Enriching Georgian Dictionary Entries with Frequency Information

Sophiko Daraselia, Serge Sharoff
University of Leeds
e-mail: mlsd@leeds.ac.uk, s.sharoff@leeds.ac.uk

Abstract

In this paper we will discuss the integration of corpus analysis into the dictionary making process for Georgian language. In general, corpus-based lexicography is not a common practice in lexicography in Georgia. This paper presents the first attempt to introduce the corpus-based dictionary wordlist, entries and examples from the Georgian web-corpus – KaWaC. This is a large web corpus of modern Georgian language covering recent 10-15 years of the language development. It contains a wide range of text types, topics and regions from the Internet, excluding translations and poetry on the assumption that the language of translation and poetry deviates from the naturally produced language, as the corpus aims to represent naturally occurring Modern Georgian language. Within this research we have defined the dictionary wordlist – 10,000 lemmas from the corpus that is a core vocabulary for the Georgian language, compiled the dictionary entries and extracted dictionary examples from the corpus.

Keywords: corpus linguistics; web-corpus; corpus-based lexicography; learner's dictionaries

1 Introduction

The aim of this paper is to demonstrate the effectiveness of corpus analysis by presenting our lexicographic approaches and comparing them with the existing Georgian dictionaries. Namely, the largest Monolingual Georgian dictionary (GMD) in 8 volumes (Chikobava 1964) including the recent editions (Arabuli 2007).

Our study is based on KaWaC – a large web-corpus of Georgian that was developed at the University of Leeds (Daraselia & Sharoff 2014). It represents the modern language within the last 10-15 years. The corpus crawled from various Internet sources consists of 618,468 webpages and over 180 million words. It was checked for balance according to the Functional Genre Dimensions (Forsyth & Sharoff 2014) by manual annotation of a randomly selected set of 1067 webpages. The prevailing genres in the corpus are:

- **A11 -** first person point of view such as personal blogs, forums: 47.5%, the largest genre, which provides a good approximation to the use of modern everyday Georgian;
- **A8 -** news: 20.5%;
- **A7 -** various texts related to instructions and tutorials: 11.2%;
- **A9 -** laws, contracts, regulations etc. 8.1%.

We have also developed a tagset for Georgian based on the MULTEXT-East Morphosyntactic Specifications (Erjavec 2010) and tagged the entire corpus based on our general approach covering the most basic features of the language (Daraselia & Sharoff 2014). To create a frequency list for the whole corpus we have lemmatised the corpus. The lemmatisation was carried out in several steps. At first, taking into consideration the morphology of the language, we have defined what a lemma stands for in our corpus. Amongst Georgian linguists there is no agreement on what a lemma for a verb can be. Some linguists think that *masdar* (verbal noun) should be considered as a lemma for verbs and masdar forms should function as headwords for verbs in dictionaries, whereas some linguists propose to put all verb paradigms as headwords. Taking into consideration the complexity of the Georgian verb, we made a decision to follow the approaches of the Georgian Monolingual Dictionary and a lemma for a verb in our corpus is as follows: a) 3rd Person, Singular, Present Tense, and b) 3rd Person, Singular, Future Tense for preverbs. For other parts of speech such as nouns, adjectives etc. a standard approach is used, in particular, Singular, Nominative Case (Daraselia & Sharoff 2015).

1.1 Corpus-based Wordlist

In the dictionary-making process the macrostructure is very important and it is particularly difficult to define a dictionary wordlist. There are many factors that should be considered in the dictionary entry selection process, such as the type of dictionary. For example, learner's dictionary focuses on the most frequent words of a particular language, whereas large monolingual dictionaries aim to document the language, to cover all language development periods, e.g. OED, GMD etc. We selected the entries to be included in a learner's dictionary of Georgian from the Corpus focusing on the most frequently used words. In other Georgian Dictionaries selection of the entries is rather intuition based. We created a frequency list for the whole corpus, in particular, extracted 18,000 words and compared them to GMD wordlist which revealed thousands of frequently used new words not given in GMD (see table 1).

Word	Transliteration	Translation	Frequency
სტატუსი	statusi	status	27983
ინტერნეტი	interneti	internet	20100
ინფრასტრუქტურა	inp'rastruk'tura	infrastructure	16247
პრეზენტაცია	prezentac'ia	presentation	15202
ვებგვერდი	vebgverdi	web-page	11814
სტრესი	stresi	stress	959
ნეგატიური	negatiuri	negative	939
ეკომიგრანტი	ekomigranti	ecomigrant	864
ინსტიტუცია	instituc'ia	institution	753
მულტიმედია	multimedia	multimedia	681

Table 1: Frequency of Some New Georgian Words not presented in GMD.

The initial frequency list of 18,000 words was corrected, for example, errors of lemmatization, proper nouns, such as personal names, names of countries, cities etc. were removed from the list. We decided to limit the word-list to 9,831 words, selecting all lemmas that appear at least 50 times in the corpus, excluding proper nouns. These 9,831 lemmas represent 66.77% of all the words of the corpus. The frequency list we have selected is the core vocabulary for the modern Georgian language and can be used as the dictionary word-list for monolingual or multilingual learner's dictionaries for Georgian, e.g. for Georgian-English Learner's Dictionary, Georgian-German Learner's Dictionary etc.

Thus, we created a totally new corpus-based dictionary word-list of 10,000 headwords that can be used for the compilation of monolingual or bilingual learner's dictionaries for Georgian. The microstructure of the dictionary, i.e. word senses and dictionary examples are also entirely corpus based. It is discussed in the next section below.

2 Comparing GMD to Senses in KaWaC

2.1 New Words and Loanwords

Further analysis of these 9,831 words revealed hundreds of new words that have entered the language in recent years, in particular, during the last 10-15 years. These new words are not presented in the Georgian Monolingual Dictionary, which is the most comprehensive dictionary for Georgian covering 113,000 dictionary entries. Despite the size of the GMD, many new words (Table 1) do not appear because the dictionary was published in 1960s and only two volumes of the revised versions are available. Our comparative study of word senses is based both on old and new editions of the Georgian Monolingual Dictionary (Chikobava 1964; Arabuli 2007).

Another aspect of the core vocabulary use that can be studied with the frequency information is the generalization of loanwords, in particular, English words in Georgian. There are Georgian equivalents available for English words. Analysing the top 3,000 frequency list we can assume that Georgian equivalents are more frequently used than the English loanwords (See Table 2).

Loanwords	Freq	Georgian Equivalent	Freq
სერვისი [servisi = 'service']	3078	მომსახურება [momsaxureba = 'service']	30086
დემონსტრირება [demonstrireba ='demonstrate']	2060	ჩვენება [č'veneba = 'demonstrate']	15990
ფორმირება [p'ormireba = 'formation']	2133	წარმოქმნა [carmok'mna = 'formation']	825
ნეგატიური [negatiuri = 'negative']	939	უარყოფითი [uarqop'it'i = 'negative']	18251
იმიტაცია [imitac'ia = 'imitate']	862	მიბაძვა [mibažva = 'imitate']	242
პერმანენტული [permanentuli = 'permanent']	759	მუდმივი [mudmivi = 'permanent']	17281
ინსტიტუცია [instituc'ia = 'institution']	753	დაწესებულება = dacesebuleba = 'institution']	32390
ლინკი [linki = 'link']	617	ბმული [bmuli = 'link']	4611
მეილი [meili = 'mail']	214	წერილი [cerili = 'mail']	48397

Table 2: Frequency of English loanwords and Georgian Equivalents.

The loanwords are preferably used in informal conversations, for expressing personal feelings and emotions on blogs and forums etc., while for official purposes the Georgian equivalents are used, e.g. on websites of the governmental or educational institutions such as ministries, presidential administration, universities etc.

2.2 New Senses

It is obvious that lexicographic applications of corpus data are very broad. Corpus data help a lexicographer not only in the process of selecting the dictionary word-list, but also with deciding on sense distinctions by viewing different occurrences of the same word or lemma on concordance lines with the surrounding contexts and even looking up the wider contexts in the corpus and enriching this with the frequency information (Kilgarriff 1993; Sinclair 1991). In particular, the frequency distribution of collocates of a particular word will provide a lexicographer with reasons for making sense distinctions. The frequency distribution of the top collocates of *guli* ('heart') in the corpus is given in Table 3, sorted by the log-likelihood score. The following senses are given in the top collocate list: ‘the inner, central part of something’, e.g.: *kverc’xis guli* = ‘yolk’; figurative usage, such as *c’xel gulze* = ‘being hot headed’; *xelis guli* = ‘palm’ etc.

We have detected 11 senses for the dictionary entry *guli* (‘heart) from the corpus, such as 1) the organ that makes blood flow around the body, e.g. *gulis daavadeba* (‘heart disease’); 2) the central part of something; e.g.: *tbilisis guli* (‘the heart of Tbilisi’) etc. In the GMD there are 13 senses for the same dictionary entry. The senses not presented in our corpus-based entry are: 1) ‘the inner and hard part of a tree’ and 2) ‘the thin plate on *č'onguri and p'anduri’* (Georgian musical instruments). These word senses do not appear in our corpus data, it can be considered as being very specific to the field and should be covered only in relevant specialized dictionaries.

Collocation	Transliteration	Translation	Joint	Freq1	Freq2	LL score
საკუთარი გული	sakutari gulič	one's own heart	2584	215364	97474	3099.65
ჩემი გული	čemi guli	my heart	2493	486693	97474	1982.2
კვერცხის გული	kverc'xis guli	yolk	811	21305	97474	1433.21
მთელი გული	mteli guli	whole heart	1324	158444	97474	1353.51
ადამიანის გული	adamianis guli	human heart	1945	510066	97474	1283.55
თავისი გული	tavisi guli	his/her/its own heart	1560	310522	97474	1223.23
ცხელ გულზე	c'xel gulze	hot minded, angry	418	15208	97474	669.97
შენი გული	šeni guli	your heart	700	114812	97474	610.38
მისი გული	misi guli	his/her/it heart	1316	707919	97474	469.2
ხელის გული	xelis guli	palm	692	240560	97474	369.62
კეთილი გული	ketili guli	kind heart	248	17879	97474	313.6
კინაღამ გული	kinağam guli	almost heart…	172	4441	97474	305.18
მაგრამ გული	magram guli	but heart..	959	624993	97474	272.21
მზის გული	mzis guli	under the Sun	259	30054	97474	267.94
მართლა გული	martla guli	really heart…	233	40769	97474	195.91
ლამის გული	lamis guli	almost heart…	151	10510	97474	193.5
სადღაც გული	sadğac guli	somewhere in heart…	153	14265	97474	174.33
დაწყდება გული	dacqdeba guli	will get hurt	55	180	97474	158.28
თქვენი გული	tkveni guli	your heart (PL)	337	153972	97474	141.43
კაცის გული	kac'is guli	man's heart	286	120384	97474	129.57

Table 3: Frequency Distribution of Collocates for *guli* (‘heart’).

For the collocations of *guli*, the corpus data provides 81 the most frequently used collocations, where, 63 collocations out of 81 do not appear in the GMD, such as the following: a) *gulis ac’rueba* (‘be discouraged from doing something’), e.g.:

(1) *sc’avlaze guli auc’ruvda* (‘discouraged from studying’).

and b) *gulis gacqaleba* (‘annoy someone by doing something’). e.g.

(2) *gamicqale guli šeni rč'evit* (‘you annoyed me with your advice‘).

However, the collocations given in the GMD are not comprehensive, in many cases the most frequently used senses are missing, such as c) *didi guli akvs* (‘have a big heart’) in the GMD only one meaning of this collocation is given: 1) ‘being mean to someone or something’, e.g.

(3) *gogostan didi guli akvt* (‘they are mean to a girl‘).

Analysing the corpus data, it becomes quite obvious that the collocation has the other meaning opposite to the given in GMD, such as: 2) ‘being kind to someone/something’ or ‘love someone/something’, e.g.:

(4) *didi guli gvak's da usazḡvro siqvaruli šegvižlia* (‘we have a big heart with endless love’).

In order to understand and evaluate the usage of these two meanings of the collocation, we analysed the frequency of the collocation in the corpus. In 100 examples extracted for this collocation the first meaning ‘being mean to someone or something’ covered by the GMD occurs 48 times and the other meaning ‘being kind to someone/something’ or ‘love someone/something’ occurs 52 times in the corpus. Thus, we can assume that both meanings are almost equally used in the Modern Georgian language and they should be reflected in Georgian dictionaries accordingly.

2.2.1 Verb Collocates

This section will be looking at the collocations for the verb ‘to do’ with the preverb lemma form *gaaketebs* (‘will do’). As we have discussed above, in preverb forms a lemma for a verb is Future Tense, 3rd Person, Singular and in English it corresponds to infinitive ‘to do’/’to make’ etc. as it is given below accordingly. We have sorted the immediate left (See Table 4) and left and right (See Table 5) collocations of the word.

Collocation	Transliteration	Translation	Joint	Freq1	Freq2	LL score
გააკეთებს კომენტარს	gaaketebs komentars	make a comment	6024	142924	90226	11274.9
გააკეთებს განცხადებას	gaaketebs ganc'xadebas	make a statement	2196	142924	214306	2061.12
გააკეთებს ყველაფერს	gaaketebs qvelap'ers	do everything	1123	142924	234690	661.53
გააკეთებს არჩევანს	gaaketebs arčevans	make a choice	793	142924	131733	547.56
გააკეთებს აქცენტს	gaaketebs akcents	put emphasis on	265	142924	4631	472.44
გააკეთებს დასკვნას	gaaketebs daskvnas	write a summary	124	142924	14643	104.89
გააკეთებს განმარტებას	gaaketebs ganmartebas	give explanations	136	142924	22233	94.76
გააკეთებს რამეს	gaaketebs rames	do something	205	142924	77552	70.32
გააკეთებს ანალიზს	gaaketebs analizs	do analyses	72	142924	12231	48.92
გააკეთებს რაიმეს	gaaketebs raimes	do something	115	142924	46444	36.58
გააკეთებს რაღაცას	gaaketebs rağac'as	do something	160	142924	111413	21.78
გააკეთებს სიკეთეს	gaaketebs siketes	do the kind thing	31	142924	5104	21.5
გააკეთებს წარწერას	gaaketebs carceras	write something	23	142924	2264	21.42
გააკეთებს ფულს	gaaketebs p'uls	make money	116	142924	78452	16.76
გააკეთებს გათვლას	gaaketebs gatvlas	to outguess something	13	142924	957	13.9
გააკეთებს საჭმელს	gaaketebs sačmels	prepare a meal	22	142924	5112	11.92
გააკეთებს მოხსენებას	gaaketebs moxsenebas	prepare a talk	7	142924	159	11.55
გააკეთებს ბლოგს	gaaketebs blogs	start a blog	60	142924	36584	10.54
გააკეთებს ჩანაწერს	gaaketebs čanacers	keep a record	22	142924	5995	10.44
გააკეთებს დავალებას	gaaketebs davalebas	do homework	8	142924	384	10.22

Table 4: Collocates for Verb ‘to do’: The Immediate Right Neighbour.

Table 4 above illustrates the immediate right neighbour collocations for the verb ‘to do’. Here we have selected the top 20 collocates, that represent the most frequently used collocates for the word and compared them to the Georgian Monolingual Dictionary. Out of 20 top collocates in Table 4 above, only 4 are given in the GMD, such as: *gaaketebs sacmels* (‘prepare a meal’), *gaaketebs moxsenebas* (‘prepare a talk’), *gaaketebs fuls* (‘make money’) and *gaaketebs daskvnas* (‘write a summary’); the collocations not given in the GMD are: *gaaketebs komentars* (‘make a comment’), *gaaketebs arcevans* (‘make a choice’), *gaaketebs ganmartebas* (‘give explanations’), *gaaketebs* *akcents* ('put emphasis on'), gaaketebs analiz ('do analyses') etc. They are the collocations that are widely used in Modern Georgian language and these corpus-derived 'word-sketches' help us analyse the word's grammatical and collocational behaviour. It also gives us information about new senses and new words entering the language, which are the things we mostly do, e.g. *gaaketebs blogs* ('start a blog'), this collocation is not given in the GMD, because the dictionary was published in 1960s and no 'blogs' existed at that time, whereas, today 'start/run a blog' is very common thing to do and a lexicographer should reflect upon the modern usage of the language when working on the modern language dictionaries. Table 5 below gives the nearest nominal neighbour collocates of the same word.

Collocation	Transliteration	Translation	Joint	Freq2	LL score
გააკეთებს კომენტარს	gaaketebs komentars	make a comment	18016	90226	44591.79
გააკეთებს განცხადებას	gaaketebs ganc'xadebas	make a statement	11271	214306	19828.22
გააკეთებს ვიდეოს	gaaketebs videos	record a video	7912	45293	18851.06
გააკეთებს არჩევანს	gaaketebs arčevans	make a choice	2990	131733	3959.77
გააკეთებს საქმეს	gaaketebs sakmes	do the job	3519	309581	3483.41
გააკეთებს განმარტებას	gaaketebs ganmartebas	give explanations	1727	22233	3353.92
გააკეთებს აქცენტს	gaaketebs akcents	put emphasis on	613	4631	1360.68
გააკეთებს პრესკონფერენციას	gaaketebs preskonperencias	prepare a press-conference	608	17976	923.04
გააკეთებს ბრიფინგს	gaaketebs brip'ings	hold a press briefing	432	9649	716.54
გააკეთებს დასკვნას	gaaketebs daskvnas	write a summary	462	14643	685.25
გააკეთებს ანალიზს	gaaketebs analizs	do analyses	296	12231	400.03
გააკეთებს ეთერს	gaaketebs eters	prepare a TV programme/show	286	18483	323.98
გააკეთებს ფულს	gaaketebs p'uls	make money	468	78452	321.03
გააკეთებს ჩანაწერს	gaaketebs čanacers	keep a record	199	5995	300.11
გააკეთებს ბლოგს	gaaketebs blogs	start a blog	296	36584	244.2
გააკეთებს სიკეთეს	gaaketebs siketes	do the kind thing	139	5104	195.93
გააკეთებს გადაცემას	gaaketebs gadacemas	prepare a TV programme/show	244	33780	188.51
გააკეთებს საჭმელს	gaaketebs sačmels	prepare a meal	113	5112	147.64
გააკეთებს პრეზენტაციას	gaaketebs prezentacias	prepare a presentation	132	14993	114.06
გააკეთებს ოპერაციას	gaaketebs operac'ias	to have an operation	139	25558	89.54
გააკეთებს რემონტს	gaaketebs remonts	to refurbish a house	56	1455	88.59
გააკეთებს პროექტს	gaaketebs proekts	prepare a project	246	115514	64.73

Table 5: Collocates for Verb 'to do': The Nearest Nominal Neighbour.

Table 5 above gives the information about the nearest nominal neighbour collocates of the verb 'to do' and, unlike Table 4, the results are filtered according to grammatical relations in which a word occurs, in this case, we have selected a noun as the nearest nominal neighbour word. It illustrates the top 22 collocations, out of which 18 collocations are not given in the Georgian Monolingual Dictionary, for example, *gaaketebs preskonferencias* 'prepare a press-conference'; *gaaketebs gadacemas* 'prepare a TV programme/show; *gaaketebs proekts* 'prepare/write a project' etc.

The corpus analyses revealed a number of new collocations, words or word senses not represented in existing Georgian monolingual dictionaries. Firstly, this is because these dictionaries are using intuition based approaches, and, secondly, the wordlist was created in 1960s. All languages change over time. We have new technologies, new industries and new products, for example, phones or the Internet did not exist in the time of Gogebashvili or Kazbegi (authors covered in Georgian Monolingual Dictionaries). These developments simply require new words that are entering the language every day and corpora and corpus analyses help us monitor these changes and describe languages better. Corpus has a wide range of applications in linguistics, especially in lexicography, as it supports many aspects of dictionary creation, including headword list development, writing individual entries, discovering the word senses and other lexical units, the collocations they participate in (Kilgarriff 2012) etc.

This section discussed the corpus-based approach we used in writing dictionary entries and discovered new word senses and collocations not given in existing Georgian monolingual dictionaries. These are the most frequently and widely used word senses and lexical units in modern Georgian language and they should be reflected in dictionaries respectively.

3 Conclusion

We created a dictionary word-list of 10,000 headwords from KaWac, a large Georgian corpus, collected from the Web from a range of websites to provide a reliable snapshot of the use of modern Georgian. The aim of this list is to provide working material for developing several dictionaries, in particular a learner's dictionary of Georgian. We also compared both the macrostructure and microstructure of senses retrieved from KaWac to the Georgian Monolingual Dictionary. From the comparison of both lexicographic approaches, we can conclude that corpus-based approach is more efficient and it gives strong empirical evidence. The dictionary entries we have compiled using the corpus analysis are more comprehensive and represent the modern Georgian language as it is used in everyday conversations, social media, journalism and for official purposes. Besides, the corpus data provides valuable information for detecting new words or word senses that we think should be included in future Georgian dictionaries.

4 References

Daraselia, S. (2015). Issues of Compilation of New Georgian-European Learner's Dictionary Using the Corpus Methodology. PhD thesis. Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia.

Daraselia, S. & Sharoff, S. (2015). Defining Web Corpus-based Lexical Frequency for Georgian. In *Bulletin of Akaki Tsereteli State University*, N01(5), pp. 165-171.

Daraselia, S. & Sharoff, S. (2015). The Main Steps of the Georgian Web-corpus Construction. In *Journal of Linguistics of Arnold Chikobava Institute*, Volume XXXVIII, pp. 52-62.

Daraselia, S. & Sharoff, S. (2015). Error Analyses in Part-of-Speech Tagging in Georgian. In *Proceedings of the Language and Modern Technologies IV Conference, 10-15 September 2015*. Tbilisi, Georgia.

Daraselia, S. & Sharoff, S. (2014). Morphosyntactic Specifications for KaWaC, a Web
Corpus for Georgian. In *Proceedings of Humanities in the Information Society II Conference, 24-26 October*. Batumi Shota Rustaveli State University, Georgia.

Daraselia S. & Sharoff S. (2014). Towards Creating a Large Corpus for Georgian, In *Proceedings of 7th Biennial IVACS Conference*, *19-21 June,* Newcastle University, UK.

Forsyth, R. & Sharoff, S. (2014). Document dissimilarity within and across languages: a benchmarking study. In *Literary and Linguistic Computing*, 29:6-22.

Erjavec, T. (2010). MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications,
Lexicons and Corpora. In *Proceedings of Language Resources and Evaluation Conference, LREC'10*, *17-23 May 2010*, Malta.

Kilgarriff, A. (1997). Putting frequencies in the dictionary. In *International Journal of Lexicography* 10 (2), pp. 135-155.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.