http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

English–Georgian Parallel Corpus and Its Application in Georgian Lexicography

Tinatin Margalitadze, *Centre for Lexicography and Language Technologies, Ilia State University, Georgia (tinatin.margalitadze@iliauni.edu.ge)*

George Meladze, *Centre for Lexicography and Language Technologies, Ilia State University, Georgia (giorgi.meladze.4@iliauni.edu.ge)*
and
Zakharia Pourtskhvanidze, *Institute of Empirical Linguistics, University of Frankfurt, Germany (pourtskhvanidze@em.uni-frankfurt.de)*

Abstract: The Georgian language, the official language of Georgia, is the only written member of the Kartvelian language family, the indigenous language family of the Caucasus region. Georgian philology and lexicography have long-standing tradition, English–Georgian lexicography being no exception.

Given the increasing use of ample electronic text corpora for lexicographical purposes, the team of Georgian lexicographers, working on the *Comprehensive English–Georgian Dictionary* (CEGD), subsequently the *Comprehensive English–Georgian Online Dictionary* (CEGOD), decided to compile an English–Georgian Parallel Corpus (EGPC). The aim of the project was to develop the methodology of building a parallel corpus of Georgian and assess its efficiency for Georgian bilingual lexicography. The work on the corpus is going on for over a decade. The ultimate aim is to create a standard for Georgian bilingual corpora that will be compiled in future.

The article describes the content and composition of the EGPC, its structure, functionalities, search engines and so on. The article also deals with various studies conducted over years in order to assess and enhance the value, applicability and efficiency of the EGPC for the automatic or semi-automatic recognition, tagging and extraction of terminology, the compilation of terminological entries, as well as the entries for the *English–Georgian Dictionary* and those for the *Georgian–English Learner's Dictionary*, etc.

Particular emphasis is laid upon the actual or potential applicability of the corpus for the lexicographical activities and for the machine translation projects. The findings of the study may be interesting for other under-resourced languages like Georgian.

Keywords: PARALLEL CORPUS, TERMINOLOGICAL ENTRIES, ENGLISH–GEORGIAN DICTIONARY, GEORGIAN–ENGLISH DICTIONARY

Opsomming: Die Engels–Georgiese parallelle korpus en die toepassing daarvan in die Georgiese leksikografie. Georgies, die amptelike taal van Georgië, is die enigste geskrewe lid van die Kartveliaanse taalfamilie, die inheemse taalfamilie van die Kaukasiese

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

streek. Die Georgiese taalwetenskap en leksikografie het 'n lang verbintenis waarvan die Engels-Georgiese leksikografie geen uitsondering is nie.

In die lig van die toenemende gebruik van uitgebreide elektroniese tekskorpora vir leksikografiese doeleindes, het die Georgiese span leksikograwe wat aan die *Comprehensive English–Georgian Dictionary* (CEGD), later die *Comprehensive English–Georgian Online Dictionary* (CEGOD), werk, besluit om 'n Engels-Georgiese Parallelle Korpus (EGPK) saam te stel. Die doel van die projek was die ontwikkeling van die metodologie vir die bou van 'n parallelle Georgiese korpus en die bepaling van die effektiwiteit daarvan vir die Georgiese tweetalige leksikografie. Daar word al meer as 'n dekade aan die korpus gewerk. Die uiteindelike doel is om 'n standaard vir Georgiese tweetalige korpora wat in die toekoms saamgestel sal word, te skep.

Die artikel beskryf die inhoud en samestelling van die EGPK, die struktuur, funksionaliteit en soekenjins daarvan, ensovoorts. Die verskillende studies wat oor die jare uitgevoer is om die waarde, toepaslikheid en effektiwiteit van die EGPK rakende die outomatiese of semi-outomatiese herkenning, etikettering en onttrekking van terminologie, die samestelling van terminologiese inskrywings asook inskrywings vir die *English–Georgian Dictionary* en die *Georgian–English Learner's Dictionary*, ens., te bepaal en te verbeter, word in die artikel uiteengesit.

Daar word spesifiek klem gelê op die werklike of potensiële toepaslikheid van die korpus vir die leksikografiese aktiwiteite en masjienvertalingsprojekte. Die bevindings van die studie mag ook van waarde wees vir ander hulpbronskaars tale soos Georgies.

Sleutelwoorde: PARALLELLE KORPUS, TERMINOLOGIESE INSKRYWINGS, ENGELS–GEORGIESE WOORDEBOEK, GEORGIES–ENGELSE WOORDEBOEK

1. History of English–Georgian Lexicography

The English–Georgian Parallel Corpus was primarily created for the *Comprehensive English–Georgian Dictionary*, in order to enrich it with entries, corpus illustrative phrases and sentences, and terminological entries. Therefore, in this chapter we will present a brief overview of English–Georgian lexicography.

The history of English–Georgian lexicography in Georgia begins in the 20th century, although there was interest of English authors towards the Georgian and its sister languages in the 18th and the 19th centuries (Margalitadze and Tchighladze 2022; Kikvidze and Pachulia 2019; Margalitadze and Odzeli 2019).

The first English–Georgian dictionary was published in Georgia in the 1940s. The 20th century saw the publication of two comprehensive dictionaries: the *Comprehensive English–Georgian Dictionary* (editor in chief Tinatin Margalitadze) and the *Comprehensive Georgian–English Dictionary* (editor in chief Donald Rayfield).

The work on the *Comprehensive English–Georgian Dictionary* (CEGD) started in the 1970s at the department of English Philology of Ivanè Javakhishvili Tbilisi State University. In the 1980s, a small team of editors embarked upon the mission of fundamentally revising, expanding and updating the dictionary in order to prepare it for publication. In the 1990s the editorial team of the dictionary started digitalization of the dictionary material and in 1995 the printed publi-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

cation of the *Comprehensive English–Georgian Dictionary* began in fascicles, on letter-by-letter basis. In 2010, the online version of the dictionary (110 000 entries) was uploaded to the Internet (CEGOD). The primary purpose of the creation of the dictionary was to facilitate the translation of English literature (both belles-lettres or fiction and specialist literature) into Georgian. This is why the dictionary includes contemporary English vocabulary, as well as obsolete, archaic words and meanings and specialist terms (Margalitadze 2012).

The *Comprehensive Georgian–English Dictionary* under editorship of Donald Rayfield was published in London in 2006 by Garnett Press (CGED). Donald Rayfield is an outstanding British Slavist and Kartvelologist. He is the author of a number of monographs on the Russian and Georgian literature. He is also a skilful translator, translating pieces of Georgian literature into English. The *Comprehensive Georgian–English Dictionary* includes contemporary, as well as Old Georgian vocabulary, the word-stock of the Georgian dialects and related Kartvelian languages, and terms from specific branches of knowledge. Donald Rayfield's dictionary contains 140 000 Georgian words and is published in two volumes.

2. English–Georgian Parallel Corpora

There are several English–Georgian parallel corpora, which were mainly developed in the context of multilingual data mining through the Web and have been processed in different ways. Three corpora are presented in this chapter as examples: CCAligned v1, CCAligned v1 and TED2020 v1. The first two are among the largest corpora in number of Georgian data, while the third parallel corpus contains translations of spoken Georgian.

CCAligned v1,[1] "A Massive Collection of Cross-lingual Web-Document Pairs" consists of parallel or comparable web-document pairs in 137 languages aligned with English. The analysis of the automatically translated English–Georgian sentence pairs reveals massive problems of alignment and translation in the Georgian part of the corpus.

Wikimedia v20210402. Wikipedia translations are published by the Wikimedia foundation and their translation system[2] (Tiedemann 2012). The WiKi-Parallel corpus contains 306 languages, including Georgian. The total number of tokens is 918.05M and total number of sentence fragments — 31.62M.

TED2020 v1.[3] This parallel corpus is interesting as it represents a spoken language and was translated by volunteers. This dataset contains a crawl of nearly 4000 TED and TED-X transcripts from July 2020 (Reimers and Gurevych 2020). The transcripts have been translated to more than 100 languages by a global community of volunteers. The parallel corpus contains 108 languages, including Georgian. The total number of tokens — 173.40M, total number of sentence fragments — 11.46M.

The study of above-mentioned, as well as other parallel corpora with the Georgian language reveals that the web-based and automatically created par-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

allel corpora have a high rate of linguistic and formatting errors of all types, particularly in a language like Georgian, which is characterized by a complex morphology (Gippert 2016; Harris 1991). For example, the whole parallel corpus of 62 languages — OpenSubtitles (Lison and Tiedemann 2016) is completely unusable for Georgian due to the formatting and coding errors.

2.1 English–Georgian Parallel Corpus of Ilia State University

The work on the EGPC started in 2011. The corpus consists of two sub-corpora: the sub-corpus of scientific and domain-specific texts and the sub-corpus of fiction (translated from Georgian into English and vice versa). From the very beginning of the project the decision was made to concentrate on the quality of translated texts, as well as the structuring of the data in it, as the primary goal of developing the EGPC was its application in English–Georgian lexicography.

The most important part of the sub-corpus of scientific texts constitute translations of professor Arrian Tchanturia, a prominent Georgian scholar, editor, translator and lexicographer (member of editorial boards of both comprehensive dictionaries: English–Georgian and Georgian–English). He was one of the first scholars to start translation of Georgian scholarly and scientific literature into English from the 1960s. His translation legacy includes hundreds of pages of translated abstracts, papers, and books from Georgian into English covering practically all fields of knowledge. The desire to transform this legacy into an English–Georgian Parallel Corpus and to apply it in the work on the CEGD gave the impetus to the development of this project (Margalitadze 2014). Later this sub-corpus was extended with other translations and grew into a sub-corpus of scientific and domain-specific texts. At the next stage, translations of literary works were added to the corpus.

2.2 The Structure of the English–Georgian Parallel Corpus

The principles of arrangement of data in the corpus databases were worked out after a long period of deliberation and aimed at the arrangement of texts in databases in a way that would enable the application of the corpus in general and specialized lexicography in future. The platform is based on the program created for the English–Hungarian parallel corpus 'HunAlign freeware tool'.[4]

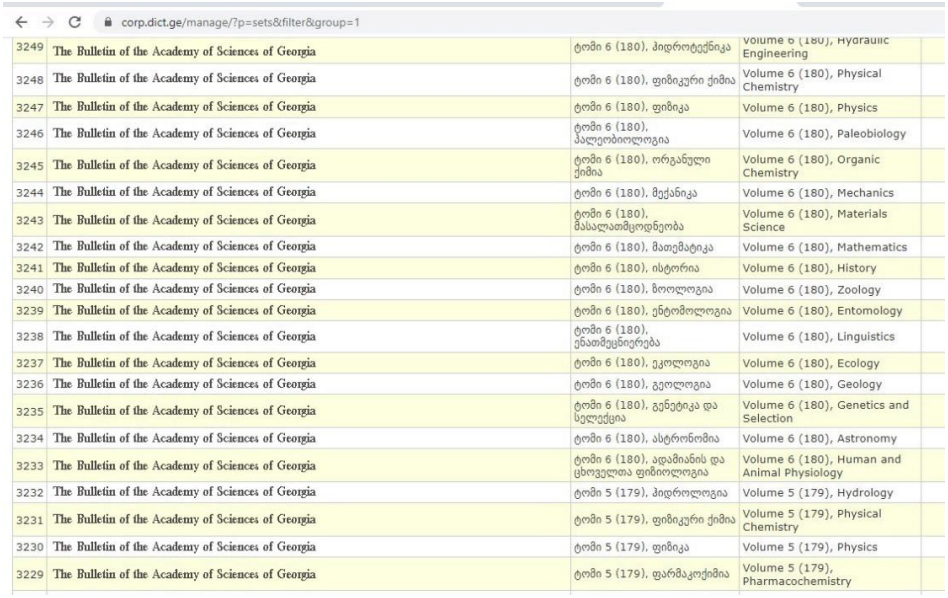
The structure of the database consists of three sections: text groups, text sets and sentence pairs. Each text group is subdivided into text sets and each text set is further subdivided into sentence pairs. Text group is the largest unit of the database and it consists of a variety of texts. At the present moment the EGPC comprises over 70 text groups of different sizes and new material is added to the corpus on a daily basis.

One of the largest text groups in the sub-corpus of scientific texts is *The Bulletin of the Academy of Sciences of Georgia*. It incorporates material from issues

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

published over a period of 24 years. This material consists of English–Georgian abstracts of scientific papers from virtually all fields of knowledge. This subcorpus also includes scholarly bilingual papers published in several bilingual scholarly journals in Georgia, e.g. Kartvelology and Kadmos. One of the text groups represents a series of publications about important archaeological excavations in Georgia. Text groups also include scholarly books, manuals of different subjects translated from English into Georgian, materials published by the Legislative Herald of Georgia, election administration, the Government of Georgia, and materials collected from different websites.

Each text group, as mentioned above, is subdivided into text sets. Text sets vary according to the type of the text group. E.g., the text group *The Bulletin of the Academy of Sciences of Georgia* is divided into volumes (with each volume containing three issues) and each volume (text set) contains abstracts of one domain: volume 6 (180) ecology; volume 6 (180) entomology; volume 6 (180) geology; volume 6 (180) human and animal physiology; volume 6 (180) mechanics; volume 6 (180) organic chemistry, etc. (see Figure 1).


corp.dict.ge/manage/?p=sets&filter&group=1

3249	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ჰიდროტექნიკა	Volume 6 (180), Hydraulic Engineering
3248	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ფიზიკური ქიმია	Volume 6 (180), Physical Chemistry
3247	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ფიზიკა	Volume 6 (180), Physics
3246	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), პალეობიოლოგია	Volume 6 (180), Paleobiology
3245	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ორგანული ქიმია	Volume 6 (180), Organic Chemistry
3244	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), მექანიკა	Volume 6 (180), Mechanics
3243	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), მასალათმცოდნეობა	Volume 6 (180), Materials Science
3242	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), მათემატიკა	Volume 6 (180), Mathematics
3241	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ისტორია	Volume 6 (180), History
3240	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ზოოლოგია	Volume 6 (180), Zoology
3239	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ენტომოლოგია	Volume 6 (180), Entomology
3238	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ენათმეცნიერება	Volume 6 (180), Linguistics
3237	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ეკოლოგია	Volume 6 (180), Ecology
3236	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), გეოლოგია	Volume 6 (180), Geology
3235	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), გენეტიკა და სელექცია	Volume 6 (180), Genetics and Selection
3234	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ასტრონომია	Volume 6 (180), Astronomy
3233	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ადამიანის და ცხოველთა ფიზიოლოგია	Volume 6 (180), Human and Animal Physiology
3232	The Bulletin of the Academy of Sciences of Georgia	ტომი 5 (179), ჰიდროლოგია	Volume 5 (179), Hydrology
3231	The Bulletin of the Academy of Sciences of Georgia	ტომი 5 (179), ფიზიკური ქიმია	Volume 5 (179), Physical Chemistry
3230	The Bulletin of the Academy of Sciences of Georgia	ტომი 5 (179), ფიზიკა	Volume 5 (179), Physics
3229	The Bulletin of the Academy of Sciences of Georgia	ტომი 5 (179), ფარმაკოქიმია	Volume 5 (179), Pharmacochemistry

Figure 1

Other text groups are structured differently. Scientific and scholarly journals are divided into text sets according to separate articles; books are divided into chapters and so on. Such organization of the database allows the sorting of the material according to domains as well as many other criteria.

Text sets are further subdivided into sentence pairs. These are aligned English–Georgian parallel sentences (see Figure 2).

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

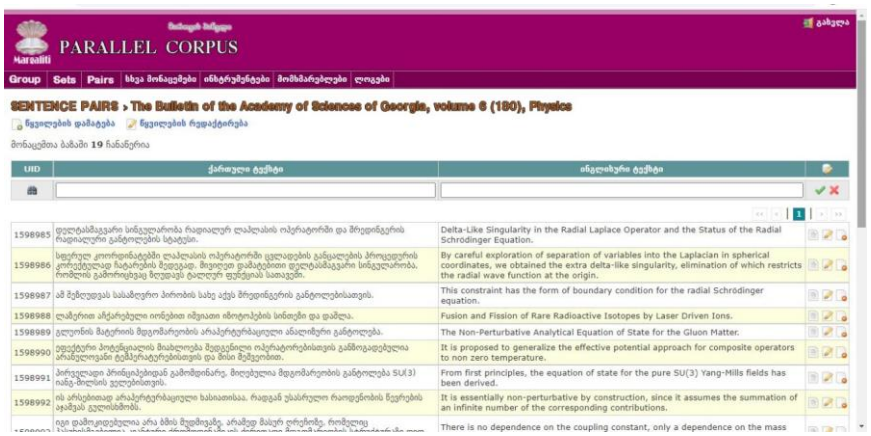


Figure 2

Text sets are uploaded to the special fields in the database, allocated to English and Georgian.

The program automatically breaks down text sets into sentence pairs (see Figure 3).

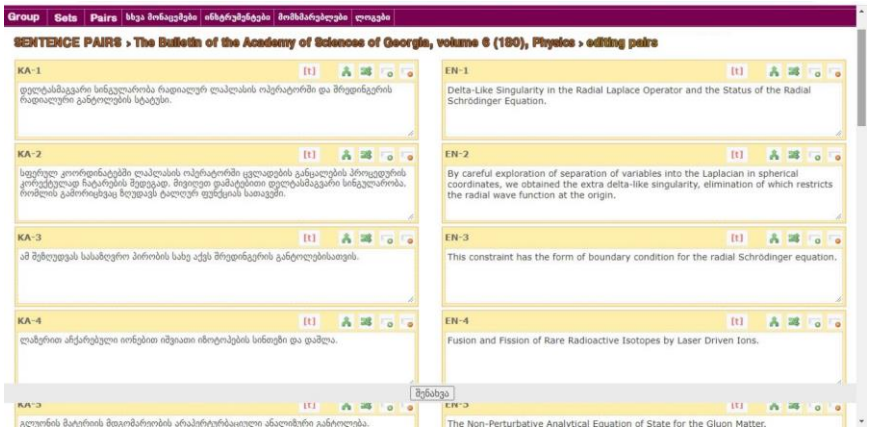


Figure 3

At the next stage, the sentences broken down automatically are manually aligned with the help of tools provided at the top right corner of each block. These tools allow one to add or delete blocks or to exchange places between two blocks. Manual alignment usually corrects minor errors, e.g. cases when one English sentence is translated by two Georgian sentences or vice versa. The result of this approach is high-quality, ideally aligned sentence pairs.

Texts uploaded to the sub-corpus of scientific texts comprise all fields of knowledge: mathematics, mechanics, geophysics, chemistry, hydrology, geol-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

ogy, palaeontology, machine building science, hydraulic engineering, electrical engineering, botany, genetics, physiology, biophysics, biochemistry, entomology, experimental morphology, experimental medicine, financing, archaeology, ethnography, Kartvelology etc. The sub-corpus of fiction contains translations of Georgian belles-lettres into English, as well as translations of English authors into Georgian. The sub-corpus of fiction also includes translations of plays.

At present, the corpus contains up to 70 text groups, 5 000 text sets, 400 000 manually aligned sentence pairs and 7 million tokens. The EGPC has an interface for searching Georgian or English words and collocations and displaying the proper text pairs containing the search results on the screen. Each sentence pair is numbered and is supplied with the information about corresponding text group and text set (see Figure 4).

Thus, unlike the English–Georgian parallel corpora, discussed in chapter 2, the EGPC of Ilia State University is characterized by the following features:

(1) high-quality translations edited by human specialists,
(2) accurate and error-free alignment of sentences, and
(3) constantly growing corpus through parallel use of human specialists and NLP.

On all three points, the *Comprehensive English–Georgian Dictionary* acts as a lexicographic source of the translation quality.

When the corpus reached 4 million tokens, studies were conducted for evaluating the efficiency of the Corpus for English–Georgian Lexicography. Three main tasks were identified for the EGPC: compiling terminological entries, compiling entries for the English–Georgian Dictionary and compiling entries for the Georgian–English Learner's Dictionary. These studies were carried out within the framework of MA and PhD programmes in lexicography with the active participation of MA and PhD students in lexicography.

Figure 4

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

2.3 Application of the English–Georgian Parallel Corpus in Terminology

The work on the elaboration of the methodology of tagging and extracting specialized terminology from the corpus started in 2015. A special module, the terminological module, was developed that allows the extraction of the previously tagged terminology from the corpus. After the development of this module, the function "Recognition of and search for the tagged terms in the corpus" was added to the existing functions of the corpus control panel, namely:

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

An advanced search function was added to the simple search functionality of the EGPC. Figure 5 shows the advanced search page which displays all fields of knowledge represented by texts of different sizes in the EGPC: aviation, archaeology, architecture, oriental studies, botany, zoology, biology, geology, ecology, ethnography, economics, banking, history, Kartvelian studies, hydrology, psychology and many others. The principles of the arrangement of corpus databases into text groups and text sets, described above, allow one to sort terminology according to domains and to extract them from the corpus for further lexicographic processing. Specialized terms are extracted from the corpus alongside their English equivalents and, significantly, collocations of terms with their respective English translations can also be extracted.

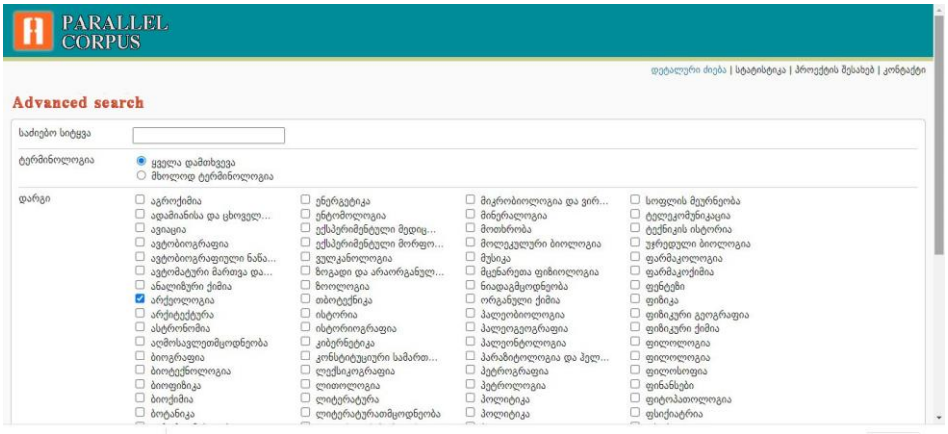


Figure 5

The analysis of terminological entries created on the basis of the EGPC revealed that the corpus is a very efficient source for the CEGOD and that it can enrich

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

the dictionary with terminology of different domains. Two cases are to be noted: some terms were not recorded in the CEGOD and were added to it from the corpus, and in some cases terminological entries of the CEGOD were improved by adding new collocations to them. For example, the financial term *direct debit* was introduced in the CEGOD with the following collocations and their Georgian translations: *direct debit order, direct debit service, direct debit transfer*. The financial terms *documentary collection* and *encashment order* were added to the dictionary macrostructure. The economic term *inflation* had been already included in the CEGOD, but the corpus material enabled the addition of the following collocations: *high inflation, the rate of inflation, high rate of inflation, a period of inflation, demand-pull inflation, cost-push inflation, to reduce the threat of inflation.* These collocations are supplied with Georgian translations from the corpus. The following collocations and their Georgian equivalents were added to the economic term *cost*: *production costs, operating costs, fixed costs, variable costs, to increase/raise costs, to reduce costs, to cut costs, rising costs, marginal costs, external costs, shipping costs, refining costs, to incur costs*.

The EGPC can also be applied in English–Georgian terminological dictionary projects, but only as one of the sources. It is unlikely to have enough translations of specialized texts in one domain to fully rely only on the parallel corpus while compiling a bilingual dictionary of one field of knowledge.

One of the recent studies conducted in the EGPC was the testing of different tools for automatic or semi-automatic recognition, tagging and extraction of terminology from the parallel corpus. Different tools were tested for this purpose, but the most efficient one proved to be *Synchroterm*, developed by a Canadian computer program company Terminotix.[5] The study will continue in this direction and the selected program will be integrated with the EGPC in order to facilitate work on the terminology.

2.4 Application of the English–Georgian Parallel Corpus for Georgian–English Learner's Dictionary

Compilation of Georgian–English Learner's Dictionary (GELD) is high on the agenda of the Centre for Lexicography and Language Technologies. The *Comprehensive Georgian–English Dictionary*, published under the general editorship of D. Rayfield, is mostly aimed at foreign scholars interested in Georgian and its sister languages, mediaeval Georgian literature, and the history of Georgia in the Middle Ages, when this country played an important role in European history. Proceeding from these considerations, the macrostructure of the dictionary includes Old and Middle Georgian words and dialectal material, which is important for the main target group of the CGED. The dictionary is more concerned with the macrostructure, reflected in the number of entries (140 000).

On the other hand, Georgian learners of English need more information about the usage of Georgian words and their rendition in English. In other words, they need a dictionary which is oriented on text synthesis, text produc-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

tion, speaking/writing and not only text analysis, i.e. understanding spoken/written text. Our decades-long experience of working on the CEGD has revealed that there is considerable semantic asymmetry between the English and Georgian languages. As a result, an English word cannot always be translated by one Georgian equivalent in various contexts and often needs different contextual equivalents to properly translate its meaning. In the CEGD our editorial team introduced two levels of equivalence in an entry: meaning equivalence and contextual/translation equivalence, which is discussed in detail in our paper presented at the XVII International Congress of EURALEX (Margalitadze and Meladze 2016). Therefore, illustrative phrases and sentences, which show the usage of an English word and its Georgian translations, are important in the CEGD entries. This is also true for the reverse Georgian–English dictionary: Georgian words should be supplied with different illustrative phrases, sentences and collocations translated into English. These considerations determined our interest in the EGPC and its efficiency for the GELD project.

The study of the effectiveness of the EGPC for the compilation of the GELD entries yielded very positive results. In many cases, the data collected from the corpus enabled editors to produce adequate dictionary entries and to identify and single out polysemous meanings of Georgian words, sometimes even more meanings than are registered in monolingual dictionaries of Georgian. The corpus data provides many illustrative phrases, collocations and sentences for Georgian words with their respective English equivalents.

For example, for the Georgian word *მტკიცე mṭḳice* two polysemous meanings are identified and each meaning is well-illustrated with the corpus examples:

> *მტკიცე mṭḳice* **1.** (*firm, solid, steady*) მტკიცე ავეჯი solid furniture; მტკიცე ქიმიური ბმები firm chemical bonds; ფანჯარა ძალიან მტკიცე მინისგან არის დამზადებული the window is made from very strong glass; განა შეიძლება შედეგი მტკიცე იყოს? Can the result be sound?; მტკიცე ფეხსაცმელი durable shoes; მტკიცე ნივთიერება enduring substance; მტკიცე და დაუძლეველი ზღუდე fast and impassable barrier; მტკიცე კარები a solid door; **2.** (*determined, decisive, resolute*) მტკიცე ხასიათი decisive character; მტკიცე ტრადიცია deep-seated tradition; მტკიცე ფასი determined price; მტკიცე ნების ადამიანი a man of hard, unbending will; მტკიცე ნებისყოფის ქონა to have an iron will; მტკიცე ოპტიმისტი a resolute optimist; მტკიცე ბიუროკრატიული კონტროლი rigid bureaucratic controls; მონარქიის მტკიცე მხარდამჭერი a staunch supporter of the monarchy; მტკიცე ნაბიჯებით with sure steps.

The corresponding entry from D. Rayfield's CEGD presents the same Georgian word in the following way:

1. Solid, firm; established; მტკიცე ნაბიჯი a decisive step; მტკიცე უარი a definite no;
2. Of good cheer (*this is an obsolete meaning of this adjective which will not be presented in a learner's dictionary*).

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

The English language abounds in synonyms. For a Georgian learner of English, it is important to know which synonym should be used in a particular context. From this point of view the EGPC provides really useful and important data about usage of Georgian words and, even more important, their translations into English.

For the Georgian verb *დაფარვა daparva* the corpus data singles out four meanings:

> *დაფარვა daparva* **1.** (*to cover*) მტვრით ხარ დაფარული you are covered in dust; მიწა თოვლით იყო დაფარული the ground was blanketed with snow; **2.** (*to keep secret, to conceal*) შიშის [მღელვარების, ნერვიულობის] დაფარვა to conceal one's fear [excitement, nervousness]; სიმართლის დაფარვა to hide the truth; მტრული დამოკიდებულების დაფარვა მეგობრობის ნიღბით to mask one's enmity under an appearance of friendliness; **3.** (*to pay debt, to compensate*) სესხის დაფარვა to pay a loan; მან ვალი დაფარა he wiped out the debt; **4.** (*to protect, to defend*) სამშობლოს მტრისგან დაფარვა to defend one's homeland from an enemy; თვალების მზისგან დაფარვა to protect one's eyes from the sun; ◊ **დაფარვის ზონა** coverage area.

The corresponding entry from D. Rayfield's CGED presents the same four meanings without providing examples of usage:

1. Covering (*with snow, clothes*); დაფარვის ზონა (*mobile phone, etc.*) coverage area;
2. Keeping hidden;
3. Paying off (debt);
4. Defence.

At present, the work is underway on the issues connected with the automation of data collection from the corpus in order to facilitate the work of lexicographers.

2.5 Application of the English–Georgian Parallel Corpus for the Comprehensive English–Georgian Dictionary

Further studies included the assessment of the corpus's efficacy for the *Comprehensive English–Georgian Dictionary*. Our aim was to assess the volume and representativeness of the EGPC by means of looking up and retrieving corpus data with respect to some pre-selected lexical units. This would enable us to find out to what extent the polysemy of these words was traceable in the parallel English–Georgian sentences represented in the corpus, and how helpful the data retrievable from the corpus could be for the composition of more or less full-fledged dictionary articles.

To that end, we chose a number of nouns, verbs, adjectives and adverbs.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

Context-based meanings retrieved from the database permitted the composition of dictionary entries with some considerable scope of polysemy.

Before proceeding to general conclusions, we would like to demonstrate the material with respect to the lexical unit *dream* (noun + verb) that was extracted from the corpus. This article is a characteristic example of dictionary articles based on the data retrieved from the EGPC:

> dream *noun* **1.** (*a vision during sleep*) სიზმარი; for a long time, I could not shut my eyes and, when I did get to sleep, I was transported by dreams დიდხანს თვალი ვერ დავხუჭე, და, რომ დამეძინა, სიზმრებმა წამიღეს; **2.** (*an aspiration, a wish to have or be something*) ოცნება; his entire poetry clearly expresses the dreams and aspirations of the Georgian people მთელი მისი პოეზია ქართველი ხალხის ოცნებებისა და მისწრაფებების ნათლად გამომხატველია; **3.** (*daydream, reverie*) ზმანება; the tender, sweet dream of a love seen once ოდესღაც ნანახი სატრფოს ნაზი, ტკბილი ზმანება; now he could know that this had truly happened and was not a dream ახლა საბოლოოდ დარწმუნდა, რომ ეს ყველაფერი ზმანება კი არა, ცხადი იყო; life is a dream სიცოცხლე ზმანებაა.
>
> dream *verb* **1.** (*to experience a dream during sleep*) დასიზმრება (<და>ესიზმრება); "is this the man I dreamt of?" she worried "ნუთუ ის კაცია, ვინც დამესიზმრაო" - წუხდა ქალი; **2.** (*to have a deep aspiration or hope*) ოცნება (ოცნებობს); the point is that many crusaders dreamed of seizing lands and becoming rich საქმე ისაა, რომ ბევრი ჯვაროსანი მიწების ხელში ჩაგდებასა და გამდიდრებაზე ოცნებობდა; he dreams of creating a library and setting up a printing press იგი ოცნებობს ბიბლიოთეკის შექმნასა და სტამბის დაარსებაზე; **3.** (*to daydream, to pass time in reverie*) ხილვის / ზმანების ქონა (აქვს); *რაიმე* ეზმანება; he only dreamed of foreign lands now and of the lions on the beach მას ახლა მხოლოდ უცხო მხარე და სანაპიროზე გამოფენილი ლომები ეზმანებოდა; **4.** (*to regard something as feasible or practical, to imagine*) უარყოფით წინადადებებში: ფიქრი (ფიქრობს), განზრახვა; the French will never dream of it ფრანგებს ეს არც დაესიზმრებათ; "I could never dream of such success in my own country," she admitted frankly "ჩემს სამშობლოში ამგვარი წარმატება არც კი დამესიზმრებოდაო" - აღიარა მან გულწრფელად.

The above entries (DREAM *noun* + *verb*) provide some interesting information about the subject under discussion. Comparing these entries with those included in the *Comprehensive English–Georgian Dictionary* (https://dictionary.ge/ka/word/dream+I/ and https://dictionary.ge/ka/word/ dream+II/) we could see that many polysemous meanings present in the entries of CEGD can be seen in corpus-based entries as well. Moreover, the third verbal meaning '*to daydream, to pass time in reverie*', is absent in the CEGD, while the same meaning could be identified based on the contexts attested in the parallel sentences retrieved from the corpus.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

On the other hand, some meanings, e.g. '*to dream up*' (to invent, concoct) which is included in the entry of the *Comprehensive English–Georgian Dictionary*, is absent from our corpus-based entry, as far as no sentences/contexts, where 'to dream (up)' would denote 'inventing or concocting something', could have been retrieved from the EGPC.

Meanwhile, the further analysis of the dictionary entries, composed using the data retrieved from the corpus, showed that some meanings of polysemous words had more hits in the corpus, while other ones were very scarce and only few occurrences thereof could be attested in the corpus database. For instance, in the case of the adjective *short*, we obtained many contexts, where *short* meant 'not lengthy', 'of short duration' or 'deficient in something' or 'lacking something', but (somewhat surprisingly), there were very few cases were *short* meant 'not long', and only one case where *short* referred to the human stature (i.e., meaning 'not high or tall'). Only one result for *short* with its semantic value referring to vowel shortness *v* length (in prosody and phonetics) came as no surprise, while the scarceness of the contexts with *short* meaning 'not long' or 'not high/tall' required some explanation. Our best guess is that a relatively large proportion of purely scientific or official texts in our corpus (*The Bulletin of the Academy of Sciences of Georgia*, legislative documents, texts related to the economic, financial and banking activities, etc.) may somehow account for the relatively scarce representation of words (*short* in this particular case) with semantic values related to everyday life and 'ordinary' situational contexts.

To summarize, we can state that our investigation has allowed us to arrive at certain conclusions. Since Georgian, as a language, is under-resourced and lacks large amounts of parallel Georgian–English texts, we cannot expect the EGPC to yield data for comprehensive dictionaries with full-size entries based on extensive polysemy. Furthermore, since approximately two thirds of the texts included in our corpus are those translated from Georgian into English, the application of the corpus-based data extracted from the corpus seems to be more appropriate for *Georgian–English Learner's Dictionary* project. It should be also mentioned that even at the present stage, the corpus proves to be very useful source for enriching the CEGD entries with additional senses or good dictionary examples. This study also showed that the development of the corpus should concentrate on texts translated from English into Georgian to provide balance and have an equal proportion of texts translated from Georgian into English and vice versa. The corpus also needs to be balanced by including more translations of literary works as opposed to translations of scientific and official texts.

3. Application of the English–Georgian Parallel Corpus for English–Georgian/Georgian–English Machine Translation Project

In 2018 our editorial team realized that we possessed the data that could be instrumental in Georgian–English/English–Georgian machine translation project (Margalitadze and Pourtskhvanidze 2019). Such a project needs: (a) a col-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

lection of software platforms and models adapted to the specifics of the Georgian language, and (b) professionally translated English–Georgian parallel sentences in the quantities and amounts as necessary to ensure quality saturation.

As a software prototype for the project, researches based on the simulation of human abilities within the framework of Artificial Intelligence were selected. DeepLearning technology has demonstrated many successful examples of becoming the leading technology and methodological framework. Out of effective models implemented within this framework, machine translation is one of the three most successful examples.

Concerning English–Georgian parallel sentences, our team possesses a database unique for the Georgian language. The base includes two sub-components: the database of the *Comprehensive English–Georgian Dictionary* mentioned above (chapter 1), and the base of the English–Georgian Parallel Corpus, discussed in Chapters 2.1 and 2.2.

For the machine translation project some additional studies were conducted on the corpus in order to evaluate it from the point of view of lexical richness (Kubát and Milička 2013; Brezina 2018). Due to its limitations in terms of digital resources, Georgian needs qualitative processing of data alongside proper structuring of databases. Balancing text types or genres is one such effort. Linguistic diversity in the corpus is represented on the basis of the lexical diversity of its components. The value of lexical diversity was obtained by automatically calculating type-token ratios (TTR) in a text. A clustered calculation for the whole corpus provided the overall picture of equal or unequal distribution of TTR values in the corpus, showing gaps in terms of the balance. Further development of the corpus will take the TTR values into account in the selection of text collections (Margalitadze and Pourtskhvanidze 2021).

At the present moment, the initial stage of the data training for machine translation is over and we are in the process of analysing the first results of the English–Georgian/Georgian–English machine translation program.[6] The training was conducted with 367 000 English–Georgian sentence pairs in which 267 000 pairs were from the EGPC and 100 000 from the CEGD. The data was trained in the OpenNMT model.[7] Although our aim is to reach up to 1 million sentence pairs, the results of this initial stage are very promising. The program has learnt even very specific vocabulary quite well, and deals particularly well with collocations.[8] From this point of view, our machine translation program, in some cases, provides more accurate translations from Georgian into English, than Google translate, which is based on the 1.3 million English–Georgian sentence pairs.[9] Below are quoted some examples which illustrate the difference in the English translations of Georgian sentences by the Google translate and our translator:

1. ღორების კოლტი ზღვაში გადავარდა:
   ġorebis ḳolṭi zġvaši gadavarda
   The Google translate: The pig colt fell into the sea.
   Our translator: A herd of swine fell into the sea.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

2. მგლების ხროვა მას ყოველი მხრიდან უტევდა:
   mglebis xrova mas ǫoveli mxridan uṭevda
   The Google translate: A herd of wolves attacked him from all sides.
   Our translator: A pack of wolves was attacking him from all sides.
3. არწივი ცაში ლივლივებდა:
   arc̣ivi caši livlivebda
   The Google translate: The eagle was flying in the sky.
   Our translator: The eagle was soaring in the sky.
4. მდინარე ტყეში მორაკრაკებდა:
   mdinare ṭǫeši morak̦rak̦ebda
   The Google translate: The river was flowing in the forest.
   Our translator: The river bubbled in the forest.
5. ფარდები ქარში ფრიალებდა:
   pardebi karši prialebda
   The Google translate: The curtains were flying in the wind.
   Our translator: Curtains fluttered in the wind.
6. ჩიტების გუნდი ერთად მიფრინავდა:
   čiṭebis gundi ertad miprinavda
   The Google translate: A team of birds flew together.
   Our translator: A flock of birds flew together.
   (see Figure 6).

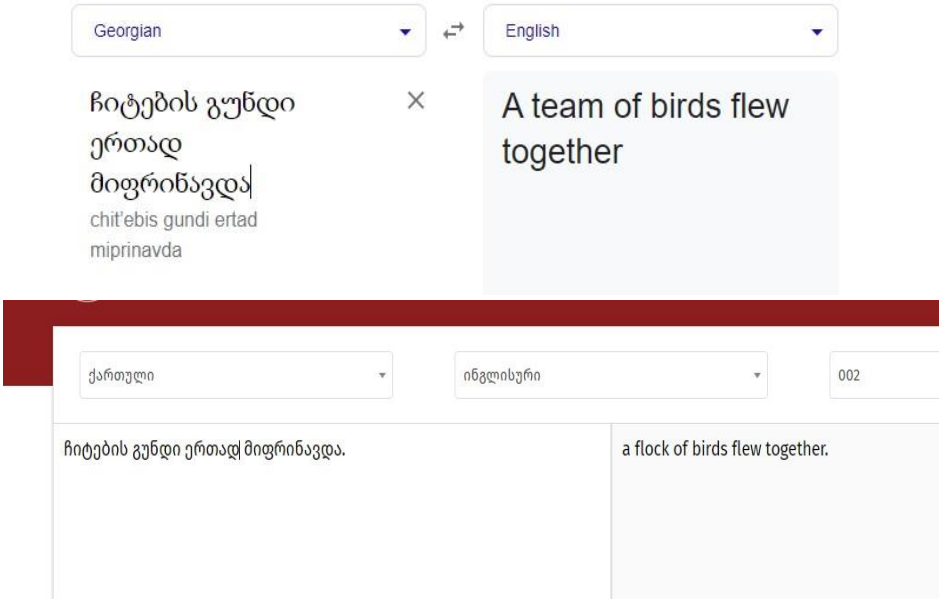


Figure 6

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

4. Conclusion

As described in above chapters, various studies were conducted in order to evaluate the applicability and efficiency of the English–Georgian Parallel Corpus (EGPC) for lexicographical and machine translation projects. These are: (a) the analysis of terminological entries created on the basis of the EGPC, which revealed that the corpus can be a very efficient source for the *Comprehensive English–Georgian Online Dictionary* (CEGOD), enriching the dictionary with terms from different domains; (b) the studies conducted in the EGPC with different tools for automatic or semi-automatic recognition, tagging and extraction of terminology from the corpus; (c) the studies intended to identify the value of the EGPC for compiling entries for *English–Georgian Dictionary* and entries for *Georgian–English Learner's Dictionary*; and (d) the studies for testing the efficacy of the EGPC for machine translation.

The wide range of research activities described above highlight the importance of well-balanced parallel corpora based on adequate, high-quality translations and thoughtfully and meticulously structured data for modern bilingual lexicography. These studies encouraged us to continue the work on the EGPC. The project will develop both quantitatively and qualitatively. From the quantitative point of view the aim is to reach up to 1 million English–Georgian sentence pairs within one year, although the work on the corpus will continue even after achieving this goal. On the other hand, we will continue testing different methods and tools for automating data collection from the corpus. The development of the EGPC will also refer to two main points of the use level: (1) the search tools that allow more granular searches and (2) the analysis tools that can structure extracted data according to different analysis criteria such as frequency, co-occurrence, word embedding, etc. This development sets up a possible move of the corpus to a new user environment.

One more direction in the development of the EGPC is adding new fields to it for other parallel corpora of Georgian with other languages. These corpora will be created and different bilingual projects will be implemented under the supervision and in cooperation with the Centre for Lexicography and Language Technologies at Ilia State University, including the framework of MA and PhD programs in lexicography at the University.

Thus our studies have revealed that parallel corpora are very useful tools for bilingual lexicography. Under-resourced languages like Georgian can balance lack of a large number of translated texts for parallel corpora by concentrating on the quality and data structure of the corpus and the lexical richness of text types and genres. It should be noted that balancing of a corpus concerns not only text genres (scientific, fiction, media), but also balanced amount of translations from a source language into a target language and vice versa. Such corpora can be conducive for compiling bilingual dictionaries, for enriching existing dictionaries with new terms, word meanings and illustrative collocations. Our study has also revealed the efficacy of high quality data of parallel

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

sentences for machine translation, achieving positive results with much less data than are required by "resource-hungry" algorithms from the field of the NLP.

The methodology and the platform of a parallel corpus, created by our team, can also be used for the composition of parallel corpora in the languages other than English and Georgian.

Endnotes

1. https://opus.nlpl.eu/CCAligned/v1/en-ka_sample.html [Accessed 20.04.2022]
2. https://dumps.wikimedia.org/other/contenttranslation [Accessed 20.04.2022].
3. https://www.ted.com/participate/translate [Accessed 20.04.2022].
4. https://github.com/danielvarga/hunalign/2
5. https://terminotix.com/index.asp?name=SynchroTerm&content=item&brand=4&item=7&lang=en https://terminotix.com/index.asp?lang=en
6. The partner of Ilia State University in this project is Vakhtang Elerdashvili, a data scientist, a PhD Student at Text Technology Lab, Goethe-University Frankfurt, Germany, https://www.texttechnologylab.org/, the author of the Georgian spellchecker (https://spellchecker.ge/).
7. https://opennmt.net/
8. At present the testing of the program is underway in a closed intranet with the access only for the members of the working team.
9. https://www.google.com/search?q=google.translate+english+to+georgian&oq=google&aqs=chrome.2.69i60j46i67i131i199i433i465j35i39j69i60l4j69i65.4480j0j7&sourceid=chrome&ie=UTF-8 [Accessed 27.04.2022].

References

Dictionaries

[CEGD] ***Comprehensive English–Georgian Dictionary.*** Vol. I–XIV. (Editor-in-chief T. Margalitadze). 1995–2012. Tbilisi: Ivane Javakhishvili Tbilisi State University, Lexicographic Centre.

[CEGOD] ***Comprehensive English–Georgian Online Dictionary.*** (Editor-in-chief T. Margalitadze). 2010. Tbilisi: Ivane Javakhishvili Tbilisi State University, Lexicographic Centre.
Available at: www.dict.ge

[CGED] ***A Comprehensive Georgian–English Dictionary.*** 2 volumes. (Editor-in-chief D. Rayfield). 2006. London: Garnett Press.
Available at: http://www.nplg.gov.ge/gwdict/index.php?a=index&d=46

[EGPC] ***English–Georgian Parallel Corpus.*** Tbilisi: Ilia State University.
Available at: http://corp.dict.ge

Other references

Brezina, V. 2018. *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge: Cambridge University Press.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

Gippert, J. 2016. Complex Morphology and its Impact on Lexicology: The Kartvelian Case. Margalitadze, T. and G. Meladze (Eds.). 2016. *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*, *6–10 September, 2016, Tbilisi, Georgia*: 16-36. Tbilisi: Ivane Javakhishvili Tbilisi State University.
Available at: http://euralex2016.tsu.ge/publication2016.pdf

Harris, A.C. (Ed.). 1991. *The Indigenous Languages of the Caucasus*: *Kartvelian. Vol. I.* Delmar, N.Y.: Caravan Press.

Kikvidze, Z. and L. Pachulia. 2019. Demetrius Rudolph Peacock and the Languages of Georgia. *General and Specialist Translation/Interpretation: Theory, Methods, Practice: International Conference Papers:* 15-22. Kyiv: Agrar Media Group.

Kubát, M. and J. Milička. 2013. Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics* 20(4): 339-349.

Lison, P. and J. Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. Calzolari, N. et al. (Eds.). 2016. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), May 23–28, 2016, Portorož, Slovenia:* 923-929. Paris: European language Resources Association (ELRA).

Margalitadze, T. 2012. The Comprehensive English–Georgian Online Dictionary: Methods, Principles, Modern Technologies. Fjeld, R.V. and J.M. Torjusen (Eds.). 2012. *Proceedings of the 15th EURALEX International Congress, 7–11 August 2012, Oslo*: 764-770. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.
Available at: http://euralex.org/category/publications/euralex-oslo-2012/

Margalitadze, T. 2014. European-Georgian Parallel Corpora for Georgian Lexicography and Translatology. *Proceedings of the International Conference 'Literary Translation — A Meeting Place for Nations and Literatures'*. Dedicated to the 100th Anniversary of a Translator, Poet and Theoretician of Literary Translation Givi Gachechiladze. Tbilisi: Ivane Javakhishvili Tbilisi State University.

Margalitadze, T. and G. Meladze. 2016. Importance of the Issue of Partial Equivalence for Bilingual Lexicography and Language Teaching. Margalitadze, T. and G. Meladze (Eds.). 2016. *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*, *Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia, 6–10 September, 2016*: 787-797. Tbilisi: Ivane Javakhishvili Tbilisi State University.
Available at: http://euralex.org/category/ publications/euralex-2016/

Margalitadze, T. and M. Odzeli. 2019. English–Georgian Dictionary *by Marjory Wardrop*. Tbilisi: Tbilisi State University Press.

Margalitadze, T. and Z. Pourtskhvanidze. 2019. The Georgian Language in AI-based Translation Models: Cooperation of Lexicographers and NLP Specialists. *EMLex Autumn Meeting and Colloquium, Tbilisi 2019, Georgia, October 8-11: Lexicography at a Crossroads.* Organized by TSU Lexicographic Centre and Consortium of European Master in Lexicography (EMLex).
Available at: https://margaliti.com/emlexweb.pdf

Margalitadze, T. and Z. Pourtskhvanidze. 2021. The Statistic-Based Mapping of the Distribution of Data Structure in a Parallel Corpus. International Conference *Languages in the Digital Age*, organized by State Language Department of Georgia, Centre for Language Technologies Tilde, under the patronage of the President of Georgia. October 2021.
Available at: http://enadep.gov.ge/uploads/Program_7_8_October_KA.pdf

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

Margalitadze, T. and S. Tchighladze. 2022. *Unknown Pages of English–Georgian Lexicography*. Tbilisi: Ilia State University Press.

Reimers, N. and I. Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP):* 4512-4525. Online: Association for Computational Linguistics.

Tiedemann, J. 2012. Parallel Data, Tools and Interfaces in OPUS. Calzolari, N. et al. (Eds.). 2012. *Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey, May 21–27, 2012* (LREC 2012): 2214-2218. Istanbul, Turkey: European Language Resources Association (ELRA).