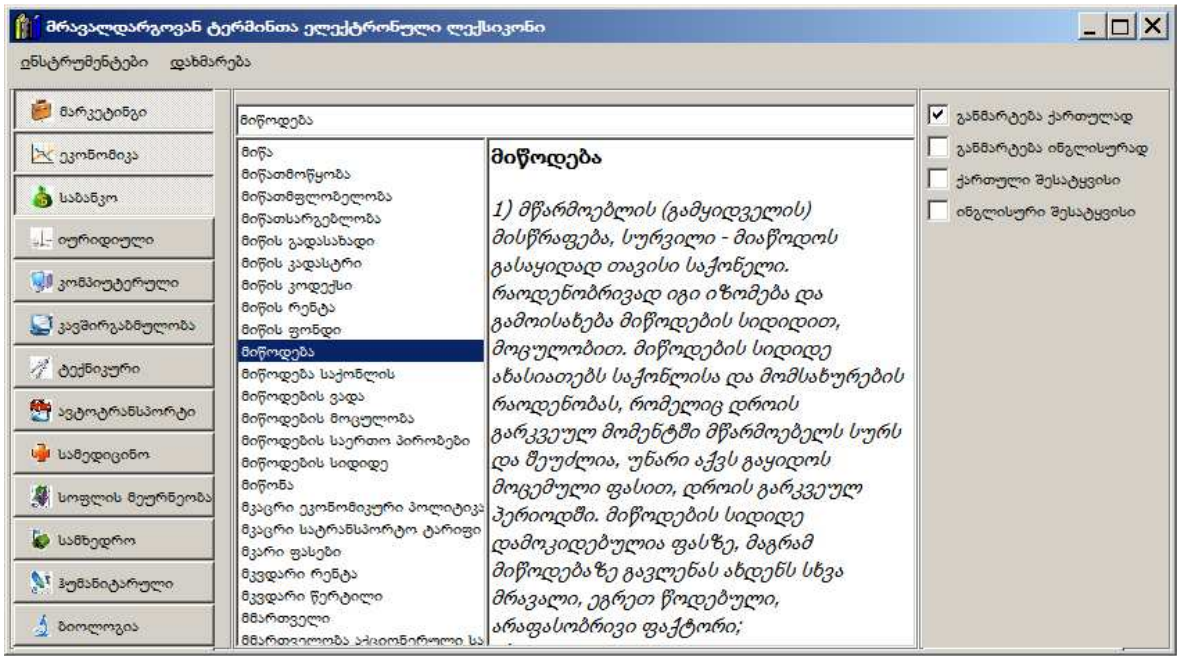


ნახ.1

მომხმარებელს საშუალება ექნება მონიშნოს მისთვის საინტერესო ერთი ან რამდენიმე ტერმინოლოგიური ლექსიკონი და მათ ფარგლებში ეძებოს სიტყვის მნიშვნელობა. ამასთანავე, შესაძლებელი იქნება როგორც სიტყვის ქართული ან ინგლისური შესატყვისის, ისე შესაბამისი განმარტების მოძიება.

1-ლ ნახატზე განხილულია შემთხვევა, როდესაც მომხმარებელი კომპიუტერულ ტერმინთა ლექსიკონში ეძებს სიტყვა “Account”-ის შესაბამის ქართულ შესატყვისსა და განმარტებას ინგლისურ ენაზე.

ძიება შესაძლებელი იქნება ასევე ქართული მეთაური სიტყვის მიხედვით. მე-2 ნახატზე მომხმარებელი ეძებს მარკეტინგის, ეკომომიკისა და საბანკო ტერმინების ლექსიკონების მიხედვით “მიწოდების“ განმარტებას ქართულად.



ნახ.2

პროგრამა შეიქმნება CodeGear™ Delphi 2009 პროგრამული გარსით. კომპიუტერული პროგრამის პროტოტიპი იქნება შესაბამისი ვებაპლიკაციაც. ვებაპლიკაციაში გამოყენებული იქნება MySQL მონაცემთა ბაზა და PHP პროგრამირების ენა. მრავალდარგოვან ტერმინთა ელექტრონული ლექსიკონის კომპიუტერული პროგრამის საინსტალაციო პაკეტი და ვებაპლიკაცია მოთავსდება მისამართზე www.ena.ge/terminology.

ნორმირებულ ტერმინთა ელექტრონული ლექსიკონი საჭირო და აუცილებელია ყველა დარგის სპეციალისტისთვის, მოსწავლისთვის, სტუდენტისთვის. დაბოლოს, უმთავრესი მნიშვნელობა ამ პროექტით შექმნილი ლექსიკონისა ისაა, რომ ქართულ ინტერნეტსივრცეში საგანგებო ვებგვერდზე მოთავსებული ტერმინოლოგიური ლექსიკონით მრავალი მომხმარებელი ისარგებლებს. იგი ხელს შეუწყობს ქართული ენის ფუნქციონირებას მთელ საქართველოში. ეს კი, გარკვეულწილად, შეცვლის თანამედროვე ქართულ ენაში დღეს დამკვიდრებულ ტერმინოლოგიურ სიჭრელეს, ქაოსს, აღადგენს კოორდინირებულ სატერმინოლოგიო მუშაობას საქართველოში.

Multi-field Electronic Terminological Dictionary

Lia Karosanidze, Rati Skhirtladze
Arn. Chikobava Institute of Linguistics (Georgia)
karosanidze@yahoo.com

The Georgian language has always been flooded with the foreign terms of different branches and fields of knowledge. This problem was of special acuity in the XI-XII centuries in relation to Greek terminology and in the XIX century in relation to Russian terminology. Today the same problem has formed in relation to the English language terminology. We owe to the XI-XII centuries Georgian scholars for establishing Georgian equivalents of many Greek terms, such as structure (agebuleba), essence (arsi), essential (arsebiti), speech (metkveleba), mind(goneba), space (sivrce), art (xelovneba), and so on. For creators of those terms it was most important to match basic Greek terms with the nature of the Georgian language.

Georgian public figures and scholars were also successful in managing to determine Georgian equivalents of the Russian terms. Today, the Georgian terminology is contaminated with foreign words, especially computer-related terminology, and even though the book-shops offer several terminological dictionaries, for the public it remains unknown which term to use instead such ones as: chat, shut down, user, restart, frame, offer, portfolio and others, and where to find Georgian correspondences.

Only standardised forms will constitute the foundation for the standardised (normative) terminological electronic base. The Internet version and the computer program of the electronic terminological dictionary will be created. It will be available to download the program. Other software tools will also be designed enabling the linguists interested in terminology to work effectively. The dictionary will be equipped with a flexible searching system. Searching can be oriented on the English and Georgian head word (or part of it) of the term as well as on the words from its definition. The user will be able to search the term according the branch of interest as well as all over the base. If the same term has different meanings in different branches, the user can get absolutely full information about it.

The primary task to receive the desired product – the electronic dictionary of standardised terms is studying and investigating several scientific issues. In the XII century, Georgian scholars carefully studied linguistic characteristics of both languages – Greek and Georgian when translating from Greek, and always would take into account the peculiarities of the native language. We can demonstrate one example: the name Athens is used in the plural in Greek, but one of the scholars of those days, Ephreme the Minor (Ephrem Mtsire) wrote: “for my language it is not natural” and thus, in Georgian we have **Ateni** for **Athens**. Each word-form, each term should be studied very carefully first in its language of origin and then in our native language. Beginning from the 90-ies

many such forms established in the Georgian language which are unacceptable for it, i.e. plural forms like **"purchases, pains"** ("gayidvebi, tkivilebi….") and so on.

Linguist-terminologists have quite a lot to do in this field. An especially urgent job is studying correspondences in English and Georgian languages. Only scientific-theoretical researches can give the foundation for creating a diversified electronic terminological dictionary.

The desktop application alongside the web application of the diversified electronic terminological dictionary will be created in the process of the project work.

Supposedly, the interface of the desktop application will be as it follows below:

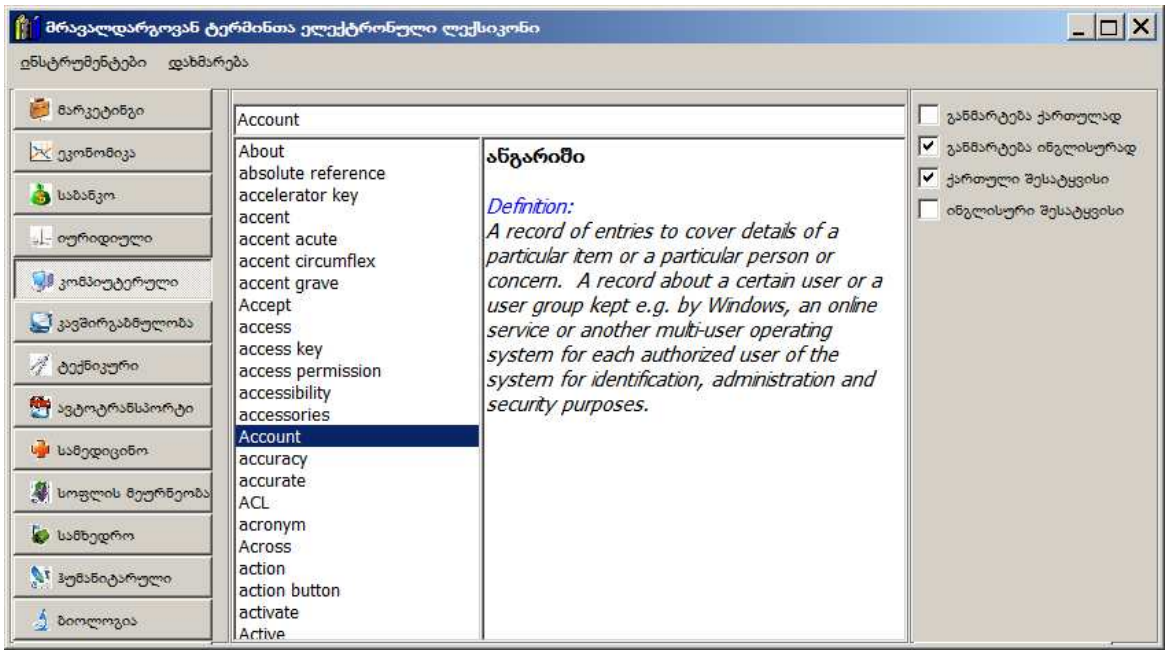


Fig.1

The user will be able to mark out one or more specialised terminological dictionaries and look up a meaning of the word in them. Besides that, the user can find Georgian as well as English correspondence of the given word and its definition.

Fig. 1 shows the example when a user looks up the word "Account" in the electronic dictionary, with its Georgian correspondence and English definition.

Searching will be possible also according the Georgian headword. Fig. 2 shows an example, when the user looks up the Georgian definition of the word "supply" in the specialised dictionaries of the

terms of marketing, banking, and economics.

It should be noted that if needed, in some terminological dictionaries, the English definition will be given, but not in all of them. The same can be said about the English correspondence - in certain cases it will be shown if needed, but sometimes the Georgian definition will suffice.

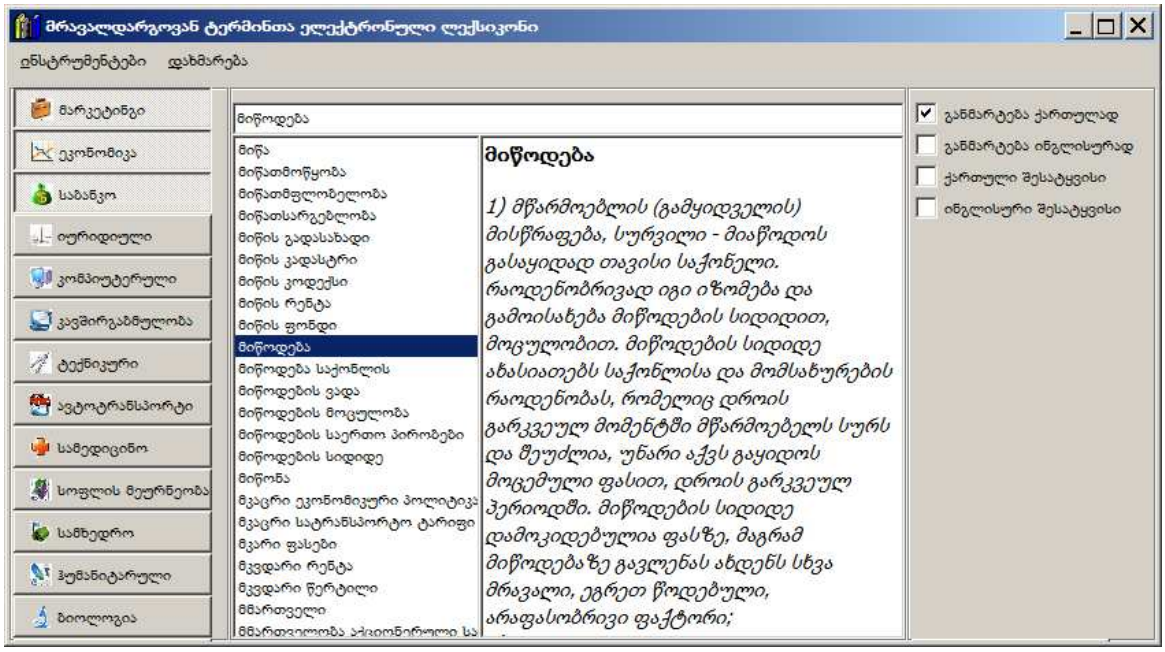


Fig .2

The program will be created in the CodeGear™ Delphi® 2009 program layer. Its prototype will be the appropriate web-application, in which MySQL database and PHP programming language. Corresponding pilot versions can be seen on the following address: www.ena.ge/orthography. The installation package of the desktop application of the diversified electronic terminological dictionary and its web application will be placed on the following address: www.ena.ge/terminology.

The electronic terminological dictionary is a useful and desirable product for the specialists of many fields as well as for teachers, students, and finally, the most important point with this dictionary is that, placed on the internet, on the special website, it will be available for a great number of users. It will facilitate functioning of the Georgian language and will at least for more instances stop the chaos existing in the terminological sphere. The electronic dictionary with standardised, correct, refined terms displayed in it, will serve as the important determinator of the future development of the Georgian language.

Resources for the Creating Interactive Bilingual Dictionaries of Daghestan Languages

Sabrina Shikhaliyeva
Daghestan Institute of language, literature and art
The Russian Academy of Sciences (Russia)
sh_shihalieva@mail.ru

An ideal language situation for Daghestan is the sustained diglossia when native speakers of Daghestan languages have already established domains of communication in both languages, Russian and Daghestan. Native speakers of Daghestan languages estimate both Daghestan and Russian languages without feeling any tension of using one language rejecting the other one. It is also important for them to transmit their native language to the younger generation, including oral and written communication of both languages.

Native speakers of Daghestan languages generally use each language in separate domains. For instance, they speak native (national) language at home and Russian at the local governmental office. They use Russian at the market and the native language at school (elementary basis). These domains are different for each language and each society member, but there is a strong tendency to draw the line between them. In Daghestan a tendency is observed to reduce the use of a native language due to the fact that Russian seems to be more advantageous and domains of the native language decrease which results in loss of knowledge and vocabulary.

According to the report of a UNESCO Expert Group on endangered languages, in the presence of a great number of languages, a language is in danger when its speakers cease to use it, use it in an increasingly reduced number of communicative domains, and cease to pass it on from one generation to the next.

The extinction of each language results in the irrecoverable loss of unique cultural, historical, and ecological knowledge. In the hope of preventing such loss five essential areas for sustaining endangered languages are distinguished: 1) basic linguistic and pedagogical training, 2) sustainable development in literacy and local documentation skills, 3) supporting and developing national language policy, 4) supporting and developing educational policy and 5) improving living conditions and respect for the human rights of speaker communities. It is obvious that literacy and education play a key role for language vitality.

In order to attain literacy in both languages in Daghestan teaching materials in native and Russian languages and resources linking these two languages must be available. A bilingual dictionary is an essential resource both for encouraging literacy and for linking languages in an area like Daghestan. A good bilingual dictionary provides a format for recording language data relating to the potentially infinite number of domains in the native and Russian languages, thus providing

information for use of both languages in more areas, as well as maintaining the vitality of the native language. These fields of knowledge that may have been once lost, can now be saved, not only for native speakers but also for the entire academic community. Specialised knowledge, like industrial vocabulary or folklore can be saved before the speakers possessing this knowledge die. A good bilingual dictionary can be a valuable device for both teachers and students. The format of online publication allows a bilingual dictionary to be accessible even during its drafting, editing, adding comments and changes, with no extra cost. The Internet dictionary has many advantages that make it more useful than the printed editions. The ability of categorizing, cross-referencing and search is potentially unlimited and is limited only by the imagination of the author. In this case we are speaking about the resources needed to create a bilingual national-Russian and Russian-national internet-dictionary. First, speaking about the Internet format, its advantages are described; second, the advantages of the approach to compiling dictionaries according to semantic spheres are explained.

There are several computer programs that create a format for the online edition of the dictionary and easily import data from a number of different vocabulary databases. One of the advantages of an online dictionary is that it can be published faster than printed, and it is easier and faster to edit. Therefore, society can see the fruits of labor of compilers while they still experience enthusiasm and inspiration from their work. This impulse grows, encouraging them to develop further materials for the local community. Moreover, online dictionaries do not restrict their compilers within any certain geographical place. Adding new articles and editing can be performed from various locations by many individuals, in view of which the compiling of the dictionary can continue until the access to the Internet is sufficient. An Internet dictionary can be published as soon as there is enough material for it.

Another advantage of the speed with which these dictionaries are compiled is that they immediately become a resource for education in national schools. A good bilingual dictionary along with good grammar textbook is the material for teachers to work out curriculum. Since vocabulary is best taught by means of semantic spheres, such a dictionary is an invaluable device, with a flexibility that enables production of highly effective sources.

The Bilingual dictionary is of great importance in a bilingual country, and it is crucial to create such a dictionary as soon and as efficient as possible. The use of online publication and collection of information on the semantic sphere not only fills the needs of every nationality, but makes the data accessible to the academic community and provides everyone with necessary resources.

Ресурсы создания интерактивных двуязычных словарей для дагестанских языков

Сабрина Шихалиева
Дагестанский Институт языка, литературы и искусства РАН (Россия)
sh_shihalieva@mail.ru

В Дагестане идеальной языковой ситуацией является устойчивая диглоссия, когда носители дагестанских языков уже установили сферы коммуникации в обоих языках (в русском и дагестанских). Носители дагестанских языков видят ценность как дагестанских, так и русского языка, и не чувствуют давления, чтобы использовать исключительно один язык, отказавшись от другого. Они также считают важным передать молодому поколению свой родной язык, включая устную и письменную коммуникацию обоих языков.

Носители языков в Дагестане, обычно, в каждой отдельной сфере используют один язык. Например, дома они говорят на родном (национальном), а в местном правительственном учреждении на русском. На базаре они используют русский, а в школе – родной (начальная база). Эти сферы бывают разными для каждого языка и каждого члена общества, но есть сильная тенденция проводить разграничение между этими сферами. В Дагестане также наблюдается тенденция снижения использования родного языка ввиду того, что русский язык представляется более выгодным, а сферы использования родного языка уменьшаются, что приводит к утрате знаний и лексики. Согласно отчету специальной группы экспертов ЮНЕСКО по языкам, находящимся в опасности, при наличии множества языков, язык считается находящимся в опасности, когда его носители перестают использовать его, используют его во все более ограниченном числе сфер коммуникации и перестают передавать его следующим поколениям. Исчезновение каждого языка влечет за собой невозвратимую потерю уникального культурного, исторического и экологического знания. В надежде предотвратить такие потери, выделяются пять основных сфер для поддержки языков, находящихся в опасности: 1) начальная лингвистическая и педагогическая подготовка, 2) устойчивое развитие грамотности и искусства документирования, 3) поддержка и развитие политики национального языка, 4) поддержка и развитие образовательной политики, и 5) улучшение жизненных условий и уважение к правам человека в языковых сообществах. Очевидно, грамотность и образование играют решающую роль в жизнеспособности языка.

Для достижения в Дагестане грамотности по обоим языкам, необходимо наличие учебных материалов на родных и русском языках, а также ресурсов, связывающих эти два языка. Двуязычный словарь является необходимым ресурсом, как для развития грамотности, так и для создания связи между языками в таком регионе как Дагестан. Хороший двуязычный словарь обеспечивает формат для записи языковых данных, относящихся к потенциально бесконечному числу сфер в родного и русского языков, тем самым предоставляя информацию для использования обоих языков в большем количестве сфер, а также, поддерживая жизнеспособность родного языка. Эти сферы знания, которые,

возможно, были однажды утрачены, теперь могут быть сохранены не только для носителей, но и для всего академического сообщества. Специальное знание, такое как отраслевая лексика или фольклор может быть сохранено прежде, чем обладающие этим знанием носители умрут. Хороший двуязычный словарь может быть ценным орудием, как для учителей, так и для учащихся. Формат интернет-издания позволяет двуязычному словарю быть доступным уже во время его составления, редактирования, добавления примечаний и изменений, причем без лишних расходов. Интернет-словарь имеет много преимуществ, которые делают его полезнее печатного издания. Способность категоризирования, использования перекрестных ссылок и поиска потенциально безгранична и ограничивается лишь воображением автора. В данном случае мы говорим о ресурсах, необходимых для создания двуязычного национально-русского и русско-национального интернет-словаря. Во-первых, говоря об интернет-формате, описываются его преимущества, а во-вторых, объясняются преимущества подхода к составлению словарей по семантическим сферам над другими подходами.

Есть несколько компьютерных программ, которые создают формат для интернет-издания словаря и с легкостью импортируют данные из ряда различных лексических баз данных. Одним из преимуществ интернет-словаря является то, что его можно издать скорее, чем печатный, а также его легче и быстрее редактировать. Поэтому общество может видеть плоды трудов составителей в то время, когда они все еще испытывают большой энтузиазм и воодушевление от своей работы. Импульс усиливается, побуждая их к дальнейшей разработке материалов для местной общины. Более того, интернет-словари не ограничивают их составителей каким-то одним географическим местом. Добавление новых статей и редактирование может совершаться из различных мест многими отдельными лицами, ввиду чего составление словаря может продолжаться, пока имеется достаточный доступ к интернету. Интернет-словарь может издаваться сразу, как только для него находится материал.

Еще одно преимущество скорости, с которой составляются такие словари, заключается в том, что они сразу же становятся ресурсом для обучения в национальных школах. Хороший двуязычный словарь вместе с хорошей учебной грамматикой представляет для учителей материал для разработки учебного плана. Поскольку лексику лучше всего учить при помощи семантических сфер, такой словарь является бесценным орудием, обладающим гибкостью, которая дает возможность более рационально произвести высокоэффективные материалы.

Двуязычный словарь имеет большое значение в двуязычном государстве, и важно создать такой словарь как можно скорее и эффективнее. Использование средств интернет-издания и сбор информации по семантическим сферам не только восполняет нужды каждой национальности, но и делает материал доступным для академического сообщества и обеспечивает всех необходимым ресурсом.

ლექსიკური ფუნქციები (LF) როგორც კვაზისინონიმური გარდაქმნების საშუალება

გიორგი ჩიკოიძე, ელზა დოკვაძე, ანა ჩუტკერაშვილი
საქართველოს ტექნიკური უნივერსიტეტი
არჩილ ელიაშვილის მართვის სისტემების ინსტიტუტი (საქართველო)
gogichikoidze@yahoo.com, annachutkerashvili@yahoo.com

მოცემული C_o ლექსიკური ერთეულის $LF(C_o)$ ფუნქციათა მნიშვნელობები ქმნიან ლექსიკური პარადიგმის მსგავს ერთობლიობას, რომლის წევრებს აქვთ მყარი ასოციაციური და კონსტრუქციული მიმართებები საწყის C_0-თან.

ერთ-ერთი სარგებელი, რომელიც ლექსიკური ფუნქციების ჩართვას ახლავს ენობრივი მოდელის სქემასა და მის ლექსიკონში, არის ის, რომ ამ ინფორმაციაზე დაყრდნობით ხშირად ხდება შესაძლებელი ტექსტში სიტყვათა სწორი კომბინირება.

გასცა ბრძანება / გააკეთა განცხადება / მიიღო გადაწყვეტილება / მისცა რჩევა /... (1)

ზმნები (1), რომლებიც წარმოადგენენ $Oper_1$ ფუნქციის მნიშვნელობების მაგალითს, ვერ შეენაცვლებიან ერთმანეთს, შესაბამისი არსებითების (C_o-ების) აშკარა სემანტიკური სიახლოვის მიუხედავად.

LF სისტემის ღირებულების მეორე ასპექტი მდგომარეობს იმაში, რომ მისი კომპონენტები უზრუნველყოფენ საფუძველს მრავალი კვაზი-სინონიმური ტრანსფორმაციისთვის. ასე მაგალითად, (1)-ის ბოლო ნიმუშში შესაძლებელია "რჩევა" განვიხილოთ როგორც "ურჩევს" ზმნის (C_o) დერივატი (Der) და მივიღოთ გარდაქმნის სქემით

$$C_o \leftrightarrow Oper_1 + Der(C_o)$$

შედეგი:

ურჩია ↔ მისცა რჩევა; (2)

შემდეგ ნაბიჯზე შეიძლება გამოვიყენოთ ის, რომ

$Oper_1$ (რჩევა) = Conv $Oper_1$(რჩევა)

და მივიღოთ:

ვიღაცამ მისცა მას რჩევა ↔ მან მიიღო რჩევა ვიღაცისგან (3)

კიდევ ერთი ბიჯი შეიძლება დაეყრდნოს ტოლობას

$Func_1$(რჩევა) = დავიდა,

რაც იძლევა (3)-ის გარდაქმნის საშუალებას:

-(3) → ვიღაცის რჩევა დავიდა მისამდე (4)

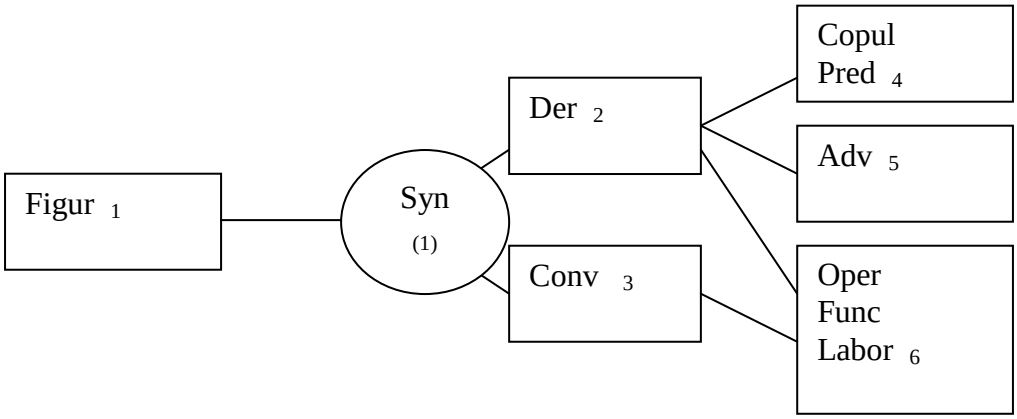
ბოლოს ისიც შეიძლება აღვნიშნოთ, რომ ზოგ პრაგმატიკულ კონტექსტში (1-4) მაგალითებში გამოყენებული ერთეული "რჩევა" შეიძლება შეიცვალოს მისი სინონიმური მწკრივის რომელიმე სხვა წევრით:

-რჩევა, მითითება, მინიშნება, ... (5)

(ეს მწკრივი შეიძლება განხილულ იქნეს როგორც Syn-ფუნქციის გარკვეული განზოგადება).

ამ კონკრეტული მაგალითების (1-5) განხილვის შემდეგ ქვემოთ მოცემულია იმ LF-ების ზოგადი სქემა, რომლებიც ყველაზე მჭიდროდ უკავშირდებიან კვაზი-სინონიმურ გარდაქმნებს (ნახ.1). კომენტარები, რომლებიც მოჰყვება ამ სქემას, ეძღვნება თითოეული

ბლოკის (ე.ი შესაბამისი LF-ის) შინაარსის ახსნასა და ილუსტრაციას რამდენიმე მაგალითის მეშვეობით. მომდევნო პარაგრაფების ინდექსაცია კი ემთხვევა იმ ბლოკების ინდექსებს, რომელთა შინაარსის აღწერას ისინი მოიცავენ.



ნახ.1 იმ LF-ების სქემა, რომლებიც ჩვეულებრივად გამოიყენება კვაზი-სინონიმურ ტრანსფორმაციებში. კომენტარები ტექსტში.

0. სქემის საწყისი (0) ბლოკი წარმოგვიდგენს LF Syn (სინონიმს), ანუ ლექსიკური ჩანაცვლების "იდეალურ" (თუმცა საკმაოდ იშვიათ) ვარიანტს, რომელიც სემანტიკურად იმდენად ახლოა შესაბამის C_o-თან, რომ შეიძლება შეენაცვლოს მას ნებისმიერ კონტექსტში და თანაც ამ კონტექსტის კონსტრუქციის ყოველგვარი ცვლილების გარეშე:

- ლინგვისტიკა ↔ ენათმეცნიერება, ტუპერკულიოზი ↔ ჭლექი, ...

1. Figur წარმოადგენს Syn-ის ფიგურალურ (მეტაფორულ, იდიომატურ) სახეობას, რომელიც მჭიდროდ არის დაკავშირებული კონტექსტთან:

- იცის: კარგად ↔ ხუთი თითივით; ჭამს: ბევრს ↔ ღორივით; ჯიუტობს: მაგრად ↔ ვირივით...

2. Der(C_o)-ის სემანტიკა ემთხვევა C_o-ისას, თუმცა სხვა მეტყველების ნაწილის ფორმით, თანაც რომელიც მოითხოვს კონტექსტის კონსტრუქციის გარკვეულ ცვლილებებს:

- სახე გაუწითლდა ↔ სახეს სიწითლე დაეტყო ↔ სახე წითელი გაუხდა ↔ წითლად გამოიყურება.

მოცემული ნიმუში იძლევა მეტყველების ნაწილთა ცვლის სრულ ჯაჭვს:

$C_o = V_o \rightarrow S_o \rightarrow A_o \rightarrow Adv_o$.

3. Conv(C_o) უჩნდება ზმნებს, რომლებიც ასახავენ პროცესს ორი მონაწილით, რომელთა შორის მოძრაობს ერთი (ან მეტი) ობიექტი, განსხვავებას კი C_o-სა და მის Conv(C_o)-ს შორის წარმოადგენს ის, რომ ისინი ასახავენ ამ მოძრაობას ხან ერთი მიმართულებით

(ვთქვათ, პირველი მონაწილის თვალსაზრისით), ხან კი მისი საპირისპირო მიმართულებით. ტიპურ მაგალითს წარმოადგენს "შესყიდვა ↔ გაყიდვა":

- ვიყიდე წიგნი ბუკინისტისგან ↔ ბუკინისტმა მომყიდა წიგნი.

4. Pred(C_o) – მე-2 ბლოკის (Der) კერძო შემთხვევაა: Copul(C_o)-თან ერთად ის ქმნის კომბინაციას C_o ↔ Copul (C_o) + C_o= Pred (C_o), რომელსაც აქვს C_o-ის სემანტიკა, მაგრამ ზმნური ფუნქციონირების ხასიათით:

- გენიოსი: ეს მწერალი გენიოსია / გენიალურია.

ამ მაგალითიდან ჩანს, რომ ქართულში Copul(C_o) შეიძლება იყოს შერწყმული C_o -თან.

5. Adv(C_o) შესაძლებელია იმ (საკმაოდ იშვიათ) შემთხვევაში, როცა სიტუაცია წარმოდგენილია ორი პარალელური პროცესით, რომელთაგან ერთ-ერთი შეიძლება იქნეს განხილული როგორც მეორის დახასიათება:

- ავირჩიე და შევცდი ↔ ავირჩიე შეცდომით ↔ ავირჩიე მცდარად ↔ გავაკეთე მცდარი არჩევანი (C_o).

6. ამ ბლოკის წევრები (Oper, Func, Labor) უზრუნველყოფენ ზოგ ტრანსფორმაციას (Der, Conv) ზმნებით, რომლებიც აუცილებელია კონტექსტის კორექტული გარდაქმნისათვის. პირველი წყვილის (Oper, Func) მაგალითები უკვე იყო მოყვანილი ზემოთ – (2), (3). მე-3 მათგანის (Labor) საილუსტრაციოდ შეიძლება გამოვიყენოთ:

- ფონდმა დააფინანსა პროექტი ↔ ფონდმა უზრუნველყო პროექტი დაფინანსებით.

ამგვარად, LF-ები წარმოადგენენ მნიშვნელოვან საშუალებასა და საფუძველს კვაზისინონიმური გარდაქმნებისთვის, რომლებიც, თავის მხრივ, ასრულებენ არსებით როლს ენობრივი მოდელების (და საერთოდ – ბუნებრივი ენის) ფუნქციონირებაში (Мельчук 1999).

Lexical Functions as the Means for Quasi-synonymous Transformation

Giorgi Chikoidze, Anna Chutkerashvili, Elza Dokvadze
Georgian Technical University
Archil Eliashvili Institute of Control Systems (Georgia)
gogichikoidze@yahoo.com, annachutkerashvili@yahoo.com

Lexical functions LF (C_0) may be considered as something similar to the lexical "paradigm" of the corresponding vocabulary unit C_0; which includes stable lexical units associated with C_0 and connected to it by some meaningful and standard semantic relations.

Particularly, the inclusion of lexical functions both in the dictionary units and in the system of a language is that this information often supplies the possibility of correct choice of a word combination in the text: *gasca brÂaneba* - ordered; *gaak'eta gancxadeba* - announced, *miiγo gadac'qvet*'ileba - decided; *misca rčeva* - advised;… (1)

The $Oper_1$ (C_0), represented by verbs of (1), cannot be replaced by each other in spite of the semantic likeness of their arguments, and thus, the correct choice of these verbs fully depends on the corresponding C_0.

On the other hand, the components of lexical functions (LF) may serve as one of the tools for quasi-synonymous transformations: e.g. in the case of the last pair of (1) the noun may be considered as derived from the corresponding verb:

Der (určevs-'advises')=rčeva ('advice'),

Which supplies the possibility of transformation: *man určia mas*('he advised him')→*man misca mas rčeva* ('he gave her advice') (2)

On the next step $Oper_1$ (rčeva) can be replaced by $Oper_2$ (rčeva)= Conv $Oper_1$ (rčeva), what makes the following transformation possible:

Man misca mas rčeva ('he gave her advice')→*man miiγo misgan/misi rčeva* ('She received advice' from him. (3)

One more transformation can be based on $Func_1$ (*rčeva*)=*davida* ('reach') *viγacis rčeva davida misamde (droze*)→ 'Somebody's advice reached her/him (on time)', etc. (4)

Finally, it can be mentioned that in some pragmatic contexts, the item used in (1-4)- *rčeva* can be replaced by any other member of its synonymous set: *mititeba (*notice), *minishneba* (imply implicitly)…..(5)

After these specific examples (1-5), below there will be given a more general scheme (Fig.1) of the LF's most closely connected with quasi-synonymous transformations. The comments which follow this scheme interpret the meaning of LF's represented by the blocks of the scheme and illustrate the roles, which they may play in transformations. The indexation of paragraphs below corresponds to the indexes of blocks which they embrace.

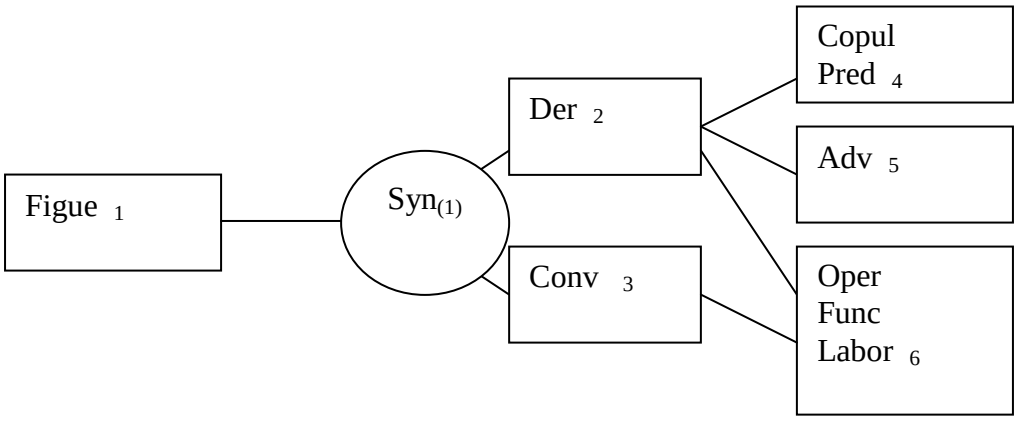


Fig.1 The Scheme of LFs, most commonly participating in quasi-synonymous transformations. Comments in the text.

0. Syn(onym), which represents the initial block (0) of the scheme is the most "ideal" (but for that the most rarely used)version of substitution for the original $C_{0:}$

lingvist'ik'a ↔enatmecniereba ('linguistics'),
t'uberk'uliozi ↔č'leki ('consumption'), etc.

The semantics of Syn (C_0) is (almost) identical to that of C_0 and its substitution for $C_{0,}$ as a rule, does not change the construction of the context.

1. fig(urative)version of Syn:

icis:k'argad ↔xuti titivit ('knows :well ↔like their own five fingers)
č'ams :bevrs ↔γorivit (eats: much↔like a swine')

Here the substitution is possible in this context (icis/čams) only and as a rule implies :Figur (C_0)=Magn (C_0).

2. Der(ivatives) have the semantics of C_0 but require the similar change of the context construction:

Saxe gauc'itlda ↔saxes sic'itle saet'qo↔saxec'iteli gauxda↔c'itlad gamoiyureba ("she/he flashed/ turned red')

Here is given the full paradigm of derivations:

$C_o=V_0\rightarrow S_0\rightarrow A_0\rightarrow Adv_0$

3.Conv(ersives) can be used, when the situation expressed by the utterance includes a process with two participants, between whom objects move : in such a case C_0 and its Conv. represent these movements from the standpoint of different, mutually opposed participants of the process. The typical example of such as situation is :
Viqide c'igni buk'inist'isgan ('I bought the book from book-seller')↔*buk'inist'ma momqida c'igni* ('Book-seller sold me the book').

4. Pred(icate) C_0 is a particular case of Der (Block 2), which transforms C_0 into a verb unit adding to its Copul:
Predicate= Copul(C_0)+ C_0

e.g. *es mc'erali geniosia/genialuria* ('This writer is a genius/genial').

5. Adv(erb) C_0 is another version of Der. Its use is possible in such (quite rare) contexts, when the situation is represented by two parallel processes and one of them may be considered as a characteristics of the other:
Avirčie da/magram ševcdi↔ mcdarad avirčie (I've chosen and/but have made an error↔ I've made the choice erroneously') .
6. The LFs of this block (Oper, Func, Labor) support the transformations based on Der or Conv functions supplying by verbs, which are necessary for the correct change of the context. Examples of Oper and Func were already given above by (2), (3). The last of them (Labor) may be illustrated by:

Fondma dafinansa proekti↔fondma uzrunvelqo proeqti dafinansebit ("The foundation financed the project↔The foundation supplied financing of the project').

Thus, the LFs represent an essential basis for quasi- synonymous transformations, which, in their turn, play the most important role in the linguistic model "Meaning↔Text" (Melchuk, 1977).

ბიბლიოგრაფია/References

Melchuk, I. The Linguistic Model "Meaning↔Text", 1977.

ქართული არანორმატიული ლექსიკის ფილტრაციის პრობლემა და მისი გადაწყვეტის გზები

ლევან ჩხაიძე, დავით ჯაში, რუსუდან ლანდია, ციცინო კვანტალიანი
საქართველოს ტექნიკური უნივერსიტეტი,
არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (საქართველო)
levan@dr.com, tsitsini.nino.kvantaliani@gmail.com

"არა მგონია, ბილწსიტყვაობა, რომელიც დღეს ასე თავმოწონებით იკიდებს ფეხს ჩვენს მწერლობაში, "მწერლის ენობრივი თავისუფლების" მაჩვენებელი იყოს! მას ... ნარკომანიაზე ნაკლები ზიანი არ მოაქვს ჩვენი ახალგაზრდობისთვის, ქვეყნისთვის!" – წერდა ჯერ კიდევ საუკუნის დასაწყისში ანა კალანდაძე. ამ ნეგატიური ტენდენციის უარყოფითი გავლენა ბევრად გაძლიერდა თანამედროვე ინფორმაციული ტექნოლოგიების განვითარებისა და გავრცელების შედეგად. ინტერნეტსივრცე განსაკუთრებით ბოლო პერიოდში წალეკა ბილწსიტყვაობამ. ეს გამოიწვია როგორც "სიტყვის თავისუფლების" არასწორმა გაგებამ, ისე ფორუმების მონაწილეთა ანონიმურობამ.

შექმნილი სოციალური ფონიდან გამომდინარე, განსაკუთრებული პასუხისმგებლობა ეკისრება ინტერნეტის ინფრასტრუქტურას, რომელმაც უნდა უზრუნველყოს მსგავსი ლექსიკის ავტომატური აღმოჩენა და მისი გამორიცხვა მომხმარებელთა მრავალრიცხოვანი არმიის თითოეული წევრისათვის. ხაზგასმით აღვნიშნავთ, რომ ამ პროცესს მომხმარებელი სათანადოდ ვერ ეწინააღმდეგება და ამიტომ, შესაბამისად, მისი თვითკონტროლი, თუნდაც ცენზურა, არარეალიზებადია, ფაქტობრივად "არ მუშაობს".

როგორც ცნობილია, ასეთი შესაძლებლობანი უკვე ჩართულია ინგლისური და რუსულენოვანი ქსელების დიდ ნაწილში. დღის წესრიგში ქართული ენისათვისაც დადგა ეკვივალენტური შესაძლებლობის პროგრამული უზრუნველყოფის საკითხი. ეს დიდი პრობლემაა და დროულ გადაწყვეტას მოითხოვს. სწორედ ამიტომ ამჟამად მუშავდება ქართული არანორმატიული ლექსიკის ფილტრაციის ავტომატური სისტემა, რომელიც ემყარება ქართული ენის სხვადასხვა კომპიუტერულ მოდელს.

ფილტრაციის სისტემა მოიცავს შემდეგ ალგორითმულ და ინფორმაციულ კომპონენტებს:

1. ქართული არანორმატიული სახელებისა და ზმნების ფუძეების მონაცემთა ბაზა; ამ შემთხვევაში გამოიყენება მართლმწერის უზრუნველსაყოფად ადრე დამუშავებული ფლექსიური ფორმაწარმოებით მიღებული სიტყვაფორმების საძიებო პროცესორი.

2. ქართულში დამკვიდრებული ნასესხები არანორმატიული ლექსიკის (პირველ რიგში, რუსულიდან) მონაცემთა ბაზა. მნიშვნელოვანია, რომ ნასესხები არანორმატიული ლექსიკა, როგორც წესი, ქართულში მკვიდრდება გაქვავებული სახით. ამიტომ იღებს უდეტრის სხვადასხვა გრამატიკულ ფორმას. შესაბამისად, ამ შემთხვევაშიც გამოიყენება ადრე დამუშავებული უდეტრების საძიებო პროცესორი.

3. როგორც ექსპერიმენტული მასალიდან ჩანს, ქართული არანორმატიული ლექსიკა ხასიათდება კომპოზიტების შექმნის საკმაოდ ინტენსიური პროცესით. ამიტომ სისტემისათვის მუშავდება კომპოზიტების გამოცნობის სპეციალური პროცესორი, რომელიც სიტყვაფორმას განიხილავს, როგორც "ტრადიციული" არანორმატიული და ენის ნორმატიული ნაწილის სახელური ფუძეებისაგან შემდგარ კომპონენტს.

4. საგულისხმოა, რომ, განხილული მასალიდან გამომდინარე, ქართული არანორმატიული ლექსიკა ხასიათდება კიდევ უფრო მძაფრი დერივაციული პროცესებით, რომლებიც ჯერ კიდევ შეუსწავლელი სოციოლინგვისტური მიზეზების გამო უფრო ინტენსიურია, ვიდრე – ნორმატიულ ლექსიკაში. პარადოქსულია, მაგრამ უფრო ინტენსიურიც კი არის, ვიდრე - პოეტურ შემოქმედებაში. მიუხედავად ამოცანის სირთულისა, დღის წესრიგში დგას დერივაციული პროცესორის შექმნის ამოცანა. პირველ რიგში, შესაძლებელი უნდა იყოს ისეთი დერივაციული ელემენტების ავტომატური შეცვლა ან დამატება, როგორიცაა ზმნისწინი, მასდარის, მიმღეობის, ზედსართავი და აბსტრაქტული არსებითი სახელების მაწარმოებელი აფიქსები და ა.შ.

5. გაცილებით რთული ამოცანაა ქართული არანორმატიული შინაარსის სინტაგმების ფორმალური მოდელის შექმნა იმ შეთხვევაში, როდესაც სინტაგმაში შემავალი არც ერთი სიტყვა თავისთავად არ ეკუთვნის არანორმატიულ ლექსიკას. ასეთი, სალაპარაკო ენაში საკმაოდ ხშირი, სტრუქტურების გამოსაცნობად აუცილებელია სინტაქსური ანალიზატორის სპეციალური ნაირსახეობის შექმნა. ყველაზე მარტივ შემთხვევაში საჭირო ანალიზი ამოიწურება ორი შერწყმული სიტყვის წყვილთა ექსპერტული მონაცემთა ბაზის შექმნით. უფრო ზოგად შემთხვევაში სინტაგმის შესაძლო სინტაქსური ანალიზატორი შეიძლება აღიწეროს შემდეგი მიმდევრობით:

– მეთაური სიტყვა, ყველაზე გავრცელებულ შემთხვევაში – უდეტერი;

– ატრიბუტული ნაწილი, ზედსართავი სახელის, ზმნისართის ან უარყოფითი ნაწილაკისგან შემდგარი (ეს უკანასკნელი აუცილებელი არ არის);

– პრედიკატული ნაწილი, ჩვეულებრივად მეშველი ან მოდალური ზმნა სხვადასხვა პირში, რიცხვში ან მწკრივში.

აღწერილი სტრუქტურის გამოსაცნობად უნდა შეიქმნას ევრისტიკულ წესებზე დამყარებული სპეციალიზებული პროდუქციული პროცესორი.

ზემოთ ჩამოთვლილი 5 კომპონენტიდან 3 უკვე შესრულებულია. არანორმატიულ სახელთა მონაცემთა ბაზა შეიცავს 200-ზე მეტ სახელურ ფუძეს, სხვა ენებიდან ნასესხები არანორმატიული ლექსიკის მონაცემთა ბაზა – დაახლოებით 50-მდე ერთეულს. რაც შეეხება კომპოზიტთა ბაზას, ის ეყრდნობა როგორც I, ისე II კომპონენტს. კომპოზიტი იქმნება ბაზაში არსებული არანორმატიული სიტყვების სხვადასხვა კომბინაციით.

IV და V კომპონენტები ამჟამად მუშავდება.

უნდა აღინიშნოს, რომ არჩეული მოდელი არ ვრცელდება ევფემურ სიტყვებსა თუ გამოთქმებზე, რაც შემდგომი კვლევის საგანს წარმოადგენს.

On the Problem of the Filtration of Georgian Nonnormative Vocabulary and the Ways to solve it

Levan Chkhaidze, Davit Jashi, Rusudan Landia, Tsitsino Kvantaliani
Georgian Technical University
Arn. Chikobava Institute of Linguistics (Georgia)
levan@dr.com, tsitsini.nino.kvantaliani@gmail.com

Ana Kalandadze wrote as early as the beginning of the century: *"I do not believe that improper swear words, which are permeating our literature, is an indicator of the lingual freedom of a writer! It harms our youth, our country not less than drug addiction!"*. The negative influence of this tendency has strengthened as a result of the development and circulation of modern informational technologies. Online space has been flooded with the improper, swear language lately. This is caused by a misunderstanding of the concept of "freedom of speech". Another reason may be anonymity of the participants of on-line forums.

Because of this social background, the internet infrastructure has to carry a special responsibility for this situation. It should automatically detect such vocabulary and delete it from each user's forum. It should also be emphasised that users do not meet their responsibilities in this domain which result in the fact that self- control or even censorship does not work.

As is known such possibilities have already been incorporated in most English and Russian networks. Similar software is on the agenda for Georgia. This is a big problem and needs an apropos solution. For this reason an automatic system of the filtration of nonnormative vocabulary, based on various computer models of the Georgian language, is being processed.

The filtration system consists of the following algorithmic and informative components:

1. The database of Georgian non-normative noun and verb stems. In this case the word-form searching processor, which was created earlier to ensure correct spelling of words derived by means of flexions, is used.
2. The database of the borrowed non-normative vocabulary of Georgian (foremost amongst which are words borrowed from Russian). It is important to note that borrowed non-normative vocabulary, as a rule, is established unchanged. This is the reason why such lexis acquires various grammatical forms of the Udetre (The parts of speech, that do not change in form). Accordingly, in such cases the search processor of Udetres is elaborated before, is used.
3. As seen from the analysed data, Georgian nonnormative vocabulary is characterised by an intensive process of producing compounds. Thus, a particular processor which will consider a word form as a component consisting of nominal stems of "traditional" non-normative and normative nominal parts of the language is currently being created.
4. It follows that Georgian nonnormative vocabulary is characterised by derivative processes that are more intensive for yet not studied sociolinguistic reasons, than the normative vocabulary.

Surprisingly, it is more intensive than in poetry. In spite of the complexity of the issue, creating a derivative processor is planned. It should be possible to change or automatically add derivative elements as preverbs, formative affixes of verbal noun (Georgian: "Masdar"), participles, adjectives, abstract nouns etc.

5. It seems more difficult to produce a formal model of Georgian nonnormative syntagmata when none of the words of the syntagma belongs to nonnormative vocabulary. In order to recognise such structures which are often used in conversational language, it is important to create a special type of a syntactic analyser. In the simplest case, the necessary analysis will produce an experimental database of combined word-collocations. In a more general cases, a possible syntactic analyzer of a syntagma may be described by the following sequence:

– A headword, mostly an Udetre;
– An attributive part consisting of an adjective, a verbal affix or a negative conjunction (the latter is not mandatory);
– A predicative part which is usually an auxiliary or modal verb of various personalities, number and screeve.

It is necessary to create a specialized productive processor based on Evristic rules to recognise the above-described structure.

Three of the above-discussed five components have already been implemented.
The database of nonnormative nouns consists of more than 200 hundred nominal stems. The database of the nonnormative vocabulary borrowed from other languages consists of about 50 units. As for the database of compounds, it is based on the first and on the second components. A compound is produced by means of various combinations of nonnormative words in the database.

The fourth and fifth components are currently being elaborated.

It should be noted that the selected model does not work for euphemistic words or phrases which will be the object of future research.

ბიბლიოგრაფია / References

ჩხაიძე ლ., ქართული ენის ელექტრონული გრამატიკული ლექსიკონის და მართლმწერი სისტემის (სპელჩეკერის) ავანპროექტი, ბუნებრივ ენათა დამუშავება, კონფერენციის მასალები, თბილისი, 2006წ.

კვანტალიანი ც., ჩხაიძე ლ., ქართული ენის ზმნური ფუძეების ელექტრონული ლექსიკონის მონაცემთა ბაზის დერივაციული კომპონენტების კვლევისათვის, ბუნებრივ ენათა დამუშავება, კონფერენციის მასალები, თბილისი, 2007წ.

ინჯია ზ., ჩხაიძე ლ., დონალდ რეიფილდის დიდი ქართულ-ინგლისური ლექსიკონის ელექტრონული ვერსია, ბუნებრივ ენათა დამუშავება, კონფერენციის მასალები, 2007წ.