https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)

# Semi-Automatic Detection of New Words in Modern Georgian*

Tamar Laluashvili, *School of Arts and Sciences, Ilia State University, Tbilisi, Georgia*
(*tamar.laluashvili.1@iliauni.edu.ge*)
(*https://orcid.org/0009-0008-3978-2333*)
and
Tinatin Margalitadze, *Centre for Lexicography and Language Technologies, Ilia State University, Tbilisi, Georgia*
(*tinatin.margalitadze@iliauni.edu.ge*)
(*https://orcid.org/0000-0001-9485-1698*)

**Abstract:** The study of neologisms in the Georgian language has gained significance due to the rapid socio-political changes in the country after the collapse of the Soviet Union and the country regaining independence. Technological advancements of the 21st century have also played a role. These developments have led to the introduction of numerous new terms and concepts into the language. However, there has been no established methodology for identifying neologisms in modern Georgian. To address this issue, a methodology was worked out at Ilia State University based on the study of existing methods applied to other languages. A corpus of the Georgian language was developed from textual materials retrieved from online platforms such as online newspapers and magazines, online media websites, websites of non-governmental organisations, and governmental agencies. Two lemmatisation tools were then applied to it to identify potential neologisms. This paper presents the methodology for the semi-automatic detection of new words in modern Georgian.

**Keywords:** NEOLOGISM, GEORGIAN LANGUAGE CORPUS, LEMMATISER, OUT-OF-VOCABULARY LEXIS, NEOLOGISM DETECTION METHODOLOGY

**Opsomming: Die semi-outomatiese opsporing van nuwe woorde in moderne Georgies.** Die studie van neologismes in die Georgiese taal het belangwekkend geraak as gevolg van die vinnige sosio-politieke veranderinge in die land nadat die Sowjetunie ineengestort het en die land onafhanklikheid herwin het. Die tegnologiese vooruitgang van die 21ste eeu het ook 'n rol gespeel. Hierdie ontwikkelings het gelei tot die ontstaan van talle nuwe terme en konsepte in die taal. Tot dusver was daar egter geen gevestigde metodologie om neologismes in moderne Georgies te identifiseer nie. Om hierdie kwessie te ondersoek, is 'n metodologie aan die Staatsuniversiteit van Ilia ontwikkel, wat gebaseer is op die bestudering van bestaande metodes wat op ander tale toegepas word. 'n Korpus van die Georgiese taal is uit tekstuele materiaal wat van aanlyn platforms soos

* A version of this paper was presented at the 6th Globalex Workshop on Lexicography and Neology (GWLN-6), held on 3 July 2024 at the University of Pretoria, Hatfield Campus, Pretoria, South Africa.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)



koerante, tydskrifte, aanlynmediawebtuistes, webtuistes van nieregeringsorganisasies, en webtuistes van regeringsagentskappe verkry is, ontwikkel. Daarna is twee lemmatiseringshulpmiddels daarop toegepas om potensiële neologismes te identifiseer. In hierdie artikel word die metodologie vir die semi-outomatiese opsporing van nuwe woorde in moderne Georgies bespreek.

**Sleutelwoorde:** NEOLOGISME, GEORGIESE TAALKORPUS, LEMMATISEERDER, BUITEWOORDESKATLEKSIS, NEOLOGISMEOPSPORINGSMETODOLOGIE

## 1. Introduction: Historical background

The present study is part of the three-year project, dedicated to the comprehensive research of neologisms in the modern Georgian language. The project is supported by the Shota Rustaveli National Science Foundation of Georgia (FR-23-4304) and its main goals are to develop a methodology for the semi-automatic detection of neologisms in modern Georgian, to create and publish an online dictionary of neologisms, and to set up a special website for the monitoring of neologisms in Georgian in the future. In this paper the methodology developed at Ilia State University for the semi-automatic detection of new words in modern Georgian is presented.

Changes taking place in a language are usually very slow and difficult to notice, and they occur over long periods of time before becoming perceptible on a synchronic level. But there are exceptions from this general tendency and contemporary Georgian is a good example of this. Currently, Georgia and the Georgian language are in an extremely interesting era from a historical point of view. After the collapse of the former Soviet Union, the country saw the emergence and rapid development of a free market economy, multiparty political system, private banking sector, the national armed forces, and many other such structures, which neither existed nor were even imaginable under the Soviet empire (Margalitadze 2020).

In 2005, Georgia joined the Bologna Process, and in 2014, the country signed an Association Agreement with the European Union. All this brought about intense relations with foreign countries on diplomatic or political, as well as on educational, cultural and economic levels, and led to the introduction of concepts, words and terms reflecting the new realities of life into the Georgian language. These include *შეზღუდული პასუხისმგებლობის კომპანია / shezghuduli p'asukhismgeblobis k'omp'ania* ('Limited Liability Company'), *ინდმეწარმე / indmets'arme* ('individual entrepreneur'), *ასპირანტი ქვეყნები / asp'iranti kveq'nebi* ('aspirant countries'), *ევრო-ატლანტიკური მისწრაფება / evroatlantik'uri mists'rapeba* ('Euro-Atlantic aspiration'), *კოლონელი / k'oloneli* ('colonel'), *ხარისხის უზრუნველყოფა / khariskhis uzrunvelq'opa* ('quality assurance'), and *სილაბუსი / silabusi* ('syllabus').

Covid-19 and the Russian-Ukrainian war also gave rise to new words in Georgian, for example *რაშიზმი / rashizmi* ('Rushism', formed by blending Russia + Fascism), *რაშისტი / rashisti* ('Rushist', a blend of Russia + Fascist), *ორკი /*

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)



*orki* ('Orc'),[1] *დრონი / droni* ('drone'), and *ატაკამსი / atak'amsi* ('ATACMS'). The recent developments in Georgia, the enactment of the so-called "Russian Law", massive demonstrations of the local population, violent attacks of the police special forces on the protest rallies, and the country's democratic backsliding "enriched" the vocabulary of Georgian with some neologisms of contemptuous meanings, for example *რობოკოპი / robok'op'i* ('Robocop'),[2] *ტიტუშკა / titushk'a* ('titushka'),[3] *ზონდერი / zonderi*,[4] and *ზონდერმოსამართლე / zondermosamartle* ('a corrupted, bought off judge').

These processes were further intensified by the increasing abundance of computer, telecommunications and mobile technologies. Consequently, all prerequisites were in place, which could cause substantial changes in the lexis and even morphology. This makes the detection and analysis of latent diachronic processes occurring in the Georgian language especially interesting at this linguistic-historical moment.

The study of neologisms is particularly relevant for lexicography, which, in addition to studying the theoretical aspects of the issue, also serves purely practical purposes. It involves updating existing dictionaries, adding new words, and assigning new meanings to existing ones. Contemporary users evaluate the quality of dictionaries by their ability to keep pace with the latest vocabulary and meanings. Dictionaries that fail to capture modern vocabulary tend to lose their appeal and popularity. Despite this, the present study reveals that many words that have entered the vocabulary of the Georgian language in the 21st century are not attested in either the explanatory or orthographic dictionaries of the language, nor in recently published Georgian bilingual dictionaries. For example, words connected to online media, such as *ონლაინმედია / onlainmedia* ('online media'), *ონლაინგაზეთი / onlaingazeti* ('online newspaper'), and *ონლაინგამოცემა / onlaingamotsema* ('online publication'), are not included even in the *Orthographic-Stylistic Dictionary of a Journalist* published in 2010 by the Institute of Linguistics of Georgia.

Additionally, increased interest in neologism research on an international level is reflected in the publication of studies examining neologisms across various languages and the development of a more systematic approach to the subject (cf. Trap-Jensen 2020; Klosa-Kückelhaus and Kernerman 2023). Conferences dedicated solely to neologisms, along with the establishment of an ongoing conference series such as GLOBALEX, also demonstrate the growing interest and commitment to this field (cf. Klosa-Kückelhaus and Kernerman 2022). Furthermore, the European Network on Lexical Innovation project, funded by the European Cooperation in Science and Technology (COST), serves as a valuable initiative that connects multiple universities around the world, including Ilia State University in Georgia. This further exemplifies the importance and relevance of this study and the project from which it stems.

## 2. An overview of neologisms

Neologisms refer to new words or meanings that appear in a language. The term

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)



*neologism* entered the English language in 1772 from French and means 'new word or expression', derived from Greek νέος (neos) 'new' + λόγος (logos) 'word'. In many cases, the reason for their emergence is new things, phenomena, or concepts that a language community learns and needs to name. They therefore enrich and add dynamism to language. A language has different registers, such as literary language, scientific language, colloquial language, slang, and dialectal language. Neologisms may appear in any register of a language, both written and oral. Georgian is no exception, and the appearance of neologisms can be observed in all registers of modern Georgian (for examples, see Section 5).

Neologisms arise in many ways. New words may be coined in a language or borrowed from other languages. Coinage of new words is interesting as their study reveals the lexical creativity of a language at a certain stage of its development. While coinage of completely new words is rare in Georgian, some cases can still be found, for example the Georgian equivalent of the English adjective *vulnerable* is a neologism *მოწყვლადი / mots'q'vladi.* It has developed from an old Georgian verb *წყვლა / ts'q'vla* ('to slay') by adding the suffix *-ადი / -adi*. Borrowing may reflect something new for speakers of the borrowing language. This is the case with the examples given in Section 1. However, borrowings may also be introduced under the influence of the prestige of another language (Trask 1996: 19). In this case, synonyms in a language consisting of foreign and native word pairs arise, such as *კონტრიბუცია / k'ontributsia* ('contribution') and *წვლილი / ts'vlili* ('contribution'); *კოლაბორაცია / k'olaboratsia* ('collaboration') and *თანამშრომლობა / t'anamshromloba* ('collaboration'), and *სტაფი / stapi* ('staff') and *თანამშრომლები / tanamshromlebi* ('staff').

Another type of neologisms is the development of new senses of existing words. Sense development is often the result of semantic borrowing and is influenced by a foreign tongue and polysemy of an equivalent word in a foreign language. There are many examples of sense development in modern Georgian due to the influence of English, with Georgian *ქიმია / kimia* ('chemistry') being a typical example. Originally, *ქიმია / kimia* ('chemistry') in Georgian meant only 'the scientific study of the structure of substances …'. The influence of the English *chemistry* and its polysemous meaning 'the relationship between two people, usually a strong sexual attraction' (Oxford Learner's Dictionary 2024) caused Georgian *ქიმია / kimia* to be used in this sense too.[5]

There are also morphological neologisms, connected to the emergence of new grammatical constructions. This is despite Georgian having a complex morphology and verbs being particularly rich in grammatical categories (see a more detailed discussion about lemmatisation issues of Georgian verbs in Section 4). Nevertheless, the development of some analytical constructions in modern Georgian verbs can be observed, for example განცხადება გაკეთდა / *gantskhadeba gak'et'da* ('a statement was made'). Such cases can be regarded as morphological neologisms.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)



## 3. Methods and tools developed for semi-automatic detection of neologisms

With the advancement of corpus linguistics, new opportunities have emerged for studying various aspects of language. Corpora have significantly contributed to the investigation of neologisms as well. Before the advent of corpus linguistics, researchers had to analyse large volumes of text manually to assess word novelty. Apart from being time-consuming, this method was subjective as it relies on the judgment of individual researchers. The advancements in modern technology have, however, expanded research possibilities. Modern electronic corpora enable the rapid and efficient processing of vast amounts of text, facilitating the study and evaluation of different linguistic phenomena, including neologisms. Although this process is largely automatic or semi-automatic, human intervention remains essential for the analysis of the material (Grochocka 2011: 62).

Different methods for the detection of neologisms have been developed in different languages, of which the following methods are commonly used: research based on the exclusion principle, application of lexical and punctuation discriminants, and statistical analysis (Janssen 2009: 69).

The exclusion principle is a well-established method for identifying neologisms, comprising two important components: the building of a study corpus for the extraction of potential neologism candidates and the creation of an exclusion list that is based on the macrostructures of a given language's dictionaries or reference corpora. Each word within the study corpus is systematically compared to the exclusion list, and words absent from the list are considered neologism candidates (Janssen 2009: 69-70; Grochocka 2011: 63). Some neologism detection tools that have been developed based on the exclusion principle are *Wortwarte* (Lemnitzer 2000) for German, *BuscaNeo* in Observatori de Neologia (OBNEO) (Cabré Castellví and Estopà Bagot 2009) for Spanish and Catalan, and *Logoscope* (Falk, Bernhard and Gérard 2014) for French.

This approach proves effective in identifying formal (or orthographic) neologisms and lexical borrowings. However, it has certain limitations, such as being incapable of detecting semantic neologisms. Moreover, the mere absence of a word from a dictionary does not necessarily indicate that it is a neologism. Online texts often contain typographical errors that should not be considered neologisms, and proper names are also excluded from the candidate list. Furthermore, there are established words that may not be documented in dictionaries but do not qualify as neologisms. For example, a well-established Georgian word *ბრძენკაცი / brdzenk'atsi* ('a wise man') is not documented in the *Explanatory Dictionary of the Georgian Language* (EDGL).

The second commonly-used method is applying lexical and punctuation discriminants for neology detection. Lexical discriminants are phrases that precede a neologism candidate and emphasise the word's novelty or unfamiliarity. Such phrases include *termed*, *so-called*, *known as*, and *defined as*. Punctuation marks such as quotation marks (single and double), italics, or parentheses can also serve as

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)



discriminants (Paryzek 2008: 165). This method suggests that the lexical unit is unfamiliar to the reader, and it is therefore presented in a different format. However, it is important to note that not all words formatted in this manner are neologisms; non-neologisms and other types of textual noise may also be marked this way (Janssen 2009: 70).

One more method for identifying neologisms is based on statistical analysis, which usually involves counting words in the study corpus and comparing the count to that of a reference corpus. Janssen (2009) identifies four approaches, of which the first relies on hapax legomena where words that occur only once in the study corpus are considered neologism candidates. This assumption formed the basis of the tool *NeoloSearch*, developed by Janicijevic and Walker (1997), which automates the process of the retrieval and analysis of neologisms. According to the second approach, a neologism is a word of the study corpus that has zero frequency in the reference corpus. The third approach is to compare the frequency of words in two corpora. If the frequency of a word in the study corpus is higher than in the reference corpus, there is a good chance that the word is a neologism, and an increase in the frequency of a word in the study corpus indicates that it may be a semantic neologism. The fourth approach counts the frequency of a word in contexts, treating a change in context as an indication that the meaning or usage of the word has changed.

Studies (for example, Cook and Stevenson 2010) show that statistical analysis can detect such complex linguistic processes as semantic change. The meanings of words change in different ways, new meanings appear, new connotative meanings develop, and a metaphorical transfer, amelioration or pejoration occur. Even though modern technologies make it easier to detect such processes, it is still semi-automatic and it is inevitable for professional linguists to intervene and exclude false candidates and validate true neologisms.

For a successful application of a combination of neologism detection methods see *Ordtrawler* ('Word Trawler'), an automatic neologism detection prototype developed by Halskov and Jarvad (2010) at the Danish Language Council. The study by Halskov and Jarvad (2010) indicates that a combination of these techniques achieves the highest level of precision, approximately 40%.

## 4. Methodology for the semi-automatic detection of new words in modern Georgian

### 4.1 Studies on neologisms in Georgia

Georgian scholars have been interested in the issue of neology and have studied it from different perspectives (cf. Beliashvili 2015; Goshkheteliani and Kikvadze 2017; Mtchedlishvili 2019; Margalitadze 2020; Rayfield 2023). While these studies are not corpus-based, they mostly investigate the topic from a lexicological point of view. Research on neology is primarily preoccupied with borrowings, especially anglicisms, and the influence of English on Georgian.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)



Kirvalidze (2017) and Davitishvili (2018) study types of and reasons for borrowing in Georgian, identifying direct borrowings, for example *ლეპტოპი / lep'top'i* ('laptop'), translation loans such as *სწრაფი კვება / sts'rapi k'veba* ('fast food'), and semantic borrowings, like *მეხსიერება / mekhsiereba* ('memory' in information technology). Some authors also analyse political neologisms in English and methods of their adoption into Georgian (cf. Mtchedlishvili 2019).

The findings point to different reasons for borrowing English words. According to Goshkheteliani and Kikvadze (2017), the reason in the majority of cases is a lack of a Georgian equivalent (e.g., laptop). Preference might also be given to a borrowed word instead of a descriptive Georgian equivalent, such as *ვერშოფი / verkshopi* ('workshop') over the descriptive equivalent *სამუშაო მეხვედრა / samushao shekhvedra* ('working meeting') (Goshkheteliani and Kikvadze 2017). Rayfield (2023) identifies the domains in which many English words are borrowed as being key fields developed in the 21st century, namely human resources, public relations, information technologies, and social media. He also highlights inconsistencies in the spelling of borrowings, noticing that one and the same word often has several spellings in Georgian, e.g. *ჩათი, ჩატი, ჩეთი, ჩეტი / chat'i, chati, chet'i, cheti* ('chat' in social media).

### 4.2 Methodology selected for the Georgian language

Despite this interest in neologisms, there was no established methodology for identifying new words in modern Georgian, which determined the decision to work out such a methodology in this project. For this purpose, existing methods of detecting new words in other languages (cf. Lemnitzer 2000; Cabré Castellví and Estopà Bagot 2009; Falk et al. 2018) were studied, after which the current approach was formulated. The present study is focused on formal (or orthographic) neologisms, as the methodology discussed below helps identify them. The detection of the development of new word senses requires different approaches, which is not addressed in this research.

For this study, the exclusion method was employed. At first, a study corpus was composed and two lemmatisers were applied to it. The first served as an exclusion source and the second was used for reducing the number of neologism candidates. This is discussed in more detail below. The corpus includes textual material from Georgian online newspapers and magazines, news and media sites, and websites of governmental and non-governmental organisations (NGOs), amounting to forty-three sites covering the last 20 years. Thirteen of these are of online magazines and newspapers, eleven of news and online media sites, nine of NGOs, and the rest of different governmental agencies. Textual material from the websites of governmental agencies was retrieved from the English–Georgian Parallel Corpus (2024).[6] The final study corpus is available online (https://neologism.iliauni.edu.ge/) and contains over 100 million tokens.

At the initial stage of the research, Georgian Wikipedia was considered for

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)



inclusion in the study corpus. Wikipedia is interesting from several points of view. It has many authors and features articles on diverse topics such as music, cinema, history, art, politics (Barbaresi 2015: 72). Wikipedia's sentences are much better structured and grammatically correct when compared to many other websites that contain grammatically incorrect texts (Yano and Kang 2008: 2). On the other hand, Wikipedia articles include many proper names, geographical names, and terms from different domains, and Georgian Wikipedia contains many typographical errors, which made it difficult to reduce the number of potential neologisms. It was thus excluded from the study.

There were five main stages involved in the data retrieval, including (1) the creation of "web crawlers", (2) retrieving the material from websites, (3) processing the material in LancsBox,[7] (4) cleaning the material from unnecessary data, and (5) uploading the material to the lemmatiser. The "web crawlers" were developed at UniLab, Cyberlaboratory of Ilia State University, and were used for downloading material from the websites.

Websites of different genres were analysed individually to identify the most productive ones concerning the creation of neologisms. The information obtained is important for the next stage of the project, in which a concept will be developed and a website will be set up for monitoring neologisms in Georgian. As mentioned above, after the data collection, unwanted items such as numbers, non-Georgian characters, symbols, and duplicates were removed, texts were tokenised, and the final lists were uploaded into the lemmatiser.

### 4.3 Issues of lemmatising Georgian words

The Georgian language belongs to the agglutinative language type. Nouns, adjectives, pronouns, and numerals are inflected and there are seven cases. Basic means for the expression of morphological categories in Georgian are prefixes and suffixes. The categorial system of verbs is especially complex. On the one hand, verbs have inflectional categories such as person, number, mood, tense, iteration, sequence of action; on the other hand, there are derivational categories — aspect, voice, version, causative, location, and direction-orientation (Shanidze 1973). Categories have their morphological markers (cf. Margalitadze 2022). For example, in terms of person, the verb can be inflected as — *v-ts'er* ('I write'), *ts'er* ('you write'), or *ts'er*-s ('he/she writes'). For number, there is — *v-ts'er* ('I write') or *v-ts'er-t* ('we write'), and for time there is — *v-ts'er* ('I write') or *da-vt'ser* ('I shall write'). Iteration is expressed through forms such as *ts'er-a* ('he/she wrote') or *ts'er-d-a* ('he/she used to write'), whereas version can be indicated by *ts'er-s* ('he/she writes' — neutral version), *i-ts'ers* ('he/she writes something for oneself' — subjective version), or *u-ts'ers* ('he/she writes something for somebody' — objective version). Location is shown through *ts'ers* ('he/she writes' — neutral) or *a-ts'ers* ('he/she writes something on something' — superessive), and voice by *ts'er-s* ('he/she writes' — active voice) or *i-ts'er-eba* ('smth. is being written' — passive voice). Finally, the causative form is represented by *a-ts'er-*

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)



*ineb-s* ('he/she has somebody write something'). The complex morphology of Georgian necessitates the lemmatisation of words in the study of neologisms.

### 4.4 Semi-automatic detection of neologisms in modern Georgian

As mentioned above, two lemmatisers were used in this research. The first is the lemmatisation tool for the Georgian language developed under the direction of Irina Lobzhanidze, a professor at Ilia State University. This lemmatiser was created based on the Finite-State Lexicon compiler (Lexc), which is a tool for creating lexical transducers. The lexicon of the morphological analyser created with this tool consists of two main components: (a) lemmas, which are presented in the form of lexicons corresponding to specific parts of speech, and (b) additional classes. In this case, a lemma is presented as an unmarked form of a word, that is the root or headword as it appears in printed or electronic dictionaries or word lists. The lemmas are based on the eight-volume EDGL and Melikishvili's Verb Index (2014).

Initially, the lexicon contained 78 000 units for nouns and 85 000 units for verbs. This lexicon has since been expanded with data obtained during the corpus testing process (cf. Lobzhanidze 2021: 64-65). Accordingly, the lemmatisation tool can only process lexical units, which are included in these resources. The lemmatiser is available on the Ilia State University website.[8] As already pointed out, a considerable segment of the vocabulary of contemporary Georgian is not represented in explanatory and orthographic dictionaries and, as a result, the said lemmatiser cannot identify them. This allowed the lemmatiser to be used as an exclusion source. The list of neology candidates was constituted by words not recognised by this tool, which are unlemmatised, out-of-vocabulary (OOV) lexis. After the tokenisation and lemmatisation of the corpus material, the OOV lexical units were subjected to analysis and the potential neologisms were sampled.

As a result of this process, the lemmatiser was able to analyse and lemmatise 278 137 words from the online media corpus data, leaving 167 070 words unlemmatised. From the NGOs' data, 61 777 words were lemmatised and 21 576 were OOV lexis. As for the data collected from websites of governmental agencies, the lemmatisation tool processed 32 821 words, and 5 923 units comprise OOV lexis (see Table 1).

While analysing OOV lexis, it was detected that they contained not only neology candidates but also undocumented lexis. These cannot be considered neologisms but are not documented in dictionaries. For instance, the example mentioned earlier, *ბრძენკაცი / brdzenkatsi* ('a wise man') is an obvious case of an undocumented word. Many undocumented derived words were encountered, such as *ბრძოლისუუნარო / brdzolisuunaro* ('combat-incapable') and *ხელნაკეთობა / khelnak'et'oba* ('a handmade item'). One of the reasons for the existence of undocumented words may be the approach of editorial boards of Georgian monolingual dictionaries, with this approach not being corpus-based. The study of

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)



this segment of vocabulary is important but requires a different methodology that could not be included within the framework of this research. To differentiate neology candidates from undocumented vocabulary, reliance had to be placed on intuition, knowledge of Georgian, and experience as lexicographers. The Georgian National Corpus was also consulted to check the dates of some words. Because part of the work was done manually, the methodology is considered semi-automatic and not fully automated.

**Table 1:** Number of lemmatised and unlemmatised words for each corpus component

| | **Lemmatised** | **OOV lexis** |
|---|---|---|
| **Online newspapers, magazines, online news sites, online media sites** | 278 137 | 167 070 |
| **NGOs** | 61 777 | 21 576 |
| **Governmental agencies** | 32 821 | 5 923 |

Undocumented words and neologisms will be added to the lexicon of the first lemmatiser so that it can recognise and lemmatise these words in the future. This tool will be used in the platform that will be developed for monitoring of neologisms in Georgian.

At the next stage of the study, to reduce the number of OOV lexis, another lemmatiser, developed by Meurer (2014) for the Georgian National Corpus (GNC, Gippert and Tandashvili 2015), was applied. The GNC contains over 200 million tokens and Meurer's lemmatiser can recognise more words than the lemmatiser applied as an exclusion source in the first phase of this research. His lemmatiser is also based on finite-state technology but contains a much larger dictionary and he developed a utility for guessing and lemmatising unknown words. In addition, lists of geographical and proper names were generated in this research, which Meurer added to his lemmatiser. After lemmatising OOV lexis with his tool, the number of unlemmatised vocabulary was reduced by 45%.

One more tool used in this research, was the Georgian word embedding platform (2022)[9] (Figure 1). The word-embedding corpus includes 1,5 billion words, making it one of the largest databases of Georgian. It allows the collection of semantically-related words. When a desired word from any language register is entered into the search bar, the word embedding presents up to thirty different words from the same register, which helps to identify more neologisms. For example, with the embedding of *ონლაინმედია / onlainmedia* ('online media'), more neologisms were extracted, such as *ინტერნეტმედია / internetmedia* ('internet media'), *ინტერნეტპორტალი / internetp'ortali* ('internet portal'), and *ინტერნეტტელევიზია / internettelevizia* ('internet television'). It should be

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)



noted that the word-embedding platform is a supplementary part of the methodology, as relying entirely on it would be time-consuming and labour-intensive.

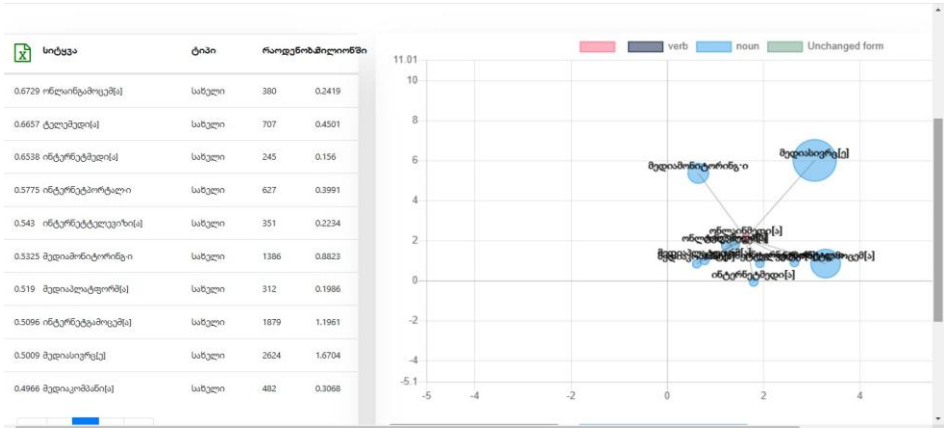


**Figure 1:** Georgian Word-Embedding Platform

## 5. Results of the study

As a result of the process described above, over 1 700 lexical neologisms in modern Georgian were identified. Most of the selected words belong to the common vocabulary of Georgian, including colloquial words and slang. Some general terms from the fields of social media, online media, and tourism were also selected. New terms from specialised fields were excluded from the study as they require detailed analysis and assistance from field professionals. The list of neologisms contains the words which entered the Georgian vocabulary during the last 20 years, and the concept of neology was expanded to words introduced in the Georgian language in the 21st century. These words cover the domains of politics, economy and finances, medicine (including esthetic medicine), tourism and hospitality, education, society and social life, agriculture, and more.

The identified neologisms will undergo thorough analysis but the preliminary study has revealed that the majority of new words in modern Georgian are anglicisms, or in other words borrowings from English. There are also hybrid words, with borrowed and Georgian roots, such as *კიბერომი / k'iberomi* ('cyberwar'), *კიბერუსაფრთხოება / k'iberusaprt'khoeba* ('cybersecurity'), *კიბერშეტევა / k'ibersheteva* ('cyber-attack'), *ონლაინგამოცემა / onlaingamotsema* ('online publication'), and *ინტერნეტგვერდი / internetgverdi* ('internet page'). The intensification of some word-forming components in Georgian were also observed, such as *ონლაინ / onlain* ('online'), *ინტერნეტ / internet* ('internet'), and *აგრო / agro* ('agro'). More examples include: *ონლაინლექსიკონი / onlainleksik'oni* ('online dictionary'),

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)



*ონლაინშეხვედრა / onlainshekhvedra* ('online meeting'), *ონლაინსწავლება / onlainstsavleba* ('online teaching'), *ონლაინგამოცემა / onlaingamotsema* ('online publication'), *ონლაინგაზეთი / onlaingazeti* ('online newspaper'), *აგრობიზნესი / agrobiznesi* ('agro business'), *აგროტურიზმი / agroturizmi* ('agro tourism'), *აგროსაწარმო / agrosats'armo* ('agro enterprise'), and *აგროლიზინგი / agrolizingi* ('agro leasing'). These word-forming components generate many new words in modern Georgian, thus enriching the language.

Georgian is a highly synthetic language with many affixes. Some of these affixes have become very productive in modern Georgian, for example the verb-forming prefix *და- / da-* and suffix *ება / eba, და-ება / da-eba*: *დაორგანიზება / daorganizeba* ('to organise'), *დასანქცირება / dasanktsireba* ('to sanction'), *დალაიქება / dalaikeba* ('to like'), *დამესიჯება / damesijeba* ('to message'), *დაფორვარდება / daporvardeba* ('to forward'), and *დაპარკინგება / dap'ark'ingeba* ('to park').

A more recent tendency is the increased number of semantic borrowings from English, as can be observed in the terminology of social media. Alongside borrowings such as *dalaikeba* ('to like'), *gashe areba* ('to share'), *peiji* ('page'), *woli* ('wall'), *prendi* ('friend'), and *poloueri* ('follower'), there developed new polysemous meanings of Georgian words, corresponding to *like*, *share*, *page*, *wall*, *friend*, and *follower*: *მოწონება / mots'oneba* ('to like'), *გაზიარება / gaziareba* ('to share'), *გვერდი / gverdi* ('page'), and *კედელი / k'edeli* ('wall'). Native Georgian words are mostly used in written or formal speech, while borrowings remain as their colloquial synonyms. For example, the borrowing *დალაიქება / dalaikeba* ('to like') coexists with its Georgian synonym *მოწონება / mots'oneba* ('to like'), and *ფრენდი / prendi* ('friend') with its Georgian synonym *მეგობარი / megobari* ('friend'). It should be noted that while these words are not new in terms of their form, they have developed semantically and have acquired new meanings, which are used in the context of social media.

Neologisms are also found in the lower register of the language. Such are, for example, slang words *ქრაში / krashi* ('crush'), *ქრინჯი / krinji* ('to cringe'), *დასტალკვა / dastalk'va* ('to stalk'), and *სლეი / slei* ('slay'). Again, it is clear that, as with other registers, the source of modern Georgian slang is also English.

## 6. Conclusion

This paper presents the methodology developed at Ilia State University for the identification of neologisms in modern Georgian, which is based on the exclusion method. A study corpus was created for the project, including textual material from Georgian-language online newspapers, magazines, news websites, media sites, and websites of NGOs and governmental agencies. The corpus contains over 100 million tokens and covers the last 20 years of the development of the Georgian language. Two lemmatisation tools created for the Georgian language and, based on the finite-state technology, were applied in the study. The first lemmatiser, developed at Ilia State University and which

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)



contains macrostructures of the existing academic dictionaries, was used as an exclusion list for the study. Another lemmatiser, developed for the Georgian National Corpus, can guess and lemmatise unknown words and its application enabled to reduce the number of OOV lexis by 45%.

Although the lemmatisation tools performed a significant amount of work, it is not without limitations. The tools cannot detect new meanings of existing words in the language. The process is solely based on word form and does not possess any kind of "semantic intellect", as termed by Lemnitzer (2000). As a result, new word meanings are "invisible" to the programme. For instance, the term *virus*, originally referring to a disease, has acquired a new meaning with the evolution of technology and designates computer viruses. In Georgian, it now has one more meaning and refers to popular videos that spread quickly and widely on the internet. Such changes cannot be detected by this approach.

When dealing with text processing on the internet, it is important to underline the existence of many typographical errors. The lemmatisers used in the study cannot handle these errors and treat them as separate lexical units. Additionally, many Georgian words are not documented in dictionaries, and thus they are perceived as new by the tool and appear in the list of OOV lexis. Since the focus of this study is mainly on non-lemmatised content, the manual selection process becomes more time-consuming.

At the next stage of the project, neologisms identified during the study will be analysed thoroughly and classified, an online dictionary of neologisms will be composed and published, and a website will be set up for monitoring of Georgian neologisms in the future. The online dictionary of neologisms will be an explanatory dictionary, in which every word will be defined, supplemented by example sentences from the study corpus, and will contain an etymology part, explaining its source in modern Georgian. The work on the dictionary is underway and it will be published online in 2026.

## Endnotes

1. The word denotes a Russian soldier.
2. Borrowing from English Robocop, a blend of robot + cop is formed. The word originates from an American science fiction film of the same title. In Georgian it is used to designate a special forces officer.
3. The word has originated from a Ukrainian name Titushko, a former sportsman who hired people to attack pro-European demonstrators in Kiev. In Georgian, it denotes a criminal hired by the Government to attack pro-European activists.
4. This word is of German origin Sonderkommandos, referring to punitive groups created by Nazi Germany. In Georgian, it means a member of a punitive group of criminals hired by the Government.
5. https://www.oxfordlearnersdictionaries.com/definition/english/chemistry?q=chemistry
6. https://enkacorpus.iliauni.edu.ge

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)



7. LancsBox (2024) is an advanced software tool designed for analysing language data and corpora, and was used for tokenising downloaded texts. It was developed by a team of researchers at Lancaster University, led by Dr. Vaclav Brezina. See: http://corpora.lancs.ac.uk/lancsbox/.
8. qartnlp.iliauni.edu.ge
9. https://wordembedding.spellchecker.ge/

## References

### Digital tools

**Georgian Word-Embedding Platform.** 2021. https://wordembedding.spellchecker.ge/ [14 November 2024]

**LancsBox.** 2021. http://corpora.lancs.ac.uk/lancsbox [14 November 2024]

**Lemmatisation Tool for the Georgian Language.** 2021. https://qartnlp.iliauni.edu.ge/upload-xls [14 November 2024]

### Other literature

**Barbaresi, A.** 2015. *Ad Hoc and General-Purpose Corpus Construction from Web Sources*. Unpublished PhD Dissertation. Lyon: École Normale Supérieure de Lyon.

**Beliashvili, T.** 2015. The Function of Neologisms in French and Georgian Media Discourse. *Spekali* 9: 1-11. Tbilisi: Tbilisi State University.

**Cabré Castellví, M.T. and R. Estopà Bagot.** 2009. Trabajar en neología con un entorno integrado en línea: la estación de trabajo OBNEO. *Revista de Investigación Lingüística* 12: 17-38.

**Cook, P. and S. Stevenson.** 2010. Automatically Identifying Changes in the Semantic Orientation of Words. Calzolari, N., K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias (Eds.). 2010. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10), Valetta, Malta, 19–21 May 2010*: 28-34. Valletta: European Language Resources Association.

**Davitishvili, N.** 2018. On the Integration of Anglicisms into Present-Day Georgian. *Studies in Literature and Language* 17(3): 22-27.

**EDGL.** 1950–1964. *Explanatory Dictionary of the Georgian Language*. Chikobava, A. (Ed.). I–VIII volumes. Tbilisi: Mecniereba.

**Falk, I., D. Bernhard and C. Gérard.** 2018. *The Logoscope: A Semi-Automatic Tool for Detecting and Documenting French New Words from the Linguistic Project to the Web Interface.* Research Report, University of Strasbourg.

**Gippert, J. and M. Tandashvili.** 2015. Structuring a Diachronic Corpus. The Georgian National Corpus Project. Gippert, J. and R. Gehrke (Eds.). 2015. *Historical Corpora. Challenges and Perspectives*: 305-322. Tübingen: Narr.

**Goshkheteliani, I. and M. Kikvadze.** 2017. The Influence of English Borrowings on the Georgian Language. *Journal of Teaching and Education* 7(1): 459-463.

**Grochocka, M.** 2011. *Lexical Creativity in English: A Corpus-Based Study*. Unpublished PhD Dissertation. Poznań: Adam Mickiewicz University.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2044 (Lexicofocus)



**Halskov, J. and P. Jarvad.** 2010. Automated Extraction of Neologisms for Lexicography. Granger, S. and M. Paquot (Eds.). 2010. *eLexicography in the 21st Century: New Challenges, New Applications. Proceedings of eLex 2009, Louvain-la-Neuve, 22–24 October 2009:* 405-410. Louvain-la-Neuve: Presses universitaires de Louvain.

**Janicijevic, T. and D. Walker.** 1997. NeoloSearch: Automatic Detection of Neologisms in French Internet Documents. Lessard, G. and M. Levison (Eds.). *The 1997 Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, Kingston, Ontario, Canada, 3–7 June 1997*. Kingston: Queen's University.

**Janssen, M.** 2009. Detección de Neologismos: una perspectiva computacional. *Debate Terminológico* 5: 68-75.

**Kirvalidze, N.** 2017. Linguo-Cultural and Pragmatic Peculiarities of the Phenomenon of Anglicisation in Georgia. *Journal of Teaching and Education* 6(2): 287-298.

**Klosa-Kückelhaus, A. and I. Kernerman.** 2022. GWLN — Globalex Workshop on Lexicography and Neology: An Overview. *Lexicala* 30: 10-13.

**Klosa-Kückelhaus, A. and I. Kernerman.** 2023. Unregistered Words, Neologisms, and Dictionaries: An Introduction. *Lexicography* 10(2): 87-93.

**Lemnitzer, L.** 2000. Einleitung und Hintergrund: Die Wortwarte — auf der Suche nach den Neuwörtern von morgen.
wortwarte.de/ [19 October 2022]

**Lobzhanidze, I.** 2021. *Principles of Morpho-Syntactic Annotation of Georgian and Morphological Analysis of the Finite State Position*. Tbilisi: Ilia University Press.

**Margalitadze, T.** 2020. Language and Ecology of Culture. *Lexicographica* 36: 225-240.

**Margalitadze, T.** 2022. Lexicography of Georgian. Hanks, P. and G.-M. de Schryver (Eds.). 2022. *International Handbook of Modern Lexis and Lexicography*: 1-24. Berlin/Heidelberg: Springer-Verlag.

**Melikishvili, D.** 2014. *Systemic Morpho-Syntactic Analysis of the Georgian Verb*. Tbilisi: Tbilisi State University.

**Meurer, P.** 2014. Morphosyntactic Analysis of Georgian. *Georgian National Corpus.*
gnc.gov.ge.

**Mtchedlishvili, M.** 2019. Political Neologisms in Global Politics and the Issue of their Transposition from English into Georgian. *Spekali* 13. Tbilisi: Tbilisi State University.

**Paryzek, P.** 2008. Comparison of Selected Methods for the Retrieval of Neologisms. *Investigationes Linguisticae* 16: 163-181.

**Rayfield, D.** 2023. The Indigestible Impact of English on the Modern Georgian Lexicon. *Digital Kartvelology* 2: 14-27.

**Shanidze, A.** 1973. *Foundations of Georgian Grammar.* Tbilisi: Tbilisi University Press (in Georgian).

**Trap-Jensen, L.** 2020. Language-Internal Neologisms and Anglicisms: Dealing with New Words and Expressions in the Danish Dictionary. *Dictionaries* 41(1): 11-25.

**Trask, R.L.** 1996. *Historical Linguistics*. Oxford: Oxford University Press.

**Yano, T. and M. Kang.** 2008. Taking Advantage of Wikipedia in Natural Language Processing.
www.cs.cmu.edu/~taey/pub/wiki.pdf [23 October 2022]