
PUBLICA • HELIAE • UNIVERSITAS •
LEGAMUS

Centre for Lexicography and Language Technologies
Ilia State University

Proceedings of the II International Conference *Lexicography in the XXI Century*

10 – 12 October, 2025
Edited by Tinatin Margalitadze

Tbilisi
2026

ლექსიკოგრაფიისა და ენობრივი ტექნოლოგიების ცენტრი
ილიას სახელმწიფო უნივერსიტეტი

საერთაშორისო კონფერენციის კრებული
ლექსიკოგრაფია 21-ე საუკუნეში

10 – 12 ოქტომბერი, 2025
რედაქტორი თინათინ მარგალიტაძე

თბილისი
2026

Scientific Committee of the Conference:

Arleta Adamska (Adam Mickiewicz University, Poznań, Poland)

Donna Farina (New Jersey City University, USA)

Rufus Gouws (Stellenbosch University, South African Republic)

Tinatin Margalitadze (Ilia State University, Georgia)

Katalin Márkus (Károli Gáspár University of the Reformed Church in Hungary)

Lars Trap-Jensen (Society for Danish Language and Literature, Denmark)

Alenka Vrbinc (University of Ljubljana, Slovenia)

Marjeta Vrbinc (University of Ljubljana, Slovenia)

Geoffrey Williams (Université de Bretagne Sud, France)

კონფერენციის სამეცნიერო საბჭო:

არლეტა ადამსკა (ადამ მიცკევიჩის სახელობის უნივერსიტეტი პოზნანში, პოლონეთი)

რუფუს გოუვსი (სტელენბოსის უნივერსიტეტი, სამხრეთ აფრიკის რესპუბლიკა)

ალენკა ვერბინცი (ლიუბლიანას უნივერსიტეტი, სლოვენია)

მარიეტა ვერბინცი (ლიუბლიანას უნივერსიტეტი, სლოვენია)

თინათინ მარგალიტაძე (ილიას სახელმწიფო უნივერსიტეტი, საქართველო)

კატალინ მარკუსი (რეფორმატული ეკლესიის კაროლი გაშპარის სახელობის უნივერსიტეტი, უნგრეთი)

ლარს ტრაპ-იენსენი (დანიური ენისა და ლიტერატურის საზოგადოება)

ჯეფრი უილიამსი (სამხრეთ ბრეტანის უნივერსიტეტი, საფრანგეთი)

დონა ფარინა (ნიუ ჯერზის უნივერსიტეტი, აშშ)

ISBN 978-9941-18-507-6

ილიას სახელმწიფო უნივერსიტეტის გამომცემლობა
ი. ჭავჭავაძის გამზირი 45, თბილისი, 0179, საქართველო

ILIA STATE UNIVERSITY PRESS
45 I. Chavchavadze ave., Tbilisi, 0179, Georgia

Contents

Foreword

The II International Conference *Lexicography in the 21st Century* was held at Ilia State University, Tbilisi, Georgia from 10 to 12 October, 2025. On behalf of the Organizing and Scientific Committees of the conference, I am pleased to present this volume of the proceedings. The volume is published online as a digital edition and will also appear in print form.

The international conference *Lexicography in the 21st century* was established in 2022 by Ilia State University. The organizer of the conference is the "Centre for Lexicography and Language Technologies". The conference is held biannually and its aim is to create a forum for the discussion of issues of modern lexicography in Georgia, to further popularize this field and to promote sustainable development of lexicography both in Georgia and the entire region.

The second edition of the conference *Lexicography in the 21st Century* was dedicated to the 300th anniversary of the death of Sulkhan-Saba Orbeliani, an eminent Georgian lexicographer, writer, diplomat and public figure.

The main theme of the conference was "Lexicography and Technologies," and it brought together 50 scholars from 16 countries. The topics of the conference included the following: Lexicography and Language Technologies; Lexicography and Corpus Linguistics; Lexicography for Specialized Languages, Terminology, and Terminography; Lexicography and Neology; Lexicography for Lesser-Used Languages; Lexicological Issues of Lexicographical Relevance; Bilingual Lexicography; The Dictionary-Making Process; Research on Dictionary Use and Presentation of Lexicographic Projects.

The conference featured 11 scientific sessions, including three plenary sessions. All submitted papers underwent a double-blind review process and were evaluated by two independent reviewers. At the conference, special attention was paid to the participation of young researchers. The conference had sessions dedicated to MA and Ph.D. students.

It should be noted that the plenary lecture delivered by Dr Lars Trap-Jensen at the First International Conference *Lexicography in the 21st Century* in 2023, devoted to the application of artificial intelligence in lexicography, made a strong impression on Georgian lexicographers, particularly students and young researchers. At the 2025 conference, three papers were presented by young Georgian researchers, including a student from the Master's programme in Lexicography, focusing on the use of AI in Georgian lexicography. Another new feature of the 2025 conference was the inclusion of neology in the conference topics.

On behalf of the organising committee of the II International Conference *Lexicography in the 21st Century*, I would like to extend our gratitude to the plenary speaker, Professor Giovanni Talarico of the University of Verona, who leads the COST action ENEOLI "European Network for Lexical Innovation" (https://www.cost.eu/ actions/CA22126/). His presentation addressed the experience of this European network and analysed the multidimensional approaches to lexical innovation.

I would also like to extend my sincere thanks to the plenary speaker, Professor Rute Costa of NOVA University Lisbon, for her in-depth exploration of the many issues related to terminology.

In her plenary address, Professor T. Margalitadze spoke about the lexicographic principles developed by Sulkhan-Saba Orbeliani at the end of the 17th century, which place the great Georgian lexicographer among the best representatives of world lexicography.

We extend our gratitude to all conference participants for their insightful and relevant presentations, to our colleagues for reviewing the abstracts and papers, and to the administration of Ilia State University for their invaluable support and assistance in organising the conference.

Professor Tinatin Margalitadze

წინასიტყვაობა

მეორე საერთაშორისო კონფერენცია *ლექსიკოგრაფია XXI საუკუნეში* ჩატარდა 2025 წლის 10-12 ოქტომბერს თბილისში, ილიას სახელმწიფო უნივერსიტეტში. კონფერენციის საორგანიზაციო კომიტეტისა და სამეცნიერო საბჭოს სახელით წარმოგიდგენთ კონფერენციის კრებულს. იგი გამოქვეყნებულია ინტერნეტში ციფრული გამოცემის სახით, თუმცა იგეგმება კრებულის წიგნად დაბეჭდვაც.

საერთაშორისო კონფერენცია *ლექსიკოგრაფია XXI საუკუნეში* დაარსდა 2022

წელს ილიას სახელმწიფო უნივერსიტეტის მიერ. კონფერენციის ორგანიზატორია ილიას სახელმწიფო უნივერსიტეტის „ლექსიკოგრაფიისა და ენობრივი ტექნოლოგიების ცენტრი“. კონფერენცია იმართება ორ წელიწადში ერთხელ და მისი მიზანია საქართველოში ფორუმის შექმნა თანამედროვე ლექსიკოგრაფიის საკითხებზე სამსჯელოდ, დარგის შემდგომი პოპულარიზაციისათვის და ლექსიკოგრაფიის მდგრადი განვითარებისათვის როგორც საქართველოში, ისე მთელ რეგიონში.

მეორე საერთაშორისო კონფერენცია *ლექსიკოგრაფია XXI საუკუნეში* მიეძღვნა დიდი ქართველი ლექსიკოგრაფის, მწერლის, დიპლომატისა და სახელმწიფო მოღვაწის, სულხან საბა ორბელიანის გარდაცვალებიდან 300 წლისთავს.

კონფერენციის მთავარი თემა იყო „ლექსიკოგრაფია და ტექნოლოგიები“ და მასში მონაწილეობა მიიღო 50 მეცნიერმა მსოფლიოს 16 ქვეყნიდან.

კონფერენციის თემატიკა მოიცავდა შემდეგ მიმართულებებს: ლექსიკოგრაფია და ენობრივი ტექნოლოგიები; ლექსიკოგრაფია და კორპუსის ენათმეცნიერება; ლექსიკოგრაფია დარგობრივი ენებისათვის, ტერმინოლოგია და ტერმინოგრაფია; ლექსიკოგრაფია და ნეოლოგია; ლექსიკოგრაფია ნაკლებად გავრცელებული ენებისათვის; ლექსიკოგრაფიასთან დაკავშირებული ლექსიკოლოგიური საკითხები; თარგმნითი ლექსიკოგრაფია; ლექსიკონთა აგების პრინციპები; ლექსიკონის მომხმარებელთა კვლევა და ლექსიკოგრაფიული პროექტების პრეზენტაცია.

კონფერენციის ფარგლებში ჩატარდა 11 სამეცნიერო სესია, მათ შორის სამი პლენარული. კონფერენციაზე წარმოდგენილმა ყოველმა მოხსენებამ გაიარა ორი ანონიმური რეცენზირება. კონფერენციაზე განსაკუთრებული ყურადღება დაეთმო ახალგაზრდა მკვლევრების მონაწილეობას, ჩატარდა სპეციალური სხდომები სამაგისტრო და სადოქტორო საფეხურების სტუდენტებისათვის. აღსანიშნავია, რომ 2023 წლის კონფერენციაზე მოსმენილმა დოქტორ ლარს ტრაპ-იენსენის პლენარულმა მოხსენებამ ხელოვნური ინტელექტის გამოყენების შესახებ ლექსიკოგრაფიაში, დიდი შთაბეჭდილება მოახდინა ქართველ ლექსიკოგრაფებზე, განსაკუთრებით სტუდენტებსა და ახალგაზრდა მკვლევრებზე. 2025 წლის კონფერენციაზე

მათ სამი მოხსენება წარმოადგინეს ხელოვნური ინტელექტის თემაზე. 2025 წლის კონფერენციაზე ასევე სიახლე იყო ნეოლოგიის შეტანა კონფერენციის თემატიკაში.

მეორე საერთაშორისო კონფერენციის *ლექსიკოგრაფია XXI საუკუნეში* საორგანიზაციო კომიტეტის სახელით მინდა დიდი მადლობა გადავუხადო პლენარულ მომხსენებელს, ვერონის უნივერსიტეტის პროფესორს ჯოვანი ტალარიკოს, რომელიც ხელმძღვანელობს მეცნიერებისა და ხელოვნების დარგში ევროპული თანამშრომლობის მოძრაობას ENEOLI „ლექსიკური სიახლეების ევროპული ქსელი“ (https://www.cost.eu/actions/CA22126/). მისი მოხსენება სწორედ ამ ევროპული ქსელის გამოცდილებას შეეხო და გააანალიზა მრავალგანზომილებიანი მიდგომები ლექსიკური სიახლეებისადმი.

ასევე დიდი მადლობა მინდა გადავუხადო პლენარულ მომხსენებელს, ლისაბონის ნოვას უნივერსიტეტის პროფესორს რუტე კოსტას ტერმინოლოგიასთან დაკავშირებული მრავალი საკითხის სიღრმისეული გაშუქებისათვის.

თავის პლენარულ მოხსენებაში, პროფესორმა თ. მარგალიტაძემ ისაუბრა სულხან-საბა ორბელიანის მიერ XVII საუკუნის ბოლოს შემუშავებულ ლექსიკოგრაფიულ პრინციპებზე, რომლებიც დიდ ქართველ ლექსიკოგრაფს მსოფლიო ლექსიკოგრაფიის საუკეთესო წარმომადგენლებს შორის უჩენს ადგილს.

დიდი მადლობა კონფერენციის ყველა მონაწილეს საინტერესო და რელევანტური მოხსენებებისათვის, კოლეგებს თეზისებისა და სტატიების რეცენზირებისათვის, დიდი მადლობა ილიას სახელმწიფო უნივერსიტეტის ადმინისტრაციას კონფერენციის ორგანიზებაში გაწეული დიდი დახმარებისა და მხარდაჭერისათვის.

პროფესორი თინათინ მარგალიტაძე

Code-Mixing Patterns in a Corpus-Based Megrelian-English Dictionary

Rusudan Gersamia,
Irina Lobzhanidze
Ilia State University
rgersamia@iliauni.edu.ge,
irina_lobzhanidze@iliauni.edu.ge

Abstract

This paper examines code-mixing patterns in a corpus-based Megrelian-English dictionary. Megrelian, an endangered Kartvelian language, is characterized by intensive contact with Georgian and, to a lesser extent, Russian. Using data from a morphologically annotated corpus of spoken Megrelian, the study analyzes the structural and functional properties of code-mixing, including insertion, alternation and lexicalized mixing.

The findings show that code-mixing is a systematic and productive phenomenon rather than random interference. Mixed forms are integrated into Megrelian morphosyntax and reflect both linguistic constraints and sociolinguistic factors such as age, region and communicative context.

The paper also demonstrates how such forms can be represented in a corpus-based bilingual dictionary using L2 tagging. This approach contributes to lexicographic practice and provides a more accurate representation of contemporary language use in endangered language documentation.

Keywords: Megrelian; code-mixing; language contact; corpus-based lexicography; endangered language documentation

1 Introduction

Megrelian (ISO 639-3: xmf) is an unwritten Kartvelian language spoken primarily in western Georgia, particularly in the Samegrelo-Zemo Svaneti region, which includes the municipalities of Zugdidi, Senaki, Martvili, Abasha, Chkhorotsku, Khobi and Tsalenjikha. It is also spoken by internally displaced populations from Abkhazia, as well as by Georgian communities residing in the Gali district of the occupied territory. According to UNESCO (2010), Megrelian is classified as an “increasingly endangered” language.

The sociolinguistic environment of Megrelian is influenced by a range of factors. These include political conditions (such as displacement from occupied territories), economic and

industrial influences (for example, the port city of Poti, which has long been a site of intensive language contact) and the impact of neighboring dialects (e.g., the influence of Imeretian dialect of Georgian in Abasha and Gurian dialect - in Poti). In addition, the dominance of Standard Georgian as the state language used in education, media and public life, plays a crucial role in language use (Makharoblidze & Gersamia 2022). Until the late 20th century, Russian also functioned as a major contact language.

As a result, Megrelian speakers are typically bilingual. Standard Georgian, as the official language, is an integral part of their daily lives. And speakers frequently mix languages in everyday communication, producing hybrid expressions and a wide range of mixed lexical forms. The contact between Megrelian and Georgian, and to a lesser extent Russian, provides a clear example of bilingual and multilingual interaction, making Megrelian an important case for the study of code-mixing and language contact.

2 Aim and Methods

The aim of this study is to examine how the outcomes of language contact are reflected in the speech of contemporary Megrelian speakers. In particular, it investigates the linguistic strategies employed in code-mixing, code-switching and borrowing.

The analysis is based on morphologically annotated texts from the Megrelian Language Corpus (MLC) (Gersamia & Lobzhanidze 2021, 2022-2025) as well as data from a corpus-based Megrelian-English dictionary available at https://xmf.iliauni.edu.ge/. The corpus consists of spoken language materials, including recorded dialogues, narratives and interviews collected through fieldwork conducted between 2021 and 2024 in urban and rural areas of Samegrelo[1].

The fieldwork followed established principles of language documentation (Austin 2006; Bowern 2008). The dataset includes 58 hours of audio-video recordings from 83 speakers, with balanced representation in terms of dialect, age and gender. The material is thematically diverse, covering topics such as traditional rituals, cuisine, oral traditions, folktales, personal narratives etc.

The corpus has been processed in the Fieldwork Language Explorer (FLEx 2024) environment and includes 150 annotated texts, comprising 97 702 tokens and 13 648 lexical and morphological units (roots and affixes), distributed across 9 255 sentences.

The analysis of language contact phenomena combines quantitative and qualitative approaches. First, non-Megrelian elements in the corpus (including roots, affixes, and lexemes) were identified and assigned an L2 tag in the FLEx database. These tagged elements were then extracted and systematically filtered to determine their frequency and distribution.

The results show that L2 elements primarily originate from Georgian (in large numbers) and, to a lesser extent, Russian, alongside a considerable number of internationalisms. Out

1 Grant No FR-21-993, Annotated Corpus of Megrelian Language with Sketch Grammar and Online Dictionary

of 13 648 unique units, 3 593 were marked as L2, including 1 150 fully lexicalized word forms that are not morphologically analyzable.

From an analytical perspective, the study examines:

- the structural integration of L2 elements into Megrelian morphosyntax;
- the types of code-mixing and switching strategies employed by speakers, and;
- the degree of lexicalization of borrowed and hybrid forms.

The analysis of L2-marked items and their integration into a lexicographic resource such as the Megrelian-English dictionary demonstrates that code-mixing is not limited to the use of loanwords or calques. Rather, it constitutes a dynamic and systematic process governed by its own lexical and grammatical principles.

Specifically, elements from the source language (roots, stems, and occasionally affixes) are incorporated into the morphostructure of the recipient language, adapting to its grammatical system.

Consequently, the identification and description of language contact phenomena form an essential component of corpus-based lexicography. Representing such mixed and hybrid forms in the Megrelian-English dictionary and in the lexicon of Megrelian morphemes allows for a more accurate and comprehensive account of contemporary language use.

3 Theoretical Framework

Code-switching is commonly defined as the systematic use of two or more languages within a single discourse (Weinreich 1953; Gumperz 1982 and others). In broad usage, the term encompasses both intersentential switching (between clauses or sentences) and intrasentential switching (within a clause or word), the latter often referred to as code-mixing. Importantly, such switching is not random, but rule-governed, reflecting both grammatical constraints and sociolinguistic conditioning (Poplack 1980 and others).

In this study, we distinguish between code-switching and code-mixing in line with common practice. Code-mixing refers specifically to the use of elements from two or more languages within a single sentence, where lexical and grammatical features from different linguistic systems co-occur (Myers-Scotton 1998; Redouane 2005 and others). By contrast, code-switching involves alternation between languages across sentence boundaries.

A central concept in the analysis is the distinction between the matrix language (L1) and the embedded language (L2) (Myers-Scotton 1993; Auer & Muhamedova 2005). The matrix language provides the grammatical framework of the utterance, while elements from the embedded language are inserted into this structure. Code-mixing may occur at multiple linguistic levels, including grammatical, phonological, lexical, and orthographic (Hoo 2007).

In intrasentential mixing, embedded language elements are typically adapted to the morphosyntactic system of the matrix language. In contrast, in alternational code-switching, speakers shift between languages at structural boundaries, producing sequences that maintain the grammatical integrity of each language.

Borrowing is closely related to these phenomena and is often treated as a special case of code-mixing or code-switching. Borrowed items are typically nativized into the recipient language, showing phonological and morphological integration and often spreading beyond bilingual speakers (Haugen 1956; Poplack et al. 1988). Over time, such forms may become fully lexicalized loanwords, filling lexical or cultural gaps and gradually losing their perception as foreign elements.

For the purposes of analysis, this study adopts a widely used typology of contact-induced phenomena based on Poplack (1980) and subsequent research. Three analytically relevant types are distinguished:

- Insertional mixing i.e. the embedding of words or phrases from a donor language within a Megrelian grammatical frame;
- Alternational switching i.e. the alternation between languages at clause or sentence boundaries;
- Lexicalized mixing i.e. the formation of hybrid structures (e.g., compounds or affixed forms) and the use of shared morphosyntactic patterns across languages.

These categories provide the analytical framework for examining the corpus data and are directly applicable to the representation of mixed forms in the Megrelian-English dictionary.

4 L2 Elements in the Megrelian–English Online Dictionary

4.1 Lexical Composition and Common Domains

The analysis of L2 elements in the Megrelian-English online dictionary reveals that code-mixing is particularly frequent in semantic domains associated with modern life, including technology, education, administration and science. In these domains, lexical items are predominantly internationalisms, which have entered Megrelian primarily through Georgian and typically lack native equivalents (See Table 1).

administ'ra͡tsia 'administration'	*glɔbaliza͡tsia* 'globalization'	*ɛksp'ɛdi͡tsia* 'expedition'	*k'ɔmpɔrt'uli* 'comfortable'
ak'adɛmia 'academy'	*ginɛk'ɔlɔgia* 'gynecology'	*ɛksp'ɔnat'i* 'exhibit'	*k'ɔmɛnt'ari* 'comment'
avt'ɔrit'ɛt'i 'authority'	*dɛk'ɔra͡tsia* 'decoration'	*ɛt'imɔlɔgia* 'etymology'	*k'ɔmunik'a͡tsia* 'communication'
ak'vat'ɔria 'aquatorium'	*dɛmɔsnt'ra͡tsia* 'demonstration'	*vɛrsia* 'version'	*k'ɔmp'iut'ɛruli* 'computer-related'
alɛrgia 'allergy'	*dikt'at'ura* 'dictatorship'	*indust'ria* 'industry'	*k'ɔmɛrsant'i* 'merchant'

alumini 'aluminium'	*ɛlɛkt'rɔpik'a͡tsia* 'electrification'	*inst'it'ut'i* 'institute'	*k'ɔɔrdinat'i* 'coordinate'
biznɛsi 'business'	*ɛksp'ɛrimɛnt'i* 'experiment'	*inʒinɛri* 'engineer'	*k'ɔnk'urɛnt'i* 'competitor'
biznɛsmɛni 'businessman'	*ɛksp'ansia* 'expansion'	*k'limat'uri* 'climatic'	*k'ɔɔrdina͡tsia* 'coordination' etc.

Table 1: International Lexical Items (L2) in Megrelian

4.2 Russian-Origin Vocabulary (Russizms)

Russian-origin lexical items (russizms) are significantly less frequent, with fewer than 100 items identified in the dataset. These elements generally belong to earlier layers of language contact and are less productive in contemporary usage (See Table 2).

agrɔt'ɛxnik'a 'agrotechnics'	*vɛdra* 'bucket'	*k'ap'ik'i* 'coin'	*sk'ɔrɔda* 'pan'
ak'ardiɔni 'accordion'	*zɔndik'i* 'umbrella'	*k'asruli* 'cooking pot'	*sp'i͡tʃka* 'match'
ak'ɔʃk'a 'window'	*zak'lɔvk'a* 'metal rivet'	*k'ɛrasink'a* 'kerosene lamp'	*sp'riɛzdɔm* 'with meeting'
baradɔk'i 'mess'	*k'uk'la* 'doll'	*k'lɛt'k'a* 'cage'	*suxɔi* 'dry'
bulk'i 'bread roll'	*k'rask'a* 'paint'	*p'advali* 'cellar'	*st'ɔlba* 'pillar'
brizɛnt'i 'canvas'	*k'amanda* 'team'	*rɛik'a* 'wooden strip'	*st'ɔik'a* 'stand'
garadɔk'i 'small town'	*k'al͡tsɔ* 'ring'	*sirɔst'i* 'dampness'	*xalt'ura* 'side job'
gradi 'hail'	*k'alink'a* 'raspberry'	*sɛt'k'a* 'net'	Etc.

Table 2: Russian-Origin Lexical Items (Russizms) in Megrelian

The distribution of russizms is strongly conditioned by sociolinguistic factors, particularly speaker age, topic of conversation and regional background. Their use is more typical of older speakers and individuals originating from Abkhazia (including internally displaced populations), reflecting historical patterns of Russian influence. Overall, the data indicate a clear decline in Russian lexical influence, in contrast to the increasing presence of Georgian-derived and international vocabulary.

4.3 Georgian Lexical Influence

Georgian constitutes the primary source of L2 elements in the corpus. Code-mixing with Georgian is especially frequent in the speech of younger and middle-aged speakers, and its occurrence increases in urban and industrial settlements, as well as in regions with closer contact with Georgian-speaking communities. Economic and professional factors also influence this process (Makharoblidze & Gersamia 2022)

Georgian-derived lexemes in the database include both:

- items without Megrelian equivalents, which fill a lexical gap, e.g. *sagani* 'item', *gak'vɛtili* 'lesson', *mas͡ts'avlɛbɛli* 'teacher' etc., and;
- items with existing Megrelian equivalents, resulting in synonymic parallelism, e.g. *adgili* Megr. equivalent *ak'ani* 'place', *axlɔmaxlɔs* Megr. equivalent *xɔlɔ-xɔlɔs* 'nearby', *gadmɔt͡sɛma* Megr. equivalent *ginɔt͡ʃama* 'transmission' etc..

In the latter case, the choice between Megrelian and Georgian forms is influenced by sociocultural and pragmatic factors, such as discourse context, speaker identity, and communicative intent.

The data further reveal a distinction between:

- Older, fully integrated borrowings, which have become part of the stable Megrelian lexicon, e.g. *gat͡ʃɛrɛba* Megrelian equivalent *gat͡ʃɛmɛba* 'stop', *gzad͡ʒvarɛdini* Megrelian equivalent *ʒir ak'artu ʃara* 'crossroad', *k'ɛtɛba* Megrelian equivalent *ɣɔlama* 'making', and;
- Recent, less conventionalized forms, which may still be perceived as foreign or are not yet widely used, e.g. *gviq'vɛbudɛs* Megrelian equivalent *mɛt͡ʃiɛbudɛs* 'someone is telling us', *gamtknarɛns* Megrelian equivalent *gɔkirɔnapuans* 'makes (someone) yawn' etc.

This contrast highlights the dynamic nature of lexical borrowing and code-mixing in Megrelian. On the one hand, the continuous influx of Georgian lexical material demonstrates the language's adaptability; on the other hand, it raises questions about the stability of the native lexical system and may be considered an indicator of language endangerment.

5 Types of Code-Switching and Code-Mixing in Megrelian and their Representation in the Dictionary

The examples extracted from the *MLC* illustrate a range of contact-induced phenomena, including code-mixing, code-switching and lexicalization, which are subsequently processed both lexically and morphologically and incorporated into the Megrelian-English dictionary. For the purposes of this study, three main types are distinguished:

- Insertional mixing, defined as the insertion of L2 elements into an L1 (Megrelian) morphosyntactic frame;

- Alternational switching, referring to the shift between L1 and L2 at the level of clauses or sentence boundaries;
- Lexicalized mixing, involving the sharing and integration of morphosyntactic features across languages, resulting in hybrid forms.

5.1 Types of Code-Mixing

5.1.1 Insertional Mixing

Insertional mixing is the most frequent type observed in the corpus and involves the embedding of Georgian, Russian or international lexical items into Megrelian sentences. Examples 1-3 illustrate different sources of insertion: Georgian, Russian and internationalisms, respectively.

(1) *baɣanɔbas brɛi mudgasiɛni vjapɛndit, magajtɔ* ***mgɛlɔbanas, daxutʃ͡ʼɔbanas.***
In my childhood, I played many games, for example, game of catch, hide-and-seek.

(2) ***skʼɔrɔda*** *kuɣududa,* ***skʼɔrɔdat*** *ɔrtəd.*
If she had a gas oven, she baked it on a gas oven.

(3) *maʒia* ***vɛrsia****, namud* ***intʼɛrnɛtʼis*** *gɛd͡ʒu, tis xɔlɔ vtkuank.*
there is another version in the Internet

Example (1) illustrates the insertion of Georgian lexical items into a Megrelian text. Example (2) shows the insertion of Russian-origin items, while Example (3) demonstrates the insertion of a borrowing and an internationalism into a Megrelian text.

The inserted lexical items follow Megrelian morphosyntactic rules, which is consistent with the principles of L2 insertion into a matrix language framework (for the principles, see Myers-Scotton 1993; Auer & Muhamedova 2005). Since the FLEx annotation system involves morpheme segmentation and tagging, the roots of non-Megrelian lexical items in examples 1-3 are marked as L2, while suffixes are analyzed as part of the Megrelian grammatical system. Consequently, dictionary entries are represented in their stem form. Vowel-final nouns are represented with their full phonemic shape in word-final position, as the nominative marker is realized as a zero allomorph. In contrast, the nominative suffix /-i/ of consonant-final nouns is not reflected in the dictionary form (4).

(4) *mgɛlɔbana* — L2 game of catch; *daxutʃ͡ʼɔbana* — L2 hide-and-seek; *skʼɔrɔda* — L2 gas oven; *vɛrsia* — L2 version; *intʼɛrnɛtʼ* — L2 internet etc.

The FLEx database also provides evidence for so-called trigger words (Clyne 2009), i.e. lexical items arising at the intersection of two linguistic systems that may trigger a shift to another language during discourse. These include: (a) proper names, (b) lexical transfers and, (c) bilingual homophones. Among these, bilingual homophones are the least frequent in Megrelian and do not typically develop into borrowings. In some cases, borrowed items coexist with Megrelian equivalents, resulting in synonymic parallelism.

(a) Proper Names. Proper names are not translated, as they retain the same phonological form in both L1 and L2. However, anthroponyms may undergo Megrelian-specific mor-

phological adaptation, which goes beyond purely phonetic processes, e.g. *Nodia – Nodixe*, *Naroushvili – Narouskua*, *Chikava – Chika*, *Lukava – Luka*, etc. (Kartozia et al. 2011). In such cases, they may be treated as instances of code-switching.

(b) Lexical Transfers and Borrowings include items that have no Megrelian equivalent; fulfill a nominative function; are frequently found in technical, administrative, or scientific domains. Borrowings may also occur in cases where a Megrelian equivalent exists, creating a parallel lexical system influenced by sociocultural and pragmatic factors, e.g. *adgili* 'place' (129), *adre* 'early' (57), *rituali* 'ritual' (58), *raioni* 'district' (42) etc.

(c) Bilingual Homophones form a distinct group. One group within this category consists of lexical items that function as elements used to organize the process of communication (Matras, 1998). These include discourse markers (words and phrases) such as *anu* 'that is', *ese igi* 'i.e.', *kho* 'you know', *nu* 'well', *aba* 'come on', *upro* 'more, mostly', *ra tkma unda* 'of course', *ki batono* 'certainly', *magalito* 'for example', *vabshe* 'basically', *karoche* 'in short', and *okey* 'okay'.

The second group consists of words that have identical orthographic and phonological forms in two different languages but differ in meaning. For example: L1 axla 'now' vs L2 *ase* 'thus'; L1 *tu* (conjunction/particle; including the discourse particle /-da/) vs L2 *tu* (conjunction); L1 *goči* 'piglet' vs L2 *tu* (conjunction). These items illustrate the complexity of cross-linguistic interaction and the role of phonological overlap in bilingual speech.

5.1.2 Alternational Code-Switching

Alternational code-switching occurs at clause or sentence boundaries, where speakers shift from Megrelian to Georgian (or vice versa) for entire segments of speech (5-6)

(5) *iprɛli gvart͡ʃkilɛ,* ***mara p'rakt'ik'uli saubari gvit͡ʃ'irs.***
Megrelian: We understand everything, **Georgian**: *but we have difficulty with practical conversation*.

(6) ***isɛti k'argi kali iq'ɔ*** *dɔ tik ma mumɔnt͡s'q'u* ***sanɛp'idst'ant͡sias samuʃaɔd.***
Georgian: She was such a good woman, **Megrelian**: and it was her who helped me find **Georgian**: a job at the **Russian**: sanitary service.

Example (5) represents a complex syntactic construction in which code-switching occurs at the boundary between the first and second clause, shifting from Megrelian to Georgian. The connective *mara* 'but', used to link the clauses, belongs to both Megrelian and Georgian dialectal lexical inventories.

In example (6), the direction of switching is reversed: the utterance begins in Georgian (*isɛti k'argi kali iq'ɔ* 'she was such a good woman'), while the subsequent segment continues in Megrelian. Such switching patterns are commonly interpreted in the literature as psycholinguistically motivated (Vogt 1954; Gumperz 1958 and others).

In this case, the use of a formulaic phrase from the dominant language (Georgian) at the beginning of the utterance contributes a particular discursive and expressive effect, add-

ing emphasis and stylistic nuance. It also reflects the speaker's conscious and intentional choice to switch codes, motivated by psycholinguistic factors.

At the end of example (6), a multilingual mixing strategy is observed. The Russian-derived lexical item *sanɛp'idst'antsia* 'sanitary service' and the Georgian form *samuʃaɔd* 'for working' represent two distinct linguistic codes within the same utterance. This combination reflects the speaker's self-identification with a multilingual environment.

As it was mentioned above, the final word *samuʃaɔd* 'for working' is both lexically and morphologically Georgian and appears in the adverbial case, marked by the suffix /-d/. In the FLEx annotation system, such forms are treated as structurally indivisible units, i.e. they are not segmented into morphemes in the analysis field and are represented in the dictionary as complete L2 forms. This principle applies to all units that are lexically and morphologically fully attributable to L2.

5.1.3 Lexicalized mixing

Lexicalized mixing results in the formation of hybrid compounds and morphologically integrated structures that have become conventionalized in Megrelian. In this study, lexicalization is understood as the sharing and integration of morphosyntactic features across languages, affecting both inflectional and derivational bound morphemes. The following discussion focuses on the interaction of such morphemes in the structure of the Megrelian verb.

The Megrelian verb is characterized by a high degree of morphological complexity (Kartozia et al. 2010; Kiria et al., 2021), with most grammatical categories expressed through a series of affixal elements. These bound morphemes are distributed both to the left of the root (prefixal domain) and to the right of the root (suffixal domain).

Prefixal Domain

Negation marker -> Perfective/affirmative marker -> Preverb -> Imperfective marker -> Evidential marker -> Person markers -> Voice / version / causative / potential markers -> R (root)

Suffixal Domain

R (root) -> Augment / thematic marker -> Voice / causative / thematic suffix -> Passive / potential / evidential marker -> Tense–aspect marker -> Paradigm marker / subordinative marker -> Person–number markers -> Emphatic vowel -> Conditional marker -> Clitics (modal particles, conjunctions)

These morphological templates provide the structural framework within which L2 elements are integrated, either through direct insertion or through adapted incorporation. As a result, lexicalized mixing in Megrelian operates not only at the lexical level but also at the morphemic and grammatical levels, demonstrating a high degree of structural integration.

In the FLEx database, the verbal lexicon reveals multiple patterns of interaction between L1 (Megrelian) and L2 (primarily Georgian) morphemes. Two main strategies can be distinguished:

(a) Direct Insertion of L2 Morphemes (6-8)

In this pattern, L2 morphemes (roots, stems or affixes) are inserted into the Megrelian verbal matrix without structural modification. This corresponds to an insertional strategy, whereby the overall morphosyntactic frame remains Megrelian.

(b) Adapted Integration of L2 Morphemes (9-13)

In this case, L2 morphemes are incorporated into the L1 matrix with phonological and morphotactic adaptation, reflecting constraints specific to Megrelian. These adaptations may involve changes in phonological shape or positional behaviour in accordance with Megrelian morphological rules.

The analysis confirms that borrowed elements and mixed forms, even when originating from another language, are systematically shaped by the morphosyntactic structure of the matrix language, in line with the Matrix Language Frame model (Myers-Scotton 1993).

5.1.3.1 Direct Insertion Strategies

The insertional strategy involves the integration of L2 morphemes into the Megrelian (L1) verbal matrix. Examples (7) and (8) demonstrate that, within the prefixal domain of L1 verbs, morphemes may appear that are shared with L2 in terms of phonetic inventory, grammatical function, and structural position (i.e. as prefixes or suffixes). In such cases, the verbal root is selected from the L2 lexicon, while the surrounding morphological structure is largely provided by L1. At the same time, towards the anlaut of the verb, L1 morphemes reappear, resulting in hybrid structures in which the matrix language remains Megrelian, despite the presence of L2 lexical material. In example (7), the root morpheme is of Georgian origin, whereas in example (8) it is derived from Russian:

(7) *a-**grd͡zɛl**-ɛn-k*
apll-**long:L2**-ts-s2
'keep going'

(8) *dɔ-v-**t'at͡ʃil**-an-k*
prv-s1-**sharpen:L2**-ts-s1
'sharpen'

In addition to L2 roots, Megrelian verbal structures may also incorporate L2 preverbs and thematic suffixes. For instance, in example (9), the Georgian preverb-containing stem ganvitar- is inserted into the Megrelian verbal matrix, while in example (10), a Georgian root combines with the thematic suffix /-eb/:

(9) *va-**ganvitar**-d-u*
neg-**develop**:**L2**-impf-s3sg
'developed'

(10) *i-**kmn-ɛb**-u-**u**-d-u*
pass-**create-ts:L2**-pass-**ts:L1-**impf-subj3sg
'to be created'

In example (10), two thematic suffixes co-occur: the Georgian /-eb/ and the Megrelian /-u/. In morphological analysis, both are identified separately. According to the shared grammatical principles of Georgian and Megrelian verbal paradigms, both thematic suffixes are removed in the second series of the TAME system, leaving the verbal roots (e.g., *kmn-, q'v-*). This indicates that the Georgian suffix retains an active grammatical function within the paradigm. For this reason, the suffix /-eb/ is included in the dictionary and marked with the tag TS:L2.

In rare cases, L2 morphemes are borrowed together with their grammatical function. Example (11) illustrates the emergence of the Georgian prefix /gv-/ in Megrelian verbal structure. In Georgian, this prefix encodes first person plural object agreement, whereas Megrelian typically employs different grammatical mechanisms to express this category. However, in bilingual speech, this prefix may appear as a functional borrowing:

(11)***gv****-i-****q'v-ɛb****-u-****u****-d-u*
O1pl:L2-appl-naratted**:L2-ts:L2**-pass-**ts**-impf-s3sg
'narrated'

In addition, Georgian preverbs (e.g., *dɑ-, ɑ-, gɑn-, t͡ʃɑ-*) may be directly incorporated into Megrelian word structure and are listed in the dictionary as L2 prefixes (preverbs), including their phonetic variants (e.g., dɑ- with variants dɔ-, dɛɛ-; gɑn- with variant gɑɑn-).

These forms are derived from examples such as (12) and (13). In example (12), the Georgian preverb /aɣ-/ is preserved unchanged, corresponding to its use in the Georgian lexeme *aɣ-sruleba* 'to carry out'. In example (13), the Georgian preverb /ga-/ (as in *ga-vrt͡s'eleba* 'to spread') undergoes phonetic adaptation within Megrelian, following the transformation sequence /ga- -> go- -> gu-/:

(12)***aɣ****-i-****srul****-in-u*
prv:L2-pass-carry out**:L2**-ts-s3sg
'was carried out'

(13)*gu-v-ɑ-****vrt͡sɛl****-i-k'ɔ*
prv-s1-apll-spread**:L2**-pm-cond
'in order to spread it'

5.1.3.2 Adapted Integration Strategies

Many borrowed or mixed items exhibit partial integration into Megrelian morphology. In such cases, Georgian roots are combined with Megrelian affixes and inflectional endings, resulting in hybrid verbal or nominal forms (14).

(14)*ɑ-tm-****ɑ****-b-****ɣ****-****niʃn****-ɛn-k*
prv-impfv-**prv:L2**-sub1-**prv:L2**-celebrate-ts-subj1sg
"I will mark"

This form combines the Georgian verb root *-niʃn-* with Megrelian inflectional morphology. Such examples demonstrate that code-mixing operates at the morphemic level, reflecting systematic grammatical adaptation rather than random alternation. Also, it represents a

more complex type of L1-L2 interaction, in which elements from the embedded language (L2) are not simply inserted into the matrix language (L1) structure, but are restructured in accordance with the morphotactic and phonotactic constraints of Megrelian.

In this example, the Georgian root *-niʃn-* appears in its preverb form *ɑɣ-niʃn-.* According to Megrelian morphotactic rules, the S1 allomorph /b-/ is metathesized into the Georgian preverb /ɑɣ-/, effectively splitting it and yielding the form *ɑ-b-ɣ-niʃn-*.

The Georgian preverb is thus desemantized, as evidenced by the addition of the Megrelian preverb /ɑ-/ in the anlaut position, which fulfills the same grammatical function. This duplication indicates that the original L2 preverb no longer carries its full functional load within the hybrid structure.

In such hybrid formations, the segments /ɑ/ and /ɣ/ should ideally be treated as components of a single underlying preverb, although this relationship is not fully recoverable in the dictionary representation. Consequently, the lexicographic treatment of such forms cannot always capture the complete morphosyntactic history of the item.

5.2 Combined Code-Mixing

The Megrelian speech of speakers from the city of Poti (Samegrelo region) can be regarded as a particularly illustrative case of combined code-mixing, code-switching and lexicalization, in which the degree of language endangerment becomes especially evident.

Poti is a Black Sea port city that has historically attracted migrant populations from different regions in search of employment. At the same time, it borders the Guria region, where the Gurian dialect of Georgian is spoken. Intensive and prolonged contact between these linguistic communities has fostered frequent code-switching and code-mixing. As a result, Megrelian has largely disappeared from everyday social use in this area and the intergenerational transmission of both the language and its associated cultural practices has been significantly disrupted.

An illustrative example from the Poti variety demonstrates the interaction of multiple types of code-mixing:

(15)***gamsazɣvrɛli*** *tɛ* ***pakt'is*** *rdu* ***samuʃaɔ adgilɛbi*** *-* ***p'ɔrt'i, sanavsadgurɔ kalaki, p'irɔbɛbi.***
'The determining factors of this act were the workplaces - the port, the harbor city, and the working conditions.'

(16)***p'irɔbɛb-i*** *i-****kmn-ɛb****-u-d-u* ***samuʃaɔ adgilɛbis****.*
'The working conditions would determine the workplaces.'

In example (15), only the demonstrative *tɛ* 'this' and the verb *rdu* 'was' are Megrelian, while the remaining lexical material is of Georgian origin. In contrast, example (16) follows Megrelian syntactic structure, but the verb root and nominal stems are derived from Georgian, with Megrelian morphological markers applied.

Such constructions illustrate how speakers combine Georgian lexical material with Megrelian grammatical templates, resulting in highly hybridized utterances. These combined

patterns of code-mixing, code-switching, and lexicalization reflect not only bilingual competence but also the advanced stage of language shift, in which the structural integrity of the minority language is increasingly reduced.

6 Conclusions

The findings of this study demonstrate that code-mixing in Megrelian is a structured and productive phenomenon, rather than a mere result of lexical borrowing. The analysis of corpus data shows that the integration of L2 elements, primarily from Georgian, is governed by systematic grammatical principles and operates across multiple linguistic levels, including the lexical, morphological and syntactic domains.

The results highlight a high degree of bilingual competence among Megrelian speakers, who are able to combine elements from different linguistic systems while maintaining overall structural coherence. In particular, the data reveal considerable morphological adaptability, whereby Georgian lexical material is incorporated into Megrelian morphosyntactic frameworks. This process is not limited to surface-level insertion, but often involves deeper integration at the morphemic level.

At the same time, the distribution of code-mixing patterns reflects clear sociolinguistic variation. Younger and urban speakers show a higher frequency of mixing, indicating the influence of social, economic, and communicative factors on language use. In certain contexts, especially in environments characterized by intensive language contact, combined forms of code-mixing, code-switching, and lexicalization may also point to an advanced stage of language shift.

From a lexicographic perspective, these findings justify the inclusion of mixed and hybrid forms as dictionary entries. The use of L2 tagging in the Megrelian-English dictionary provides a systematic way to document the origin and structure of such items, thereby offering a more accurate representation of contemporary language use. This approach reflects the linguistic reality of Megrelian, in which multilingual interaction forms an integral part of speakers' everyday communication and identity.

Finally, the data suggest that certain morphemes shared between Megrelian (L1) and Georgian (L2) may be analysed as bilingual homophones, particularly in cases where morphosyntactic markers are phonologically identical, functionally equivalent, and occupy corresponding positions within the morphological structure, for example in prefixal or suffixal chains. This observation further underscores the depth of structural convergence between the two languages.

Overall, the study contributes to the understanding of code-mixing in endangered languages by demonstrating that such phenomena are not merely indicators of language erosion, but also reflect systematic linguistic processes that can be effectively captured through corpus-based and lexicographic methods.

References

Auer, P. & Muhamedova, R. (2005). Embedded Language and Matrix Language in Insertional Language Mixing: Some Problematic Cases." In *Italian Journal of Linguistics* 17 (1), pp. 35-54.

Austin, P. K. (2006). Data and Language documentation. In *Essentials of language documentation*, by N. P. Himmelmann, & U. Mosel (Eds.) J. Gippert, pp. 87-113. Berlin & New York: Mouton de Gruyter.

Bowern, C. (2008). *Linguistic fieldwork: A practical guide.* New York: Palgrave Macmillan.

Clyne, M. (2009). *Community Languages: The Australian Experience.* Cambridge University Press.

Gersamia, R. & Lobzhanidze, I. (2021). *Mingrelian Converter.* 03 17. https://xmf.iliauni.edu.ge/converter/converter.html.

Gersamia, R. & Lobzhanidze, I. (2022-2025). *Megrelian Language Corpus (MLC).* Accessed 02 28, 2026. https://xmf.iliauni.edu.ge/.

Gumperz, J. (1982). *Discourse Strategies.* Cambridge: Cambridge University Press.

Haugen, E. (1956). *Bilingualism in the Americas: A Bibliography and Research Guide.* Alabama: American Dialect Society.

Hoo, Y. (2007). Code-mixing: Linguistic Form and Socio-cultural Meaning. In *The International Journal of Language, Culture and Society,* 21, pp. 23-30.

International, SIL. (2024). *SIL Feildworks Language Explorer (FLEx), Version 9.1.* January 01. https://software.sil.org/fieldworks/.

Kartozia, G., Gersamia, R., Lomia, M. & Tskhadaia, T. (2010). *megrulis lingvisturi analizi [Linguistic analysis of Megrelian].* Tbilisi: Meridiani.

Kiria, Ch., Ezugbaia, L., Memishishi, O. & Chukhua, M. (2021). *Laz-Megrelian Grammar. I. Morphology [lazur-megruli gramatika. I. morp'ologia].* Tbilisi: Meridiani.

Makharoblidze, T. & Gersamia R. (2022). "On Mountainous Samegrelo: General description and linguistic situation." *Vitalité sociolinguistique des langues des massifs montagneux: Alpes et Caucase. 17. Sezione IV. Sociolinguistica*, pp. 303-324.

Matras, Y. (1998). Utterance Modifiers and Universals of Grammatical Borrowing. In *Linguistics,* 36-2, pp. 281-333.

Myers-Scotton, C. (1998). *Codes and Consequences: Choosing Linguistic Varieties.* New York: Oxford University Press.

Myers-Scotton, C. (1993). *Social Motivations for Codeswitching: Evidence from Africa.* Oxford: Clarendon Press.

Poplack, Sh. (1980). Sometimes I'll Start a Sentence in Spanish y Termino en Espanol: Toward a Typology of Code-switching. In *Linguistics* 18 (233-234), pp. 581-618.

Poplack, Sh., Sankoff, D. & Miller, Ch. (1988). The Social Correlates and Linguistic Processes of Lexical Borrowing and Assimilation." In *Linguistics* 26 (1), pp. 47-104.

Redouane, R. (2005). Linguistic Constraints on Codeswitching and Codemixing of Bilingual Moroccan Arabic-French Speakers in Canada. In J. Cohen, K. T. McAlister, K. Rolstad, & J. MacSwan (Eds.), ISB4:. 2005. "Linguistic Constraints on Codeswitching and Codemixing of Bilingual Moroccan Arabic-French Speakers in Canada." *Proceedings of the 4th International Symposium on Bilingualism.* Arizona: Cascadilla Press. 1921-1933.

Vogt, H. (1954). Language Contacts. In *Word* 10(2−3), pp 365−374.

Weinreich, U. (1953). *Languages in Contact.* The Hague: Mouton.

Peculiarities of Linguistic Modelling of English and Georgian Scientific Terminology

Tamar Kvitsiani
Ivane Javakhishvili Tbilisi State University
E-mail: tamunakviciani@yahoo.com

Abstract

In today's rapidly evolving scientific and technological landscape, the study of terminological issues has gained increasing significance. Terminology is a fundamental prerequisite for the development and effective functioning of scientific disciplines and professional domains. Inconsistent, ambiguous, or inadequately standardized terminology may impede progress within a given field. Therefore, greater attention should be paid to terminological work in Georgia.

As T. Cabré notes, scientific and technological development is predominantly concentrated in economically advanced countries, where the production of knowledge is most intensive. As a result, other, economically less powerful countries largely acquire scientific knowledge through these centres, which leads to extensive borrowing of technical and scientific vocabulary. Within terminology studies, this situation is often associated with asymmetrical knowledge transfer and linguistic dependency. Consequently, such countries are faced with the need to regulate terminological borrowing in order to avoid uncontrolled lexical influx that may weaken the structural integrity and functional stability of their languages (Cabre, 1992). Georgian linguists and subject-field experts note that terminological development in Georgia faces a number of challenges. A significant proportion of terms across various disciplines are borrowed, which is generally considered an undesirable tendency.

The importance of this issue prompted an in-depth investigation into term-formation methods in modern Georgian terminology, comparing them both with methods used in the twentieth century and with English term-formation patterns. This paper presents the study's findings.

Keywords: terminology; Georgian term-formation methods; term-formation methods in English; etymology; the influence of Russian

1 Introduction

In the context of ongoing scientific and technological advancement, the study of terminological issues is of particular importance. Numerous new concepts emerge on a daily basis,

which creates the need to designate them, that is, to assign names to them that comply with international standards as well as the specific linguistic norms of a given language (ISO 704:2022: 83–104). Consequently, the regulation of term-formation processes is crucial for the survival and proper development of any language.

Georgian linguists and field specialists note that greater attention should be paid to terminological work in the country. Many terms across various fields are mainly borrowed words, a tendency that is generally seen as undesirable and negative (Gogia 2023; Margalitadze 2019; Karosanidze 2012).

To examine this tendency, we conducted a quantitative study in 2022–2023 based on a dataset of 1,600 terms drawn from various subfields of biology. The primary aim was to determine the proportion of borrowed terminology in modern scientific domains, with genetics, immunology, and biotechnology selected as the focus areas. A total of 800 terms from these fields were examined, and the results showed that approximately 75–80% of them are loanwords. This finding was then compared to terminology from more traditional disciplines, reflecting the term formation patterns in the 20th century. For this purpose 800 terms from botany, zoology, and anatomy were selected. The analysis revealed a radically different tendency, as nearly 80% of the examined terms were found to be of Georgian origin, formed on the basis of Georgian language resources (Kvitsiani 2023).

The results of this study motivated an in-depth investigation of term-formation methods in Georgian terminology. The research was conducted in two stages. In the initial stage, term-formation methods in more traditional fields (botany, zoology, and anatomy) were examined. In the second stage, term-formation methods in fields that developed relatively later (immunology, biotechnology, and genetics) were analysed and compared with those used in the twentieth century. The study of term-formation methods also included a comparison with equivalent English and Russian terminology. English was included because modern Georgian terminology is largely influenced by the English language, while Russian was considered due to its strong impact on Georgian terminology in the twentieth century.

The results of the first stage of the research have been published in the electronic bilingual scholarly journal *Spekali* (Kvitsiani, 2024). The study revealed that the primary source of Georgian botanical, zoological, and anatomical terminology is the Georgian language itself, accounting for approximately 80% of the analysed terms. These findings further demonstrate that, in the greater part of the twentieth century, terminology was developed in accordance with Georgian linguistic norms, making extensive use of the resources of the Georgian language.

Georgian terminologists most frequently employed structural borrowing in the process of term formation. It has also been shown that, due to the political context of the country, twentieth-century Georgian terminology was significantly influenced by Russian, and many terms were created on the basis of Russian structural models. Nevertheless, this process ultimately resulted in transparent, easily understandable, and readily interpretable Georgian terms. In addition, Georgian terminologists actively drew on Latin models for structural borrowing. The study has also revealed that Georgian terminologists created new terms from common words through metaphorical extension, assigning them specialized terminological meanings. Word formation is another frequently used method. Direct borrowings

in these fields are relatively rare, accounting for approximately 20% of the analysed terms. The sources of these borrowings were primarily Russian and Latin.

The study of English terms from the fields of botany, zoology and anatomy showed the characteristic features of their modelling. For example, most of the one-word terms are formed from Latin and sometimes Greek roots. Latin and Greek borrowings express the feature which is one of the characteristics of the respective concept. Analytical terms, by contrast, are predominantly formed using English lexical resources. By English resources, we refer to words of Germanic origin, also internal creations in English, as well as words borrowed from French or other languages in earlier stages of the development of English that are now fully integrated into its lexicon. In this case, the term-formation process is likewise motivated by the inherent characteristics of the concept. In the process of term-formation Latin names of plants and animals are also used applying the method of structural borrowing (Kvitsiani, 2024).

2 Data and Methodology

The data for the second stage of the research were based on the dataset of 800 terms drawn from the abovementioned study (Kvitsiani, 2023), from which a random sample of 100 terms was selected, representing three subfields of biology: immunology, biotechnology, and genetics. To examine traditional Georgian term-formation methods, the study relied on the approaches and principles outlined in the monograph by the Georgian terminologist R. Ghambashidze, *Georgian Scientific Terminology and the Main Principles of its Compilation* (1986). To analyse the key characteristics of modern terminology, reference was made to international terminological standards, specifically ISO 704:2022, *Terminology Work – Principles and Methods*.

As noted above, the study employed a comparative approach involving three languages – Georgian, Russian, and English. In order to obtain a more in-depth understanding of English term-formation processes, we also applied the etymological method of analysis. The etymological origins of English terms were identified with the help of the *Oxford English Dictionary of Historical Principles* (OED) and the *Online Etymology Dictionary* available at the URL etymonline.com.

3 Tendencies in the Modelling of English and Georgian Terms: The Case of Immunology, Biotechnology, and Genetics

The randomly selected 100 terms from the domains of immunology, biotechnology and genetics were analyzed from the point of view of term-formation methods. The in-depth study of these terms has revealed the extent of the influence of the English language on Modern Georgian terminology. English terms are borrowed directly into Georgian or Georgian terms are loan translations of their English equivalents. There are cases when even the analytical terms are borrowed into Georgian which is a very negative tendency for modern Georgian terminology. Below are given some examples of analysed terms (1-15):

(1) Affinity – an English immunological term of the Latin origin.

The Georgian equivalent of this term is *აფინობა / apinoba*.

The Russian equivalent of this term is *родство*.

***

As the comparison of these three terms shows, the English term is a word of the Latin origin. This word is polysemous in English. It is used in other domains and has migrated to the field of immunology at a later stage. The Georgian term is borrowed from English. The Russian term is formed from a Russian word, it is not a borrowing.

(2) Anergy – an English immunological term that means the absence of a normal immune response to a particular antigen. This English term is of the Greek origin (OED) and denotes one of the characteristics of the concept.

The Georgian equivalent of this term is *ანერგია / anergia*.

The Russian equivalent of this term is *толерантность*.

***

The comparison of these three terms demonstrates that the English term is formed from an English word of the Greek origin. The Georgian term is borrowed from English. The Russian equivalent is not formed from English.

(3) Antigenic shift – an English immunological term. It is an analytical term consisting of the word of a German origin *antigen* and native English *shift.*

The Georgian equivalent of this term is *ანტიგენური შიფტი / ant'igenuri shipt'i*.

The Russian equivalent of this term is *антигенный сдвиг.*

***

As the analysis of these three terms reveals, the English term is formed by the combination of English words, one of which is of the German origin. The Georgian term is borrowed from English; as for the Russian term, it is a loan translation, presumably from English.

(4) Bivalent – an English immunological term. Its components *bi-* and *valent* are of the Latin origin.

The Georgian equivalent of this term is *ბივალენტური / bivalent'uri*.

The Russian equivalent of this term is *бивалент*.

***

As the comparison of these three terms shows, English term is formed from English elements of the Latin origin, whereas the corresponding Russian and Georgian terms are borrowed from English.

(5) Cognate – an English immunological term of the Latin origin.

The Georgian equivalent of this term is *კოგნატური, მსგავსი / k'ognat'uri, msgavsi*.

The Russian equivalent of the word is *похожий; сродный.*

***

A comparison of these three terms shows that English term is formed from an English word of the Latin origin. It is a legal term, and has non-terminological meanings as well. In this case this legal term migrated into the domain of immunology. The Georgian equivalents of this term consist of both, a borrowing and a Georgian word. Russian equivalents are formed from Russian words.

(6) Combining site – an English immunological term. It is formed by the combination of words of the French origin *combine* and *site* (OED).

The Georgian equivalent of this term is *კომბინატორული საიტი / kombinatoruli saiti*.

The Russian equivalent of this term is *связывающий сайт*; *комбинаторный сайт.*

***

Comparing these three terms demonstrates that English term is formed from English words of the French origin. The Georgian term is a borrowing, while the Russian equivalent terms are loan translations from English.

(7) Complement – an English immunological term, formed from an English word of the Latin origin.

The Georgian equivalent of this term is *კომპლემენტი / komplementi*.

The Russian equivalent of this term is *комплемент*.

***

As can be seen from the comparison of these three terms, English term is based on an English word of the Latin origin. This word is used in different fields in the English language. Georgian and Russian equivalents are borrowed from English.

(8) Constant domain – an English immunological term, formed by the combination of *constant* and *domain*, words of the Latin origin.

The Georgian equivalent of this term is *კონსტანტური დომენი / konstanturi domeni*.

The Russian equivalent of this term is *константная область; С-домен.*

***

As the analysis of these three terms demonstrates, English term is formed from English words of the Latin origin, that have several polysemic meanings including terminological. The Georgian analytical term is borrowed from English, whereas the Russian equivalents are calques.

(9) Effector cell – an English immunological term that consists of *effector* and *cell,* words of the Latin origin.

The Georgian equivalent of this term is *ეფექტორული უჯრედი / epekt'oruli ujredi.*

The Russian equivalent of this term is *клетка-эффектор*.

***

The analysis of these three terms shows that the English term is formed from English words of the Latin origin. As for the Georgian and the Russian equivalents, they are calques from English.

(10) Follicular dendritic cell – an English immunological term, that consists of the words *follicle, cell* (Latin origin) and *dendrite* (Greek origin).

The Georgian equivalent of this term is *ფოლიკულური დენდრიტული უჯრედი / polik'uluri dendrit'uli ujredi*.

The Russian equivalent of this term is *фолликулярная дендритная клетка.*

***

As illustrated by the comparison of these three terms, English term is formed from English words of the Latin and Greek origins. Georgian and Russian equivalents are calques of the English term.

(11) Germinal centre – an English immunological term formed from words of the Latin origin *germinal* and *centre*.

Georgian equivalent of this term is *გერმინაციული ცენტრი / germinatsiuli tsent'ri*.

Russian equivalent of this term is *зародышевый центр.*

***

The English analytical term is formed from words of the Latin origin. The Georgian equivalent is a borrowing, while the Russian term is a calque.

(12) High endothelial venule – an English immunological term that is formed from English words of Germanic (*high*), German (*endothelioma*) and Latin origin (*venule*).

The Georgian equivalent of this term is *მაღალი ენდოთელიალური ვენულა / maghali endotelialuri venula*.

The Russian equivalent of this term is *наружная эндотелиальная венула.*

***

As can be seen from these terms, Georgian and Russian terms are only partly calques of the respective English term, as only the first components of this analytical term are replaced by Georgian and Russian words.

(13) Homing – an English immunological term formed from a native English word.

The Georgian equivalent of this term is *ჰუმინგი/ houmingi*.

The Russian equivalent of this term is *хоуминг.*

***

The English term is formed from a native English word. It is a polysemous word and has several terminological meanings. This term migrated to immunology from other domains. Georgian and Russian terma are borrowed from English.

(14) Natural killer cell – an English immunological term that consists of English words of French origin (*natural, cell*) and a native English word *kill*.

The Georgian equivalent of this term is *ბუნებრივი კილერული/მკვლელი უჯრედი / bunebrivi k'ileruli/mk'vleli ujredi*.

The Russian equivalents of this term are *нормальная клетка-киллер; натуральная клетка-киллер*.

***

As the analysis of these three terms demonstrates, English term is formed from English words of native English and French origin, whereas Georgian and Russian equivalents are calques of the English term.

(15) Southwestern blot – an English biotechnological term that consists of native English words *south, western* and *blot*.

The Georgian equivalent of this term is *სზერნ-ვესტერნ ბლოტინგი / sazern-vest'ern blot'ingi*.

The Russian equivalent of this term is *саузерн блот анализ*.

***

As the comparison of these three terms reveals, English term is formed from native English words. As for its Georgian and Russian equivalents, they are borrowings from English.

4 Discussion

The analysis of terms from the fields of immunology, biotechnology, and genetics revealed the significant influence of English on contemporary Georgian terminology. A comparison of the Georgian and Russian equivalents of English terms showed that Georgian terms are formed under the influence of English rather than Russian. They are either calques of English terms or direct borrowings from English.

The study of contemporary English term-formation patterns has revealed that one-word terms are often polysemous and may possess multiple terminological meanings. Most one-word terms in English arise either through the migration of common words into specialised terminology or through the transfer of terms from one domain to another. This indicates that terminologists frequently employ metaphorical extension and semantic transfer in the process of term formation. In the case of one-word terms, words of Latin and Greek origin are particularly common. The designations of concepts typically reflect one of the characteristics of the concept concerned.

Most terms are multi-word (analytical) terms. In English, they are formed from lexical units of Greek, Latin, French, and occasionally German or other origin. Having been borrowed at

earlier stages of the language's development, these words have become an integral part of the English lexicon. As with one-word terms, the formation of analytical terms is generally based on the selection of a characteristic feature of the concept denoted.

The analysis of contemporary English terms revealed a high degree of transparency in the terminology of these fields. This transparency results from the use of common words and terms borrowed from other domains in the term-formation process. Multi-word terms are likewise transparent, as they are formed from lexical units that either already exist in the language or were borrowed from other languages at earlier stages of the development of English. A contrasting situation can be observed in Georgian, where English terms are frequently introduced through direct borrowing. In some cases, even multi-word terms are transliterated, and their components are not replaced by Georgian equivalents. As a result, many contemporary Georgian terms are neither transparent nor semantically motivated.

The analysis of the terms from the fields of immunology, biotechnology and genetics revealed that English language has significant influence not only on Georgian, but on Russian as well. The study of the terms from these fields proved that most of the Russian terms are formed according to the patterns of English terms. In most cases Russian terms are either transliterated from English or terminologists apply the method of structural borrowing to form terms. It can be said that some Georgian terms of these fields may be introduced into Georgian not directly from English, but from Russian and modern Georgian terminology may still be under the Russian influence.

5 Conclusion and Recommendations

The study has demonstrated that traditional (20th-century) and modern term-formation methods in Georgian differ significantly. Traditional Georgian terminology, exemplified by terms from the fields of botany, anatomy, and zoology, is predominantly based on the resources of the Georgian language and follows its grammatical and word-formation rules. As a result, the terminology is generally transparent and easily comprehensible. The research has also shown that, in the 20th century, Georgian terminologists employed a range of methods in term creation, including structural borrowing, semantic borrowing, metaphorical extension, and word-formation processes. Overall, the term-formation process was strongly influenced by the Russian language, and the majority of Georgian terms were formed by the analogy of respective Russian terms.

In contrast, modern Georgian terminology, examined on the basis of terms from the fields of immunology, biotechnology, and genetics, is strongly influenced by the English language and predominantly consists of direct borrowings and calques from English. There is a tendency to borrow even multi-word terms that are neither concise nor transparent or easily comprehensible. In many cases, the norms of the Georgian language are not consistently observed. Unlike this, English terminology is generally clear and transparent, as it is based either on common lexical items that have been transferred into specialised usage or on terms that migrate from one domain to another. Thus, English actively exploits existing linguistic resources in the formation of contemporary terminology.

Proceeding from the results of our research we came up with some recommendations for terminology work in Georgia:

1. If an English one-word term is created through the metaphorical extension of a common word and its transfer into specialised terminology, the method of semantic borrowing can be applied to create a Georgian equivalent by analogy with the English term. The history of Georgian terminology demonstrates that Georgian terminologists have skillfully transferred words from the general vocabulary into specialised terminological use. They have also employed the method of semantic borrowing from Russian. We believe that these approaches should be preferred to direct borrowing in the process of term formation, particularly in the case of terms based on common words. The application of this method may lead to polysemy; however, the polysemy of terms is now recognised as one of the characteristic features of modern terminology and is acknowledged in international terminological standards.

2. In the process of introducing multi-word terms into Georgian, the method of structural borrowing should be applied. This approach was employed by Georgian terminologists and translators in the Middle Ages, as well as in the 20th century and has also been widely used in various European languages, including Russian. The transliteration of analytical terms should be regarded as unacceptable. According to our study, the number of analytical terms has increased in modern terminology. The application of structural borrowing would enrich modern Georgian terminology with transparent and readily comprehensible terms.

3. Georgian word-formation means can be successfully applied in the formation of Georgian terminology.

4. Developing a terminological policy and regulating terminological work are priorities in many countries, where language and terminological policies are often governed by legislation. This should also become a priority in Georgia. The process of terminology creation should involve government agencies, terminologists, linguists, and subject-field specialists working together to coordinate and oversee terminological work, which is currently rather fragmented. When introducing terms into Georgian, priority should be given to transparency, motivation, and comprehensibility, ensuring that terms are easily understood by specialists, students, and the wider Georgian-speaking public.

References

Cabre, M. T. (1992). *Terminology – Theory, Methods and Application*. Volume 1. Amsterdam: John Benjamins publishing company.

English-Georgian Online Biology Dictionary. Accessed at: http://bio.dict.ge/ka/ [18.11.2025].

Ghambashidze, R. (1986). Georgian Scientific Terminology and Main Principles of its Compilation. Tbilisi: "Metsniereba".

Gogia, N. (2023). Terminological Inconsistences and Translation Problems in Legislative and Other Institutional Documents. *Proceedings of the I International Conference "Lexicography in the 21st Century".* Tbilisi: Ilia State University Press.

ISO 704: 2022. Terminology work – Principles and Methods.

Karosanidze, L. (2012). The Difficulties of Georgian Terminology (History and Modernity). *II International Symposium in Lexicography.* Batumi: Batumi University Press.

Kumar, Sh. (2007). *Modern Dictionary of Genetics.* Deep & Deep Publications.

Kvitsiani, T. (2024). Methods of Georgian Term Formation: On the Example of Botanical, Zoological and Anatomical Terms. A Scientific Bilingual Peer-Reviewed Online Journal *Spekali.* Tbilisi: Ivane Javakhishvili Tbilisi State University Press.

Kvitsiani, T. (2023). Georgian Terminology: Tradition and Present State. *Proceedings of the I International Conference "Lexicography in the 21st Century".* Tbilisi: Ilia State University Press.

Lawrence, E. (2008). *Henderson's Dictionary of Biology* (14th edition). Benjamin Cummings.

Margalitadze, T. (2019). Towards Issues of Terminological Policy in Georgia. *Proceedings of International Conference "Humanities in the Information Society – III".* Batumi: Batumi Shora Rustaveli State University Press. ISSN 1987-7625 ISBN 978-9941-462-86-3.

Martin, E., Hine, R. S. (2008). *A Dictionary of Biology* (6th edition). Oxford: Oxford University Press.

Online Etymology Dictionary. Accessed at: *https://www.etymonline.com/* [18/11/2025].

Oxford English Dictionary, second edition on CD-ROM. Version 3.1.

Singleton, P.& Sainsbury, D. (2006). *Dictionary of Microbiology and Molecular Biology* (Third Edition, Revised). Wiley.

Semantic and Formal Classification of Modern Georgian Neologisms

Tamar Laluashvili
Ilia State University
E-mail: tamar.laluashvili.1@iliauni.edu.ge

Abstract

This paper presents a semantic and formal classification of Georgian neologisms based on the analysis of a dataset of 1,700 new lexical units extracted from the Corpus of Georgian Neologisms. These neologisms were identified in a previous study that developed a semi-automatic method for detecting neologisms in Modern Georgian (Laluashvili & Margalitadze 2025). New lexical units appear in various fields and typically reflect social, cultural, political, and other processes within a language community. In this study, a total of 14 thematic groups were identified, illustrating the diverse contexts in which new words emerge: technologies and internet; business and finance; society and social life; politics; arts and entertainment; media; education; healthcare / medicine; infrastructure; food and kitchen appliances; environmental protection; beauty and law. With regard to the formal classification of modern Georgian neologisms, four main classes were identified: borrowings (direct borrowings, morphologically adapted borrowings, and loan translations), word formation (compounding and derivation), shortening, and semantic change. These classifications help us understand how new words are formed in modern Georgian. They also show how the language changes and adapts to contemporary communicative needs.

Keywords: neologisms; thematic groups; borrowing; compounding; derivation; affixes; affixoids

1 Introduction

New words or neologisms, that appear in a language, are crucial for its evolution. These new lexical units emerge across various fields and typically reflect the social, cultural, political, and other processes occurring within a language community. The appearance of neologisms indicates that fields, in which these new words are emerging, are evolving and developing. It is exactly these changes that keep a language alive and dynamic. In recent years, a range of new words has emerged in the Georgian language due to various developments in the country and around the world. A methodology for the semi-automatic detection of new words in Modern Georgian was developed at Ilia State University, which enabled us to collect 1,700 neologisms from the

Corpus of Georgian Neologisms (Laluashvili & Margalitadze 2025). The present paper discusses the results of the semantic and formal classification of these neologisms.

Analysed linguistic innovations span multiple fields. Based on research conducted, several productive areas were identified in which these innovations have become established. The following 14 fields were distinguished in the Georgian language: Technologies and Internet (22%); Business and Finance (13%); Society and Social Life (11%); Politics (10%); Arts and Entertainment (6%); Media (6%); Education (5%); Healthcare / Medicine (5%); Infrastructure (3%); Food and Kitchen Appliances (3%); Environmental Protection (2%); Beauty (2%); Law (2%) and the category "Other" (10%), in which we have grouped together those neologisms that do not belong to any specific field.

According to the thematic groups identified in the Georgian language, the percentage distribution of neologisms is shown in Figure 1.

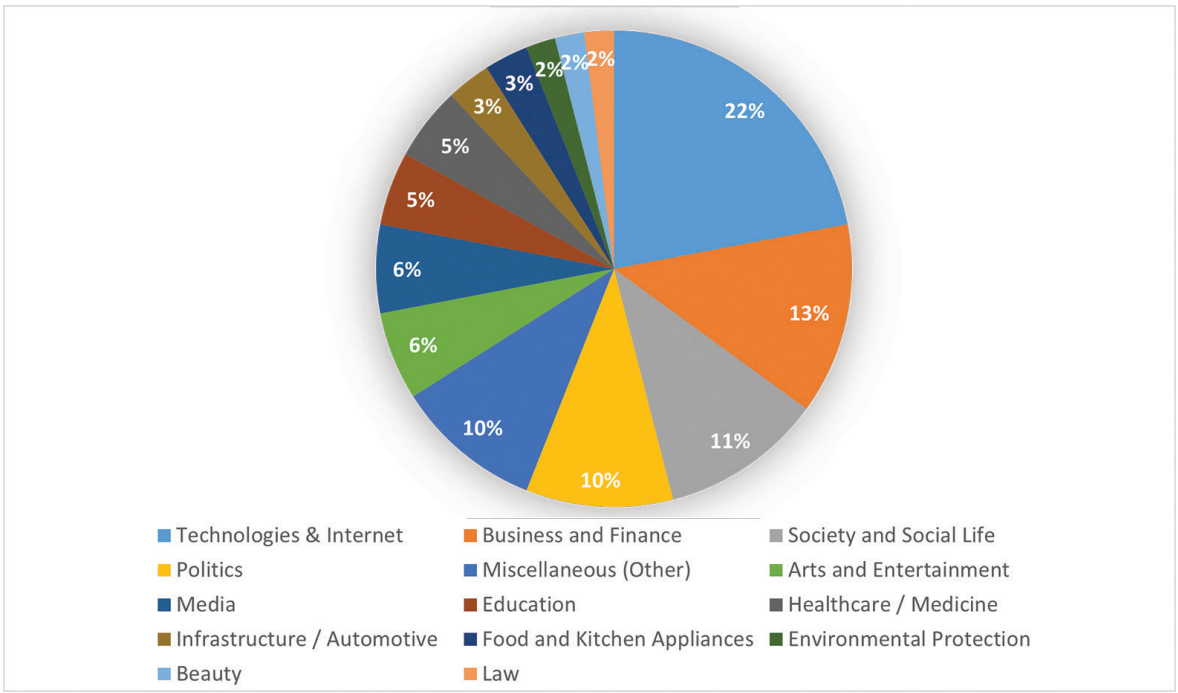


Figure 1: Percentage Distribution of Neologisms in Georgian.

Based on similar studies, thematic groups have also been identified in other languages, which in many cases coincide with those identified in Georgian. In the Spanish Dictionary of Neologisms (Sánchez Manzanares et al. 2016), which contains 2,400 words, 38 fields are identified. These range from common sectors like Economics and Law to highly specific areas such as Bullfighting (Corrida), Enology (Winemaking), and Astronomy.

The German (Lemnitzer 2010) classification closely follows the Georgian model. A notable distinction is the inclusion of the "Language" category, which comprises neologisms specifically related to linguistic issues.

The French dictionary (Logoscope 2018) shows some similarities to Georgian as well. Major spheres such as Technology, Education, Politics, Medicine/Health, and Entertainment coincide across both languages.

In the present paper, fields more relevant to modern Georgian will be discussed. These include politics, medicine, and education due to recent developments in the country.

2.1 Neologisms in Politics

Georgian political neologisms accurately reflect the processes that have taken place in the country from its independence in 1991 to its current phase of democratic backsliding.

Following the signing of the Association Agreement between Georgia and the European Union in 2014, terminology related to European integration became more widespread. Productive prefixes expressing attitudes toward this process also gained prominence. For example, the prefix pro- that indicates support for a person or initiative, became quite productive in word-formation. Examples include terms such as *პროევროპული / p'roevrop'uli* 'pro-European', *პროქართული / p'rokartuli* 'pro-Georgian', *პროდასავლური / p'rodasavluri* 'pro-western', *პროუკრაინული / p'rouk'rainuli* 'pro-Ukraininan'; however, during the recent developments in the country, i.e. political turmoil and country's democratic backslide intensified the usage of prefix ანტი- / anti-, which resulted in the production of new words such as *ანტიევროატლანტიკური / ant'ievroat'lant'ik'uri* 'anti-Euro-Atlantic', *ანტიევროპული / ant'ievrop'uli* 'anti-European', and *ანტიდასავლური / ant'idasavluri* 'anti-western'. Additionally, the prefix euro- is frequently used in terms like *ევროგაერთიანება / evrogaertianeba* 'European Union', *ევროინტეგრაცია / evroint'egratsia* 'Euro-integration', and *ევროსტანდარტი / evrost'andart'i* 'euro-standard'.

The recent developments in Georgia, the enactment of the so-called "Russian Law" and suspension of Euro-integration, 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', *ტიტუშკა / titushk'a* 'titushka', *ზონდერი / zonderi* 'sonder', and *ზონდერმოსამართლე / zondermosamartle* 'a corrupted, bought off judge'. These words refer to violent groups that are subordinate to or hired by state structures to instill fear and violence, thereby hindering democratic processes (Laluashvili & Margalitadze 2025). During the protests, people began to act independently, without any leader. This led to the emergence of a term with new political meaning: *თვითორგანიზებული / tvitorganizebuli* 'self-organized'. This term now refers to a collective group of citizens united by a common goal, such as protest rallies. These groups are characterized by the absence of a single leader or party mandate, with participants collaboratively distributing roles and responsibilities among themselves.

Following the Russia-Ukraine war that began in 2022, a number of new terms, including military terms, emerged in international politics which were borrowed into Georgian, for example, *რაშიზმი / rashizmi*, which is a blend word, formed by *Russia* and *Fascism*, resulting in Rushism, *რაშისტი / rashisti,* 'Rushist', a blend of Russia and Fascist, *ორკი / orki* 'Orc,' which originates from J.R.R. Tolkien's literature, where it refers to a fictional race known for

being aggressive, destructive, and evil. Since 2022, following the Russia-Ukraine war, the term has taken on a new meaning. In both Georgian and international socio-political discourse, it is now used as a powerful metaphor for occupying and aggressive forces, looters, and military aggression (Laluashvili & Margalitadze 2025). The ongoing processes gave rise to a new multiword expression as well. For example, *გლობალური ომის პარტია / globaluri omis p'art'ia* 'Global War Party' is a conspiracy theory or political metaphor denoting influential international circles alleged to be behind the provocation of conflicts and wars in various countries to advance their own interests. The main goal of the 'Global War Party' is to 'open a second front' *მეორე ფრონტის გახსნა / meore pront'is gakhsna* denoting the involvement of Georgia (or another neutral country) in the ongoing military conflict with Russia. Historically, it was a military term that transformed into propaganda in Georgian politics.

These linguistic examples illustrate the country's political development, emphasizing current issues and its aspirations in the global political landscape. The increased use of word-forming elements such as *pro-, anti-,* and *Euro-* indicates a highly polarized society, characterized by two primary directions: European / Western or Russian. There is a noticeable trend where the use of the prefix anti- is increased in relation to the European / Western world (*anti-Euro-Atlantic, anti-European, anti-Western*), which is caused by recent political processes. Furthermore, some political neologisms undergo semantic shifts. A representative example is *Arkipo*, a character originally appearing in Chabua Amirejibi's novel *Data Tutashkhia*. In the current political context, the term's meaning has expanded to describe a demagogic, tyrannical, and manipulative leader.

2.2 Neologisms in Medicine

Healthcare represents one of the most significant fields to produce new words, primarily driven by recent global events. In 2020, the world was overtaken by the COVID-19 pandemic, which posed a new challenge not only for medical professionals but also for linguists. These processes led to the creation of numerous new terms across various languages, and Georgian was no exception.

Medical terminology has entered everyday usage. Since COVID-19 was an entirely new phenomenon for society, the description of new concepts became essential for effective communication. The pandemic proved to be a crisis not only for the medical field; it permeated daily life and nearly every sector of society, including economics, law, politics, business, education, and entertainment (Trap-Jensen & Lorentzen 2022: 826).

The study revealed that part of the COVID-19-related vocabulary consists of the word-forming element covid-. Examples include *კოვიდინფიცირებული / k'ovidinpitsirebuli* 'covid infected', *კოვიდრეგულაცია / k'ovidregulatsia* 'covid regulation', *კოვიდსიტუაცია / k'ovidsit'uatsia* 'covidsituation', *კოვიდტესტი / k'ovidt'est'i* 'covidtest' and others. We also encounter other medical terms, compounds such as *საწოლფონდი / sats'olpondi* 'bed capacity/fund', *რისკჯგუფი / risk'jgupi* 'high risk group', *იმუნორეგულაცია / imunoregulatsia* 'immunoregulation', *პისიარ-ტესტი / p'isiar-t'est'i* 'PCR test', *ლოკდაუნი / lok'dauni* 'lockdown', etc.

Completely new words directly related to the spread of the coronavirus also appeared, such as *სწრაფი ტესტი / sts'rapi t'est'i* 'rapid test' that is a virus detection test, which makes it possible to determine within a short time whether a person is infected. Another example is the so-called *მწვანე პასპორტი / mts'vane p'asp'ort'i* 'green passport' - that is *კოვიდსაშვი / k'ovidsashvi* 'covid pass', 'covid certificate,' or 'health pass,' which was mainly used for travel and contained information about a citizen's health status (vaccination, positive or negative test results).

Following the first wave of COVID-19, new waves emerged as a result of viral mutations, bringing with them new medical and immunological designations: the Alpha, Beta, Delta, and Omicron variants (strains). Additionally, new types of vaccines were introduced, such as *Pfizer (ფაიზერი), AstraZeneca (ასტრაზენეკა),* and *Sinopharm (სინოფარმი).* Furthermore, it became possible to receive an additional control dose of the vaccine *ბუსტერი / bust'eri* 'booster'

There are terms or collocations that already existed in the language but were previously used infrequently; however, as a result of the COVID-19 pandemic, these words gained new prominence, that is, their frequency of use increased significantly: *თვითიზოლაცია / tvitizolatsia* 'self-isolation', *სოციალური დისტანცია / sotsialuri dist'antsia* 'social distance', *კარანტინი / k'arant'ini* 'quarantine', *პანდემია / p'andemia* 'pandemic', *ეპიდემია / ep'idemia* 'epidemic'. The term *უსიმპტომო / usimp't'omo* 'asymptomatic' came to denote a person infected with COVID-19 who does not exhibit symptoms.

Beyond the medical sphere, COVID-19 significantly impacted social life and various other sectors. This led to the emergence of compound words such as *ონლაინშეხვედრა / onlainshekhvedra* 'online meeting', *ონლაინსწავლება / onlainsts'avleba* 'online learning', *დისტანციური სამუშაო / dist'antsiuri samushao* 'remote work', *ტელესკოლა / t'elesk'ola* 'teleschool'. To avoid physical contact, the use of various virtual platforms (Zoom, Teams, Google Meet, Skype) became frequent for conducting work meetings, conferences, and lectures.

In addition to the linguistic innovations prompted by COVID-19, the fields of aesthetic medicine and psychology also saw significant development. Procedures related to self-care and rejuvenation emerged, along with their corresponding terms: *ანტიეჯინგი / ant'iejingi* 'anti-aging', *აუგმენტაცია / augment'atsia* 'augmentation', *ბიორევიტალიზაცია / biorevit'alizatsia* 'biorevitalization', *ბოტოქსი / bot'oksi* 'Botox', *ბრაშინგი / brashingi* 'brushing', *ლიფტინგი / lipt'ingi* 'lifting', *ფილერი /pileri* 'filler'. Similarly, with the growth of interest in psychological health, some terms gained popularity in Georgian as well: *არტთერაპია / art'terap'ia* 'art therapy', *ბიპოლაური / bip'olaruli* 'bipolar', *მენტალური / ment'aluri* 'mental', *პოზიტიური ფსიქოლოგია / p'ozit'iuri psikologia* 'positive psychology', *პოსტტრავმული / p'ost't'ravmuli* 'post-traumatic', *ქოუჩინგი / kouchingi* 'couching' and others.

2.3 Neologisms in Education

Neologisms have also emerged in the field of education in Modern Georgian. This trend became especially noticeable after 2005, when Georgia joined the Bologna Process and,

as a result of reforms, the entire education system was transformed. These changes were accompanied by shifts in terminology, as Soviet-era terms were gradually replaced by new ones: *ბაკალავრიატი / bak'alavriat'i* 'bachelor's degree', *მაგისტრატურა / magist'rat'ura* 'master's degree', *დოქტორანტურა / dokt'orant'ura* 'PhD program', *კოლოკვიუმი / k'olok'viumi* 'colloquium', *სილაბუსი / silabusi* 'syllabus', *კურიკულუმი / k'urik'ulumi* 'curriculum', *ჯიპიეი / jip'iei* 'GPA', *კრედიტების სისტემა / k'redit'ebis sist'ema* 'ECTS Credits' etc.

Furthermore, in 2020, the spread of the coronavirus pandemic also affected the field of education and influenced the formation of new words. The neologisms that appeared for this reason were mainly related to technologies and online learning formats. New terms in the language included *ონლაინსწავლება / onlainsts'avleba* 'online learning', *ვიდეოგაკვეთილი / videogak'vetili* 'video lesson', *ვიდეოლექცია / videolektsia* 'video lecture', *ვებინარი / vebinari* 'webinar', *ჰიბრიდული სწავლება / hibriduli sts'avleba* 'hybrid learning' and others.

3. Formal Classification of Modern Georgian Neologisms

With regard to the formal classification of modern Georgian neologisms, four main classes have been identified. The most prevalent category is borrowing (57%), followed by word formation (38%), shortening (4%), and semantic shifting (1%). In this paper, only loanwords and patterns of word formation are discussed, as these two classes have proven to be the most productive and numerous (see Figure 2).

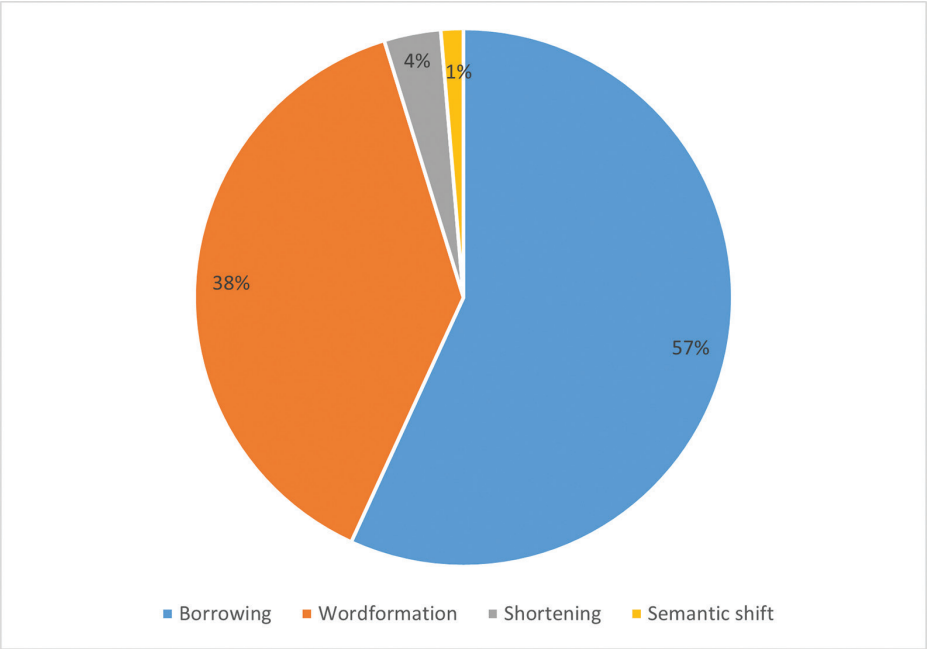


Figure 2: Percentage Distribution of Formal Neologisms.

3.1 Borrowings

3.1.1 Direct Borrowings

As mentioned above, borrowing is one of the most productive ways of forming neologisms in modern Georgian. With globalization, technological advancements, and the rise of Internet communication, the number of lexical units borrowed from other languages, particularly English, has increased significantly.

Within the study, 967 loan words have been identified, which fall into three types of borrowing: direct borrowing, morphologically adapted borrowing, and calques (loan translations). Direct borrowing refers to the process in which one language borrows words from another language unchanged or with minimal changes. There are many examples of such words in Georgian across various fields, with modern technologies being particularly productive in this respect. Direct borrowings are also found in media, business, environment, cooking, entertainment, and other areas. Examples of direct borrowings include, *ბრაუზერი / brauzeri* 'browser', *ბიტკოინი / bit'k'oini* 'bitcoin', *ბლოკჩეინი / blok'cheini* 'blockchain', *გრინვოშინგი / grinvoshingi* 'greenwashing' and others.

The reasons for direct borrowing may include a lexical gap in the recipient language, the need to reflect non-existent realities, or naming new concepts (e.g., blockchain, webinar). Another factor is linguistic prestige, specifically the current global influence and status of the English language, which leads to an increase in Anglicisms. For instance, *workshop (ვორკშოპი)* may sound more fashionable than *სამუშაო შეხვედრა / samushao shekhvedra* 'workshop', or *event (ივენთი)* instead of *ღონისძიება / ghonisdzieba.*

Furthermore, linguistic economy can encourage direct borrowing, as foreign lexical units are often rendered in Georgian through multi-word expressions. Consequently, the linguistic community prioritizes conciseness and flexibility, opting for *დაგუგლვა / daguglva* 'to google' instead of *ინფორმაციის მოძიება / inpormatsiis modzieba* 'searching for information' or *სელფის გადაღება / selpis gadagheba* 'taking a selfie' instead of *საკუთარი ფოტოსურათის გადაღება / sak'utari pot'osuratis gadagheba* 'taking a photograph of oneself'. Ultimately, such words fill the vacuum created by the emergence of new technologies, objects, or phenomena. In these instances, the language adopts ready-made terms from a donor language, striving to keep pace and evolve alongside the modern world.

3.1.2 Morphologically Adapted Borrowings

The flexible grammatical system of the Georgian language facilitates the integration of loanwords into the language. Its rich morphology allows for the adaptation and "Georgianization" of foreign words. When such words are transferred into Georgian, they acquire local characteristics. Georgian affixes or affixes borrowed into the Georgian language at an early stage of its development are added to foreign roots. For example: *digit'al**uri*** ('digital'), ***da**tag**va*** ('to tag'), ***da**blok'**va*** ('to block').

Classification of morphologically adapted borrowings in Georgian was based on Filipović's (1980) transmorphemization framework. Hereby, three types of transmorphemization have been singled out: zero transmorphemization, partial transmorphemization, and complete transmorphemization.

Zero transmorphemization happens in the case of direct borrowings, when a loan word is not modified or is modified with minimal changes. Borrowed words that end in a consonant automatically trigger the native Georgian nominative marker *-i / -ი* (e.g., *ბლოკჩეინ**ი** / blok'chein**i,*** 'blockchain', *გრინვოშინგ**ი** / grinvoshing**i*** 'greenwashing', *ინფლუენსერ**ი** / inpluenser**i*** 'influencer', therefore, words ending with the suffix *-i* can be categorized as part of zero transmorphemization. Real representatives of zero transmorphemization can be found in vowel-final loans, where the recipient language does not add any morphemes to the words e.g. *ალუმნი / alumni* 'alumni', *სელფი / selpi* 'selfie' *ემოჯი / emoji* 'emoji', *სტორი / stori* 'story', etc.

Partial transmorphemization happens when a loan word retains its original suffix, while it adds some Georgian inflections as well. For instance, anglicisms with suffix *-er* producing singular nouns in the donor language, when borrowed, form their plural with Georgian morphemes, as in *ჰეით**ერი** / hei**teri*** 'hater' (*sg*.) and *ჰეით**ერებ**ი / hei**terebi*** 'haters' (*pl*.) or *იუთუბ**ერი** / iutub**eri*** 'youtuber' (*sg*.) and *იუთუბ**ერებ**ი / iutub**erebi*** (*pl*.) 'youtubers', *ინფლუენსერ**ი** / inpluenser**i*** 'influencer' and *ინფლუენსერ**ებ**ი / inpluenser**ebi*** 'influencers' (*pl*.). Since these words end in consonant, they require to add Georgian native nominative case suffix *-i / -ი.*

In the case of *complete transmorphemization* morphemes of donor language are discarded or replaced by synonymous native (in this case Georgian) affixes. In modern Georgian, the examples of complete transmorphemization can be found in verbal neologisms such as *დალაიქება / dalaikeba* 'to like', *დაფოლოვება / dap'oloveba*; 'to follow', *დაშეარება / dashearеba* 'to share', *დათაგვა / datagva* 'to tag', *დაგუგლვა / daguglva* 'to google', etc., producing *და-...-ება / da-...-eba* or *და-...-ვა / da-...-va* verbal frames; adjectival neologisms undergo an identical processes; foreign suffixes, such as English *-al, -ic, -ive, -ful,* etc., are rejected in favor of the native adjectival suffixes *-ური / -uri* or *-ული / -uli,* producing words like *კრეატიული / k'reat'iuli* 'creative', *ტოქსიკური / t'oksik'uri* 'toxic', *ვირუსული / virusuli* 'viral', *დიგიტალური / digitaluri* 'digital', *სტრესული / st'resuli* 'stressful'. Complete transmorphemization happens in the case of nouns as well. In Georgian nouns foreign root can be inflected by Georgian affixes producing new lexical units like *ბლოგერობა / blogeroba* 'the state of being a blogger' utilizing the native suffix *-ობა /-oba; შოპინგობა / shop'ingoba* 'the act or routine of going shopping'.

3.1.3 Loan Translations

Loan translations, or calques, represent one of the productive sources of lexical innovation in modern Georgian. A calque is a literal translation of foreign lexical units, reproducing their structure in the recipient language element by element using native lexical material. This method enriches the language with new lexical units formed through its internal resources, thereby reducing the need for direct borrowing (Trask 2015: 19).

For example, the English word *download* consists of two elements: *down* and *load*. In English, the morpheme *down* indicates 'movement from top to bottom, down, or below,' which corresponds to the Georgian morpheme *ჩამო- / chamo-* . The Georgian equivalent of the second component *load* is *ტვირთი / tvirti.* Consequently, to describe this concept, a new

lexical unit emerged in Georgian: *ჩამოტვირთვა / chamotvirtva.* This term preserves the semantic and functional content of the source language, representing a successful morphological calque that has become firmly established and actively used in the technological and digital domains.

The word *upload* in English is formed in the similar way, breaking down into the morphemes *up* + *load*. Since *up* denotes movement upward, it produces the word *ატვირთვა / atvirtva*, where the morpheme *ა-* is the equivalent of English *up*. Similarly, *reload* and *restart* are translated as ***გადა***ტვირთვა / ***gada***tvirtva, where the English prefix *re-* corresponds to the Georgian morpheme *გადა- / gada-,* indicating the repetition or renewal of an action.

This method is also employed in the formation of analytical or multiword terms. For example, the Georgian equivalent of the English term *emotional intelligence* is *ემოციური ინტელექტი / emotsiuri int'elekt'i.* The term consists of two lexical units with exact Georgian equivalents; their translation results in the phrase *ემოციური ინტელექტი / emotsiuri int'elekt'i*, a term firmly established in the field of psychology. Similarly, terms like *საწყისი გვერდი / sa'tsq'isi gverdi* 'homepage', *სოციალური ქსელი / sotsialuri k'seli* 'social network, *კლიმატური ცვლილება / k'limat'uri tsvlileba* 'climate change', and *გენდერული თანასწორობა / genderuli tanasts'oroba* 'gender equality' have been adopted through direct translation.

Figure 3 below shows the percentage distribution of borrowings in Georgian. As can be seen from the graph, the great majority of loanwords are direct borrowing (60%), which comes from English, i.e. Anglicisms. If we compare the results to the studies conducted in other languages, we will see that English is also the main source for borrowings, but the numbers are different. In Danish, 27% of neologisms are borrowings, out of which 21% are from English (Trap-Jansen 2020); in Spanish, 29% of new words are borrowings and 27% are Anglicisms (Sánchez Manzanares, Azorín Fernández, & Santamaría Pérez 2016); and in German, 38% of new words are borrowings and 35% are Anglicisms (Neologismenwörterbuch, IDS Mannheim).

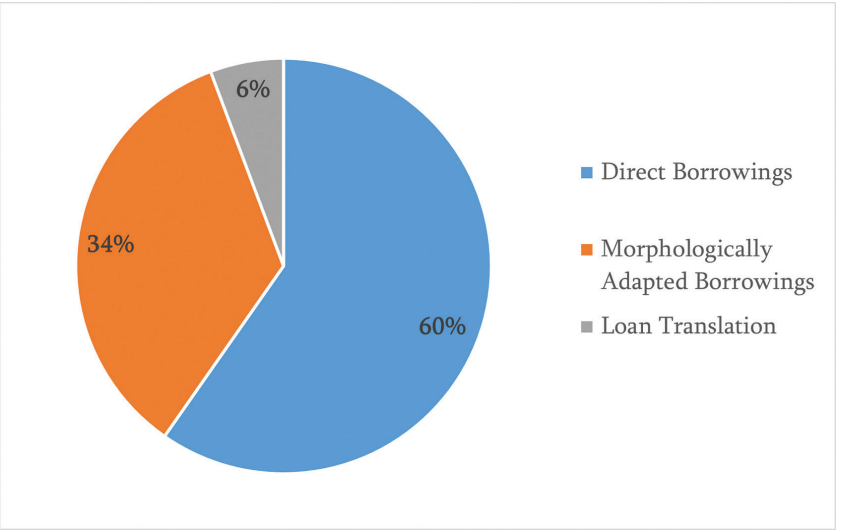


Figure 3: Percentage Distribution of Borrowings.

3.2 Word Formation

Word formation is another productive method for creating new words. Two main directions are identified in this process: compounding and derivation. Compound words are formed by combining two or more independent roots, each with its own meaning. For example: *მედია + გარემო / media + garemo* 'media environment', *ვიდეო + თვალი / video + tvali* 'video surveillance', *ინტერნეტ + ბანკი / int'ernet' + bank'i* 'internet banking'. Some lexical units have been transformed into components that form compound words. These include *ვიდეო- / video-*, *ინტერნეტ- / int'ernet'-*, *მედია- / media-*, *ონლაინ- / onlain-*, which are commonly used in creating terms related to the internet and the digital world, for example: *ვიდეოთამაში / videotamashi* 'video game', *ინტერნეტმაღაზია / int'ernet'maghazia* 'online store', *მედიარეგულაცია / mediaregulatsia* 'media regulation', *ონლაინლექსიკონი / onlainleksik'oni* 'online dictionary', etc. Another word-forming element, *biznes-*, which is used in the field of business and finance, e.g., *ბიზნესალღო / biznesalgho* 'business intuition'.

Derived words can be categorized into three types based on word formation processes: prefixation, suffixation, and a combination of both (prefixation and suffixation). In case of prefixation, in addition to prefixes, prefixoids are used in word-formation. This term – prefixoid – was introduced into Georgian linguistics by Rogneda Ghambashidze (Ghambashidze 1986: 142). These are structural elements, mainly of Greek and Latin origin, that once had independent meanings (and may still retain traces of those meanings today), but over time became constituent parts of complex words and acquired a word-forming function similar to that of affixes. Hence its name - prefixoid (prefix + Greek *eidos*), similar to a prefix, prefix-like (Ghambashidze 1986: 142).

The most productive prefixes and prefixoids identified during the study are the following: *აგრო- / agro-, ანტი- / ant'i-, არა- / ara-, ბიო- / bio-, ევრო- / evro-, ეკო- / ek'o-, ვებ- / veb-, თანა- / tana-, თვით- / tvit-, კიბერ- / k'iber-, მულტი- / mult'i-, პრო- / p'ro-.*

The prefixoid *agro-* is primarily used in the field of agriculture (*აგრობიზნესი / agrobiznesi* 'agro business', *აგროექსპერტი / agroeksp'ert'i* 'agro expert'). The prefixes *anti-*, *euro-*, and *pro-* are mostly found in politics, denoting an attitude toward a specific process or idea (*პროევროპული / p'roevrop'uli* 'Pro-European', *ანტიდასავლური / ant'idasavluri* 'Anti-Western', *ევროინტეგრაცია / evro'int'egratsia,* 'European integration'); The prefixes *ეკო- / eco-* and *ბიო- / bio-* are used to create neologisms in the fields of environment and ecology. Prefixes such as *არა- / ara-* 'non-', *თანა- / tana-* 'co-', and *თვით- / tvit-* 'self-' are of Georgian origin; their use is not restricted to any single field and is characterized by a broader application, as is the prefix *მულტი / multi-*, e.g. *არაბინარული / arabinaruli* 'non-binary', *თანახელმოწერა / tanakhelmots'era* 'co-signature', *თვითცენზურა / tvittsenzura* '*self-censorship', მულტიეთნიკური / mult'iet'nik'uri* 'multi-ethnic', etc. Finally, the prefixes or prefixoids *კიბერ- / k'iber-* 'cyber-' and *ვებ- / veb-* 'web-' are used to produce terms emerging in the world of technology and the internet, e.g. *კიბერომი / k'iberomi* 'Cyber-war', *ვებგამოცემა / vebgamotsema* 'web publication'.

The productive suffixes that are identified within the study include: *-ადი / -adi, -ობა / -oba, -ური / -uri, -ული / -uli, -იანი / -iani, -(იზ)აცია / -(iz)atsia*. The suffixes *-ადი / -adi, -ური /*

-uri, -ული / -uli, -იანი / -iani produce adjectives like *მოწყვლადი / mots'q'vladi* 'vulnerable', *ბიუჯეტური / biujet'uri* 'budget', *ჰაკერული / hak'eruli* 'hacking', *სამგანზომილებიანი / samganzomilebiani* 'three dimentional', while *-oba* creates nouns or verbal nouns such as *ღიაობა / ghiaoba* 'openness', *შოპინგობა / shop'ingoba* 'shopping'; the suffix *-(iz)atsia* produces nouns, as in *არქივიზაცია / arkivizatsia* 'archiving', *აფილიაცია / apiliatsia* 'affiliation'.

From the analysis of examples of prefixation and suffixation, the following combinations were noted: ანტი- / *ant'i- + -ური/-uri / -ული/-uli*, e.g. *ანტიგენდერული / ant'igenderuli* 'anti-gender', *ანტიევროატლანტიკური / ant'ievroat'lant'ik'uri* 'anti-Euro-Atlantic', *არა- / ara- + -ული /-uli / -ადი/adi,* e.g. *არასანქცირებული / arasanktsirebuli* 'unsanctioned', *არაპროვოცირებული / arap'rovotsirebuli* 'non-provoked', and these examples are all adjectives, since the suffixes determine the part of speech labels; the combination of *დე- / de- + -(იზ)აცია / -(iz)atsia* creates nouns, e.g. *დეკრიმინალიზაცია / dek'riminalizatsia* 'decriminalization', *დელეგიტიმაცია / delegit'imatsia* 'delegitimization', *დეოლიგარქიზაცია / deoligarkizatsia* 'deoligarchization', and as for და- / *da- + -ება/ვა / -eba/va* combination, it produces verbs, usually when the root of the word is foreign, e.g. *დაგუგლვა / daguglva* 'to google', *დაბუსტვა / dabust'va* 'to boost'; the combination of the *სა- /sa-/* prefix with the *-ო /-o/* and *-ე /-e/* suffixes contributes to the enrichment of Georgian vocabulary with more adjectives, e.g. *სამოტივაციო / samot'ivatsio* 'motivational', *საპარკინგე / sap'ark'inge* 'parking', etc.

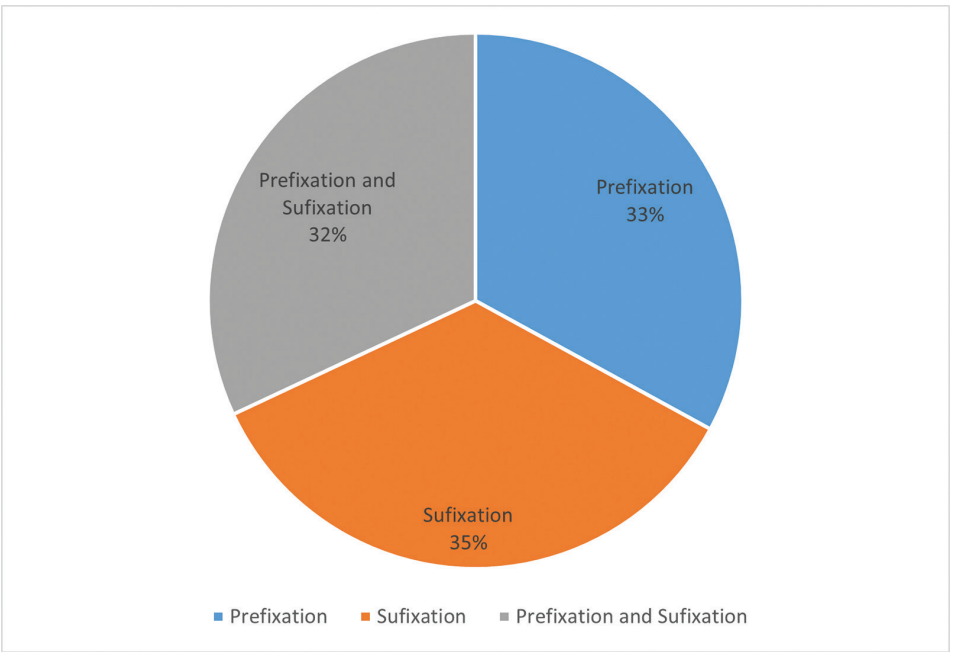


Figure 4: Percentage Distribution of Derivation.

4 Conclusion

As a result of the semantic analysis of the obtained neologisms, it became clear that new words in modern Georgian appear mostly in areas that are constantly in the process of development and progress. In addition, the processes taking place in recent years have contributed to the emergence of political neologisms (10%). The Russia-Ukraine war and ongoing political events in Georgia have significantly influenced the formation and activation of new lexical units. Following the signing of the Association Agreement with the European Union, numerous new words entered the Georgian language. This indicates the country's aspiration to align with European values and standards, and the language is adopting appropriate terminology to better understand and comprehend this development. Lexical innovations have also appeared in the Georgian education system (5%) after its inclusion in the Bologna process. As a result of the reforms, the old system was replaced by a new one, and accordingly, new terminology has appeared here as well. Lexical innovations have also occurred in the medical field (5%) during the global pandemic. The influx of new terms across multiple languages became a subject of linguistic research worldwide.

Neologisms in Georgian appear within academic and literary registers, as well as in entertainment and slang. This demonstrates that the language easily assimilates concepts related to new realities, assigning them names and borrowing from other languages across all registers as needed.

Formal classification reveals that borrowing is the most active method in the process of forming new lexical units. The majority of terms, 967 words (57%), consist of either direct borrowings or units built upon a borrowed root through transmorphemization rules. English serves as the primary source for these processes. Identical trends can be observed in other languages as well: in Danish, Spanish, and German. Even French absorbs anglicisms, despite the country's strict policy and linguistic purism. This result is entirely expected and logical. In the contemporary era, English functions as the global lingua franca, serving as the primary language of the internet, digital technologies, and other scientific advancements.

While borrowing remains the dominant influence, the study clearly demonstrates that the internal resources can also generate neologisms. Alongside external global influences, the Georgian language exploits its own potential. Some existing Georgian words have undergone a profound shift in meaning, acquiring entirely new, highly specific definitions (e.g., *არქიფო / arkipo* from literary character to authoritarian, manipulative leader; *გვერდი / gverdi* 'page' from one side of paper to webpage or social media page). Georgian morphology also plays a crucial role in generating new lexical units and "Georgianization" of foreign roots (e.g., *დაგუგლვა / daguglva* 'to google', *დასქრინვა / daskrinva* 'to create a screenshot'). Internal resources are also used in loan translations. While the concept might be foreign, the Georgian language uses native lexical material to create equivalents that convey exactly the same meaning (e.g., *გონებრივი იერიში / gonebrivi ierishi* 'brainstorm', *მიზნობრივი ჯგუფი / miznobrivi jgupi* 'target group', etc.).

The second most active method for the production of neologisms is word formation. Within the study, 653 words (38% of the total list) were identified in this direction. As a result of this process, compound words (words with two or more stems) and derived words were

identified. Derivation is further subdivided into prefixation (33%), suffixation (35%), and the combination of both - prefixation and suffixation (32%). This method has proven highly flexible, as attaching affixes to a single root can yield diverse new terms.

In conclusion, the process of creating new words in Georgian occurs in several ways. New vocabulary is either adopted from foreign languages or generated using internal resources. However, native resources are mostly limited to loan translations (calques) and semantic shifts. The lack of original, purely Georgian-derived words indicates a significant gap in vocabulary growth that requires further study and attention.

References

Falk, I., Bernhard, D., & Gérard, C. (2017). *néologismes*. Retrieved from logoscope.unistra.fr: https://logoscope.unistra.fr/definition_neologisme.html

Filipović, R. (1980). Transmorphemization: Substitution on the Morphological Level Reinterpreted. *SRAZ: Studia Romanica et Anglica Zagrabiensia, 25*(1-2), 1-8. doi:https://hrcak.srce.hr/121799

Laluashvili, T., & Margalitadze, T. (2025). Semi-Automatic Detection of New Words in Modern Georgian. *Lexikos, 35*(1), 399-413. doi:https://doi.org/10.5788/35-1-2044

Lemnitzer, L. (2010). *Einleitung und Hintergrund: Die Wortwarte - auf der Suche nach den Neuwörtern von morgen*. https://wortwarte.de/

Logoscope. (2018). Retrieved from Logoscope Web Site: https://logoscope.unistra.fr/dbLogoscope.html

Neologism Dictionary (2006ff.), in: OWID – Online Vocabulary Information System German, ed. by Leibniz Institute for the German Language, Mannheim, http://www.owid.de/wb/neo/start.html. (n.d.).

Sánchez Manzanares, C., Azorín Fernández, D., & Santamaría Pérez, M. I. (2016). *NEOMA. Dictionary of neologisms of current Spanish.* Murcia: Editum.

Trap-Jensen, L. (2020). Language-Internal Neologisms and Anglicisms. *Journal of the Dictionary Society of North America, 41*(1), 11-25.

Trap-Jensen, L., & Lorentzen, H. (2022). Recent neologisms provoked by COVID-19 – in the Danish Language and in The Danish Dictionary. *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*, 825-832.

Trask, L. (2015). *Historical Linguistics* (3rd ed.). (R. M. Millar, Ed.) New York: Routledge.

Ghambashidze, R. (1986). *Georgian Scientific Terminology and the Main Principles of Its Compilation.* Tbilisi: Metsniereba. (in Georgian)

AI in Lexicography for Low-Resource Languages: The Case of the Georgian Language

Ketevan Lomidze
Ilia State University
lomidze.ketevani29@gmail.com

Abstract

The emergence of AI has accelerated the integration of chatbots into various fields, reducing the need for human involvement in many areas. Lexicography has been no exception, and since 2022 numerous studies have explored the role of AI in dictionary compilation. However, the role of AI in lexicography for low-resource languages such as Georgian has remained largely outside this discussion, as most existing research focuses on English – a high-resource lingua franca that may be regarded as the "mother tongue" of chatbots. Consequently, the performance of chatbots in low-resource languages remains underexplored, particularly with regard to whether AI can outperform human lexicographers on a broader scale.

The aim of this research is to examine the performance of ChatGPT in the context of a low-resource language such as Georgian and to analyse its lexicographic capabilities. This qualitative study investigates the extent to which ChatGPT can generate dictionary entries for a Georgian explanatory dictionary, including headwords, definitions, illustrative examples, grammatical information, usage labels, and etymological data.

Keywords: Georgian Language; lexicography; ChatGPT; artificial intelligence; dictionary-making

1 Introduction

Artificial intelligence has become an integral part of many aspects of modern life. ChatGPT, Claude, Gemini, and other chatbots continue to evolve in response to users' growing expectations and demands. As users increasingly seek to employ AI across a wide range of fields to complete tasks more quickly and efficiently or simply to test its ability to replace human labour, interest in its potential applications continues to expand. It is therefore unsurprising that lexicography, a field traditionally associated with labour-intensive and time-consuming work, has also turned its attention to chatbots and begun to explore their effectiveness. As with many new technological developments, lexicographers initially approached artificial intelligence with scepticism and questioned its relevance to the field. On the other hand, some scholars speculate that chatbots may eventually transform or significantly reshape

lexicographic practice. However, it is important to emphasise that most studies on AI performance have been conducted in English, and their findings cannot be straightforwardly generalised to other languages. It is often overlooked that, although ChatGPT and similar systems are accessible worldwide, their performance varies significantly across languages. This raises the question of whether AI systems such as ChatGPT can perform equally well in both high-resource and low-resource languages in lexicographic tasks. The case of Georgian, a South Caucasian language spoken primarily in Georgia, has so far received little attention in this debate, making it particularly relevant to examine whether AI can meaningfully support or even replace human lexicographers in the production of Georgian dictionaries.

2 Literature Review

Lexicography is not new to the use of technology in the dictionary-making process. The emergence of corpora, corpus query tools (e.g. Sketch Engine and AntConc), annotating and parsing tools, dictionary writing and management systems, as well as terminological and lexical databases, has made lexicographers' work more efficient, faster, and easier. Against this background, the emergence of chatbots – particularly ChatGPT – has attracted considerable interest among researchers. As ChatGPT is still too recent to be fully understood in terms of its potential and reliability, lexicographers differ in their assessments.

Artificial intelligence appears to be key to reducing what Johnson (1755) famously described as "protracted drudgery" for lexicographers. Fuertes-Olivera (2024) notes that lexicographers must welcome technologies and move on from language-centred traditions to increase the productivity of lexicography. Chatbots like ChatGPT have the potential to quicken the lexicographic process, reduce costs, and boost efficiency. AI can readily replicate lexicon styles and generate well-crafted dictionary entries, challenging the idea that definition-writing is "a rare elite skill" (Lew 2023). In addition, ChatGPT also seems to generate good dictionary examples (Merx *et al*.2024) and deals fairly well with IPA pronunciation (de Schryver 2023; Lars 2024).

On the other hand, there is also evidence of conflicting views in the literature. According to other research, ChatGPT significantly lacks proficiency in most aspects of lexicography and is, therefore, disappointing. McKean & Fitzgerald (2023) emphasise that ChatGPT is inconsistent with IPA. While it may create quite well-written definitions, it sometimes deduces "false polysemy" and creates additional senses (Jakubíček & Rundell 2023). Lew (2023) and Jakubíček & Rundell (2023) add that generated examples cannot be considered good and usable. ChatGPT also makes errors in grammatical information and usage labels, although some researchers consider its performance superior to that of other systems (Lars 2024). Most importantly, Jakubíček & Rundell (2023) note that Kilgarriff's (2007) opinion about the disadvantages of using data found on Google for linguistic purposes still stands; ChatGPT uses unknown sources to produce prompts, and it cannot be considered fully credible. ChatGPT has access to the open web, which underrepresents minorities and can introduce bias (McKean & Fitzgerald 2023).

Most of the above-mentioned researchers note in their articles that their findings cannot be generalised, as they may vary depending on the language in which the prompts or questions are formulated. Merx *et al*. (2024) emphasise that ChatGPT's efficiency vividly degrades for low-resource languages. Additionally, Jakubíček & Rundell (2023) explain that as ChatGPT is mostly trained in English, it tries to answer questions through English data and transfers English word senses to other languages. This brings us to the central question of how ChatGPT performs in languages other than English, particularly in low-resource languages.

ChatGPT ("Generative Pre-trained Transformer") is a widely known artificial intelligence developed by OpenAI (Brown *et al*. 2020). It is a type of neural network which, in order to work and generate responses, is trained on large amounts of unannotated corpora from different sources online (Wikipedia, news, books, and others). It is not a human or a thinking being; therefore, it does not infer. It learns sentence patterns from the data and generates the most likely word sequences (Jakubíček & Rundell 2023). In technical terms, ChatGPT uses generative language modelling, which, given a prompt (a set of semantics), produces sentences (natural language texts) that match the context the user has asked for (Lian 2024). According to many sources, this model, and, therefore, ChatGPT, can deal with many different languages in the same manner (McGiff & Nikolov 2025).

Since ChatGPT's performance is influenced by the quantity of data it is trained on, languages are categorised according to the availability of data. Therefore, there are high-resource and low-resource languages. And while a plethora of people use ChatGPT on a daily basis and are quite pleased, concerns about linguistic inequality in natural language processing (NLP) and AI performance in other languages seem to be a problem.

The term high-resource language refers to languages that have abundant text data and corpora, a significant number of native speakers, and technological support to develop natural language processing. On the other hand, low-resource languages lack data; they are understudied and underrepresented (Ferger 2020). Nigatu *et al*. (2024) deduce that this distinction depends on socio-political aspects, resources, artefacts (technological infrastructure and data), and community agency (the involvement and interest of the speakers).

"This framing of high- versus low-resource languages resembles Zeno's Achilles paradox: 'high-resource languages' function as the tortoise, having been given a head start in the research community and continuing to receive the majority of attention, while 'low-resource languages' correspond to Achilles." (Nigatu *et al*. 2024)

Consequently, English, the "mother tongue" of AI, is a high-resource language. It has been given a head start, and ChatGPT is expected to perform far better in English than in other languages. As a matter of fact, ChatGPT-3 was trained on 93% English data (Walsh 2025). OpenAI has already launched ChatGPT-5 on August 7, 2025, and we can only predict that this percentage may have increased rather than decreased.

On the other hand, the Georgian language is the "Achilles," as Nigatu *et al*. (2024) refer to it. Georgian is a South Caucasian agglutinative language, making language processing trickier. As there is not much demand for Georgian and it lacks available data, it belongs to extremely low-resource languages, the "left-behinds" (Joshi *et al*. 2020). Thus, it is unsur-

prising that ChatGPT-3 received only 0.00060% of its training data in Georgian. Accordingly, it is important to examine how well ChatGPT performs in this language and whether its performance can approach that of English.

3 Methodology

This research uses a qualitative method to investigate how ChatGPT-5 performs in extremely low-resource languages like Georgian, with a focus on its lexicographic abilities.

ChatGPT-5 was recently launched on August 7, 2025. It can now process up to 400K tokens for input and output and has a better ability to deal with longer texts. It outperforms other versions by following instructions and integrating various tools to solve tasks. It is also thought to be better in terms of reduced hallucinations and mistakes. ChatGPT-5 is 45% less likely to produce factual errors than ChatGPT-4, and this rate is even lower in its thinking mode (Designveloper 2025).

The article will examine and evaluate the performance of ChatGPT-5 regarding generating:

- headwords (with an emphasis on identifying neologisms and new entries yet to be recorded in dictionaries);
- definitions (with a focus on the ability to detect polysemy);
- illustrative examples;
- grammatical information and labels;
- etymology – for a Georgian explanatory dictionary.

For evaluation purposes, the generated texts will be compared to the entries of the Georgian online monolingual dictionary (Tsotsanidze, Loladze and Datukishvili 2014). In addition, as ChatGPT-5 was launched during the writing of this article, some responses generated by ChatGPT-4 will also be included for comparison.

The chosen methodology partly draws on the studies of Trap-Jensen (2024) and Lew (2023) concerning the English language.

4 Results

ChatGPT-5 was assigned the role of a professional Georgian lexicographer tasked with compiling a monolingual Georgian dictionary. The *Online Georgian Dictionary* (Tsotsanidze, Loladze & Datukishvili, 2014) was used as a model, and ChatGPT-5 was provided with selected entries to analyse and internalise its style. It was then given a series of tasks designed to evaluate its lexicographic performance.

4.1 Headwords

The aim of this research was to determine whether ChatGPT-5 can identify headwords that have not yet been recorded in dictionaries. Accordingly, in order to detect new headwords and neologisms, it was first prompted in Georgian as follows:

„მომეცი იმ სიტყვების სია, რომელიც ამ ლექსიკონში არ არის შეტანილი. ეს სიტყვები უნდა იყოს საჭირო და ხშირად გამოყენებული, რომ ლექსიკონებში დაემატოს. ეს სიტყვები შეიძლება ნეოლოგიზმიც იყოს (Give me a list of words that are not recorded in this dictionary. These words must be necessary and frequently used to be added to dictionaries. These words may also be neologisms)."

Of the 20 headwords suggested by ChatGPT-5, only 8 appear to be potentially useful for our purpose. These words are *სელფი / selpi* 'selfie', *მიკროინფლაცია / mik'roinplatsia* 'microinflation', *ვირტუალური რეალობა / virt'ualuri realoba* 'virtual reality', *სტრიმინგი / st'rimingi* 'streaming', *ბლოკჩეინი / blok'cheini* 'blockchain', *ვებინარი / vebinari* 'webinar', *გეოლოკაცია / geolok'atsia* 'geolocation' and *დრონი / droni* 'drone'. Most of the other generated headwords are already attested in dictionaries (*გლობალიზაცია / globalizatsia* 'globalisation', *ინტერაქტიული int'erakt'iuli* 'interactive', *ოსტატობა / ost'at'oba* 'craftsmanship/mastery' and *მოზაიკა / mozaik'a* 'mosaic'). Some of them are hallucinations influenced by English., such as *ეკოლოგიური ნაკერი / ek'ologiuri nak'eri* 'eco-stitch' and *არაგამომდინარო პროდუქცია / aragamomdinaro p'roduktsia* 'sustainable product'.

To test ChatGPT-5's abilities further, ChatGPT's thinking mode was used, and it was given specific examples. The second attempt produced better results. Many suggested words can be considered neologisms, as they are commonly used by Georgians and appear frequently in search engines. However, only time will tell whether they will remain relevant and eventually be included in dictionaries. Interestingly, it is evident that most of the suggested headwords are direct borrowings from English (e.g: *სტრიმინგი / st'rimingi* 'streaming', *ლაიქი / laiki* 'like', *ლაივი, ლაივსტრიმი / laivi, laivst'rimi* 'live', *ინფლუენსერი / inpluenseri* 'influencer', *სქროლვა / skrolva* 'scrolling', *კრაუდფანდინგი / k'raudpandingi* 'crowdfunding' and others). Some other interesting neologisms are *გამომწერი* / gamomts'eri 'subscriber', *გამოწერა / gamots'era* 'subscription', *ვებგვერდი / vebgverdi* 'website' and *ელექტრონული საფულე / elekt'ronuli sapule* 'e-wallet'.

Furthermore, it was interesting to examine whether ChatGPT can identify neologisms which are not borrowed from English and detect headwords that may have been omitted from dictionaries. It struggled more than in the second attempt and hallucinated non-existent words such as *წვდომადობა / ts'vdomadoba* 'the quality of being accessible', *გარემოცვიანობა / garemotsvianoba* 'the quality of being surrounded', and *დასარგადობა / dasargadoba* 'planting-ness'. One of the useful headwords included *ჩამოტვირთვა / chamot'virtva* 'download' and *თანამოსაუბრე / tanamosaubre* 'fellow speaker', which were not mentioned before.

During the research process, a similar query was conducted using ChatGPT-4 before the release of ChatGPT-5. Interestingly, ChatGPT-4 produced a higher number of hallucinations and generated more implausible examples such as *თეთრხალათიანობა / tetrkhalatianoba* 'wearing of white gown', *ზაფხულწათვისობა / zaphkults'atvisoba* 'summer-ness', and

ვაშლიანობა / vashlianoba ‘appleness’. ChatGPT-5 hallucinated less and generated only a small number of questionable words. However, ChatGPT-4 was more creative and less repetitive; it rarely repeated already given answers, unlike ChatGPT-5.

4.2 Definitions

Five of the suggested headwords above (*გამომწერი / gamomt’seri* ‘subscriber’, *გამოწერა / gamot’sera* ‘subscribe’, *ვებგვერდი / vebgverdi* ‘website’, *ინფლუენსერი / inpluenseri* ‘influencer’, and *აპლიკაცია / ap’lik’atsia* ‘application’) were used to evaluate AI-generated definitions and examples (which will be discussed below). Initially, ChatGPT was provided with examples of word entries from the *Online Georgian Dictionary* (Tsotsanidze, Loladze, Datukishvili 2014). It was instructed to analyse the structure of its entries and generate similar definitions and examples for the new headwords.

According to the results, ChatGPT-5 is generally capable of producing descriptive and clear definitions. Most of the time, the definitions are semantically well-formed, but they may contain minor grammatical errors and awkward wordings. Furthermore, ChatGPT-5 exhibited inconsistent handling of polysemy. It generated additional senses for words with a single concrete meaning or failed to identify all senses.

გამოწერა

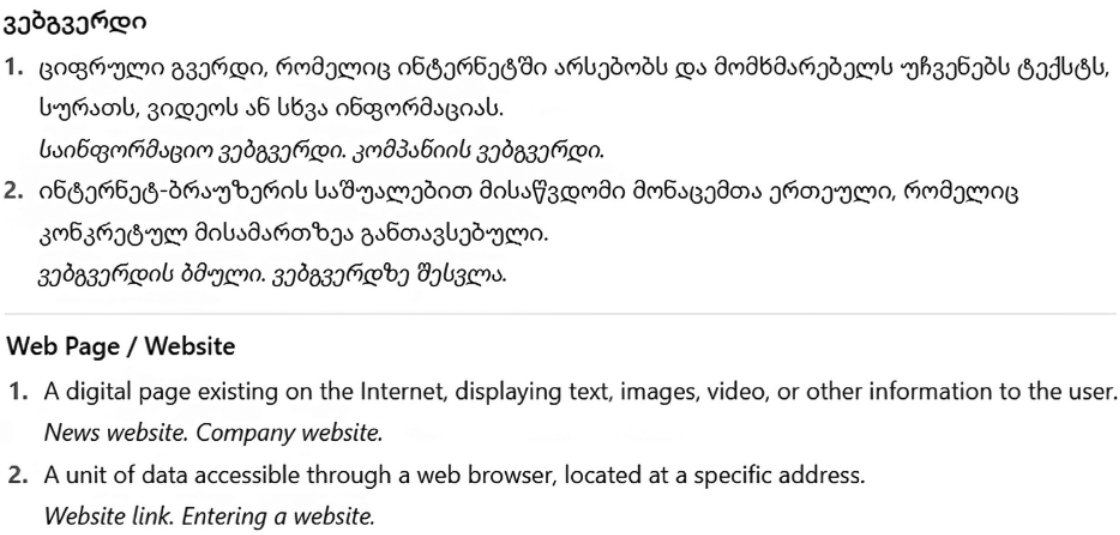
1. რაიმე გამოცემის (გაზეთის, ჟურნალის და სხვ.) ან მომსახურების რეგულარული მიღებისათვის დარეგისტრირება და გადახდა.
   მან გაზეთი გამოიწერა. სტუდენტებმა ჟურნალი გამოიწერეს.
2. რაიმეს დაწერა, ჩამოწერა, ჩანიშვნა (ხშირად ოფიციალურ-ადმინისტრაციული დოკუმენტის სახით).
   ექიმმა რეცეპტი გამოიწერა. ბუღალტერმა ქვითარი გამოიწერა.
3. იურიდიული ან ადმინისტრაციული დოკუმენტის გაფორმება, ხელმოწერა.
   დირექტორმა ბრძანება გამოიწერა. ხელშეკრულება გამოიწერა ორ მხარეს შორის.

gamots’era

1. To subscribe to a publication (newspaper, magazine, etc.) or service by registering and paying to receive it regularly.
   He subscribed to a newspaper. The students subscribed to a journal.
2. To write out, prescribe, or record something (often as an official or administrative document).
   The doctor wrote a prescription. The accountant issued a receipt.
3. To draft, formalize, or sign an official or legal document.
   The director issued an order. A contract was signed between the two parties.

Figure 1: Generated entry for *გამოწერა* ‘subscribe’.

As shown in Figure 1, the generated definitions of *გამოწერა / gamot'sera* 'subscribe' are not fully accurate. According to the *Online Georgian Dictionary* (Tsotsanidze, Loladze, Datukishvili, 2014), the first generated meaning, "to subscribe to a publication," is correct but narrow. In the dictionary, it is combined with ChatGPT-5's second sense, "to prescribe or record an official document". The third sense, "to draft, formalise, or sign an official or legal document," is also incorrect. Additionally, ChatGPT-5 missed another meaning of *გამოწერა,* which is "to discharge from a hospital."



ვებგვერდი

1. ციფრული გვერდი, რომელიც ინტერნეტში არსებობს და მომხმარებელს უჩვენებს ტექსტს, სურათს, ვიდეოს ან სხვა ინფორმაციას.
   საინფორმაციო ვებგვერდი. კომპანიის ვებგვერდი.
2. ინტერნეტ-ბრაუზერის საშუალებით მისაწვდომი მონაცემთა ერთეული, რომელიც კონკრეტულ მისამართზეა განთავსებული.
   ვებგვერდის ბმული. ვებგვერდზე შესვლა.

Web Page / Website

1. A digital page existing on the Internet, displaying text, images, video, or other information to the user.
   News website. Company website.
2. A unit of data accessible through a web browser, located at a specific address.
   Website link. Entering a website.

Figure 2: Generated entry for *ვებგვერდი* 'website'.

ChatGPT-5 hallucinated polysemy in words that have only a single sense, such as *ვებგვერდი / vebgverdi* 'website' and *ინფლუენსერი / inpluenseri* 'influencer'. The Figure 2 shows that both of the generated definitions: 1. A digital page on the internet and 2. A unit of data accessible on a web browser, are the same with different formulations.

აპლიკაცია

1. კომპიუტერული პროგრამა ან პროგრამული პროდუქტი, რომელიც განკუთვნილია კონკრეტული ამოცანის შესასრულებლად კომპიუტერზე ან მობილურ მოწყობილობაზე. *მობილური აპლიკაცია. საბანკო აპლიკაცია. სასწავლო აპლიკაცია.*
2. განაცხადი, წერილობითი ფორმა ან ოფიციალური მიმართვა რაიმე მომსახურების, სამსახურის, სასწავლო ადგილის და სხვ. მიღების მიზნით. *აპლიკაციის შევსება. აპლიკაციის შეტანა უნივერსიტეტში.*
3. ხელოვნებაში, ხელსაქმესა და დეკორატიულ ხელოვნებაში — ორნამენტის ან გამოსახულების დამაგრება სხვა ზედაპირზე (ქსოვილზე, ქაღალდზე და სხვ.). *ტანსაცმლის აპლიკაცია. დეკორატიული აპლიკაცია.*

Application

1. A computer program or software product designed to perform a specific task on a computer or mobile device. *Mobile application. Banking application. Educational application.*
2. An application form, written request, or official submission made to receive a service, a job, a place at an educational institution, etc. *Filling in an application. Submitting an application to a university.*
3. In arts, crafts, and decorative work – an ornament or image attached to another surface (fabric, paper, etc.). *Clothing applique. Decorative applique.*

Figure 3: Generated entry for *აპლიკაცია* 'application'

ChatGPT-5 generated the best entry for *აპლიკაცია / ap'lik'atsia* 'application'. It correctly distinguished three separate meanings: (1) an app, (2) a formal request, and (3) art. Formulation of definitions was clear, precise and natural.

Additionally, two Georgian words were tested, *მგრძნობიარე / mgrdznobiare* 'sensitive' and *ამღვრევა / amghvreva* 'to cloud/blur', which are of native Georgian origin and not borrowed from English. The aim was to determine whether ChatGPT-5 would perform similarly in this case. Interestingly, the results were consistent, and its performance did not deteriorate. The word senses were identified correctly, and ChatGPT-5 successfully recognised the polysemy of *ამღვრევა / amghvreva*, even highlighting its figurative meaning. However, the generated Georgian sentences were somewhat awkward and occasionally incoherent.

4.3 Examples

Concerning examples of *გამომწერი / gamomt'seri* 'subscriber', *გამოწერა / gamot'sera* 'subscribe', *ვებგვერდი / vebgverdi* 'website', *ინფლუენსერი / inpluenseri* 'influencer', *აპლიკაცია / ap'lik'atsia* 'application', *მგრძნობიარე / mgrdznobiare* 'sensitive' and *ამღვრევა / amghvreva* 'to cloud/blur', ChatGPT made very few mistakes. Most of the examples were correct, frequently used in Georgian, and accurate. As noted above, ChatGPT-5 was instructed to generate examples similar to those found in the *Online Georgian Dictionary*, including both common collocations, phrases and full sentences.

Some of the good examples were კომპანიის ვებგვერდი / k'omp'aniis vebgverdi 'company webpage', ვებგვერდზე შესვლა / vebgverdze shesvla 'to access a website', მოდის ინფლუენსერი / modis inpluenseri 'fashion influencer', მობილური აპლიკაცია / mobiluri ap'lik'atsia 'mobile application' and others.

Examples such as პოლიტიკური ინფლუენსერი / p'olit'ik'uri inpluenseri 'political influencer', ტექნოლოგიური ინფლუენსერი / t'eknologiuri inpluenseri 'technology influencer', and the sentence მან ამ წერილს ბევრი გამომწერი მოიპოვა / man am ts'erils bevri gamomts'eri moip'ova 'the letter gained many subscribers' appear to be generated via English. The first two terms are not used in Georgian, and the last sentence is grammatically incorrect and barely understandable. ChatGPT meant to say "the letter gained many subscribers." But it literally said, "He made this letter, many subscribers gained". Another clear illustration of the influence of English-language data is the example of მგრძნობიარე თემა / mrdzgnobiare tema 'sensitive topic' provided for the headword *მგრძნობიარე / mgrdznobiare*, which would be more naturally expressed in Georgian as დელიკატური თემა / delik'at'uri tema 'delicate topic'.

ChatGPT-5 attempted full-sentence examples, mainly with verbs *გამოწერა / gamotsera* and *ამღვრევა / amghvreva*. They were less useful, grammatically awkward and incoherent. For example, these sentences are difficult to understand in Georgian, and the highlighted parts require thorough revision: "*ექიმმა რეცეპტი* ***გამოიწერა***"; "***მან*** *ამ* ***წერილს*** *ბევრი გამომწერი* ***მოიპოვა***" and "*ფიქრმა გონება* ***ამღვრია***".

4.4 Grammatical Information

Regarding grammatical information, ChatGPT was tested using the words mentioned above, along with 20 additional words covering all ten parts of speech in Georgian: noun, adjective, numeral, verb, adverb, postposition, conjunction, particle and interjection.

Overall, ChatGPT correctly assigned parts of speech to most words. Nouns such as *გამომწერი / gamomts'eri* 'subscriber', *ვებგვერდი / vebgverdi* 'website', *ინფლუენსერი / inpluenseri* 'influencer', and *აპლიკაცია / ap'lik'atsia* 'application' were correctly identified. Adjectives like *ფაქიზი / pakizi* 'delicate', *ღია / ghia* 'open', *მგრძნობიარე / mgrdznobiare* 'sensitive', and *ამღვრეული / amghvreuli* 'blurred' were also accurately labelled. Numerals were correctly distinguished as cardinal, ordinal, etc., e.g., *ხუთ-ხუთი / khut-khuti* 'five each', *ასი / asi* 'hundred' and *მეცხრე / metskhre* 'ninth'. Pronouns such as *შენი / sheni*

'your', *ვინ / vin* 'who', and *რომელი / romeli* 'which', as well as adverbs (*გულუხვად / gulukhvad* 'generously' and *სასიამოვნოდ / sasiamovnod* 'pleasantly'), conjunctions (*მაგრამ / magram* 'but', *როცა / rotsa* 'when' and *ან / an* 'or'), and particles (*აბა / aba* and *მაშ / mash*) were also correctly identified.

For verbs, ChatGPT-5 was provided with masdars (verbal nouns) because Georgian dictionaries typically list masdars as headwords. Initially, ChatGPT labelled them as nouns. After correcting and clarifying that they also function as verbs, all given verbs (*კითხვა / k'itkhva* 'read', *სიარული / siaruli* 'walk' and *ყურება / q'ureba* 'watch') were labelled as "noun, verbal noun." However, it correctly labelled *მიშველება / mishveleba* 'help' and *ჩაძაღლება / chadzaghleba* 'die as a dog') as verbs, even though they were provided in masdar forms.

Interjections were also correctly identified, but the grammatical label was inconsistent. ChatGPT labelled them as *შესძახილი / shesdzakhili* instead of the proper Georgian term *შორისდებული / shorisdebuli* (*შესძახილი / shesdzakhili* is a misspelling of *შეძახილი / shedzakhili*, which is a literal equivalent of 'interjection'). The tested interjections were *ვაიმე / vaime* 'oh dear!', *აუუ / auu* 'oh', and *არიქა / arika* 'quick!'.

An interesting case occurred with *თანდებული / tandebuli* 'postposition'. ChatGPT-5 did not initially recognise the Georgian term and refused to classify words like *გარდა / garda* 'except', *-ში / -shi* 'in', *-ზე / -ze* 'on', *-გან / -gan* 'from', and *მიერ / mier* 'by'. However, after some explanation, it eventually labelled them correctly using the English term "postposition," transliterated as *პოსტპოზიცია / p'ost'p'ozitsia*.

By contrast, ChatGPT-4 produced more inconsistent results. It created non-existent or mixed parts of speech, such as *ზმნიზედა არსებითი სახელი / zmnizeda arsebiti sakheli* 'adverb noun', *ზმნებითი სახელი / zmnebiti sakheli* 'verb-noun' and *ზმნიზედა სახელი / zmnizeda sakheli* 'adverb name'). Interestingly, it identified verbs better than ChatGPT-5, but struggled with adjectives.

4.5 Labels

In lexicography, various labels are used to indicate style or register, temporal relevance, regional variation, terminology, formality, and other features. This study primarily focused on whether ChatGPT-5 could identify domains of terms; less attention was given to style and formality. The terms were specifically selected to be of Georgian origin and not borrowed from English, in order to minimise responses influenced by English usage (see Table 1).

Terms	Domain labels
ჭანჭიკი (bolt)	ტექნოლოგია (technology), მექანიკა (mechanics)
მოკლე ჩართვა (short circuit)	ელექტროტექნიკა (electrical engineering), ტექნოლოგია (technology)
სამარხი (tomb)	არქეოლოგია (archaeology), ისტორია (history)
მგეზავი ატომი (tracer)	ბირთვული ფიზიკა (nuclear physics), ქიმია (chemistry)
თავდასხმა (attack/assault)	სამართალი (law), სამხედრო საქმე (military science), სოციალური მეცნიერებები (social sciences)
წყლის ნიშანი (water sign)	ასტროლოგია (astrology)
ანფასი (en face, frontal view)	ხელოვნება (art), ფოტოგრაფია (photography)
პერსპექტივა (perspective)	ხელოვნება (art), არქიტექტურა (architecture)
წილადი (fraction)	მათემატიკა (mathematics)
დედაპლატა (motherboard)	ტექნოლოგია (technology), კომპიუტერული ტექნიკა (computing)
მყარი დისკი (hard disk)	ტექნოლოგია (technology), კომპიუტერული ტექნიკა (computing)
მინდვრის ბოლობეჭედა (hen harrier)	ბოტანიკა (botany)
აცენტრული (acentric)	ქიმია (chemistry), კრისტალოგრაფია (crystallography), ფიზიკა (physics)
ყორღანი (borrow)	არქეოლოგია (archaeology), ეთნოგრაფია (ethnography), ისტორია (history)[1]

Table 1: List of terms and their domain labels.

Out of 14 words, 11 were correctly labelled. Some words received additional labels corresponding to marginal fields, which were counted as partly correct. For example, *ჭანჭიკი / ch'anch'ik'i* 'bolt' was also labelled as technology, *სამარხი / samarkhi* 'tomb' was addition-

1 These terms were selected and their accuracy was verified using the following terminological dictionary: http://multiterm.iliauni.edu.ge/.

ally labelled as history, and *თავდასხმა / tavdaskhma* 'attack/assault' was labelled as social sciences.

ChatGPT-5 made only a few mistakes. *მინდვრის ბოლობეჭედა / mindvris bolobech'eda* 'hen harrier', a zoological term, was incorrectly translated as "field speedwell." Other hallucinations included *აცენტრული / atsent'ruli* 'acentric', which belongs to biology and genetics, and *მგეზავი ატომი / mgezavi atomi* 'tracer', a term in biotechnology, was incorrectly understood as "moderator atom."

Words	Style/register label
შტერი (idiot, dumb)	ვულგარული, შეურაცხმყოფელი (vulgar, offensive)
გასიებული (swollen, bloated, fat)	ნეიტრალური (neutral), არაფორმალური ან შეურაცხმყოფელი (საუბარში) (informal or offensive (colloquial))
პონტი (situation)	არაფორმალური, ჟარგონი (informal, slang)
ჩაძაღლება (to die like a dog)	ვულგარული, უხეში (vulgar, harsh)
კაცი (man)	ნეიტრალური (neutral), ზოგჯერ ოფიციალური (sometimes formal)
კაცმაცუნა (tiny man, unmanly man)	არაფორმალური, ირონიულ-შემამცირებელი (informal, ironic-diminutive)
ბალღი (child, kid)	ნეიტრალური, დიალექტური, არაფორმალური (neutral, dialectal, informal)
მიჯნური (lover, beloved)	წიგნიერი, პოეტური (literary, poetic)
დაბრძანება (to take a seat)	წიგნიერი, ოფიციალურ-მაკადამური (literary, formal-honorific)
მშობელი (parent)	ნეიტრალური, ოფიციალური (neutral, formal)

Table 2: List of words and their style/register labels.

Regarding style and formality, ChatGPT-5 performed well in labelling words as formal (*დაბრძანება / dabrdzaneba*), informal (*ბალღი / balghi)*, vulgar (*ჩაძაღლება / chadzaghleba*), offensive (*გასიებული / gasiebuli*), neutral (*კაცი / k'atsi*), slang (*პონტი / p'ont'i*), literary and poetic (*მიჯნური / mijnuri)*. Interestingly, *გასიებული* can be neutral when referring to something being "swollen," but it can also be offensive in colloquial usage when

describing a person. ChatGPT-5 clearly captured this distinction and gave correct secondary labels to words.

Only a few of the given labels were ambiguous. For example, *კაცი / katsi* 'man' was also marked as "sometimes formal," *ბალღი / balghi* 'child, kid' was labelled as both neutral and informal, and *მშობელი / mshobeli* 'parent' was categorised as both neutral and formal. Additionally, incorrectly generated label names were observed, such as *ირონიულ-**შემამცირებელი*** and *ოფიციალურ-**მაკადამური***, in which the second component does not exist in Georgian (see Table 2).

4.6 Etymology

ChatGPT frequently generated etymological information even when it was not explicitly requested, especially for words borrowed from English. To evaluate its performance, ChatGPT-5 was instructed to provide etymologies for several types of words, such as *სპამი / spami* 'spam', *ინფლუენსერი / inpluenseri* 'influencer', and *აპლიკაცია / ap'likatsia* 'application', which are direct borrowings from English; *ვებგვერდი / vebgverdi*, a semantic calque of 'webpage'; *წილადი / ts'iladi* 'fraction' and *მზე / mze* 'sun', which are of native Georgian origin; and *ფანჯარა / panjara* 'window', a borrowing from Persian.

სპამი

სესხება ინგლისურიდან *(spam)*. თავდაპირველად აღნიშნავდა დაკონსერვებულ ხორცის პროდუქტს (მარკა SPAM), მოგვიანებით ინტერნეტ-სლენგში გადაიტანეს ზედმეტად გამეორებული, შემაწუხებელი შეტყობინებების აღსანიშნავად.

Borrowed from English *(spam)*. Originally referred to a canned meat product (the SPAM brand); later, in internet slang, it came to mean excessively repeated, annoying messages.

Figure 4: Generated etymology of *სპამი* 'spam'.

ინფლუენსერი

ქართული სიტყვა ინფლუენსერი სესხებულია ინგლისურიდან *(influencer) — influence* + სუფიქსი *-er*.

The Georgian word ინფლუენსერი *is borrowed from English (influencer) — influence + the suffix -er.*

Figure 5: Generated etymology of *ინფლუენსერი* 'influencer'.

აპლიკაცია

სესხებაა ლათინურიდან (*applicatio* – მიდგმა, მიჭვრა) შუამავალი ენებით (რუს. „аппликация“, ინგლის. application). თანამედროვე ქართულში უფრო მეტად დამკვიდრებულია კომპიუტერულ-ტექნოლოგიური მნიშვნელობით.

Borrowed from Latin (*applicatio* – “attachment, fastening”) through intermediary languages (Russian “аппликация”, English application). In modern Georgian, it is more commonly established with a computer-technological meaning.

Figure 6: Generated etymology of *აპლიკაცია* ‘application’.

The generated etymology examples for borrowed words from English are well-written, descriptive, and accurate. Following the examples in figures 4, 5 and 6, ChatGPT-5 correctly identifies the source language from which the words are borrowed, explains the word-formation process and describes how polysemy developed (as in figure 4). It also provides the original Latin form where it is relevant (as in figure 6).

მზე

ქართული სიტყვა. წარმოშობა ანალოგიურია პროტოკავკასიურ ფუძესთან *(mze)*, რაც ნიშნავს „მზეს“. არაა სესხებული; გამოყენება ენაში რეგისტრირებულია უძველესი დროიდან.

Native Georgian word. Origin traces to the Proto-Caucasian root *(mze)*, meaning “sun.” Not borrowed; attested in Georgian from ancient times.

Figure 7: Generated etymology of *მზე* ‘sun’.

წილადი

წარმოებულია ქართული ფუძიდან „წილი“ („წაწილი, მომდევნო წაწილი“) + სუფიქსი -ადი. ტერმინოლოგიაში ნიშნავს „ფრაქციას, წილს“.

Derived from the Georgian root წილი (“part, portion”) + suffix -ადი. In terminology, it means “fraction” or “share.”

Figure 8: Generated etymology of *წილადი* ‘fraction’.

Regarding other Georgian words, ChatGPT-5 was inconsistent and made several errors. The etymologies of *მზე / mze* 'sun' (Figure 7) and *წილადი / ts'iladi* 'fraction' (Figure 8) were somewhat well-written and partially correct. For *mze*, the proposed proto-Caucasian form "mze" is incorrect. As for *წილადი / ts'iladi*, it was incorrectly stated that it means "myth". Aside from these issues, the generated etymologies were mostly accurate, though some Georgian wordings were awkward or incoherent. The only fully incorrect etymology was for *ფანჯარა / panjara*, which is actually borrowed from Persian (*panjareh*), not Italian, as ChatGPT suggested.

5 Analysis and Discussion

This research examined the performance of ChatGPT as a lexicographic tool for the low-resource Georgian language. After evaluating its ability to generate headwords, definitions, examples, grammatical information, labels, and etymology, it can be concluded that chatbots can serve as valuable assistants for lexicographers.

ChatGPT-5 can effectively identify headwords and detect neologisms. However, its performance is more accurate when dealing with words borrowed from English. In addition, it often hallucinates non-existent Georgian words due to the influence of English data. Therefore, lexicographers must carefully review the generated headwords.

Regarding definitions, ChatGPT is inconsistent. Although ChatGPT-5 can identify word meanings and polysemy, it often hallucinates additional senses. While some definitions are well-constructed, others lack precision, indicating inconsistency in performance. A similar pattern is observed with generating examples. ChatGPT-5 performs well with collocations and short illustrative sentences, but its performance declines when generating full-sentence examples for verbs, which are often awkward or ungrammatical.

In terms of grammatical information, ChatGPT-5 performs very well. It correctly identifies eight out of ten parts of speech in Georgian: noun, adjective, numeral, pronoun, adverb, conjunction, particle and interjection. However, it struggles with verbs when analysing masdar forms and occasionally misclassifies them as nouns. With appropriate guidance, its performance improves significantly. It also requires additional instruction for interjection and postposition, as these terms are unknown to ChatGPT-5. While it can identify interjections, it incorrectly labels them as *შესძახილი / shesdzakhili*, and it fails to classify postpositions under the Georgian term *თანდებული / tandebuli*.

ChatGPT-5 performs well in assigning domain and style/register labels, with only minor errors, typically arising from English data interference or mistranslations. Its etymological explanations are particularly strong for words borrowed from English (e.g., *სპამი / spami* 'spam', *ინფლუენსერი / inpluenseri* 'influencer' and *აპლიკაცია / aplikatsia* 'application'). However, for native Georgian words, the results are less consistent and require careful revision. In this study, it correctly identified the etymology of *წილადი / ts'iladi* 'fraction' and *მზე / mze* 'sun', while producing an incorrect etymology for *ფანჯარა / panjara* 'window', which is in fact an older borrowing from Persian.

Overall, this study demonstrates that ChatGPT is a useful tool for lexicographic work, even in low-resource languages. It can significantly increase efficiency and provide useful insights,

definitions, and examples. However, it cannot replace human lexicographers, as it remains prone to errors. The findings suggest that ChatGPT does not process Georgian as reliably as high-resource languages such as English. Due to limited training data, it may generate ungrammatical, incoherent, or unnatural sentences in Georgian. This is an issue which is not typically observed in high-resource languages. Therefore, no such problems have been reported in other researchers' studies on the same topic.

These findings highlight the broader issue that low-resource languages remain at a disadvantage compared to high-resource languages. The effectiveness of chatbots decreases as the availability of linguistic data declines. Furthermore, chatbots tend to be biased toward American English (Trap-Jensen 2024). As a result, English-based patterns are often transferred into Georgian contexts where they do not make sense. ChatGPT-5 exhibits a higher tendency to hallucinate in Georgian and often relies on English structures, leading to nonexistent words or meanings that reflect English usage rather than authentic Georgian. Its performance improves with simpler tasks, while more complex, language-specific prompts increase the likelihood of errors. In lexicography, where precision and contextual accuracy are essential, such limitations are significant.

Nevertheless, chatbots continue to improve. Compared to earlier models such as ChatGPT-4, ChatGPT-5 demonstrates clear progress, including reduced hallucination and improved processing. Rather than dismissing chatbots as unreliable, integrating them into lexicographic work is a more productive approach. Even for low-resource languages, ChatGPT-5 can provide valuable perspectives and assist in generating content that may not be immediately apparent to human researchers.

ChatGPT can be guided and adapted to specific needs. It can learn preferred dictionary styles and follow structured instructions. Therefore, for low-resource languages, users need to provide clear guidance and incorporate relevant linguistic knowledge. With proper revision, ChatGPT can become a highly effective supporting tool in lexicographic practice.

6 Conclusion

Artificial intelligence is developing rapidly and can be readily integrated into many fields, including lexicography. Chatbots such as ChatGPT can accelerate the process of dictionary compilation by generating headwords, definitions, examples, grammatical information, labels and even etymology. However, their performance is not always consistent. They may hallucinate information and introduce biases derived from unreliable or unknown internet sources.

This study has shown that ChatGPT exhibits more limitations when working with low-resource languages, such as the Georgian language. ChatGPT-5 is inconsistent in its performance in Georgian, and its errors become particularly evident in ungrammatical sentences. However, it still remains an efficient and practical tool, as it can generate valuable lexicographic information. Although its outputs require revision and editing, the generated entries can still be meaningful and creative. Given the growing demand for such technologies, it is likely that chatbots, and ChatGPT in particular, will continue to improve and become more effective in handling low-resource languages. Therefore, lexicographers should test

the capabilities of artificial intelligence, learn how to use it, and incorporate it into their work.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. *et al*. (2020). Language models are few-shot learners. *Advances in neural informational processing systems*, 33, 1877–1901.

de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4), 355-387.

Designveloper (2025). ChatGPT-4 vs ChatGPT-5: Full comparison breakdown. https://www.designveloper.com/blog/chat-gpt-4-vs-5/ (Accessed: 30.08.2025)

Ferger, A. (2020). Processing Language Resources of Under-Resourced and Endangered Languages for the Generation of Augmentative Alternative Communication Boards. *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, 2644–2648.

Fuertes-Olivera, P. A. (2024). Making Lexicography Sustainable: Using ChatGPT and Reusing Data for Lexicographic Purposes. *Lexikos*, 34(1), 123-140. https://doi.org/10.5788/34-1-1883

Jakubiček, M. & M. Rundell (2023). The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-Editing Lexicography? M. Marek, M. Měchura, C. Tiberius, I. Kosem, J. Kallas and M. Jakubiček (eds.). *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century*. Brno: Lexical Computing, 508-523. https://www.youtube.com/watch?v=8e52vvDpdfQ (Accessed: 31.01.2024)

Johnson, S. (1755). Preface to the Dictionary. S. Johnson (ed.). *A Dictionary of the English Language*. London: Strahan.

Joshi, P., Santy, S., Budhiraja, A., Bali, K. & M. Choudhury (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (eds.). *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 6282–6293. doi:10.18653/v1/2020.acl-main.560

Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, 33(1), 147–151. https://doi.org/10.1162/coli.2007.33.1.147

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10, 704. https://doi.org/10.1057/s41599-023-02119-6 (Accessed: 31.01.2024)

Lian, Y., Tang, H., Xiang, M. & X. Dong (2024). Public attitudes and sentiments toward ChatGPT in China: A text mining analysis based on social media. *Technology in Society*, 76, 102442.

McGiff, J. & Nikola S. Nikolov (2025). Overcoming data scarcity in generative language modelling for low-resource languages: A systematic review. *arXiv*. https://doi.org/10.48550/arXiv.2505.04531

McKean, E. & W. Fitzgerald (2023). The ROI of AI in Lexicography. *Proceedings of the 16th International Conference of the Asian Association for Lexicography: „Lexicography, Artificial Intelligence, and Dictionary Users“*. Seoul: Yonsei University, 10-20.

Merx, R., Vylomova, E. & K. Kurniawan (2024). Generating bilingual example sentences with large language models as lexicography assistants. *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*. Canberra: Association for Computational Linguistics, 64–74.

Nigatu, H. H., Tonja, A. L., Rosman, B., Solorio, T. & M. Choudhury (2024). The Zeno's Paradox of 'Low Resource' Languages. Y. Al-Onaizan, M. Bansal, and Y. Chen (eds.). *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami: Association for Computational Linguistics, 17753–17774. https://aclanthology.org/2024.emnlpmain.983

Trap-Jensen, L. (2024). The Best of Two Worlds: Exploring the Synergy between Human Expertise and AI in Lexicography. T. Margalitadze (ed.). *Proceedings of the I International Conference "Lexicography in the XXI Century"*. Tbilisi: Ilia State University Press.

Tsotsanidze, G., Loladze, N. & Datukishvili, K. (2014). *Georgian Dictionary*. Tbilisi: Sulakauri Publishing. https://www.ganmarteba.ge/

Walsh, M. (2025). ChatGPT Statistics — The Key Facts and Figures. *Style Factory*. https://www.stylefactoryproductions.com/blog/chatgpt-statistics (Accessed: 04.04.2026)

Semantics of Eating in Afrikaans, Northern Sotho and Georgian: Cross-linguistic Variation in Metaphor

Ketevan Lomidze, Gvantsa Nadiradze, Natia Sadghobelashvili
Ilia State University
ketevani.lomidze.1@iliauni.edu.ge, gvantsa.nadiradze.3@iliauni.edu.ge,
natia.sadghobelashvili.1@iliauni.edu.ge

Abstract

The study of conceptual metaphor is of particular importance in the theory of cognitive semantics (Geeraerts 2010: 203-222). Building on the theory of Lakoff and Johnson (1980), metaphor is viewed not merely as a linguistic feature but as a conceptual and culturally grounded phenomenon. Following the work of Taljard and Bosman (2014) on Afrikaans and Northern Sotho, this research investigates metaphors related to the semantics of eating in Georgian and compares the findings cross-linguistically. The study is corpus-based and draws on data from the Georgian National Corpus, the English-Georgian Parallel Corpus, and Araneum Georgianum. Keywords associated with eating (e.g., "eat," "swallow," "digest") were used to identify metaphorical expressions, which were then categorised into agent-oriented, patient-oriented, and mixed groups. The analysis shows that Georgian shares 12 of the metaphorical categories previously identified in Afrikaans and Northern Sotho. Additionally, Georgian demonstrates four unique metaphorical extensions only depicted in the Georgian context. The findings support the idea that conceptual metaphors are largely universal, grounded in shared human experience, while still allowing for language-specific variations. This study highlights both cross-linguistic convergence and cultural nuance in metaphorical systems.

Keywords: conceptual metaphor; cognitive semantics; cross-linguistic comparison; semantics of eating; corpus linguistics

1 Introduction

Conceptual metaphor theory views metaphor as a fundamental cognitive mechanism through which abstract concepts are structured in terms of embodied human experience. One of the most productive source domains in metaphorical mapping can be the universal activity of eating, which involves hunger, consumption, satisfaction, and digestion. Previous research, particularly by Taljard and Bosman (2014) and Newman (2009), has shown that eating-related metaphors are widespread across languages and often extend into domains such as emotional satisfaction, deception, possession, and interpersonal relations. Follow-

ing Taljard and Bosman's research, "The Semantics of Eating in Afrikaans and Northern Sotho: Cross-Linguistic Variation in Metaphor", it is interesting to see how the Georgian language correlates with the depicted metaphorical mappings. While Afrikaans and Northern Sotho have been previously studied in this regard, Georgian provides a contrasting linguistic perspective that allows broader cross-linguistic comparison. The aim of this research is to identify shared and language-specific metaphorical mappings and to examine whether these patterns are primarily motivated by universal embodied experience or influenced by cultural factors. By analysing naturally occurring linguistic data, the study highlights similarities and differences in how eating is used metaphorically across the three languages.

2 Literature Review

According to Lakoff and Johnson (1980), the conceptual system of humans is metaphorical in character. It contains a metaphorical system which emerges from embodied experiences, ontological concepts and structured activities. Lakoff and Johnson classify metaphorical concepts into three categories: orientation metaphors, ontological and structural metaphors.

Geeraerts (2010) considers conceptual metaphor within the broader scope of cognitive semantics. As noted by Geeraerts, cognitive semantics is defined by these core ideas: the contextual, pragmatic flexibility of meaning, the idea that meaning is a cognitive phenomenon which exceeds the boundaries of the word, and lastly, meaning involves perspectivization. In this regard, metaphor is considered a deep-seated cognitive mechanism which influences our ideas, thoughts and language.

Conceptual metaphor theory within the framework of cognitive semantics constituted the theoretical foundation of Taljard and Bosman's (2014) study of eating metaphors in Afrikaans and Northern Sotho. The study highlights two main ideas: embodied experience and cultural models. As noted by Taljard and Bosman (2014), human experiences with the physical world shape our understanding of abstract ideas. In addition, culture plays a key role in creating distinct metaphorical mappings of embodied experiences. The primary aim of the study was to investigate whether conceptual metaphors related to eating were shared by Afrikaans and Northern Sotho and to examine the extent to which they were influenced by cultural factors.

Taljard and Bosman provide a brief overview of the languages under research. Afrikaans is an Indo-European language of the West Germanic branch spoken in South Africa. Northern Sotho is a Bantu language belonging to the Sotho group of South Eastern Bantu languages. As there are distinct cultural and linguistic differences, Taljard and Bosman (2014) claim that comparing these two languages creates an opportunity to research whether the metaphors of eating are universal or culture-specific.

The research method of Taljard and Bosman is based on a corpus-based approach. The researchers implemented a "source-domain vocabulary approach", which means searching for lexical items that belong to a specific semantic field, in this case eating, and then analysing the contexts where the word appears. In addition to the main verb for "eating", they investigated semantically related verbs such as "swallow", "chew", "devour", "digest", "gulp

down" and others. The corpora used in the study were the *University of Pretoria Afrikaans Corpus* (UPAC), containing 15,452,173 tokens and the *University of Pretoria Sepedi Corpus* (UPSC), containing 7,424,335 tokens. It should be noted that neither corpus was annotated. WordSmith tools were used to generate concordances, and then the researchers manually analysed lines to find metaphorical meanings.

The researchers identified two main categories: internalisation (agent-oriented) and destruction (patient-oriented), which indicates that these metaphors are universal. Under the agent-oriented category, there were metaphors like "Intellectual satisfaction is eating" (in Afrikaans "she devours one book after another" and in Northern Sotho "to gulp down the moisture of learning"). Under the patient-oriented category, there was "Diminishing is eating", which appeared in both languages: "the 18% tax is continuously eating into our hard-earned pension", and in Northern Sotho the reference was to an expensive program which "eats a lot of money". Interestingly, some metaphors were found to be characteristic of only one particular language. For example, in Northern Sotho, the following categories were found: "Emotional satisfaction is eating", "Sex is eating", "Accumulating possessions is eating" and "Killing is eating". As for Afrikaans, they discovered: "Understanding is eating", "Humiliation is eating", "Defeating is eating" and "Endearing is eating". It is worth mentioning that conceptual metaphors were not influenced by culture or customs.

3 Methodology

In order to identify conceptual metaphors related to the semantics of eating in the Georgian language and to compare them with corresponding metaphorical models analysed in Afrikaans and Northern Sotho, we employed a corpus-based methodology, following the approach used by Taljard and Bosman (2014). Corpus-based methodology is a research approach that investigates language on the basis of attested linguistic data found in natural texts, rather than relying on the researcher's intuition or artificially constructed examples.

When the object of study is a specific lexical domain, such as, in our case, words, units, or fixed expressions denoting the semantics of eating, the process is relatively straightforward, as it allows for the direct retrieval of relevant units. This approach makes it possible to identify both literal instances of eating and conceptual metaphors associated with eating semantics.

To collect the empirical data, we used the *Georgian National Corpus*[1], the *English-Georgian Parallel Corpus*[2], and the Georgian web corpus *Araneum Georgianum*[3], created by Vladimír Benko at the Slovak Academy of Sciences.

Given the aims of the study, the search for metaphorical examples related to eating semantics did not focus on frequency; instead, only those examples were selected in which eating

1 http://gnc.gov.ge/gnc/page

2 https://enkacorpus.iliauni.edu.ge/

3 http://unesco.uniba.sk/aranea/

was used metaphorically and did not refer to actual food consumption. More specifically, all keywords were selected based on the semantic field of "eating" and included synonyms, hypernyms, and other related units. These included verbs such as: *ჭამა / to eat, (ch'ama), ყლაპვა to swallow (q'lap'va), კვება to feed (k'veba), ძღომა to be satiated (dzghoma), დაგემოვნება to taste (dagemovneba), ხრამუნი to crunch (khramuni), გამოხვრა to gnaw (gamokhvra), შთანთქმა to engulf (shtantkma), ღეჭვა to chew (ghech'va), მონელება to digest (moneleba), ძიძგნა to rip (dzidzgna), სკდომა to burst (sk'doma) and სანსვლა to wolf down (sansvla).*

The extracted lexical units were classified into three groups of metaphors: subject-oriented, object-oriented, and mixed metaphors. Subject-oriented metaphors refer to cases where the focus is on the internal or external state of the agent performing the act of eating (e.g., "to swallow the bait"). Object-oriented metaphors focus on what happens to the object itself (e.g., "to devour with one's eyes"). The mixed group includes metaphors in which both the subject and the object are affected (e.g., "you have to chew it for them so they understand what you are saying").

The study by Taljard and Bosman on Northern Sotho and Afrikaans identified some of the following conceptual metaphors:

- "Emotional satisfaction is eating" – emotional fulfilment is conceptualised as the pleasure derived from physical eating;
- "Intellectual satisfaction is eating" – intellectual achievement or the acquisition of knowledge is perceived as satisfying mental hunger, i.e., "consuming" knowledge;
- "Understanding is eating" – understanding something is conceptualised as grasping or "digesting" it;
- "Intellectual activity is eating" – intellectual activity is described as a feeding process, where thinking, analysis, or learning resembles "mental eating";
- "Sex is eating" – sexual activity is conceptualised through the metaphor of eating;
- "Disappearance / Absence is eating" – sudden disappearance, loss, or absence of something, just like food disappears while eating;
- "Torment (Physical and psychological) is eating" – pain, both physical and psychological, is conceptualised as an internal force that "eats" a person from within;
- "Defeating is eating" – defeating an opponent is perceived as swallowing or devouring them;
- "Conceding is eating" – agent is forced to yield or admit defeat by swallowing their words;
- "Endearment is eating" – expressions of affection are sometimes metaphorized through eating (e.g., "you're so cute I could eat you");
- "Humiliation is eating" – humiliation is conceptualised as the act of consuming something unpleasant (e.g., "to swallow one's shame").

At the next stage of the research, the Georgian data collected within our study were systematically compared with the findings from Afrikaans and Northern Sotho. The comparison revealed that, out of the sixteen metaphorical classes identified by Taljard and Bosman, twelve were also attested in Georgian. In addition, the Georgian language exhibited types of metaphors that were not found in Afrikaans or Northern Sotho.

4 Data Analysis

Obtained Georgian conceptual metaphors were analyzed according to Taljard and Bosman's metaphor mapping (2014). Georgian examples were examined to see if they correlated with Afrikaans and Northern Sotho. Additionally, they were grouped according to agent-oriented, patient-oriented and mixed cases.

Table 1 below summarises the domain mappings found in Afrikaans, Northern Sotho, and Georgian.

Domain Mapping	**Afrikaans**	**Northern Sotho**	**Georgian**
Internalisation (Agent-Oriented)			
"Emotional Satisfaction Is Eating"	—	√	√
"Intellectual Satisfaction Is Eating"	√	√	√
"Understanding Is Eating"	√	—	√
"Intellectual Activity Is Eating	√	√	√
"Sex Is Eating"	—	√	√
"Uncritical Acceptance of Ideas Is Eating"	√	√	—
"Humiliation Is Eating"	√	—	√
Destruction And/Or Elimination (Patient-Oriented)			
"Accumulating Possessions Is Eating"	—	√	—
"Diminishing Is Eating"	√	√	√
"Killing Is Eating"	—	√	—
"Disappearance/Absence Is Eating"	√	√	√
"Torment (Physical And Psychological) Is Eating"	√	√	√
Mixed Cases			
"Defeating Is Eating"	√	—	√
"Conceding Is Eating"	√	√	√
"Endearment Is Eating"	√	—	√
"Charming/Submission Is Eating"	√	—	—

Table 1: Domain Mappings in Afrikaans, Northern Sotho and Georgian.

4.1 Internalisation: Agent-Oriented Expressions

"Emotional Satisfaction is Eating"

According to Taljard and Bosman (2014), the process of eating entails partaking of food that usually brings pleasure and contentment to the agent. Therefore, metaphorically, the internalisation aspect of eating involves emotional well-being and feelings of joy.

In Northern Sotho, the link between eating and emotional joy is especially strong, as the verb "*ja*" (eat) also has a lexicalised sense of "enjoy."'

(1) O *jele* bjang mafelelong a beke/Keresemose? "How did you *enjoy* the weekend/Christmas?" (Lit. "How did you *eat* during the weekend/Christmas?")

This metaphorical sense is not as prominent or easily found in Georgian as in Northern Sotho, but examples of emotional satisfaction were still identified:

(2) ამ სახლს სიყვარულით *კვებავენ* მესხეთელები (am sakhls siq'varulit *k'vebaven* meskhetelebi). "This house is *nourished* with love by Meskhetians."

Additionally, the physical process of "eating" is also associated with religious nourishment that leads to similar pleasurable sensations. Such spiritual experience is vivid in Jewish and Christian rituals surrounding Passover and Lent, respectively. With regard to partaking of the Christian Holy Communion, the lexical item *ja* "eat," in Northern Sotho is used to express the notion of "partake of, participate in" (Taljard and Bosman 2014: 233). This metaphorical domain is evident in Northern Sotho:

(3) Re tlo *ja* Selalelo se Sekgethwa ka Sontaga. "We will *partake* of Holy Communion on Sunday." (Lit. "We will *eat* Holy Communion on Sunday.")

Georgian also has similar religious metaphors with the verb *kveba* (eat, nourish), which can be motivated by the fact that Georgia's religion is Orthodox Christianity.

(4) ფიზიკურ საკვებზე მეტად *სულიერი საკვები* გვესაჭიროება (pizik'ur sak'vebze met'ad sulieri *sak'vebi* gvesach'iroeba). "The *nourishmen*t of the soul is more necessary than physical *nourishment*."

"Sex is Eating"

Sexual intercourse relates to the same emotional domain, especially emotional satisfaction. As Newman (2009) notes, it is a pleasurable experience. As the agent seeks to satisfy hunger and attain contentment, sexual desire (conceptualised as hunger) similarly requires satisfaction.

Interestingly, this link is very prominent in Northern Sotho, where *ja* (eat) has been lexicalised to mean "having sex." Therefore, there are many metaphorical expressions, such as "a man who does not eat at home" (he is unfaithful) or "eat chips" (to engage in sexual activity).

While Afrikaans does not exhibit this metaphorical domain, Georgian shows a vivid link between eating and sex:

(5) უმზერს ლამაზ ქალბატონს და თვალებით *ჭამს* (umzers lamaz kalbat'ons da tvalebit *ch'ams*). "He is looking at a beautiful woman and *wants to have sex* with her." (Lit. "He is watching a beautiful woman and *eating* her with his eyes.")

(6) ქალის *ჩახრამუნება* (kalis *chakhramuneba*). "*To have sex* with a woman." (Lit. "To *eat/crunch/munch* a woman.")

"Understanding is Eating"

This metaphor is used, and must be interpreted as "to understand", not "derive intellectual satisfaction". It clearly expresses the aspect of internalising food (figuratively, ideas) and digesting the ideas (understanding them) (Taljard & Bosman 2014).

In Afrikaans, the verb *verteer* ("to digest") is interpreted as "to understand." Examples from the corpus include expressions such as "to digest a message" and "to digest luxury and eccentricity." On the other hand, in Georgian, the verb *ღეჭვა* ghech'va (*to chew*) is used to express a similar meaning.

(7) უნდა *დაუღეჭო*, რომ გაიგოს რას ეუბნები (unda *daughecho* rom gaigos ras eubnebi).

"You have to *chew* it for them so they can understand what you're saying".

"Humiliation is Eating"

This mapping is grounded in experiential reality, namely the fact that not everything consumed results in a pleasant experience, and that individuals may even be compelled to consume something that is distinctly unpleasant. The Afrikaans lexical item that introduces this mapping is *sluk* ("to swallow"), which evokes the conceptual image of bitter medicine or another similarly unpalatable, typically liquid substance, being ingested against one's will.

In Georgian metaphorical cognition, the verb *ჭამა ch'ama (to eat)* and *ყლაპვა q'lap'va (to swallow)* are employed to convey this conceptual meaning.

(8) ქართულმა ფეხბურთმა სირცხვილი *ჭამა* (kartulma pekhburtma sirtskhvili *ch'ama*). "Georgian football *suffered a shame*". (Lit. "Georgian football *ate* their shame.")

(9) სირცხვილი *გადაყლაპა,* რისხვა კი უდუღდა (sirtskhvili *gadaq'lap'a*, riskhva k'i udughda). "He endured the humiliation, but anger was boiling inside him." (Lit. "He *swallowed* the humiliation, but anger was boiling inside him.")

"Intellectual Satisfaction is Eating"

This mapping is grounded in experiential reality, namely the universal human need for intellectual stimulation and the gratifying experience that follows the consumption and processing of knowledge. Through reading, listening, or observing, information is taken in and mentally processed, leading to the satisfaction of intellectual curiosity. In the context of agent-oriented extensions (internalisation), the eater (reader/listener/observer) actively consumes knowledge as if it were food (Taljard & Bosman, 2014: 233).

The Afrikaans lexical item that introduces this mapping is *verslind* ("to devour"), which evokes the conceptual image of someone eagerly and hungrily consuming books one after another. In Northern Sotho's metaphorical cognition, the verb *phaka* ("to gulp down") is

employed to convey this conceptual meaning. In Georgian, the verb *ყლაპვა q'lap'va* ("to devour" or "to swallow") is used.

Afrikaans

(10) Sy *verslind* die een boek na die ander ("She devours one book after another.")

Northern Sotho

(11) Go *phaka* monola wa thuto ("To gulp down the moisture of learning.")

Georgian

(12) წიგნებს *ყლაპავს* (tsignebs *ylapavs*). "Devours books."

4.2 Patient-Oriented Expressions

"Disappearance/Absence is Eating"

The domain of disappearance and absence is evoked by how food is visible at the beginning but becomes invisible as it is chewed and swallowed. Here, the focus does not fall on killing or death; rather, this metaphor underscores the visual aspect of something quickly disappearing, often for unknown reasons.

This domain is present in all three languages. Additionally, the examples show that the causes of disappearance are also somewhat similar across these languages. Interestingly, this metaphorical domain is particularly prominent in Georgian, where many similar figurative usages can be found.

Afrikaans

(13) Dooie swart sonne wat lig *insluk*. "Dead black suns that *swallow* light."

Northern Sotho

(14) Leswiswi la ba *metša* bjang? "How did the darkness then *swallow* them?"

Georgian

(15) ცამ *ჩაყლაპა* (tsam *chaq'lap'a*). "The sky has *swallowed* it."

(16) მიწამ *ჩაყლაპა* (mits'am *chaq'lap'a*). "The ground has *swallowed* it."

"Torment (Physical or Psychological) is Eating"

A distinction is made between physical and psychological torment. Torture sometimes leads to the complete destruction of the object, indicating an overlap between this and other mappings distinguished in this section.

In Northern Sotho, examples include "eating someone with a fist," resulting in physical harm. The verb to eat is used to convey the destructive, corrosive effect of physical force on the patient.

In Afrikaans, examples found in the corpus for physical torment are caused by agents such as fists, acid, cancerous growths, the sun, or drought that eat the patients. These agents gradually consume, wear down, or destroy the affected entity, whether a person or an object (Taljard & Bosman, 2014: 238).

In Georgian, corpus data confirm the use of the verb to eat in cases referring to both physical and psychological torment.

(17) საშინელი დაავადება, რომელმაც შიგნიდან გამოხრა (sashineli daavadeba, romelmats shignidan *gamokhra*). Lit. "A terrible disease that ate him/her from the inside."

(18) ტკივილმა *გამოუხრა* გული (tkivilma *gamoukhra* guli). Lit. "The pain ate his/her heart."

These metaphors express the slow, destructive process by which an external or internal agent consumes the patient, causing gradual erosion, suffering, or annihilation.

4.3 Mixed Cases

"Endearment is Eating"

An interesting case is identified in the domain of endearment. While the process of eating typically entails the destruction of the patient, in this domain, both the agent and the patient experience emotional satisfaction. Taljard and Bosman (2014) regard this as a somewhat unusual metaphor found in Afrikaans, noting that its motivation is difficult to determine. It is seen as a part of the broader category of emotional satisfaction.

(19) Soos Adam wat aan Eva sê, à la Hennie Aucamp: "Ek kan jou *opeet* van liefde." "Like Adam telling Eve, à la Hennie Aucamp: 'I can *eat you up* out of love.'"

It is noteworthy that a similar example is also found in Georgian colloquial usage of the verb "swallow," where the agent, driven by affection, metaphorically "eats" or "devours" the patient.

(20) ისეთი სასაცილო და საყვარელია, *ჩავყლაპავ*! (iseti sasatsilo da saq'varelia, *chavq'lap'av!*). "She is so funny and lovely that I will *swallow her / eat her up*."

"Conceding is Eating"

Across all three languages, the conceptualisation of concession is confirmed through the metaphor "Conceding *is Eating"*. This metaphor illustrates situations in which an individual is compelled to "swallow" their own words, that is, to retract or take back what has been said. In this sense, both Afrikaans and Northern Sotho employ the same lexical item denoting "swallowing": in Afrikaans, *sluk/insluk* ("to swallow"), and in Northern Sotho, *metše* ("to swallow"). In Georgian, a comparable metaphorical expression is also attested, namely, *ჩაყლაპვა chaq'lap'va (*"to swallow"), which similarly conveys the idea of forced concession and emotional restraint.

(21) დამცირებას *ყლაპავს* (damtsirebas *q'lap'avs*). "*Swallowing* humiliation."

(22) გინებას *ყლაპავს* (ginebas *q'lap'avs*). "(He/She) *swallows* cursing."

"Defeating Is Eating"

The metaphor of eating as a representation of defeat is attested in both Afrikaans and Georgian. In Afrikaans, the verb *eet* ("to eat") is employed, as illustrated in the expression "Ons glo ons kan die Reds *eet"* ("We believe we can *eat*/*defeat* the Reds"), which conveys complete domination over an opponent and confidence in victory. A comparable metaphor in Georgian is expressed through the construction *აჭამა ach'ama*" ("to feed"), which likewise denotes the subject's victory over an opponent. This metaphor, however, is not attested in Northern Sotho.

(23) მარცხი *აჭამა* (martskhi *ach'ama*). "He *fed him* defeat."

(24) ცივი წყალი გადაასხა და მარცხი *აჭამა* (tsivi ts'q'ali gadaaskha da martskhi *ach'ama*). "He poured cold water on him and *fed him* defeat."

4.4 Georgian metaphorical extensions

Georgian Metaphor Domain Mapping
"Deceiving is Eating"
"Neutralising Negative is Eating"
"Infusing with Thoughts/Feelings is Eating"
"Depletion/Exhausting Resources/Ceasing is Eating"
"Squandering is Eating"

Table 2: Exclusive Domain Mapping in Georgian.

"Deceiving is Eating"

The Georgian agent-oriented domain of "Deceiving is Eating" is related to the concept of "Uncritical Acceptance of Ideas is Eating". This idea is motivated by the action of "swallowing" or "nourishing," in which the agent is not active and either passively swallows false information without chewing or digesting it, or is being fed. Just as food quickly disappears in one's mouth while eating, false information is also quickly received by the agent.

(1) სატყუარა *გადაყლაპა* / ანკესი *გადაყლაპა* (sat'q'uara *gadaq'lap'a* / ank'esi *gadaq'lap'a*). "He/she *believed* it." (Lit. "He/she *swallowed* the bait / hook").

(2) დაპირებებით *გვკვებავს* (dap'irebebit *gvk'vebavs.*). "He/she is *lying* to us by making promises." (Lit. "He/she is *feeding* us with promises").

(3) ილუზიებით *კვებავს* (iluziebit *k'vebavs*). "He/she is *deceiving* him/her." (Lit. "He/she is *feeding* him/her with illusions").

"Neutralising Negative is Eating"

This patient-oriented mapping is motivated by the idea of food being destroyed and digested. This metaphor is connected to the domain of "Diminishing or Destroying is Eating." In Georgian, the focus is on neutralising and eliminating negative experiences such as defeat or stress, which can be conceptualised as digesting something unpleasant that one has eaten. This metaphor is typically expressed by the verb "digest."

(4) მარცხი *მოინელა* (martskhi *moinela*). "He/she *got over* a loss." (Lit. "He/she *digested* the loss").

(5) სტრესი *მოინელა* (st'resi *moinela*). "He/she *got over* stress." (Lit. "He/she *digested* stress").

(6) საქართველოს აღარ ძალუძს ამდენი სამოქალაქო დაპირისპირების და რევოლუციის *მონელება* (sakartvelos aghar dzaludzs amdeni samokalako dap'irisp'irebis da revolutsiis *moneleba*.) "Georgia can no longer *neutralize* so many civil conflicts and revolutions." (Lit. "Georgia can no longer *digest* so many civil conflicts and revolutions").

"Infusing with Thoughts/Feelings is Eating"

This specific domain does not have similarities with the conceptual metaphors Taljard and Bosman identified. In this case, the agent is being fed or infused with thoughts or feelings which are often negative. This metaphorical mapping can be explained by the imagery of a person (usually in a passive state) being fed by another, more dominant agent, with something that must be accepted, swallowed, and digested. Therefore, the following examples typically involve the verb "feed/nourish" and negative thoughts or feelings as "food".

(7) გაღიზიანება განცდებს *კვებავს* (gaghizianeba gantsdebs *k'vebavs*). "Irritation affects feelings." (Lit. "Irritation *nourishes* feelings").

(8) სიძულვილით *კვებავს* (sidzulvilit *k'vebavs*). "He/she is *making* them feel hatred." (Lit. "He/she *nourishes* him/her with hatred").

"Depletion/Exhausting Resources/Ceasing is Eating"

Following the domains of destroying/diminishing and killing, in Georgian, metaphors of eating also refer to the exhaustion of one's resources and the eventual end of existence. This domain is motivated by the imagery of eating, where the agent starts with a full plate of food and ends up with an empty plate. Rather than immediate disappearance, the examples suggest a gradual process. Notably, the entity that ceases to exist is the agent itself, which has consumed its own resources.

(9) წყობილებამ თავისი დრო *მოჭამა* (ts'q'obilebam tavisi dro *moch'ama*). "The regime *ceased to exist*." (Lit. "The regime has *eaten up* its time").

(10) *მოჭამა* თავისი პოლიტიკური რესურსი *(moch'ama* tavisi p'olit'ik'uri resursi). "He/she *ran out of* political resources." (Lit. "He/she *ate up* his/her political resources").

"Squandering is Eating"

In relation to possession or money, in Northern Sotho, the mapping of "Accumulating Possession is Eating" refers to gaining possession, while in the case of Afrikaans, the domain "Destroying/Diminishing is Eating" reflects the element of spending money. In Georgian, a prominent domain of "Squandering is Eating" relates to squandering money, savings, or grants. Just as the domain of disappearance is motivated by the image of food being quickly swallowed and becoming invisible, the same motivation applies to squandering. In these examples, money is metaphorically "swallowed" and wasted without justification.

(11) თავდაცვის სამინისტრო ფულის *ყლაპვას* შეუდგა (tavdatsvis saminist'ro pulis *q'lap'vas* sheudga). "The Ministry of Defence is *embezzling/squandering* money." (Lit. "The Ministry of Defence started *swallowing* money").

(12) გრანტი-*ყლაპია* (grant'i-*q'lap'ia*). "Someone who *squanders* funds." (Lit. "Grant/fund *swallower*").

5 Conclusion

It is clear from the analysis that metaphors of eating are shared across languages, and three distinct languages, Afrikaans, Northern Sotho, and Georgian, exhibit similar figurative associations connected to eating.

Out of Taljard and Bosman's 16 domain mappings (2014), Georgian corresponds to 12 of them. Georgian does not seem to include "Uncritical Acceptance of Ideas is Eating," "Accumulating Possessions is Eating," "Killing is Eating," and "Charming/Submission is Eating." However, the reasons for these correspondences cannot be specifically related to cultural norms or aspects of Georgian society. The reason for that is that the Georgian language has five additional metaphorical mappings closely related to the ones above. For example, "Uncritical Acceptance of Ideas is Eating" is similar to the Georgian domain "Deceiving is Eating," which is more narrowly defined and emphasises the accidental acceptance of lies without fully processing the information. "Accumulating Possessions is Eating" mainly concerns money, and the Georgian domain of "Squandering is Eating" also includes metaphors connected to money and possessions. "Depletion/Exhausting Resources/Ceasing is Eating" is partly connected to the domain of killing and destroying; therefore, no clear cultural differences can be identified here either. As for "Charming/Submission is Eating," Georgian does not seem to have similar expressions. However, several examples in Georgian (especially in the Georgian domain of "Infusing with Thoughts/Feelings is Eating") show how the act of eating is used to infuse others with certain thoughts or feelings. While this usage does not necessarily imply submission, a similar interpretation can also apply.

Our study supports the view that eating is not strongly linked to specific cultural customs or beliefs. Conceptual metaphors appear to be motivated by the experience of eating itself and arise from the process of consumption, involving hunger, chewing, swallowing, and digestion. The focus is on both the agent (eater) and the patient (food). As the process of eating is generally shared across the globe and human physiology plays an important role, it is not surprising that distinct languages such as Afrikaans, Northern Sotho, and Georgian

have similar metaphors. As Deignan and Potter (2004) and Newman (2009) state, some metaphors are, in fact, universal.

Therefore, our research underscores that, in terms of conceptual metaphors related to eating, notable similarities and only a few differences were observed among these three languages. It is evident that cultural factors do not play a primary role in establishing figurative expressions of eating, and they are mainly influenced by shared human psychological and embodied experience.

References

Deignan, A. & L. Potter. (2004). A Corpus Study of Metaphors and Metonyms in English and Italian. In *Journal of Pragmatics*, 36, pp. 1231–1252.

Geeraerts, D. (2010). *Theories of Lexical Semantics.* Oxford: Oxford University Press.

Lakoff, G. & Johnson, M. (1980). The Metaphorical Structure of the Human Conceptual System. In *Cognitive Science,* 4 (2), pp. 195-208.

Newman, J. (2009). A Cross-Linguistic Overview of 'Eat' and 'Drink'. In *J. Newman (Ed.), The Linguistics of Eating and Drinking*. Amsterdam: John Benjamins.

Taljard, E. & Bosman, N. (2014). The Semantics of Eating in Afrikaans and Northern Sotho: Cross-linguistic Variation in Metaphor. *Metaphor and Symbol*, 29 (3), pp. 224-245. DOI: 10.1080/10926488.2014.924306.

Appendix

List of Georgian verbs relating to the eating process that have been used for data collection.

ჭამა — ch'ama (eat)

ყლაპვა — q'lapva (swallow)

ხრამუნი — khramuni (munch, crunch)

კვება — kveba (feed)

გამოხრა — gamokhra (gnaw, chew)

სანსვლა — sansvla (wolf down)

ღეჭვა — ghechva (chew)

ძიძგნა — dzidzgna (rip)

მონელება — moneleba (digest)

დაგემოვნება — dagemovneba (taste)

ძღომა — dzghoma (be satiated)

შთანთქმა — shtantkma (engulf)

სკდომა — sk'doma (burst)

Harnessing AI for Creating Definitions of Georgian Military Terminological Neologisms in the context of the Russia-Ukraine War

Ketevan Mchedlishvili
Ilia State University
E-mail: ketevani.mchedlishvili.3@iliauni.edu.ge

Abstract

The outbreak of the Russia-Ukraine war was followed by a rapid increase in drone-related terminology in the Georgian language. However, these terms remain absent from existing bilingual and encyclopedic Georgian military dictionaries to date. This paper explores whether Generative AI tools can be integrated into the terminological workflow, particularly to streamline the process of crafting draft definitions for Georgian military neoterms. For this, 10 drone-related multiword terminological neologisms were selected from Georgian comparable corpora, MiliCorpKa, consisting of analytical blog materials regarding the Russia-Ukraine war. Next, draft definitions were generated with five LLMs (ChatGPT, Grok, Claude, Gemini, DeepSeek) using two prompting strategies: a so-called basic prompt and an ISO-based prompt. The outputs were assessed against a self-developed evaluation scheme. Preliminary results of this study emphasize the importance of fine-tuning prompts and the differences in effectiveness among 5 models for creating definitions in Georgian. The results of this exploratory, qualitative research offer a glimpse of the extent to which Generative AI Chatbots can be a helpful tool for post-editing terminological work and for creating draft definitions for low-resourced languages, such as Georgian.

Keywords: LLMs; GenAI; Georgian military neoterms; draft definitions; Russia-Ukraine war

1 Introduction

Given the central role of technological innovation in contemporary armed conflicts, it is not surprising that these developments are accompanied by linguistic change, manifested in the creation of new terms and the revival of previously unrecorded ones (Georgieva 2015: 42; Struk, Semiginovskaya & Sitko 2022). The outbreak of the Russia-Ukraine war, which has already been assessed as new generation warfare by military experts (Sorge 2023), caused a rapid influx of drone-related terminology into the Georgian military language.

Although drone-related terminology is well-established and recorded in English terminological resources, this is not the case for Georgian specialized dictionaries. The preliminary

analysis of the Georgian military term დრონი / *droni* ('drone') in existing Georgian military dictionaries (2009, 2017) and in Georgian general-language corpora revealed its absence in lexicographic resources and its recent emergence in the latter[1]. Based on these facts, drone-related military terms can be considered candidates of Georgian Neoterms, new terminological units "specifically coined for a given general concept" (Cabre 1999: 205; ISO 1087 2019).

Given the absence of drone-related terms in Georgian military terminology and the importance of the ongoing conflict for Georgia within the framework of NATO partnership, it is crucial to analyze, define, and document these terms in updated versions of military dictionaries. One of the essential components of terminological entries is definitions. However, the process of their crafting can be laborious and time-consuming (San Martin 2024: 1).

Recently, Generative AI (GenAI), with ChatGPT and other large language models (LLMs) under the spotlight of public and academic interest, has been assessed as a potential game-changer for many professionals, including linguists. A growing body of literature focusing on employing GenAI tools, such as Chatbots, for various lexicographic and terminological tasks (Lew 2023; Trap-Jensen 2024; Sabanés & Cunha 2025) has opened the possibility of integrating LLMs into lexicographic/terminological workflows. According to Trap-Jensen (2024: 30), the majority of researchers agree that "ChatGPT is doing a fine job as a defining lexicographer". The following statement is further supported by San Martin (2024), who proposes ways to use ChatGPT to craft draft terminological definitions. According to him, GenAI tools especially those powered by LLMs, can be linguistically promising and with varying degrees of reliability can reduce time and effort needed for creating terminological definitions (San Martin 2024: 1). Therefore, they have the potential of changing the existing terminological workflow and heralding the beginning of new age of post-editing or AI-assisted terminography (San Martin 2024: 4).

Despite a growing body of scientific papers mainly dedicated to applying LLMs, particularly ChatGPT, to lexicographic/terminological tasks in English (de Schryver 2023; Nunzio, Gallina, & Vezzani 2024; Heinisch 2025; de Schryver 2025), research on applying GenAI tools to Georgian is thin on the ground. This topic has only recently begun to attract scholarly interest, namely, 3 papers were presented at the II International Conference *Lexicography in the XXI Century* (Lomidze 2025; Chighladze 2025; Mchedlishvili 2025). This paper expands on the talk presented during the above-mentioned conference (Mchedlishvili 2025). It aims to evaluate the effectiveness of large language models in crafting draft definitions for Georgian drone-related military neologisms by comparing outputs from 5 freely available LLMs (ChatGPT 5.0, Gemini 2.5 Flesh, DeepSeek V3-1, Grok 3.0 Fast, Claude Sonnet 4.0) and 2 prompting strategies (basic prompt vs. ISO-based prompt). Additionally, their potential as a support tool for crafting draft definitions and evaluating the role of the human editors will be further explored.

Chapter 2 presents a brief overview of AI-assisted lexicography and terminography, with a key focus on crafting definitions. In the 3rd chapter, the methodology adopted for this

1 Available at: http://aranea.juls.savba.sk/aranea/run.cgi/rst?corpname=AranGeor&reload=1[accessed 17/02/2026]

study is discussed, while chapter 4 presents results and interpretations per-prompt and per-model. Finally, the article concludes by summarizing the findings, discussing the limitations of the current study, and suggesting future directions.

2 AI in Lexicography and Terminography: a Brief Overview

As de Schryver (2023: 358) states: "Lexicographers have always been at the forefront of dreaming up ways to harness the latest technology in order to compile better dictionaries". Starting from building a corpus to applying various corpus-based methods for linguistic analysis, lexicographers contemplated whether the full automation of dictionary compilation without further human assistance is possible (Rundell & Kilgariff 2011). It should be noted that the field of artificial intelligence, alongside its subfield natural language processing, has been well-established and is not a novelty for the field of lexicography. However, recent developments, particularly the public availability of AI chatbots, so-called LLMs, have led to a renewed interest in exploring ways to use new technology to the benefit of scientific disciplines.

OpenAI launched ChatGPT in November 2022, and since then, other companies, including xAI, Anthropic, and Google, have rapidly entered the field with their own generative AI models. Nowadays, various models, such as Gemini (Google), GPT (OpenAI), Claude (Anthropic), DeepSeek (High-Flyer), Grok (xAI), Qwen (Alibaba), Gemma (Google), and others, have become widely available not only for domain experts but also for laypeople. After the lecture given by Gilles-Maurice de Schryver (de Schryver and Joffe 2023) in Tokyo, the Center for Open Data in the Humanities (CODH) in February 2023 titled "The End of Lexicography, Welcome to the Machine", the study of GenAI has gained momentum.

The first scientific articles dedicated to exploring the application of LLMs, particularly focused on the usage of ChatGPT as an innovative linguistic tool (de Schryver 2023; Lew 2023). In his review paper, de Schryver (2023) critically overviews the existing 10 papers on the usage of ChatGPT. The author points out that although views vary, and concerns towards this technology are present, "one should jump on the wagon and use the technology, to free up time for us humans to do more useful things" (de Schryver 2023: 361). He further claims that in-the-flesh lexicographers are still crucial for the lexicographic process, though, human intervention is better to be saved for the so-called vetting stage (de Schryver 2023: 361). This view is supported by follow-up studies (Barret 2023, as cited by de Schryver 2023: 361; San Martin 2024; Heinisch 2025) requiring human involvement both in the beginning to create an optimal prompt and at the end, to review and refine output.

Notwithstanding varying views and uncertainties around LLMs, they have been successfully applied to different lexicographic tasks, including but not limited to generating headword lists, illustrative examples, equivalents in target languages, and sense division (Jakubíček & Rundell 2023; Lew 2023; Trap Jensen 2024; de Schryver 2025). Another topic that has been in the limelight of academic interest is the creation of definitions (McKean & Fitzgerald 2023, as cited by de Schryver 2023: 369; Lew 2023; Rundell 2023; Trap-Jensen 2024). Although results vary, Trap Jensen (2024: 30) states that most researchers agree that ChatGPT is capable of crafting concise, eloquent definitions that require minimal lexicographic scru-

tiny at the post-editing stage. The results can be further enhanced by feeding specific illustrative examples, whole corpora, or specifying and refining prompts (Barret 2023, as cited by de Schryver 2023: 362; McKean & Fitzgerald 2023, as cited by de Schryver 2023: 369).

Although the advantages of using LLMs, including user-friendliness, quick access, and potentially lightening the load of lexicographic work, are hard to overlook, scholars indicate important limitations to applying new technology. Among these disadvantages, one of the most important ones is the lack of reliability and replicability (Trap-Jensen 2024: 35). The absence of information regarding sources and constant evolution of the existing models, which in turn, results in different outputs, can be deemed as a deal-breaker for scientific research (Rundell 2023; Trap-Jensen 2024). Additionally, the same prompts can generate completely different answers that decrease inter-agreement and comparability of scientific results (Trap-Jensen 2024). Moreover, the enthusiasm for applying LLMs to lexicographic tasks in English cannot be carried over to other languages, especially low-resourced or endangered ones (Trap-Jensen 2024: 35; de Schryver 2025: 397). The fact that over 90% of the training data is in English creates bias, and all the prompts given in any other language are answered via English (de Schryver 2023: 370). Thus, it is too soon or overly enthusiastic to predict "heralding the end of lexicography" (de Schryver 2023: 356).

Similar tendencies toward integrating AI in the workflow have been observed in terminology. Likewise, terminologists started to apply with different levels of efficacy GenAI tools to various terminological tasks, including term coinage, term extraction, generation of equivalents, terminological variants, definitions, building concept systems, etc. (Łukasik 2023; Nunzio, Gallina & Vezzani 2024; Sabanes & Da Cunha 2025; Heinisch 2025). Similar to lexicographic practice, the same views persist regarding human participation (Heinisch 2025; Sabanes & Da Cunha 2025). Henisch (2025: 68) points out the importance of terminologists' expertise in LLM literacy to "assess whether an LLM is suitable for a particular terminology task, domain, or language". Moreover, the restrictions in terms of languages are particularly evident when it comes to terminological work. As Henisch (2025: 76) states: "hallucinations are more likely in niche domains, rapidly developing domains, or newly emerging domains, as not all current data can be present during training of the model". Along the same line, Łukasik (2023) in his research on the low-resource Kashubian language notes that ChatGPT's performance is significantly worse compared to traditional, corpus-based methodology.

It is noteworthy that the creation of terminological definitions has been under the spotlight of terminologists, too (San Martin 2024; Nunzio, Gallina, & Vezzani 2024; Heinisch 2025). In his paper, San Martin (2024) presents various methods for the crafting and refinement of useful prompts for generating definitions. Some of his chosen techniques include simple, basic questions that yield encyclopedic definitions as well as more fine-tuned approaches. For instance, preparing definitional templates, specifying the type of terminological definitions needed, for example intensional definitions (Nunzio, Gallina & Vezzani 2024: 3232), feeding illustrative contexts from corpora that are especially helpful in new or emerging domains due to the lack of training data (San Martin 2024; Heinisch 2025: 71). Notwithstanding the existing challenges, the authors agree upon the inevitability of adopting LLMs into terminological tasks as the "distinction between human-created and AI-created content will become increasingly blurred" (San Martin 2024: 7). One thing remains unchanged,

the necessity of human editor involvement and assigning the role of assisting tool to GenAI (Łukasik 2023; San Martin 2024; Heinisch 2025).

The studies reviewed here so far indicate that "a new age, that of the successful application of generative AI in lexicography, has dawned" (de Schryver 2023: 355). Thus, in view of all that has been mentioned, it is crucial to scientifically analyze the potential of integrating GenAI tools into Georgian lexicographic/terminological work. In this paper, my main questions of interest are:

1. How do different freely available LLMs (ChatGPT, Grok, Claude, Gemini, DeepSeek) vary in their abilities to generate accurate, contextually appropriate draft definitions for drone-related Georgian military neologisms?

2. How do different prompts (basic, ISO-based) influence the quality, precision, and terminological adequacy of crafted definitions?

3. What are the advantages and disadvantages of using specific LLM models for creating draft definitions in Georgian?

4. What is the role of a human editor in assessing, refining, and validating generated definitions?

To answer these questions, I will discuss the methodology adopted for this study and further explain results per-prompt and per-model in the subsequent chapters.

3 Methodology

This exploratory, qualitative study uses a multistage methodology, including (1) using Georgian specialized comparable corpora (MiliCorpKa) to extract modern Georgian military multiword terms; (2) employing corpus-based methods (keywords, collocation and concordance analysis) for the semi-automatic extraction of the multiword terminology; (3) defining and adopting criteria for the concept of terminological neologism; (4) selecting 10 drone-related neoterms for the case study; (5) choosing 5 freely available LLMs for evaluating their effectiveness to generate draft definitions; (6) creating two prompts, namely a basic one, and ISO-based for further analysis; (7) developing evaluation scheme for the generated outputs.

The first stage of analysis was based on the English and Georgian military comparable corpora (MiliCorpEng, MiliCorpKa) previously compiled within the framework of PhD research (Mchedlishvili 2024: 247; Mchedlishvili 2025: 112). The Georgian corpus consisted of 426,362 tokens, 53,405 types, 9 sources related to military-analytical blog materials covering the Russia-Ukraine war. Within the framework of a PhD dissertation, using the #Lancsbox6.0 (Brezina et al. 2021) software tool and employing contrastive collocation analysis (Vicente 2012; Ďurčo 2020; Cabezas-García 2024) 27 drone-related English-Georgian multiword terms were extracted.

At the next stage, the Georgian terms were further analyzed to determine whether they could be considered Georgian terminological neologisms i.e. neoterms candidates. For this purpose, we followed Cabre's (1999: 205) approach, namely their absence from existing

Georgian military encyclopedic dictionary (2017) and English-Georgian military dictionaries (2009, 2017) coupled with the criteria of their recent emergence. In order to check the latter one, we used two Georgian general language corpora: KaWak[2], which includes texts up to 2013, and Aranea[3], a Georgian comparable general language corpus that consists of texts crawled from 2014 onwards, including the latest 2023-2024 crawl covering the Russia-Ukraine war. The absence from the dictionaries, combined with the presence of terminological units in Aranea, was considered a strong indicator of terminological neologisms. The analysis revealed that all Georgian drone-related terms emerged mostly after 2018. For this case study, 10 drone-related Georgian multiword terms were randomly selected (see Figure 1).

A	B	C
Georgian Term	**Transliteration**	**English Equivalent**
კამიკაძე დრონი	k'amik'adze droni	kamikaze drone
სადაზვერვო დრონი	sadazvervo droni	reconnaissance drone
თვითმკვლელი დრონი	tvitmk'vleli droni	suicide drone
FPV (ხედვა პირველი პირისგან	FPV (khedva p'irveli p'irisgan) (	FPV (First Person View) drone
დრონების ხროვა	dronebis khrova	swarm drone
"ლანცეტის" ტიპის დრონი	lantset'is t'ip'is droni	Lancet drone
დრონის ოპერატორი	dronis op'erat'ori	drone operator
წყალქვეშა დრონი	ts'qalkvesha droni	underwater drone
კვადროკოპტერის დრონი	k'vadrok'op't'eris droni	quadcopter drone
დამრტყმელი დრონი	damrt'qmeli droni	attack drone

Figure 1: selected terminological neologisms for the case study.

The next step involved selecting freely available LLM versions operating in the Georgian language. Given that underlying models evolve rapidly, it is important to mention that experiments related to this study were conducted between September 2025 and October 2025, particularly with these models: ChatGPT 5.0,[4] Gemini 2.5 Flesh,[5] DeepSeek V3-1,[6] Grok 3.0 Fast[7], Claude Sonnet 4.0.[8] Following San Martin (2024), two prompting strategies were selected. The first one was the so-called basic prompt which only involved asking to define a particular term. The second one was named ISO-based prompt in which it was specified that the terminological definition style should follow ISO-704 Standard (2022) for

2 Available at: https://cqpweb.lancs.ac.uk/kawac200mw [accessed: 20/02/2026]

3 Available at: http://aranea.juls.savba.sk/aranea [accessed 20/02/2026]

4 https://chatgpt.com [accessed 20/02/2026]

5 https://gemini.google.com/app [accessed 20/02/2026]

6 https://www.deepseek.com [accessed 20/02/2026]

7 https://grok.com [accessed 20/02/2026]

8 https://claude.ai [accessed 20/02/2026]

intensional definitions, and include a generic concept as well as differentiating criteria. All instructions were written and asked in Georgian. Accordingly, outputs were provided in the Georgian language. Definitions for both prompts were generated in multiple independent sessions in September 2025 using the web interface of the above-mentioned models. Table 1 below shows the exact formulation of the prompts in the original Georgian version, with an English translation provided.

Prompt type	Georgian Instruction	English translation
Basic prompt	შექმენი ქართული სამხედრო ტერმინის [კამიკაძე დრონი] ტერმინოლოგიური განმარტება	Generate terminological definition for the Georgian military term [კამიკაძე დრონი / kamikadze droni (Kamikaze drone)]
ISO-704 Based	ISO-704 სტანდარტის მიხედვით, შექმენი ქართული სამხედრო ტერმინისთვის [კამიკაძე დრონი] შინაარსის დამაზუსტებელი დეფინიცია (intensional definition). ტერმინოლოგიური დეფინიცია უნდა შეიცავდეს გვარ-სახეობით მიმართებას: უშუალო გვარეობით ცნებასა და გამმიჯნავ მახასიათებლებს	Craft an intensional definition of the Georgian military term [კამიკაძე დრონი / kamikadze droni (Kamikaze drone)] based on the ISO-704 standard. The terminological definition must include generic relation: a generic concept (superordinate) and delimiting characteristics

Table 1: Basic and ISO-704 based prompts in Georgian with English translation.

To evaluate the outputs, a self-developed scheme was designed. The process was carried out in a Microsoft Excel worksheet. In each case, a 3-point scale was used for both prompts for the 2 main criteria. The first one was named terminological and structural consistency (TSC) with the following specifications for assigning points: (3 points good) - fully structured, genus and differentiating features present; terminology precise and consistent; (2 points partial) - some structural elements present, minor encyclopedic information and redundancy; some terms correct but minor errors or inconsistencies remain; (1 point poor) – definition lacks ISO structure, genus, or differentiating features; terminology is incorrect or inconsistent. The second one was named content accuracy (CA) with the following scores: (3 points good) – fully accurate, facts correct, matches military context; (2 points – partial) - mostly accurate, minor factual/contextual errors; (1 point – poor) - incorrect facts or concept misinterpreted. In addition, special codes were designed to indicate error types in generated definitions: *O (omission of essential characteristics), I (inclusion of irrelevant information), G (genus term missing), V (vague/non-specific language), E (encyclopedic style definitions instead of terminological).* The average TSC and CA scores were calculated for all 10 terms under analysis, per-model and per-prompt (See Figure 2).

Kamikadze drone	ISO template based	Groc 3.0 fast (Free)	კამიკაძე დრონი არის უპილოტო საფრენი აპარატი (genus, superordinate), რომელიც განკუთვნილია ერთჯერადი გამოყენებისთვის თვითგანადგურებით, რათა მიაყენოს მაქსიმალური ზიანი სამიზნეს (deliminating characteristic).	A kamikaze drone is an unmanned aerial vehicle (genus, superordinate) designed for single-use self-destruction to inflict maximum damage on a target (deliminating characteristic).	2- language in georgian is clumsy	3	no error, a bit clumsy	Provides targeted, not encyclopedic information. All essential information is present.

Figure 2: extract of assessment from Microsoft Excel worksheet.

4 Results and Discussions

This chapter presents the results obtained from the outputs generated in response to individual prompts and offers an interpretation of the findings. It further examines the advantages and limitations of using these models in Georgian. Finally, it analyzes the role and extent of human editors' contribution to AI-assisted terminological work.

4.1 Basic Prompt – general assessment

As mentioned before, the first two questions in this study sought to determine how effective the selected 5 LLMs are at generating terminologically and linguistically accurate draft definitions of Georgian military terms. Additionally, my point of interest was whether differences in prompts can affect the accuracy and terminological adequacy of the generated draft definitions.

For the basic prompt, each model was asked to define the selected Georgian drone-related neologisms (see Table 1 above). Table 2 below presents the general assessment of the cumulative average TSC and CA scores across all 10 terms. Although in this subchapter only several representative examples are discussed per-model, the same procedure was applied to the analysis of all selected terms.

Model	Average TSC	Average CA	General errors
ChatGPT 5.0	2,1	3	E, I
Claude Sonnet 4.0	2,6	3	E
Gemini Flesh 2.5	2,1	3	E, I
Grok 3.0 Fast	2,1	3	E, I
DeepSeek V3-1	1,6	2,7	E, I, O

Table 2: Average scores per model with basic prompt.

As shown in Table 2 above, although content accuracy was generally acceptable across all models, the average terminological and structural consistency varied. Further analysis showed that, for a basic prompt, all LLMs tended to generate more encyclopedic than terminological definitions, resembling those of explanatory articles. Moreover, the main tendency observed in terms of typical errors was the inclusion of lots of irrelevant information, for instance, the detailed list of all types of suicide drones, or operations in which they were used (see Figure 3 below). What is striking about the data in this table is that the DeepSeek model in particular, occasionally (3 terms out of 10 terms) missed a generic concept and delimiting characteristics. Notwithstanding differences in average scores across LLMs, the syntactic and morphological errors in all outputs in Georgian were persistent. Figure 3 below illustrates examples of the generated draft definitions for the neoterm kamikaze drone of the Claude which had the highest average TSC score, and DeepSeek which had the worst result among others.

A	B	C	D
Model	**Definition in Georgian**	**Definition translated in English**	**Errors Indicated**
DeepSeek V.3-1	კამიკაძე-ტიპის დრონი, ასევე ცნობილი როგორც **"მოხმხუბარი საბრძოლო საშუალება" (Loitering Munition)**, არის ჰიბრიდული სისტემა, **რომელიც აერთიანებს უპილოტო საფრენ აპარატის (UAV) და მართვადი რაკეტის თვისებებს.**	A kamikaze-type drone, also known as a **"fighting munition" (Loitering Munition)**, is a hybrid system that combines the features of an unmanned aerial vehicle (UAV) and a guided missile.	Incorrect terminology, definition is general and vague. Does not provide essential characteristics of the concept
Claude Sonnet 4.0	**პილოტის გარეშე საჰაერო ხომალდი**, რომელიც განკუთვნილია ერთჯერადი გამოყენებისათვის და **მიზნის განადგურებისათვის თვითდაკოცნებით**. ეს უპილოტო საფრენი აპარატი ატარებს საბრძოლო ტვირთს და პროგრამირებულია ან **დისტანციურად მართულია მიზნისკენ მიფრენისა და მასთან პირდაპირი შეჯახებისათვის.**	**Pilotless aerial vehicle** intended for single use and for **destroying the target through self-kissing**. This unmanned flying apparatus carries a combat payload and is programmed or remotely controlled to fly toward the target and directly collide with it.	Clumsy sentence structure in Georgian, terminological errors in Georgian

Figure 3: illustrative examples of draft definitions generated by Claude and DeepSeek.

In these illustrative examples, even the best average scoring model made lots of mistakes in terms of using correct Georgian terminology, explaining kamikaze drone as *პილოტის გარეშე საჰაერო ხომალდი/p'ilot'is gareshe sahaero khomaldi* ('pilotless aerial vehicle'), or explaining that targets are being destroyed by self-kissing (*თვითდაკოცნით/ tvitdak'otsnit*), rather than self-destruction. As for the DeepSeek model, there is no information on the essential characteristic of a kamikaze drone, beyond the vague claim that it is a hybrid system that combines features of a guided missile and an unmanned aerial vehicle. In this case, the model used unusual and pretty comical equivalent for loitering munitions *მოხმხუბარი საბრძოლო საშუალება/ mochkhubari sabrdzolo sashualeba* ('fighting munitions'). However, it should be noted that none of the models could provide or coin a correct equivalent in Georgian for the loitering munitions, possibly because it was absent from the training data.

A closer analysis of illustrative examples revealed that the use of a basic prompt without further specifications yields very long, verbose, encyclopedic explanations. Thus, it increases the terminologists' post-editing burden. This finding is consistent with that of San Martin (2024: 5) where he highlights that giving specific instructions provides better results in terms of crafting terminological, rather than encyclopedic definitions. Additionally, limita-

tions of all these models in terms of handling the generation of the definitions in Georgian accords with earlier observations made by Trap-Jensen (2024: 35) and Henisch (2025: 76), the lack of training data in Georgian, or the fact that some terms are emerging cause terminological errors. For instance, DeepSeek's worst performance may be caused by less material available to the Chinese LLM for Georgian.

4.2 ISO-704 based prompt – general assessment

If we now turn to the ISO-704-based prompt, each model was instructed to provide an intensional terminological definition and it was further specified to include generic concept and essential characteristics of the concept. The most significant finding compared to basic prompt results was the generation of generally shorter targeted terminological definitions that better aligned with terminological expectations. However, the effect of using an ISO-based prompt was model-dependent. As shown in Table 3 below, the average TSC and CA scores improved significantly for ChatGPT and Grok.

Model	Average TSC	Average CA	General errors
ChatGPT 5.0	2,8	3	syntactic, morphological errors
Claude Sonnet 4.0	2,6	3	V
Gemini Flesh 2.5	2,1	3	O, V, syntactic errors
Grok 3.0 Fast	2,1	3	minor syntactic, morphological errors
DeepSeek V3-1	1,6	2,7	incorrect terminology, incorrect grammar

Table 3: Average scores per model with ISO-704 based prompt

A closer analysis of the generated outputs revealed that ChatGPT and Grok produced concise, accurate, terminologically acceptable definitions that had a clear structure consisting of a generic concept and delimiting characteristics. The results for Gemini and DeepSeek remained largely unchanged, still struggling with correct terminology and, occasionally the definitions crafted in Georgian were hard to grasp. Claude showed the most unexpected performance, in 4/10 terms definitions became vaguer and less clear than those produced by basic prompt. Claude's performance might be caused by the model's general knowledge of what an intensional definition is, and by suggesting broader phrasing for a generic concept and delimiting characteristics. This finding is in line with prior research by San Martin (2024: 5), suggesting that, at times, for greater clarity, models perform better when provided with context or definitional templates rather than simply being instructed to generate intensional definitions. Figure 4 below illustrates comparative examples: the ChatGPT model performed better, and the Claude model's definitions became vaguer.

Model	Term	Basic prompt	ISO-704 based prompt
ChatGPT 5.0	დრონების ხროვა (swarm drone)	სამხედრო ტექნოლოგიაში აღნიშნავს ერთობლივად მოქმედ უსაფრთხო აპარატთა (დრონების) ჯგუფს, რომლებიც კოორდინირებულად, როგორც ერთი სისტემა, ახორციელებენ დავალებებს. ასეთი ხროვა შეიძლება იყოს სხვადასხვა ზომის და დანიშნულების დრონებისგან შემდგარი და მისი მთავარი მახასიათებელია კოლექტიური მოქმედება, მათ შორის თვითმართული მანევრები, სამიზნეების ერთდროული აღმოჩენა, შეტევა ან თავდაცვითი ოპერაციები. **დრონების ხროვის გამოყენება ზრდის ოპერაციების ეფექტიანობას, ართულებს მოწინააღმდეგის თავდაცვით ზომებს და უზრუნველყოფს მოქნილ, დეცენტრალიზებულ სამხედრო მოქმედებებს.**	დრონების ხროვა არის **უპილოტო საფრენი აპარატების ჯგუფი,** რომელიც მოქმედებს **კოორდინირებულად საერთო ამოცანის** შესასრულებლად
ChatGPT 5.0	დრონების ხროვა (swarm drone)	English Translation - In military technology, it denotes a group of pilotless apparatuses (drones) operating jointly, which in a coordinated manner, as one system, carry out tasks. Such a swarm may consist of drones of different sizes and purposes, and its main characteristic is collective action, including self-directed maneuvers, simultaneous detection of targets, attack, or defensive operations. **The use of a drone swarm increases the effectiveness of operations, complicates the opponent's defensive measures, and ensures flexible, decentralized military actions.**	English Translation - Swarm Drone is a **group of unmanned aerial vehicles** that **operates in a coordinated manner** to **accomplish a common task.**
Claude Sonnet 4.0	თვითმკვლელი დრონი (suicide drone)	უპილოტო საბრძოლო საშუალება, რომელიც განკუთვნილია ერთჯერადი გამოყენებისთვის სამიზნის დასაზიანებლად ან **განსანადგურებლად** მისი ფიზიკური **შეჯახების ან აფეთქების შედეგად**	თვითმკვლელი დრონი არის უპილოტო საჰაერო ხომალდი, რომელიც განკუთვნილია **მიზნის გაუქმებისთვის საკუთარი განადგურების საშუალებით.**
Claude Sonnet 4.0	თვითმკვლელი დრონი (suicide drone)	English Translation - Unmanned munition, which is intended for one-time use **for damaging or destroying a target as a result of its physical collision or explosion.**	English Translation - Suicide drone is an unmanned aerial vehicle, which is intended for **elimination of a target by means of its own destruction.**

Figure 4: basic prompt vs. ISO-based prompt generated by ChatGPT and Claude.

These examples (Figure 4) indicate that when it comes to ChatGPT, a long encyclopedic-style definition that contains information about types of operations and models transforms into a terminological one with an ISO-based prompt. In this instance, for the term swarm drone we have a clearly defined generic concept – unmanned aerial vehicle, and a delimiting characteristic – coordinated work. As for the Claude Sonnet model, the basic prompt definition is more specified, explaining that the suicide drone is being destroyed after physical collision with the target. On the other hand, in the ISO-based definition we have a vaguer phrasing that a suicide drone is cancelling a target by self-destruction, without explaining how or what follows this self-destruction. Despite varying model performance, the main problem of morphological, syntactic and/or lexical errors persists to varying degrees across all models.

The results obtained from the use of the ISO-704-based prompt support the findings of other studies in the area of terminological definitions, which claim that specifying the type of definition leads to better, more targeted outputs (San Martin 2024; Heinisch 2025; Nunzio, Gallina & Vezzani 2024: 3232). On the one hand, the findings highlight the necessity of post-editing by a human lexicographer/terminologist as well as the importance of the in-flesh-terminologist to create and get optimal prompts to get satisfying output (Barret 2023, as cited by de Schryver 2023: 362; Heinisch 2025: 68). On the other hand, the sentence clumsiness and errors in the Georgian language indicate the bias towards English and the fact that all models answer via English, thus, resulting in unnatural sentence structures in Georgian (Jakubíček & Rundell 2023; Trap Jensen 2024).

Finally, if we return to the first two research questions, we can deduce that freely available large language models differ in their capacity to generate accurate, terminologically coherent definitions. Their performance is primarily dependent on the correct, detailed prompt

given. That is why an ISO-based prompt yields better results when crafting draft terminological definitions. In the following subchapter, I will briefly summarize the advantages and limitations of the models tested in this case study.

4.3 Assessment per-model: a brief summary

Following the cumulative results of a basic and ISO-704-based template, the tested LLMs were assigned the following places in terms of their effectiveness in handling the generation of draft definitions: 1. ChatGPT 5.0; 2. Grok 3.0 Fast; 3. Claude Sonnet 4.0; 4. Gemini Flesh 2.5; 5. DeepSeek V3-1.

ChatGPT and Grok shared most of the advantages and limitations. Their common strengths included considerably strong performance in Georgian, a good response to a more structured ISO-based prompt, a mainly consistent terminology and lexis. Notably, the most interesting advantage was the minimal human effort required for post-editing in the case of an ISO-based prompt. The distinguishing feature of Grok that deserves mention was the provision of English equivalents and some Georgian synonyms without specific request, which made this information particularly useful. However, this feature was not investigated systematically here. As for the common disadvantages, both of them still made occasional syntactic and morphological errors in Georgian.

As for the Claude Sonnet 4.0 model, one of the striking features of its output was the accuracy of its etymological information for Georgian terms. This information was provided right away without further request. For example, in the case of the equivalent for *swarm drone*, it was highlighted that the term ხროვა/*khrova* (‘pack’) alluded by analogy to the same semantic category as swarm. It should be noted that this feature was not explored in a systematic fashion within this paper. Moreover, this model performed worse with the ISO-based template, perhaps because it required more nuanced information or templates to craft definitions.

The worst performance was given by Gemini Flesh 2.5 and DeepSeek V3-1. The syntactic, contextual, and lexical errors were sometimes so abundant that it became hard to grasp the terms defined. In the majority of cases, one should resort to crafting definitions from scratch rather than putting effort into post-editing the definitions generated by these two models.

Based on this general overview and assessment of models, one could claim that ChatGPT and Grok are the first LLMs that should come to a lexicographer’s/terminologist’s mind for using its assistance. However, it should be mentioned that the main limitation of this study is the lack of reproducibility of the results. Because models are non-deterministic, the outputs given vary (Trap Jensen 2024: 36). In addition, models are frequently updated. Therefore, the results that are true for Gemini 2.5. Flesh are not valid when it comes to, for instance, Gemini 3.0 Thinking versions. As a result, findings cannot be generalized across systems without specifying prompts and model context.

Consequently, answering the fourth research question whether generated draft definitions can be readily used and where a human lexicographer/terminologist stands in this loop we should pay attention to the material we have. Even with the best models in this par-

ticular case study, ChatGPT 5.0 and Grok 3.0 Fast one still needs to validate and linguistically refine definitions when it comes to an ISO-based prompt. However, with the basic prompt, the work becomes time-consuming and laborious. Therefore, this research supports claims made previously by Barret (2023, de Schryver 2023: 361), San Martin (2024), Heinisch (2025) that in-flesh-linguist will still be necessary at both the prompt-creation and, especially, the post-editing phase.

5 Conclusions

This paper explores the effectiveness of using LLMs to assist in crafting draft definitions for Georgian drone-related terminological neologisms. The case study of the selected 10 terms, using two prompting strategies and comparing 5 different models, provided a practical view of how to integrate GenAI tools into terminological work. To the best of our knowledge, no previous study has conducted a comparative analysis of definitions generated by different LLMs, either in general or specifically for Georgian. This study seeks to address that gap by examining and comparing the definitions produced by different models in the context of Georgian terminography and lexicography.

The findings of this case study, on the one hand, indicate that LLMs can generate usable draft definitions for Georgian drone-related terminology, but with significant variation across models and with different prompts. On the other hand, the results demonstrated that all outputs should be checked and manually refined by terminologists, as Georgian syntax errors persist even in Grok and ChatGPT. Thus, GenAI tools should be assigned the role of co-terminologist or AI-assistant, rather than completely replacing human terminologists, especially in specialized domains, emerging terms, or low-resource languages.

While this study provides insights into AI-assisted terminological work, it is important to acknowledge its limitations. First of all, this case study was carried out on a limited set of drone-related terms. The evaluation relied on a self-developed scheme, which was assessed by only one person, potentially increasing bias. Additionally, only freely available LLMs were used. Paid versions or updated models may yield different results, thereby hindering the replicability of results.

Notwithstanding the above-mentioned limitations, this study provides two clear contributions. On the one hand, it makes an applied contribution by drafting terminological definitions for drone-related terminology that will be integrated into the existing multidomain English-Georgian dictionary[9]. On the other hand, it makes a methodological contribution, by assessing the effectiveness and potential integration of Generative AI for under-resourced languages, such as Georgian, and specialized terminology. This work also fosters the idea that chatbots should be used as a supporting tool and not as a replacement for human editors, since critical evaluation and validation of outputs are essential.

Future research should focus on a larger set of Georgian terms from different domains, particularly those formed using exclusively Georgian resources rather than borrowings or structural calques. In addition, it is advisable to have several terminologists to assess the

9 Available at: https://multiterm.iliauni.edu.ge (accessed 21.02.2026)

generated outputs against the existing evaluation scheme, as inter-annotator agreement can enhance the validity of the results. Further work may also focus on using different templates and prompt-generating strategies to yield better outputs.

Overall, we anticipate that research findings will contribute to the use of large language models in assisting with various terminological tasks, particularly in creating draft definitions in the Georgian language.

References

Anthropic Claude. 2025. Claude Response to Keti Mchedlishvili. 3 September. https://claude.ai/share/64cea118-86d3-4250-9015-3a5e708d69ef [accessed 21/02/2026]

Barrett, G. (2023). 'Defin-O-Bots: Challenging A.I. to Create Usable Dictionary Content.' Paper presented at the *24th Biennial Conference of the Dictionary Society of North America*. *Boulder, CO, USA, 31 May – 3 June 2023.*

Benko, V. (2024). The Aranea Corpora Family: Ten+ Years of Processing Web-Crawled Data. Nöth, E., A. Horák, and P. Sojka (Eds.). 2024. *Text, Speech, and Dialogue. TSD 2024. Lecture Notes in Computer Science:* 1-16. Cham: Springer. https://doi.org/10.1007/978-3-031-70563-2_5 [accessed 21/02/2026]

Brezina, V., P. Weill-Tessier & T. McEnery 2021. #LancsBox 6.x [software]. Available at: http://corpora.lancs.ac.uk/lancsbox. [accessed 21/02/2026]

Cabezas-García, M. (2024). *The Pragmatics of Multiword Terms. The Impact of Context*. New York: Routledge. DOI: https://doi.org/10.4324/9781003390091 [accessed 21/02/2026)

Cabré, M.T. (1999). Terminology: Theory, Methods and Applications. Amsterdam/ Philadelphia: John Benjamins. DOI: https://doi.org/10.1075/tlrp.1. [accessed 21/02/2026]

Daraselia, S. (2015). *Issues of Compilation of New Georgian-European Learner's Dictionary Using the Corpus Methodology*. Unpublished Ph.D. thesis. Tbilisi: Ivane Javakhishvili Tbilisi State University. (In Georgian) https://cqpweb.lancs.ac.uk/kawac200mw/ [accessed 21/02/2026]

de Schryver, G.-M. & D. Joffe. (2023). 'The End of Lexicography, Welcome to the Machine: On How ChatGPT Can Already Take over All of the Dictionary Maker's Tasks.' Paper presented at the *20th CODH Seminar. Center for Open Data in the Humanities, Research Organization of Information and Systems, National Institute of Informatics, Tokyo, Japan, 27 February 2023*. https:// youtu.be/watch?v=mEorw0yefAs [accessed 21.02.2026]

de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Us ing ChatGPT. In International Journal of Lexicography, 36(4), pp. 355-387. https://doi.org/10.1093/ijl/ecad021 [accessed 06/02/2026]

de Schryver, G.-M. (2025). Customised GPTs for Lexicography. In *Lexikos*, *35*(2), pp. 397-422. https://doi.org/10.5788/35-2-2104 [accessed 06/02/2026]

Ďurčo, P. Multiword Expressions in Comparable Corpora. Pastor G. C and J. Colson (Eds.). 2020. *Computational Phraseology*: 136-150. Amsterdam/Philadelphia: John Benjamin's Publishing Company. DOI: https://doi.org/10.1075/ivitra.24 [accessed 21/02/2026]

English-Georgian Military Dictionary. https://mil.dict.ge/en/ [accessed 21/02/2026]

Georgieva, V. (2015). *Military English: from Theory to Practice.* Sofia: Georgi Stoikov Rakovski National Defence College https://www.researchgate.net/publication/363582336_MILITARY_ENGLISH_FROM_THEORY_TO_PRACTICE [accessed 21/02/2026]

Google Gemini. 2025. Gemini Response to Keti Mchedlishvili. 3 September. https://gemini.google.com/app/4774657b6e333c9c [accessed 21/02/2026]

Heinisch, B. (2025). Next-gen Terminology: Transforming Terminology Work with Large Language Models. In *Across Languages and Cultures*, 26 (S), pp. 64-80. DOI: 10.1556/084.2025.01061. [accessed 11/02/2026]

High-Flyer. DeepSeek. 2025. DeepSeek Response to Keti Mchedlishvili, 3 September. https://chat.deepseek.com/share/8wi9r7dj36cs8lenx4 [accessed 21/02/2026]

ISO 1087: 2019. (2019). Terminology Work and Terminology Science – Vocabulary. Geneva: ISO

ISO 704:2022. Terminology work — Principles and methods. Geneva: ISO

Jakubíček, M. & M. Rundell. (2023). 'The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-Editing Lexicography?' In Medveď, Marek, Michal Měchura, Carole Tiberius, Iztok Kosem, Jelena Kallas and Miloš Jakubíček (eds), *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century*. Brno: Lexical Computing, 508–523. https://elex.link/elex2023/wp-content/uploads/102.pdf [accessed 21/02/2026]

Lew, R. (2023). ChatGPT as a COBUILD Lexicographer. *Humanities and Social Science Communications, 10(1), 704*, pp 1-10. https://doi.org/10.1057/s41599-023- 02119-6 [accessed 21/02/2026]

Lomidze, K. (2025). AI in Lexicography for Low-Resource Languages: The Case of the Georgian Language. *In II International Conference Lexicography in the XXI century, book of abstracts*. Tbilisi: Ilia State University, pp. 97-101. https://lexicography21.iliauni.edu.ge/wp-content/uploads/2025/10/Book-of-Abstracts-2025.pdf [accessed 21/02/2026]

Łukasik, M. (2023). Corpus linguistics and Generative AI Tools in Term Extraction: a Case of Kashubian – a Low-resource Language. In *Applied Linguistics Papers* 27 (4), pp. 34-45. DOI: 10.32612/uw.25449354.2023.4.pp.34-45 [accessed 11/02/2026]

Mchedlishvili, K. (2024). Comparable Corpora: Compilation Methods and Areas of Application. *Kadmos. A Journal of the Humanities*, (16), 237–253. https://doi.org/10.32859/kadmos/16/237-253 (In Georgian) [accessed 21/02/2026]

Mchedlishvili, K. (2025). Harnessing AI for Creating Definitions of Georgian Military Terminological Neologisms in the context of the Russia-Ukraine War. *In II International Conference Lexicography in the XXI century, book of abstracts*. Tbilisi: Ilia State Univer-

sity, pp. 112-116. https://lexicography21.iliauni.edu.ge/wp-content/uploads/2025/10/Book-of-Abstracts-2025.pdf [accessed 21/02/2026]

McKean, E & W. Fitzgerald. (2023). 'The ROI of AI in Lexicography' *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users".* Seoul: Yonsei University, 10–20. DOI:10.1558/lexi.27569 [accessed 21/02/2026]

Medzmariashvili, E. (Ed.). 2017. *Georgian Military Encyclopedic Dictionary*. Tbilisi. (In Georgian)

Military-Explanatory Dictionary and Graphic Symbols. FM (Field Manual) 1-02. (2017). Tbilisi: e Ministry of Defense of Georgia, Command of Training and Military Education, Doctrine Development Center. (In Georgian)

Nunzio, G.M., Gallina, E., & Vezzani, F. (2024). UNIPD@SimpleText2024: A Semi-Manual Approach on Prompting ChatGPT for Extracting Terms and Write Terminological Definitions. In G. Faggioli, N. Ferro, P. Galuscáková, A. García Seco de Herrera (eds.) *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9–12 September 2024*. CEUR Workshop Proceedings 3740. CEUR-WS, pp. 3230–3237. DOI: 2-s2.0-85201605270 [accessed 21/02/2026]

OpenAI ChatGPT. 2025. ChatGPT Response to Keti Mchedlishvili, 3 September. https://chatgpt.com/share/e/6999bb7b-47c0-8011-a5d9-809019d9eced [accessed 21/02/2026]

Rundell, M. & A. Kilgarriff. (2011). 'Automating the Creation of Dictionaries. Where Will It All End?' In Meunier, Fanny, Sylvie De Cock, Gaëtanelle Gilquin and Magali Paquot (eds), *A Taste for Corpora*. *In Honour of Sylviane Granger. In Studies in Corpus Linguistics 4,* pp. 257-281 Amsterdam: John Benjamins. https://doi.org/10.1075/SCL.45.15RUN [accessed 21/02/2026]

Rundell, M. (2023). 'Automating the Creation of Dictionaries: Are We Nearly There?' *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users"*. Seoul: Yonsei University, 1–9. https://www.hltmag.co.uk/feb24/automating-the-creation-of-dictionaries [accessed 21/02/2026]

Sabanés, L.V. & Da Cunha I. (2025). AI as a Resource for the Clarication of Medical Terminology. An Analysis of its Advantages and Limitations. In *Terminology, Computational terminology, 31(1)*, pp. 37-71. https://doi.org/10.1075/ term.00083.vid [accessed 21/02/2026]

San Martín, A. (2024). What Generative Artificial Intelligence Means for Terminological Definitions. In F. Vezzani, G. M. Di Nunzio, B. Sánchez-Cárdenas, P. Faber, M. Cabezas García, P. León Araúz, A. Reimerink & A. San Martín (eds.) In *Proceedings of the 3rd International Conference on Multilingual Digital Terminology Today (MDTT 2024), Granada, Spain, 27–28 June 2024*. Granada: CEUR-WS, pp. 1-37. Available at: https://ceur-ws.org/Vol-3703/paper1.pdf [accessed 21/02/2026]

Sorge, H. (2023). War in Ukraine: Unleashing Innovation and Unseen Frontiers. Policy Center of the New South. https://www.policycenter.ma/publications/war-ukraine-unleashing-innovation-and-unseen-frontiers [accessed 21/02/2026]

Struk I. V, T. H. Semyhinivska & A. V. Sitko. (2022). Formation and Translation of Military Terminology. *Scientific notes of V. I. Vernadsky Taurida National University, Series: Philology. Journalism Vol. 2. No. 2:* 29-33. http://www.philol.vernadskyjournals.in.ua/journals/2022/2_2022/part_2/5.pdf [accessed 21/02/2026]

Tchighladze, K. (2025). Integrating AI into the Term Creation Process: Case Study of English and Georgian Terms of Emerging Technologies. *In II International Conference Lexicography in the XXI century, book of abstracts*. Tbilisi: Ilia State University, pp. 149-155. https://lexicography21.iliauni.edu.ge/wp-content/uploads/2025/10/Book-of-Abstracts-2025.pdf [accessed 21/02/2026]

Trap-Jensen, L. (2024). The Best of Two Worlds: Exploring the Synergy between Human Expertise and AI in Lexicography. In T. Margalitadze (eds.) *Proceedings of the I International Conference Lexicography in the XXI Century, Tbilisi, 10-12 November 2023.* Tbilisi: Centre for Lexicography and Language Technologies Ilia State University, pp. 26-39. Available at: https://lexicography21. iliauni.edu.ge/wp-content/uploads/2024/06/03_Lars-Trap-Jensen.pdf. [accessed 21/02/2026]

Vicente, L. S. (2012). Approaching Secondary Term Formation Through the Analysis of Multiword Units: An English–Spanish Contrastive Study. Neology in Specialized CommunicationTerminology. *International Journal of Theoretical and Applied Issues in Specialized Communication*. *No. 18(1)*: 105-127, John Benjamins Publishing Company. DOI: 10.1075/term.18.1.06san [accessed 21/02/2026]

X Grok. 2025. Grok Response to Keti Mchedlishvili. 3 September. https://grok.com/share/c2hhcmQtNA_3060c25e-5479-44a0-be4a-59479efeffda [accessed 21/02/2026]

Past voices, present tools. Citizen Science and the Hesse-Nassau Dictionary

Nathalie Mederake, Robert Engsterhold
Marburg University, Research Center Deutscher Sprachatlas
E-mail: mederake@uni-marburg.de, engsterhold@uni-marburg.de

Abstract

This paper examines the role of Citizen Science in dialect lexicography through the case of the Hesse-Nassau Dictionary (HNWb). It argues that participatory practices have long been integral to dialect documentation, particularly through large-scale questionnaire surveys and the involvement of local speakers as linguistic informants. These historically established forms of collaboration constitute both the material and methodological foundation of the HNWb archive, which comprises thousands of questionnaires and hundreds of thousands of handwritten paper slips.

In light of this context, the paper conceptualises Citizen Science in dialect lexicography not as a methodological innovation introduced by digital media, but as an epistemic reconfiguration of historically embedded participatory practices. Building on this legacy, it analyses how contemporary digital tools allow for a systematic continuation and formalisation of participation under present-day conditions. Two dedicated applications support citizen contributors in the transliteration and annotation of historical dialect materials that cannot be efficiently processed using automated handwritten text recognition.

Through these workflows, contributors move from the role of informants to that of collaborators who actively participate in the archive's transformation, structuring, and enrichment. The paper argues that incorporating Citizen Science into lexicographic workflows enhances the sustainability, accessibility, and societal relevance of long-term dialect dictionary projects. Digital infrastructures thus allow past voices to be reactivated and reinterpreted, strengthening the preservation and contemporary usability of regional linguistic heritage.

Keywords: Citizen Science; Dialect Lexicography; Digital Archives

1 Citizen Involvement in Dialect Documentation

Citizen Science provides a useful conceptual framework for understanding the long-standing involvement of non-academic contributors in dialect lexicography. Rather than being a recent innovation, the systematic inclusion of lay participants has been a defining feature of many large-scale dictionary projects, particularly those concerned with regional and

non-standard varieties. Although general historical dictionaries such as the *Oxford English Dictionary* are often cited as early examples of public participation in lexicography[1], dialect documentation practices represent an even more consistent culturally and structurally embedded form of citizen involvement (cf. Spiekermann 2016).

From the early nineteenth century onwards, dialectological research relied heavily on local speakers (e.g., Schmeller 1821) and large-scale undertakings such as linguistic atlases would have been inconceivable without this distributed network of contributors (cf. Wenker 1889-1923; Gilliéron 1902-1914). Particularly in the context of dialect dictionaries, citizen participation was not merely supportive but constitutive. The empirical basis of these projects depended almost entirely on contributions from non-academic collaborators, who acted as knowledgeable insiders to local speech communities, contributed lexical knowledge, usage examples, and metalinguistic insights through questionnaires, word lists, and correspondence. Their role corresponds closely to what is now described in Citizen Science research as the co-production of knowledge: contributors did not simply deliver raw data but shaped the scope, depth, and regional differentiation of the resulting lexicographic descriptions.

The Hesse-Nassau Dictionary (hereinafter also HNWb) is not merely an example but a paradigmatic condensation of these participatory principles. As one of the most innovative dialect dictionaries of the early twentieth century, it was the first to systematically integrate questionnaire data into individual word entries, thereby establishing the principle of word geography (cf. Berthold 1938; Knoop et al. 1982; Lameli 2014). Information on location thus functions as an indicator of a lemma's geographical scope of validity, foregrounding spatial distribution as an epistemically relevant dimension of lexical meaning. This approach embedded citizen-generated data directly into the analytical structure of the dictionary. Seen from a contemporary perspective, the HNWb archive thus exemplifies an early, large-scale Citizen Science infrastructure avant la lettre, in which expert knowledge and local linguistic competence were tightly interwoven.

In today's digital environment, this historical model of participation is being reinterpreted and extended. Citizen contributors increasingly move beyond the role of informants to become active collaborators who participate in the transcription, annotation, interpretation, and contextualisation of dialect data (Dorn et al. 2024; Dücker/Fischer 2025). This shift reflects a broader development in Citizen Science toward more reflexive, sustainable, and mutually beneficial forms of collaboration and highlights the continuing relevance of participatory approaches for dialect lexicography and regional linguistic heritage.

1 As early as 1859 the Philological Society laid the groundwork for the collective enterprise, that would become the Oxford English Dictionary (cf. Proposal for the Publication of a New English Dictionary by The Philological Society. London: Trübner). Especially Murray (1879) explicitly mobilized the English- and American-speaking public to collect usage evidence; hundreds of volunteers supplied quotation slips that underpinned definitions and preliminary semantic structures.

2 Defining Citizen Science

Against the historical backdrop of sustained public involvement in dialectological and lexicographic research, the concept of Citizen Science provides a differentiated, albeit contested, analytical framework. As Haklay et al. (2021) argue, the field is resistant to a single definition, as criteria often depend on geographical, disciplinary, and political contexts. Generally, however, in the humanities it is understood as the active involvement of non-professional participants in one or more stages of the research process, ranging from data generation to interpretation and dissemination (Pettibone/Ziegler 2016). Crucially, participation is not conceived as an undifferentiated activity but as structured engagement embedded in scientific workflows.

For projects such as the HNWb, Haklay's typology, which distinguishes four levels of Citizen Science according to the depth of citizen involvement (Haklay 2013, 2018), is a useful classification scheme, but comes with limitations. At Level 1 (Crowdsourcing), citizens contribute resources or automatically collected data without being cognitively involved in scientific interpretation. Level 2 (Distributed Intelligence) marks the threshold at which active participation begins: citizens perform intellectually relevant tasks such as observing phenomena, categorising data, or transcribing texts. Level 3 (Participatory Science) involves citizens in shaping research questions and data collection strategies, while Level 4 (Extreme or Collaborative Citizen Science) is characterised by an egalitarian collaboration in which professional researchers and citizen scientists jointly negotiate all stages of the research process.

This model is advantageous for classifying citizen involvement in dialect lexicography, as it allows historical and contemporary practices to be assigned to different levels of participation. Traditional dialect documentation relied heavily on what would today be classified as Level-2 activities: local speakers systematically answered questionnaires, compiled word lists, and provided usage examples, thereby contributing expert linguistic knowledge rooted in everyday practice. In some cases, long-term informants or regional correspondents, many of whom occupied socially and institutionally hybrid positions, such as teachers, acted as knowledgeable insiders and contributed material in a sustained and locally focused manner, thereby shaping the empirical profile of the documented data.[2] Although dialect informants might superficially resemble Level-1 contributors acting as linguistic 'sensors', the cognitive effort involved in translating lived linguistic experience into structured responses justifies their classification as Level-2 contributors. Rather than constituting participatory research in the sense of Haklay's Level 3, such contributions can be understood as proto-participatory practices whose epistemic relevance remained largely implicit. The decisive

2 The prominent role of teachers as contributors highlights a hybrid form of participation that challenges contemporary Citizen Science definitions. While teachers were not professional linguists, their involvement was institutionally embedded and partially aligned with their occupational roles. This blurs the often-assumed distinction between 'professional' and 'citizen' participation, which in many modern CS frameworks is implicitly tied to a separation between paid work and voluntary leisure activity. This historical constellation complicates typologies that rely on a strict binary between professional researchers and voluntary citizen contributors (cf. Haklay 2013).

difference from contemporary Citizen Science thus lies not in the principle of participation itself but in its explicit conceptualisation, formalisation, and infrastructural embedding.

Digital infrastructures make historically implicit forms of collaboration explicit by embedding them into transparent, structured, and reproducible workflows. Activities such as transcription, annotation, and categorisation can thus be systematically implemented as Level-2 (Distributed Intelligence) practices, while at the same time creating conditions for more reflexive and sustained forms of engagement (Dücker/Fischer 2025). Contributors are no longer confined to isolated acts of data provision but gain insight into the epistemic context, decision-making processes, and research goals that shape the archive.

Understanding Citizen Science as a graded and historically embedded practice therefore provides a conceptual bridge between analogue dialect documentation and contemporary digital lexicography. The HNWb illustrates that Level-2 activities in the humanities are not an "inferior" preliminary stage but constitute the epistemic backbone of long-term lexicographic research. Haklay's typology is thus not employed as a normative scale of participatory "success", but as a descriptive model for analysing the division of labour between academic research, public participation, and digital infrastructures in long-term dialect dictionary projects.

From this perspective, digital Citizen Science in dialect lexicography represents both a continuation and a reconfiguration of earlier participatory practices. Its point of departure is not an abstract research design but the material legacy of historical collaboration. In the case of the Hesse-Nassau Dictionary, the extensive archive of paper slips and questionnaires embodies both the outcome of earlier citizen involvement and the foundation for renewed participation under digital conditions.

3 Starting points for Citizen Science: The Archive

As a material legacy of early participatory research, the archive of the Hesse-Nassau Dictionary forms the natural starting point for contemporary Citizen Science initiatives. The archive is not only a repository of data but also a repository of participatory practices. It is fed by two distinct but closely interrelated data classes.

(1) The **slip archive** constitutes its core component. It comprises more than 300,000 paper slips on which meanings and usage information were excerpted, classified, and lemmatized. These data stem from unsolicited public submissions that were actively encouraged through postcard surveys as well as calls published in journals and calendars between 1914 and 1926 (Sitzungsberichte 1920: 132). A single slip may contain different types of linguistic information; in broad terms, these can be distinguished between meaning descriptions, phonetic notes, contextualised word attestations (sentence examples), form attestations, and morphological or syntactic information.

(2) The second data class consists of a **questionnaire corpus**, made up of responses provided by local contributors (i.e. Gewährspersonen), predominantly schoolteachers. Its core comprises 79 questionnaire rounds in parts of the dictionary area, amounting in total to approximately 19,000 completed questionnaire pages. The responses

cover a wide thematic spectrum, ranging from family life and traditional customs to flora and fauna and local knowledge, and frequently include structured elicitation tasks, such as naming the parts of illustrated objects (e.g. a loom) or listing regional expressions for predefined concepts (Werth et al. 2021, 203). While the questionnaire material represents a substantial and systematically collected body of data, it was only rudimentarily excerpted by the dictionary editors, leaving much of its analytical potential unrealised.

A central motivation behind the questionnaire surveys – and indeed behind the Hesse-Nassau Dictionary as a whole – was the transfer of the dialect-geographical method developed at the Deutscher Sprachatlas to lexicography (the so-called "Marburg School"; cf. Knoop et al. 1982; Lameli 2014). In this tradition, linguistic forms are not primarily conceived as abstract system elements but as spatially distributed phenomena whose geographical patterning constitutes an essential part of their linguistic description. Consequently, information on the place of attestation does not merely function as a source reference or documentary metadata. Rather, it serves as an indicator of the geographical scope of validity of a lexical item and thus becomes an integral component of its semantic and pragmatic characterisation. Lexicographic description, in this sense, is inseparable from the mapping of linguistic variation.[3]

Together, these materials were gathered with the involvement of more than 1,000 local contributors and amount to roughly 1.5 million dialectal references. The richness of this material stands in marked contrast to its limited integration into the printed dictionary since large parts of the archive remain accessible only through indirect references. The complete digital preservation of the archive therefore not only secures a unique historical resource but also exposes its latent analytical potential. Unlocking this potential requires a systematic transformation of heterogeneous, analogue materials into structured, machine-readable data – a process that cannot be achieved by editorial expertise alone but calls for renewed forms of distributed, participatory engagement. A further obstacle to systematic evaluation lies in the material and graphic characteristics of the sources themselves. Most paper slips and questionnaire answers are written in a range of historical handwriting systems, including cursive scripts such as Kurrent and Sütterlin, which are no longer in everyday use. Owing to their graphic variability and palaeographic complexity, these scripts cannot yet be transcribed efficiently using current Handwritten Text Recognition (HTR) technologies such as Transkribus (https://www.transkribus.org) or eScriptorium (https://github.com/UB-Mannheim/escriptorium). As a result, large parts of the archive continue to depend on manual transliteration, making renewed forms of distributed, human-based participation not merely desirable but methodologically indispensable.

3 According to the principle of "word geography" proposed by Berthold (1938), the geographical distribution of the keywords described is illustrated with the aid of maps. As the first dictionary project of its kind, the HNWb attempted to consistently apply this principle from the outset (an example that many parallel projects have followed), and its word articles and maps provide a wealth of ideas and results for further research in the field of geographical comparative analysis.

4 From Archival Data to Digital Participation

Against this background, the digital Citizen Science infrastructure of the Hesse-Nassau Dictionary is designed to operationalise the archive's latent analytical potential by transforming historically grown materials into structured, machine-readable data through renewed forms of public participation. Two complementary applications support this process, addressing different core components of the archive and enabling distinct but interrelated modes of participation. In addition to their epistemic relevance, the applications contribute to the preservation of the archival material. Many original questionnaires are in a critical physical condition due to the high wood content of the paper and advanced ageing. The digital workflow reduces the need for physical handling and thus helps to safeguard the material for future research. The two applications presented here are implemented as high-fidelity prototypes. Their functional scope corresponds to that of intended production systems, and parts of the infrastructure are publicly accessible (https://uni-marburg.de/bNlc7L). At the same time, the tools are currently in a pilot phase: they have so far been tested by a limited number of users and do not yet provide access to the entire digitised archive.

From a technical perspective, both applications are embedded in a shared digital infrastructure. Built as a progressive web app (https://developer.mozilla.org/en-US/docs/Web/Progressive_web_apps, https://quasar.dev) with a FastAPI-based service endpoint (https://fastapi.tiangolo.com) and backed by a PostgreSQL (https://www.postgresql.org) database, the system transforms the heterogeneous archival holdings into structured, analysable data. In addition, the integration of IIIF-compliant image services (https://iiif.io) ensures that digitised cultural assets – such as questionnaire pages, handwritten slips, maps, or other archival objects – can be delivered and displayed in a standardised manner via web interfaces. Users can interact with the applications of the website without requiring personal registration. A token-based system ensures that individual contributions can be persistently associated with anonymous contributor profiles, enabling continuity and transparency. This architectural design not only supports the current Citizen Science workflows but also provides a sustainable foundation for future extensions, interoperability, and reuse of the data.

4.1 Document Editor

While questionnaire surveys formed the empirical backbone of early dialect documentation, the smallest and most analytically condensed units of the HNWb archive are the individual paper slips. The document editor is dedicated to their systematic digitisation, structuring, and enrichment.

4.1.1 Concept: From Card Index to Database

The conceptual aim of the document editor is the in-depth annotation of the slip archive. The HNWb is based on around 300,000 slips, many of often handwritten slips, each documenting a lexical item together with information on meaning, usage in context, and place of attestation. In their original analogue form, these slips were organised in a physical card

index that embodied decades of editorial practice but remained inaccessible to systematic computational analysis.

In contrast to the questionnaire material, which is primarily organised by locality, the slip archive follows a word-oriented logic. The document editor is therefore designed to enable contributors to work on clusters of slips associated with individual lemmas. Its overarching goal is to transform the analogue card index into a digital, searchable corpus that can serve as the empirical foundation for both printed and digital dictionary entries.

4.1.2 Structure and Functionality

The document editor is implemented as a specialised interface optimised for the efficient processing of small-scale data units. At its centre, a digital image window displays the scan of the original slip. As many slips are no larger than a small note card, the viewer is designed for rapid readability and focused inspection.

Figure 1: User interface of the document editor. The image shows the digital capture of a handwritten note (ID: 1057) on the lemma "Eselsohr" (dog-ear). On the left is the original document in cursive script, and on the right is the corresponding line-by-line transcription and annotation in the database mask.

Adjacent to the image, the editor interface allows contributors to transliterate the handwritten content and to assign structured metadata using colour-coded and numerical tags. These tags capture information such as lemma assignment, meaning, regional spelling,

place of attestation, and local informant. Users will be able to work through entire sets of slips associated with a given lemma in a continuous workflow, closely resembling – but digitally enhancing – the traditional practice of sorting and annotating paper slips. In its current implementation, the document editor operates on a selected subset of the digitised slip archive. While the core functionalities reflect the design of the intended end product, access to the complete corpus will be enabled in later development stages.

4.1.3 Involvement of Citizen Scientists

The granularity of the document editor allows the processing of single paper slips as bite-sized tasks. This lowers the threshold for participation and allows volunteers to contribute flexibly, even in short intervals. At the same time, the platform supports a degree of thematic and regional specialisation. Citizen scientists may deliberately select lemmas that are of interest and thus contribute situated linguistic knowledge to the annotation process. Local expertise is particularly valuable when interpreting usage notes, regional spellings, or semantically opaque expressions. Quality assurance is supported through review mechanisms that allow entries to be checked by other contributors or by expert editors. Through this form of collaborative philology, citizen scientists participate directly in the transformation of the “raw data” of linguistic history into structured evidence. Their work provides the indispensable groundwork for scholarly analyses of lexical variation, semantic change, and regional distribution.

To ensure the scholarly reliability and analytical usability of citizen-generated data, the HNWb infrastructure implements a multi-layered quality assurance framework that is embedded directly into the digital workflows. At the level of data entry, structural constraints within the interfaces – such as predefined tags, controlled vocabularies, and guided input fields – reduce formatting inconsistencies and limit interpretative variance. In addition, a collaborative validation model is employed: for complex or palaeographically ambiguous documents, multiple independent transliterations or annotations can be created for the same source, allowing discrepancies to be identified through comparative inspection. Cases of significant ambiguity or conflict could also be systematically flagged and subjected to expert review by the editorial team. Crucially, all digital entries remain persistently linked to their original manuscript sources via IIIF-compliant image services, ensuring transparent provenance and verifiability at every stage of the lexicographic process. This combination of structured input, collaborative validation, and expert oversight balances the openness of Citizen Science with the methodological standards of digital lexicography.

4.2 Questionnaire Application

While the document editor operates on the smallest analytical units of the archive, the questionnaire application addresses the large-scale survey material that forms the historical backbone of the HNWb.

4.2.1 Concept and Objectives

The overarching goal of the questionnaire application is the digitisation and transliteration of the historical questionnaire corpus of the Hesse-Nassau Dictionary. At its core lies a collection of approximately 6,500 handwritten questionnaires completed between 1914 and 1926 by voluntary local contributors. These questionnaires represent an extensive and systematically collected source of dialect data in the archive.

From a scientific perspective, the application responds to the fact that, during the original compilation of the dictionary, only selected parts of the questionnaire material were manually excerpted and transferred onto paper slips. Large portions of the original responses were therefore never fully evaluated. By converting the handwritten content into machine-readable text through manual transliteration, the application aims to unlock these previously unexplored linguistic resources and make them available for integration into modern linguistic databases.

4.2.2 Structure and Functionality

Similar to the document editor, the questionnaire application is implemented as a modular web-based environment designed to support focused transliteration work. The entry point is a questionnaire catalogue page that provides an overview of the available material. On the left-hand side, users see a list of questionnaire pages available for processing; on the right-hand side, the corresponding localities are displayed as points on a map. Visual indicators signal the processing status: an orange outline around a location point or an orange background behind a page icon indicates that a transliteration already exists. Users may nevertheless initiate additional transliterations at any time.

Selecting a page icon opens an individual questionnaire page that is uniquely associated with a specific locality and questionnaire series. The working view is again organised as a split-screen interface. On the left-hand side, a digital viewer displays a high-resolution scan of the original handwritten questionnaire. On the right-hand side, the editor interface provides input fields for the letter-by-letter transliteration of the dialect data, supplemented by auxiliary tools such as special-character palettes for phonetic symbols.

Contributors may transliterate an entire questionnaire page or focus on individual questions, for example a prompt eliciting local expressions for *Wacholder* (*Juniperus communis*). A token-based session mechanism stores the progress persistently without requiring a formal registration process, thereby combining usability with data protection. Although the application is technically capable of handling large-scale questionnaire material, it currently provides access only to a defined portion of the digitised corpus. This phased approach allows for iterative testing and refinement of the workflow under real research conditions.

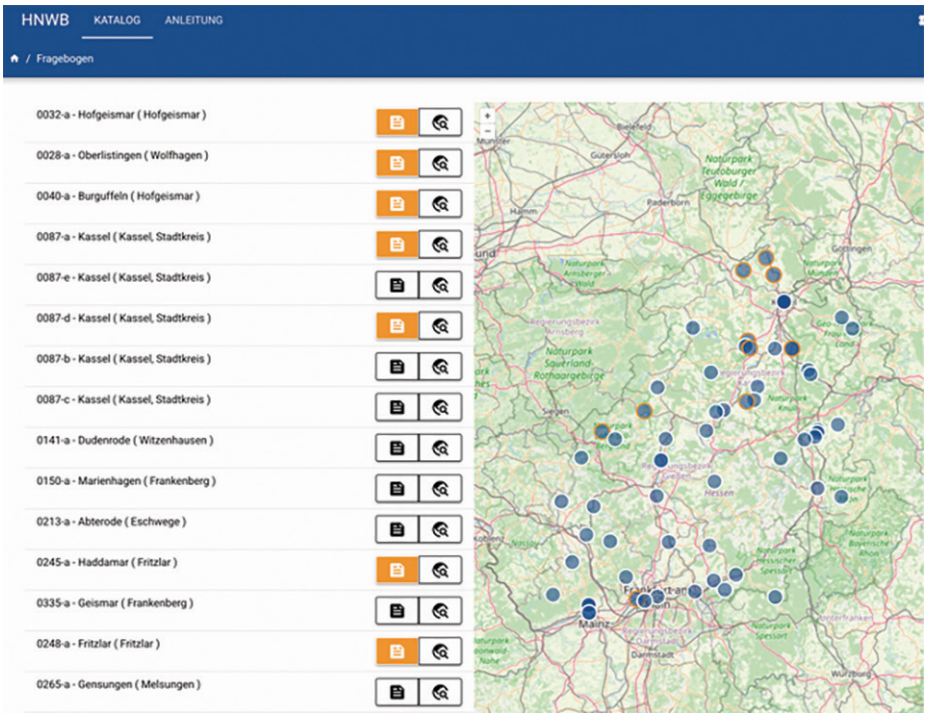


Figure 2: Questionnaire catalogue. The interactive overview map and document list enable targeted access to dialect data via a list or direct selection of locations in the Hesse-Nassau region.

4.2.3 Involvement of Citizen Scientists

The questionnaire application represents a classic Citizen Science setting that depends on broad public participation. Its design deliberately lowers barriers to entry in order to engage interested laypersons, dialect speakers, and regional historians. The anonymous token system allows contributors to begin working immediately and to resume their contributions at a later point.

At the methodological level, the platform implements transliterations through distributed intelligence. Given the volume of handwritten material and the palaeographic challenges it presents, systematic processing would not be feasible without the involvement of citizen scientists. Their contributions enable the deciphering and digitisation of handwriting that cannot be reliably processed using automated methods.

At the same time, the platform facilitates knowledge transfer. Contributors engage actively with the linguistic history of their region, while the application provides detailed guidelines and instructional materials to ensure data quality suitable for scholarly analysis. In this way, the questionnaire application functions as a digital bridge between an analogue archival legacy and contemporary linguistic research, extending historical participatory practices into the digital age.

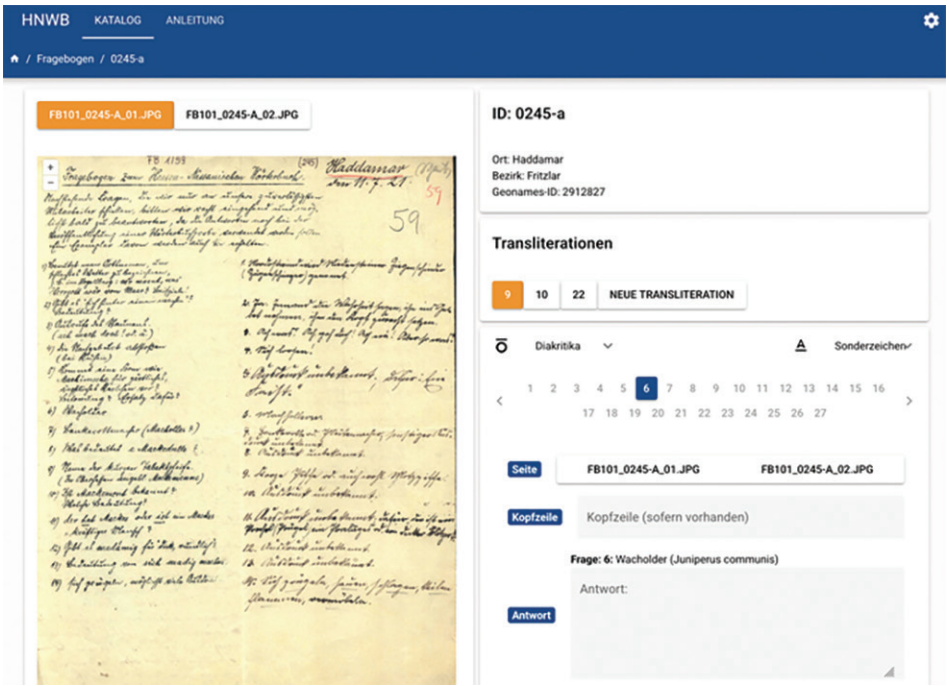


Figure 3: Questionnaire App. The image shows the work interface for transliterating questionnaire 0245-a from Haddamar (Fritzlar district). On the left is the original document, while on the right is the input mask for the transliteration of the answers – here, for example, for question 6 ("juniper").

5 From Informants to Collaborators

In earlier stages of dialect lexicography, the contributions of local respondents were indispensable to the empirical foundation of dictionary projects, yet largely confined to the phase of data generation. While questionnaires, correspondence, and unsolicited submissions supplied the material basis of dialect lexicographic work, decisions concerning selection, structuring, interpretation, and presentation remained firmly within the domain of academic expertise. Citizen participation was thus constitutive but epistemically backgrounded, operating as an implicit prerequisite or proto-participatory rather than as an explicitly theorised component of knowledge production.

Within the framework of digital Citizen Science, this historically established division of labour is being reconfigured. As demonstrated by the applications presented above, citizen contributors are no longer restricted to providing linguistic material but participate directly in the transformation, structuring, and enrichment of historical sources. Through activities such as the transliteration of handwritten questionnaires and the annotation of archival slips, they intervene at stages that decisively shape the form, granularity, and analytical usability of the data. In this sense, they function as collaborators embedded within clearly defined lexicographic workflows.

Crucially, this shift does not imply a dissolution of established epistemic roles, nor does it amount to participatory research in the strong sense of Haklay's highest levels. At the same time, it invites a reconsideration of where citizen contributions are situated within the lexicographic process as traditionally described in lexicographic theory (Klosa-Kückelhaus/Tiberius 2024). Rather, it reflects a reassessment of tasks that were traditionally classified as preparatory or technical. In the framework of the lexicographic process, such tasks form integral stages linking data collection, data modelling, and the construction of dictionary entries; they are not auxiliary to lexicographic work, but constitutive of how lexical knowledge is selected, structured, and ultimately represented. Decisions taken during transliteration, annotation, and categorisation – such as the interpretation of ambiguous handwriting, the identification of regional variants, or the assignment of annotations – are epistemically consequential precisely because they intervene at these processual interfaces. They pre-structure the organisation of the data and condition the kinds of linguistic generalisations that can subsequently be drawn. Digital Citizen Science thus renders visible a layer of interpretative labour that has always been present in lexicographic practice but rarely acknowledged as such.

From this perspective, Haklay's typology can be productively aligned with the dialect-geographical tradition of the Marburg School. The spatial principle underlying the Hesse-Nassau Dictionary – according to which geographical distribution constitutes a core dimension of lexical meaning – was only realisable through the distributed intelligence of its contributors. The historical reliance on locally embedded actors exemplifies a form of situated expertise: Rather than representing a merely logistical solution, this hybrid constellation constituted an institutionalised epistemic arrangement that prefigures contemporary models of collaborative knowledge production.

The prototype status of the digital applications plays a central methodological role in this reconfiguration. As high-fidelity yet experimental systems, they allow collaborative workflows to be observed, evaluated, and iteratively refined. Patterns of participation, interpretative strategies, and quality-related challenges become empirically accessible before large-scale implementation. In this way, prototyping functions not simply as a technical development stage, but as an epistemic instrument for reflecting on the conditions and consequences of Citizen Science in lexicographic research.

Taken together, the transition from informants to collaborators does not mark a rupture with earlier dialectological practices (Fischer 2024). Rather, it makes the epistemic significance of participatory labour that has long underpinned dialect lexicography explicit. By embedding this labour into transparent, structured, and traceable digital workflows, contemporary Citizen Science re-articulates historical practices under new media conditions and establishes them as a reflexive component of lexicographic methodology (cf. Dorn et al. 2024).

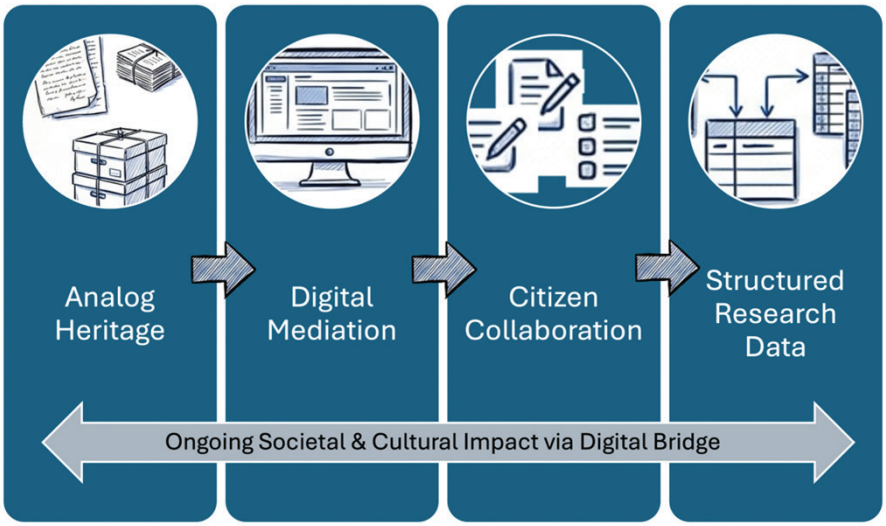


Figure 4: The Digital Bridge model. The diagram illustrates the transformative workflow from analogue archival heritage to structured research data, mediated by digital infrastructures and powered by the epistemic collaboration of citizen scientists to ensure ongoing societal impact.

6 Conclusion and outlook

Building on the epistemic reconfiguration outlined in Section 5, this paper has argued that Citizen Science in dialect lexicography is best understood not as an external innovation, but as a reflexive framework for analysing and reorganising long-standing forms of cooperative knowledge production. The case of the Hesse-Nassau Dictionary shows that the participatory logic of dialect documentation predates contemporary Citizen Science discourse by more than a century. What has changed is the media, infrastructures, and conceptual tools through which participation is made explicit, formalised, and sustained.

From a meta-theoretical perspective, the study contributes to current discussions of Citizen Science in the humanities by foregrounding the role of archives, long-term projects, and historically layered data. The HNWb exemplifies a form of distributed and situated expertise that corresponds closely to what is classified as Level-2 participation according to Haklay (2013, 2018), while at the same time challenging normative hierarchies that equate deeper participation exclusively with co-design or full collaboration. In the context of lexicography, distributed intelligence does not represent a preliminary or subordinate mode of engagement, but constitutes the epistemic backbone of the enterprise.

Methodologically, the two digital prototypes presented here – the Document Editor and the Questionnaire Application – demonstrate how participatory workflows can be operationalised within a robust digital infrastructure. These tools function not merely as technical interfaces, but as epistemic environments in which historical sources are transformed into structured, analysable data. By integrating quality assurance, provenance tracking, and collaborative validation directly into the workflows, the infrastructure reconciles the openness

of Citizen Science with the methodological standards of digital lexicography. At the same time, the prototype-based approach enables a reflective mode of implementation in which participatory practices themselves become objects of empirical observation and methodological adjustment.

Looking ahead, the long-term sustainability of dialect lexicography will depend on the successful integration of such participatory infrastructures. This includes not only the maintenance of technical systems, but also the cultivation of enduring relationships with citizen contributors and regional communities. Beyond their immediate scientific value, digital Citizen Science initiatives foster shared ownership of linguistic heritage and create opportunities for cultural rediscovery that extend well beyond academic contexts.

Ultimately, the transformation of the Hesse-Nassau Dictionary's archive into a digital, participatory infrastructure underscores a central insight of this study: the "citizen" has always been at the heart of dialectology. By making participatory labour explicit and epistemically accountable, digital tools do not replace historical practices, but enable past linguistic knowledge to be reinterpreted, recontextualised, and integrated into contemporary research. In this sense, Citizen Science provides not only a practical framework for data processing, but a conceptual lens through which the continuity, adaptability, and societal embeddedness of dialect lexicography can be understood anew.

References

Berthold, L. (1938). Die Wortkarte im Dienste der Bedeutungslehre. In *Zeitschrift für Mundartforschung 14*, Stuttgart: Franz Steiner, pp. 101–106.

Dorn, A., Stocker, R. & Stöckle, P. (2024). Old dialect words through the ages – the ABCs of dialect project. In *ARPHA Proceedings 6,* pp. 213-216. Accessed at: https://doi.org/10.3897/ap.e126110 [19/01/2026].

Dücker, L. & Fischer, H. (2025). Sprachvariation dokumentieren, transkribieren, reflektieren – Citizen Science in der Regionalsprachenforschung. In E. Wyss, et al. Gruß & Kuss – Briefe digital. Bürger*innen erhalten Liebesbriefe. Studien zur Liebesbriefforschung und zu Citizen Science in den Geistes- und Kulturwissenschaften. *Digital Philology | Working Papers in Digital Philology 04|2025*. Darmstadt: TUPrints. Accessed at https://doi.org/10.26083/tuprints-00030732 [19/01/2026].

Fischer, H. (2024): Dialektologie als Citizen Science. Perspektiven und Tools für eine bürgernahe Wissenschaft. In Ch. Ferstl, P. Kaspar, L. Zehetner (eds.): *Dialekt unterwegs. Varietäten im Zeichen von Globalisierung und Migration*. Tirschenreuth: Johann-Andreas-Schmeller-Gesellschaft.

Gilliéron, J. & Edmont, E. (1902-1914). *Atlas linguistique de la France*. 10 vols. Paris: Champion.

Haklay, M., Dörler, D., Heigl, F., Manzoni, M., Hecker, S. & Vohland, K. (2021). What Is Citizen Science? The Challenges of Definition. In K. Vohland, et al. *The Science of Citizen Science*. Cham: Springer, pp. 13-33. Accessed at: https://doi.org/10.1007/978-3-030-58278-4_2 [19/01/2026].

Haklay, M. (2013). Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation. In D. Sui, S. Elwood, M. Goodchild (eds.) *Crowdsourcing Geographic Knowledge*. Dordrecht: Springer. Accessed at: https://doi.org/10.1007/978-94-007-4587-2_7 [19/01/2026].

Haklay, M. (2018). Participatory citizen science. In S. Hecker, et al. (eds.) *Citizen Science: Innovation in Open Science, Society and Policy*. London: UCL Press, pp. 52-62. Accessed at: https://doi.org/10.14324/111.9781787352339 [19/01/2026].

Klosa-Kückelhaus, A. & Tiberius, C. (2024). 3 The Lexicographic Process. In A. Klosa-Kückelhaus (ed.) *Internet Lexicography: An Introduction*. Berlin, Boston: De Gruyter, pp. 61-96. Accessed at: https://doi.org/10.1515/9783111233758-004 [19/01/2026].

Knoop, U., Putschke, W. & Wiegand, H. E. (1982). Die Marburger Schule: Entstehung und frühe Entwicklung der Dialektgeographie. In W. Besch, U. Knoop, W. Putschke, H. E. Wiegand (eds.) *Dialektologie. Ein internationales Handbuch zeitgenössischer Forschung*. Bd. 1. Berlin: de Gruyter, pp. 38-92.

Lameli, A. (2014): *Schriften zum Sprachatlas des Deutschen Reichs*. Bd. 3. Erläuterungen und Erschließungsmittel zu Georg Wenkers Schriften. Hildesheim: Olms.

Murray, J. A. H. (1879). *An appeal to the English-speaking and English-reading public to read books and make extracts for the Philological Society's new English Dictionary*. Oxford: Clarendon Press.

Pettibone, L. & Ziegler, D. (2016). Citizen Science: Bürgerforschung in den Geistes- und Kulturwissenschaften. In K. Oswald, R. Smolarski (eds.) *Bürger Künste Wissenschaft: Citizen Science in Kultur und Geisteswissenschaften*. Gutenberg: Computus Druck Satz & Verlag, pp. 57-69.

Proposal for the Publication of a New English Dictionary by The Philological Society (1859). London: Trübner.

Schmeller, J. A. (1821). *Die Mundarten Bayerns grammatisch dargestellt*. München: Thienemann.

Sitzungsberichte der Preussischen Akademie der Wissenschaften. Philosophisch-historische Klasse. Jahrgang 1920. Berlin: Verlag der Akademie der Wissenschaften.

Spiekermann, H. H. (2016). 73. Kulturwissenschaftliche Orientierung in Dialektologie / Sprachgeographie. In L. Jäger, et al. (eds.) *Sprache - Kultur - Kommunikation / Language - Culture - Communication: Ein internationales Handbuch zu Linguistik als Kulturwissenschaft*. Berlin, Boston: De Gruyter Mouton, pp. 701-707. Accessed at: https://doi.org/10.1515/9783110224504-074 [19/01/2026].

Wenker, G. (1889-1923). Sprachatlas des Deutschen Reichs. Marburg: Forschungszentrum Deutscher Sprachatlas. Published as Digitaler Wenker-Atlas (DiWA) at https://regionalsprache.de/atlanten-und-karten.aspx [19/01/2026].

Werth, A., Vielsmeier, B. & Aumann, St. (2021). Hessen-Nassauisches Wörterbuch (HNWb). In A. N. Lenz, Ph. Stöckle (eds.). *Germanistische Dialektlexikografie zu Be-*

ginn des 21. Jahrhunderts. Stuttgart: Steiner, pp. 201-221. Accessed at: https://doi.org/10.25162/9783515129206 [19/01/2026].

Acknowledgements

Thanks are due to Tinatin Margalitadze, organiser of the II International Conference on *Lexicography in the XXI Century*, for insightful feedback and constructive discussion. We also would like to thank Selina Berus for her careful reading of the manuscript; any remaining errors are our own.

ენაში ანალიტიზმის წარმოქმნის მიზეზები. დეგრადაცია თუ განვითარება? (თანამედროვე ქართული ენისთვის დამახასიათებელი ზოგიერთი ტენდენციის მაგალითზე)

გიორგი მელაძე
ილიას სახელმწიფო უნივერსიტეტი
giorgi.meladze.4@iliauni.edu.ge

აბსტრაქტი

ენების განვითარებაზე და სინთეზური, მორფოლოგიურად მდიდარი და მრავალფეროვანი ტიპიდან ანალიზურ, თითქოსდა მორფოლოგიურად „გამარტივებულ“ ტიპზე მათი გადასვლის შესახებ ენათმეცნიერებაში ოდითგანვე ურთიერთგამომრიცხავი შეხედულებები გამოითქმებოდა.

რომანტიკული მიმდინარეობის წარმომადგენლები: იაკობ და ვილჰელმ გრიმები, იოჰან გ. ჰერდერი და სხვანი (Rastorguyeva, 1983, pp. 20-21), მიიჩნევდნენ, რომ ძველი ინდოევროპული და გერმანიკული ენების დიაქრონიული განვითარება და თანამედროვე ენებად მათი ჩამოყალიბების პროცესი არსებითად მათს გადაგვარებას და დეგრადაციას წარმოადგენდა. ამგვარი დეგრადაციის მკაფიო გამოხატულებად მათ ამ ენების გრამატიკული ფორმების, ბრუნებისა და უღლების მრავალფეროვნებისა და სიმდიდრის დაკარგვა ესახებოდათ. თუ ერთმანეთს შევადარებთ ანგლოსაქსურ (ძველინგლისურ) და თანამედროვე ინგლისურ, ლათინურსა და იტალიურ, ძველ ზემოგერმანულსა და თანამედროვე გერმანულ, ანდა პროტოგერმანიკულთან ძალზე ახლოს მყოფ გოთურსა და, ვთქვათ, თანამედროვე შვედურ ენებს, თვალნათლივ დავინახავთ, რას ეფუძნებოდა ლინგვისტიკის „რომანტიკოსთა“ ამგვარი მიდგომა.

მეორე მხრივ, ცნობილი დანიელი ენათმეცნიერი ოტო იესპერსენი მიიჩნევდა, რომ ენაში მიმდინარე დიაქრონიული პროცესები და ცვლილებები არსებითად მის დახვეწაში და პროგრესში მდგომარეობს და რომ ეს პროცესები ენის ეფექტიანობის უცილობელ ზრდას იწვევს. ითვლება, რომ მის ამგვარ შეხედულებებზე გარკვეული ზეგავლენა ჩარლზ დარვინისეულმა ევოლუციის თეორიამაც მოახდინა. იესპერსენის ფუნდამენტური ნაშრომის: „ენა, მისი ბუნება, განვითარება და წარმოშობა“ XVII თავს ასევე ეწოდება: „პროგრესი თუ გადაგვარება?“ (Progress or Decay?) (Jespersen, 1922, pp. 319-337). სწორედ ამ თავმა და მისმა საერთო სულისკვეთებამ შთაგვაგონა, წინამდებარე ნაშრომი ამგვარად დაგვესათაურებინა.

თავად ჩვენი ნაშრომის მიზანია, განვიხილოთ ანალიზური კონსტრუქციები, რომლებიც საკმაოდაა მომრავლებული თანამედროვე ქართულში, თუმცა ტრადიციული ქართული ენათმეცნიერება მათ ან ვერ ამჩნევს, ან მათ ქართული ენის ნორმების დარღვევად, კალკებად და შეცდომებად მიიჩნევს და შეძლებისდაგვარად ებრძვის. სტატიაში შევეცდებით გავაანალიზოთ, რა ენობრივი თუ საკომუნიკაციო აუცილებლობა განაპირობებს ამგვარი ანალიზური კონსტრუქციების გაჩენასა და გამრავლებას კონკრეტულად ქართულში და ზოგადად ენებში, როგორც ასეთში, და გამოვთქვამთ ჩვენს მოსაზრებასაც იმის თაობაზე, ენაში ამგვარი პროცესების მიმდინარეობა და საზოგადოდ ასეთი (ანალიზური) მიმართულებით ენის განვითარება მის გადაგვარებას და დეგრადაციას იწვევს, თუ პირიქით, უფრო ქმედითსა და გამომსახველს ხდის მას და აზრის მკაფიოდ და გასაგებად გამოთქმის აუცილებლობითაა განპირობებული.

საკვანძო სიტყვები: ანალიტიზმი; დიაქრონია; ლექსიკოგრაფია; NLP; კორპუსული ლინგვისტიკა

The Formation of Analyticity in Language. Degradation or Development? (On the example of certain trends in the modern Georgian language)

George Meladze
Ilia State University
giorgi.meladze.4@iliauni.edu.ge

Abstract

The issue of the development of languages and their transformation from the synthetic, morphologically rich and multiform type to the analytic one, which may appear morphologically "simplified", has always been extremely controversial in linguistic circles.

The representatives of the romantic trend in linguistics, such as Jacob and Wilhelm Grimms, Johann G. Herder and others (Rastorguyeva, 1983, pp. 20-21) held that the diachronic development of the Old Indo-European and Germanic languages and the process of their transformation into modern languages constituted in fact their decay and degradation. They saw the loss of the variety and richness of the grammatical forms, declension, and conjugation paradigms by these languages as a clear example of such degradation. If we compare the Anglo-Saxon (Old English) language with the modern English, or Latin with Italian, Old High German with the modern German, or Gothic (almost identical to Proto-Germanic) with, say, the modern Swedish language, we will clearly see what this approach of the Romanticists was based upon.

Otto Jespersen, a well-known Danish linguist on the other hand believed that the diachronic processes and changes occurring in the languages represent in fact their improvement and progress and that the said processes cause the constant increase in the efficiency of language. It is believed that his views on the subject were also influenced by Charles Darwin's evolutionary theories. Chapter XVII of Jespersen's masterpiece "Language: Its Nature, Development, and Origin" is titled "Progress or Decay?" (Jespersen, 1922, pp. 319-337). The title of our proposed paper was inspired by this very chapter of Jespersen's book and its general spirit.

The purpose of our paper itself is to discuss analytic constructions which have become quite abundant in the modern Georgian language, although the traditional Georgian linguistics either ignores them, or regards them as the violation of the set norms of the Georgian language, as *calques* or mistakes, trying hard to eradicate them. In our paper we will try to analyse, what linguistic or communicative circumstances necessitate the appearance

and spread of such analytic constructions in languages generally and in the Georgian language in particular; we will also express our opinion as to whether such analytic trends in diachronic linguistic change constitute ineed the decay and degradation of languages, or, on the contrary, such developmjent only make language more efficient and expressive and are caused by the necessity to clearly and understandably express one's thoughts and ideas verbally.

Keywords: analyticity; diachrony; lexicography; NLP; corpus linguistics

1 შესავალი

როგორც ცნობილია, ენების განვითარების ძირითადი ხაზია სინთეზური ტიპიდან ანალიზურ ტიპზე გადასვლა. მრავალი ენის თანამედროვე ვარიანტები უფრო მეტი ანალიტიზმით გამოირჩევა, ვიდრე ამავე ენების ძველი, 300, 400, 500 და ა.შ. წლის წინანდელი ვარიანტები, რომლებშიც, თავის მხრივ სინთეზური ფორმები ჭარბობს. ძველი ზემოგერმანული და და თანამედროვე გერმანული, ძველი ინგლისური (ანგლოსაქსური) და თანამედროვე ინგლისური, ანდა ლათინური და მისგან განვითარებული ნებისმიერი თანამედროვე რომანული ენა (იტალიური, ესპანური, ფრანგული ან რუმინული), ძველსლავური და თანამედროვე ბულგარული ამის კარგი მაგალითია.

დიაქრონიულ პროცესში სინთეზური, მორფოლოგიურად მდიდარი და მრავალფეროვანი ტიპიდან ანალიზურ, თითქოსდა მორფოლოგიურად „გამარტივებულ" ტიპზე ენათა გადასვლის შესახებ ენათმეცნიერებაში ოდითგანვე ურთიერთგამომრიცხავი შეხედულებები გამოითქმებოდა. რომანტიკული მიმდინარეობის წარმომადგენლები: იაკობ და ვილჰელმ გრიმები, იოჰან გ. ჰერდერი და სხვანი (Rastorguyeva, 1983, pp. 20-21), მიიჩნევდნენ, რომ ძველი ინდოევროპული და გერმანიკული ენების მიერ დროთა განმავლობაში განცდილი ცვლილებანი და თანამედროვე ენებად მათი ჩამოყალიბების პროცესი არსებითად მათს გადაგვარებას და დეგრადაციას წარმოადგენდა. ამგვარი დეგრადაციის მკაფიო გამოხატულებად მათ ამ ენების გრამატიკული ფორმების, ბრუნებისა და უღლების მრავალფეროვნებისა და სიმდიდრის დაკარგვა ესახებოდათ.

თუ (ზემომოყვანილ ენათა მაგალითების გარდა) ერთმანეთს შევადარებთ პროტოინდოევროპულთან ძალზე ახლოს მყოფ სანსკრიტს, პრუსიულსა და ლიტვურს, პროტოგერმანიკულთან ძალზე ახლოს მყოფ გოთურს, ერთი მხრივ და, ვთქვათ, თანამედროვე მაკედონიურ, ნორვეგიულ, შვედურ ენებს და თუნდაც თანამედროვე ჰინდი ენას, მეორე მხრივ, თვალნათლივ დავინახავთ, რას ეფუძნებოდა ლინგვისტიკის „რომანტიკოსთა" ამგვარი მიდგომა.

მეორე მხრივ, ცნობილი დანიელი ენათმეცნიერი ოტო იესპერსენი ("Otto Jespersen", 2022; McElvenny, 2013; Rastorguyeva, 1983, p. 292) მიიჩნევდა, რომ ენაში მიმდინარე დიაქრონიული პროცესები და ცვლილებები არსებითად მის დახვეწაში და პროგრესში მდგომარეობს და რომ ეს პროცესები ენის ეფექტიანობის უცილობელ ზრდას იწვევს. ითვლება, რომ მის ამგვარ შეხედულებებზე გარკვეული ზეგავლენა ჩარლზ დარვინისეულმა ევოლუციის თეორიამაც მოახდინა. იესპერსენის ფუნდამენტური ნაშრომის: „ენა, მისი ბუნება, განვითარება და წარმოშობა" XVII თავს ასევე ეწო-

დება: „პროგრესი თუ გადაგვარება?“ (Progress or Decay?) (Jespersen, 1922, pp. 319-337). სწორედ ამ თავმა და მისმა საერთო სულისკვეთებამ შთაგვაგონა, წინამდებარე ნაშრომი ამგვარად დაგვესათაურებინა.

მართლაც, თუ ერთმანეთს შევადარებთ გოთურსა და ანგლოსაქსურს, სადაც ზმნას ორი გრამატიკული დრო, აწმყო და ნამყო ჰქონდა ერთი მხრივ, და თანამედროვე ინგლისურს, სადაც ანალიზური ფორმების განვითარების წყალობით დაახლოებით 16 გრამატიკული დრო (tenses) დასტურდება, გაგვიჭირდება ინგლისური ზმნის სისტემის მიერ განცდილ ცვლილებებს რეგრესი და დეგრადაცია ვუწოდოთ. თანამედროვე ინგლისურში ზმნის უღლება მხოლობითი და მრავლობითი რიცხვის პირველი, მეორე და მესამე პირის მიხედვით მნიშვნელოვნადაა რედუცირებული, სამაგიეროდ მოქმედების დასრულებულობისა და განგრძობითობის სხვადასხვა უმცირესი ნიუანსების გამოხატვის თვალსაზრისით თანამედროვე ინგლისური ზმნა თავის ანგლოსაქსურ „წინაპართან“ შედარებით გაცილებით უფრო ეფექტიანი და გამომსახველია. ვერც იმას ვიტყოდით, რომ არსებითებისა და ნაცვალსახელების ბრუნების რედუცირებას და წინადადების წევრთა შორის ურთიერთმიმართებათა „ანალიზურად“ გამოხატვას წინდებულებისა და წინადადებაში სიტყვათა თანამიმდევრობის მეშვეობით ინგლისური ენის გამომსახველობა რაიმენაირად დაეკნინებინოს ან შეემცირებინოს.

ამდენად, პირადად ჩვენც, პრაქტიკულად, ოტო იესპერსენის მოსაზრებისკენ ვიხრებით და მიგვაჩნია, რომ დროთა განმავლობაში ენა იხვეწება და უფრო გამომსახველი ხდება. ენაში ამგვარი პროცესების მიმდინარეობა და ანალიზური მიმართულებით მისი განვითარება, ჩვენი აზრით, ენის გადაგვარებას და დეგრადაციას კი არ იწვევს, პირიქით, მას უფრო ქმედითსა და გამომსახველს ხდის და, როგორც ასეთი, აზრის მკაფიოდ და გასაგებად გამოთქმის აუცილებლობითაა განპირობებული. პრინციპში, მხოლოდ ლოგიკურია, რომ ეს ასეც უნდა იყოს. წინააღმდეგ შემთხვევაში ძნელი წარმოსადგენია, რატომ უნდა იყოს ენა, ნიშანთა სისტემა, რომელსაც ადამიანები საურთიერთოდ და ინფორმაციის გასაცვლელად იყენებენ, ასე ვთქვათ, დეგრადაციისა და ენტროპიისაკენ მიდრეკილი.

ვაღიარებთ, რომ ჩვენ პირადად ლათინური, გოთური, ანგლოსაქსური და ძველსკანდინავიური ლინგვისტური თვალსაზრისით გაცილებით უფრო მრავალფეროვნად, საინტერესოდ და მიმზიდველად გვეჩვენება, ვიდრე თანამედროვე ინგლისური, იტალიური, ნორვეგიული ან შვედური, მაგრამ ენის განვითარება, როგორც ჩანს, პრაგმატული მოსაზრებებითაა განპირობებული, და არა ლინგვისტური მრავალფეროვნებისა და გრამატიკული ეგზოტიკის მაძიებელ მკვლევართა სურვილებით.

2 ანალიტიზმის გაჩენა, როგორც ენის განვითარების ისტორიული კანონზომიერება

როგორიც არ უნდა იყოს ჩვენი დამოკიდებულება ენობრივი ანალიტიზმის მიმართ, ძნელია იმ ფაქტის უგულებელყოფა, რომ ენათა განვითარება სინთეზური ტიპიდან ანალიზურ ტიპზე თანდათანობითი გადასვლის გზით ხდება. ამაში უბრალოდ ენათა განვითარებაზე დაკვირვება და ენების დიაქრონიულ ჭრილში განხილვა გვარწმუნებს.

საინტერესოა, რომ დაკვირვებულმა თვალმა ანალიზური კონსტრუქციები ისეთ ენებშიც შეიძლება აღმოაჩინოს, რომლებიც სამართლიანად ითვლება საკმაოდ არქაულად და სინთეზური აგებულებით გამორჩეულად.

2.1 განგრძობითი დროის „ჩანასახები“ გოთურსა და ანგლოსაქსურში

ვისაც ვულფილას გოთური სახარების რამდენიმე თავი მაინც წაუკითხავს, ალბათ შემჩნეული ექნება შემდეგი ტიპის კონსტრუქციები: jah þan was miþ im Paitrus standands jah warmjands sik (იოან. XVIII.18); iþ Seimon Paitrus was standands jah warmjands sik (იოან. XVIII, 25); jah was sitands miþ andbahtam jah warmjands sik at liuhada (მარკ. XIV.54) (*Wulfila Project*, 1997). ეს ფრაგმენტები სახარების იმ ცნობილი ეპიზოდიდანაა, როდესაც იესოს დაკავების შემდეგ პეტრე მოციქული ცეცხლთან თბება და მას მღვდელმთავრის ერთ-ერთი მსახური იცნობს.

აქ **standands**, **sitands** და **warmjands** გოთური ზმნების **standan** („დგომა“), **sitan** („ჯდომა“) და **warmjan** („გათბობა“) აწმყო დროის მიმღეობებია და ეს კონსტრუქციები თანამედროვე ინგლისურად სიტყვა-სიტყვით რომ ვთარგმნოთ, მივიღებთ: **was standing/sitting and warming himself** (!). ამ გრამატიკულ დროს თანამედროვე ინგლისურში ნამყო განგრძობითი დრო (past continuous tense) ეწოდება.

შეიძლება აქ სახარების ბერძნული ორიგინალის გავლენასთან გვქონდეს საქმე: ბერძნულ ტექსტშიც (რომელიც დაახლ. ჩვ. წ. IV საუკუნეში შესრულებული გოთური თარგმანის წყაროდ ითვლება) და ლათინურ ვულგატაშიც (რომელსაც ვულფილა, როგორც ჩანს, არ იყენებდა), იოანეს სახარების შესაბამის ადგილებში აწმყო მიმღეობებია ნახმარი: შესაბამისად ἦν δὲ καὶ ὁ πέτρος μετ› αὐτῶν ἑστὼς καὶ θερμαινόμενος; erat autem cum eis et Petrus stans, et calefaciens se (*იოან*. XVIII.18) და ἦν δὲ σίμων πέτρος ἑστὼς καὶ θερμαινόμενος; erat autem Simon Petrus stans, et calefaciens se (*იოან*. XVIII.25), სადაც ἑστὼς, θερμαινόμενος და stans, calefaciens (*Wulfila Project*, 1997). აწმყო დროის მიმღეობებია. ოდნავ სხვანაირადაა საქმე მარკოზის სახარების ციტირებულ ფრაგმენტში (*მარკ*. XIV.54), სადაც ბერძნულ ტექსტში აწმყო მიმღეობებია ნახმარი: καὶ ἦν **συνκαθήμενος** μετὰ τῶν ὑπηρετῶν καὶ **θερμαινόμενος** πρὸς τὸ φῶς; ხოლო ლათინურში – ზმნის პირიანი ფორმები: et **sedebat** cum ministris ad ignem, et **calefaciebat** se *(ხაზგასმა ყველგან ჩვენია)* (*Wulfila Project*, 1997; *STEPBible*, 2022).

გავიმეორებთ, შეიძლება აქ ბერძნული ორიგინალის გავლენასთან გვქონდეს საქმე, მაგრამ საკმაოდ არქაულ გოთურ ენაში თანამედროვე ინგლისური past continuous-ის, ნამყო განგრძობითი დროის, ზუსტი ანალოგის ხილვა საყურადღებო და რამდენადმე მოულოდნელია.

მსგავსი მაგალითის მოყვანა ანგლოსაქსურიდანაც შეიძლება, თანაც ტექსტში, რომელიც თავად ანგლოსაქსური პერიოდის ინგლისშია შექმნილი დაახლოებით ჩვ. წ. VIII საუკუნეში. ამბავში პოეტ **კედმონის** (Cædmon) შესახებ, რომელიც ანგლოსაქსი თეოლოგისა და მემატიანის, **ღირსი ბედას** *(ძვ. ინგლ*. Bæda, *ინგლ*. Bede the Venerable) მიერ დაწერილ „ინგლისელი ხალხის საეკლესიო ისტორიაში“ (Bede the Venerable, 1907) შედის, ვხვდებით ასეთ პასაჟს: “he forlet þæt hus þæs ȝebeorscipes ond ut wæs ȝonȝende to neata scipene, þara heord him wæs þære neahte beboden” („და-

ტოვა სახლი, სადაც ნადიმი იმართებოდა და გაემართა ბოსლისკენ, სადაც მას იმ ღამით საქონლის ჯოგისთვის უნდა მიეხედა“). **wæs ȝonȝende** აქ ზუსტად შეესატყვისება თანამედროვე ინგლისურს **was going** (ზმნის to go „სიარული, სვლა“ ნამყო განგრძობითი დროის მესამე პირის მხოლობითი რიცხვის ფორმა).

იმავე ტექსტში, რამდენიმე სტრიქონით ქვემოთ ვკითხულობთ: eft he cwæð, se ðe wið hine sprecende wæs („შემდეგ თქვა მან, ვინც მას [*კედმონს*] ელაპარაკებოდა“). აქაც **wæs sprecende** ზუსტად შეესატყვისება თანამედროვე ინგლისურს **was speaking** (ზმნის to speak „ლაპარაკი“ ნამყო განგრძობითი დროის მესამე პირის მხოლობითი რიცხვის ფორმა).

ამ მაგალითებზე დაყრდნობით შეგვიძლია დავასკვნათ, რომ ანალიზური კონსტრუქციები (განგრძობითი დროის ფორმები ამ კონკრეტულ შემთხვევებში) საკმაოდ ადრეულ პერიოდში (შეგახსენებთ, რომ *დაახლ*. IV და VIII საუკუნის, შესაბამისად, გოთურ და ანგლოსაქსურ ძეგლებთან გვაქვს საქმე) შეიძლება გამოჩნდეს სინთეზურობით გამორჩეულ ენებშიც კი.

3 ანალიზური კონსტრუქციები თანამედროვე ქართულში. ვებრძოლოთ თუ ვაღიაროთ?

თავად ჩვენი ნაშრომის თემა, უნდა ითქვას, არა იმდენად დიაქრონიული ლინგვისტიკის საკითხების კვლევაა, ან იმის ანალიზი, ენებში ანალიტიზმის გაჩენა განვითარება და ენის სრულყოფაა, თუ მისი გადაგვარება და დეგრადაცია (აქ ჩვენთვის, როგორც ზემოთ მოკლედ აღვნიშნეთ, ყველაფერი ნათელია), არამედ უფრო იმის გარკვევა, თუ როგორი უნდა იყოს ჩვენი დამოკიდებულება ამავე ანალიზური ტიპის კონსტრუქციების მიმართ, რომლებიც (გოთურის, ანგლოსაქსურის, ინგლისურის, იტალიურისა და შვედურის დარად) სულ უფრო ხშირად ჩვენს მშობლიურ ქართულ ენაშიც ჩნდება (ეს ჩვენ, გასაგები მიზეზების გამო, უპირატესად ლექსიკოგრაფიული თვალსაზრისით გვაინტერესებს, თუმცა ამაზე უფრო დაწვრილებით ქვემოთ, წინამდებარე ნაშრომის დასკვნით ნაწილში ვისაუბრებთ).

საქმე ისაა, რომ ანალიზური კონსტრუქციები თანამედროვე ქართულში საკმაოდაა მომრავლებული, თუმცა ტრადიციული ქართული ენათმეცნიერება მათ ან ვერ ამჩნევს, ან მათ ქართული ენის ნორმების დარღვევად, კალკებად და შეცდომებად მიიჩნევს და შეძლებისდაგვარად ებრძვის. ყოველ შემთხვევაში, ამგვარი კონსტრუქციების სწორი მეცნიერული ანალიზი, ჩვენი აზრით, არ ხდება (განცხადებები ტიპისა ***„ქართულ ენას ასეთი კონსტრუქციები არ სჭირდება!“*** მეცნიერულ ანალიზად და ინტერპრეტაციად ვერ ჩაითვლება).

ქვემოთ შევეცდებით მოკლედ გავაანალიზოთ, რა ენობრივი თუ საკომუნიკაციო აუცილებლობითაა განპირობებული ამგვარ ანალიზურ კონსტრუქციათა გაჩენა და გამრავლება კონკრეტულად ქართულში და ვეცდებით ისიც განვსაზღვროთ, რა პოტენციური რისკებისა და უარყოფითი შედეგების შემცველია ქართულ ენაში მორფოლოგიური თუ ლექსიკური სიახლეების იგნორირება და/ან მათთვის ბრძოლის გამოცხადება „ქართული ენის სიწმინდის დაცვის“ ლოზუნგით.

უნდა ითქვას, რომ ყოველთვის ვიყავით დაინტერესებული დიაქრონიით და კონკრეტულად იმით, თუ როგორ და რატომ ხდება ენებში ისეთი პატარ-პატარა ცვლილებები, რომლებიც თანდათან გროვდება და გარკვეული ხნის შემდეგ იმგვარ განსხვავებებს იწვევს, როგორიცაა განსხვავება ანგლოსაქსურსა და თანამედროვე ინგლისურს, ანდა ლათინურსა და იტალიურს შორის.

ამდენად, თუკი ქართულ ენაში რაიმე საინტერესო სტრუქტურულ სიახლეს, რაიმე ახალ ტენდენციაზე მიმანიშნებელ თუნდაც მცირე, მაგრამ, ჩვენი აზრით, არსებით და ენობრივად საყურადღებო ინოვაციას შევამჩნევდით, ამას, ყოველთვის ჯეროვნად აღვნიშნავდით და ხშირად ჩავიწერდით კიდევაც.

ამიტომ არ გაგვჭირვებია არგუმენტირებულად შეპაექრება, როდესაც რამდენიმე წლის წინ ჩვენი „ლექსიკოგრაფიული ცენტრის“ ერთ-ერთ ფეისბუქ-გვერდზე ენობრივი საკითხებისადმი მიძღვნილი პოსტის ერთ-ერთ ფრაზას: ***«გასულ კვირას, ჩვენი თანამშრომლების გასვლით სამუშაო შეხვედრაზე წლიური ანგარიშის განხილვისას, რამდენიმე საგულისხმო საკითხი გამოიკვეთა და განცხადება გაკეთდა ახალი სამოქმედო გეგმის შესახებ»*** შემდეგი კომენტარები მოჰყვა, როგორც ჩანს, კონსერვატიული ლინგვისტურ-სტილისტური შეხედულებების მქონე ინტერნეტ-მომხმარებელთაგან: „განცხადება გაკეთდა-სტილისტური ხარვეზი“ და სწორი ვერსიაა «განაცხადეს» (სტილი დაცულია, გ.მ.).

ამ შენიშვნებს მაშინ ჩვენი **დიდი ინგლისურ-ქართული ონლაინლექსიკონის** ბლოგზე გავეცით პასუხი. რამდენადაც ეს ბლოგი მოცემული მომენტისთვის სხვადასხვა მიზეზების გამო არ ფუნქციონირებს, გვინდა აქ მოკლედ გავიმეოროთ ის არგუმენტები, რომლებიც მაშინ საკმაოდ გულდასმით ჩამოვაყალიბეთ, და, გარდა ამისა, დავუმატოთ ზოგიერთი მოსაზრება იმასთან დაკავშირებით, თუ, ერთი მხრივ, რატომ ჩნდება ქართულში ანალიზური კონსტრუქციები ტიპისა „განცხადება გაკეთდა“, „ცვლილება მოახდინა“, „გაკეთებული მაქვს“ და *ა.შ.*, და, მეორე მხრივ, იმის თაობაზე, თუ რატომ არ ღირს ქართულ ენაში გაჩენილი ყოველგვარი სიახლისა და ინოვაციის წინააღმდეგ გალაშქრება.

ამდენად, მოდით კიდევ ერთხელ განვიხილოთ ვნებითი გვარის ანალიზური კონსტრუქცია „განცხადება გაკეთდა“, რომელიც მაშინ კრიტიკის საგანი გახდა. არა იმიტომ, რომ რამდენიმე წლის წინანდელ ხანმოკლე პოლემიკას დავუბრუნდეთ, არამედ იმის გამო, რომ აღნიშნული ტიპის კონსტრუქციები და ენაში მათი გაჩენის ლოგიკა წინამდებარე სტატიის ერთ-ერთი ძირითადი თემაა.

„განცხადება გაკეთდა“ ზმნის ვნებითი გვარის ფორმაა. ვნებითი გვარი ანუ პასივი იმ დროს გამოიყენება, როდესაც წინადადებაში აზრობრივად მთავარი და წინა პლანზე წამოსაწევი არის მოქმედების ობიექტი (ამ შემთხვევაში „განცხადება“). მოქმედების ობიექტის განსაკუთრებულად გამოყოფა იმით მიიღწევა, რომ პასივში მდგომი წინადადების ქვემდებარედ მოქმედების ობიექტია ხოლმე გამოტანილი, თავად ზმნა კი ვნებით გვარში იხმარება („განცხადება გაკეთდა“). ვნებით გვარს საზოგადოდ მაშინ იყენებენ, როდესაც მოქმედების რეალური შემსრულებელი (ჩვენს შემთხვევაში განცხადების გამკეთებელი) ან ცნობილი არ არის, ან რაიმე მოსაზრებებიდან გამომდინარე, მისი დასახელება ან არ სურთ, ან საჭიროდ არ მიაჩნიათ. მაგალითად, უკანასკნელი ცნობების გამოშვებაში ხშირად შეიძლება გავიგონოთ დაახლოებით შემდეგნაირი ფრაზა: **„ამ თემაზე ცოტა ხნის წინ სპეციალური გან-**

ცხადება გაკეთდა საგარეო საქმეთა სამინისტროში“. მთავარი აქ თავად განცხადება, მისი შინაარსი და საგარეო საქმეთა სამინისტროა. ეს განცხადება კი, როგორც ასეთი, შეიძლება ვინმე ნაკლებად მნიშვნელოვან პრეს-სპიკერს ან სამინისტროს რიგით თანამშრომელს წაეკითხა ჟურნალისტებისთვის და ტელერეპორტიორებისთვის და მისი ვინაობა შესაძლოა არც ჟურნალისტებისთვის არ ყოფილიყო ცნობილი, და არც პრეს-კონფერენციაზე არ დასახელებულიყო.

ჩვენი მაშინდელი ფეისბუქ-გვერდის კომენტირებულ ფრაზაშიც ასეთივე შემთხვევასთან გვქონდა საქმე: ფრაზაში **«გასულ კვირას, ჩვენი თანამშრომლების გასვლით სამუშაო შეხვედრაზე წლიური ანგარიშის განხილვისას, რამდენიმე საგულისხმო საკითხი გამოიკვეთა და განცხადება გაკეთდა ახალი სამოქმედო გეგმის შესახებ»** ასევე არ არის დასახელებული მოქმედების კონკრეტული შემსრულებელი, ანუ აღნიშნული არ არის, ვინ იყო განცხადების ავტორი. აზრობრივად მნიშვნელოვნად მიჩნეულია და წინა პლანზეა გამოტანილი **„გამოკვეთილი საგულისხმო საკითხები“** (გამოკვეთილი, სავარაუდოდ, კოლექტიური განხილვისა და მსჯელობის პროცესში) და **„განცხადებები“**, ხოლო ფაქტობრივი მოქმედი პირები არც ცნობილი არ ჩანს, და არც მათი დასახელება არ არის საჭიროდ მიჩნეული. ამ ტიპის იმპერსონალურ წინადადებას მოქმედებით გვარში ვერაფრით ვერ გადავიყვანთ. (განსაკუთრებით ძნელი წარმოსადგენია, როგორ შეიძლება მოქმედებით გვარში გადმოიცეს აზრი ფრაზისა

„ ...რამდენიმე საგულისხმო საკითხი გამოიკვეთა“).

მოკლედ აღვნიშნავდით, რომ მოქმედებით გვარშიც „განცხადება გააკეთა“ ან, ვთქვათ, „ცვლილება განახორციელა“ ყოველთვის ვერ ჩანაცვლდება „მარტივი“ ზმნით ტიპისა „განაცხადა“ ან „შეცვალა“. მაგალითად, წინადადება **„ბიზნესმენმა მოულოდნელი, სენსაციური და სკანდალური განცხადება გააკეთა“** ვერ გარდაიქმნება იმგვარად, რომ „განცხადების გაკეთების“ ნაცვლად „განაცხადა“ ვთქვათ. თანაც „განაცხადა“ ყოველთვის მოითხოვს ობიექტს, დამატებას, ანუ იმის დაკონკრეტებას, თუ რა განაცხადა მავანმა, მაშინ როდესაც „განცხადება გააკეთა“ თვითკმარი, აზრობრივად დასრულებული ფრაზაა და რაიმე დამატებით ინფორმაციას არ საჭიროებს. ასევე, წინადადება **„მინისტრმა თავისი ქვეყნისთვის მნიშვნელოვანი და აუცილებელი არაერთი რადიკალური ცვლილება განახორციელა“** ვერაფრით ვერ შეიცვლება წინადადებით, რომელშიც შემასმენელი „შეცვალა“ იქნება (**„მნიშვნელოვნად და აუცილებლად რადიკალურად შეცვალა ბევრი რამ“** კომიკურად ჟღერს, იმედია ამგვარი „ტრადიციული“ და „სტილისტურად უხარვეზო“ კონსტრუქციების აგებას არავინ ეცდება). ზოგადად, ენებში (აქ მარტოოდენ ქართულს არ ვგულისხმობთ) ანალიზური კონსტრუქციების გაჩენას სწორედ ამგვარი რთული აზრობრივი ნიუანსების გადმოცემის აუცილებლობა იწვევს ხოლმე, სხვათა შორის.

ამ მაგალითებიდან კარგად ჩანს, რომ ანალიზური კონსტრუქცია „ცვლილება განახორციელა“ და *მისთ*. სინტაქსურად უფრო „ღია“ და მოქნილია და ატრიბუტების შეუზღუდავად დამატების საშუალებას იძლევა, რაც შეიძლება გარკვეულ კონტექსტებში აზრობრივად აბსოლუტურად აუცილებელი იყოს, მაშინ როდესაც „შეცვალა“ და „განაცხადა“ ტიპის მარტივი შემასმენლების შემცველ წინადადებებში ამგვარ ატრიბუტებს ვერ ჩავრთავთ. სიტუაციებიც, რომლებშიც ამგვარი ანალიზური კონსტრუქციების ხმარება აუცილებელია, საკმაოდ სტანდარტული და

ხშირია, მათ გვერდს ვერ ავუვლით და თუ რაიმე მეტ-ნაკლებად რთულ და საინტერესო თემაზე აზრის მკაფიოდ და ადვილად გასაგებად გამოთქმა გვსურს, ასეთი კონსტრუქციები აუცილებლად დაგვჭირდება.

ამდენად, ამ ტიპის ანალიზური კონსტრუქციები კალკები და სტილისტური ხარვეზები კი არ არის, არამედ ენობრივი, საკომუნიკაციო აუცილებლობის შედეგად წარმოქმნილი შესიტყვებები, რომლებიც აზრის მკაფიოდ და გასაგებად გამოხატვის აუცილებლობითაა ნაკარნახები. მათი გამოყენება ზემოთ აღწერილ სიტუაციებში აუცილებელი და უალტერნატივოა და ენაში მათს გამოჩენასა და ხმარებასაც თავისი ლოგიკა აქვს.

კიდევაც რომ დავუშვათ, რომ ასი ან ორასი წლის წინ ქართულში ნაკლებად ხშირი იყო იმპერსონალური წინადადებებისა და ვნებითი გვარის ხმარება, ეს სულაც არ ნიშნავს იმას, რომ დღეს მათი გამოყენება რაიმე სახის „სტილისტურ ხარვეზად“ ან შეცდომად შეიძლება ჩაითვალოს. სულხან-საბა ორბელიანის, დავით გურამიშვილისა და გრიგოლ ორბელიანის დროიდან საუკუნეები გავიდა, მსოფლიოც შეიცვალა და ქართული ენაც აღარაა ისეთი, როგორიც მათ დროს იყო.

ჩვენ მიერ აღწერილის მსგავსმა სიტუაციებმა გამოიწვია სხვა ენებშიც ვნებითი გვარის, ფრაზული ზმნების და მრავალი სხვა ტიპის ანალიზური კონსტრუქციების გამოჩენა და დამკვიდრება. ასეთი კონსტრუქციები თავის დროზე არც გოთურში, ძველსკანდინავიურში, ძველ ზემოგერმანულში და ანგლოსაქსურში არ იყო, მაგრამ უფრო გვიანდელი ფორმაციის ენებში ისინი დროთა განმავლობაში სწორედ იმიტომ გაჩნდა, რომ ეს ენობრივმა აუცილებლობამ მოითხოვა, და არა იმიტომ, რომ რომელიმე ეს ენა დეგრადირდა და გადაგვარდა. ანალოგიურ სიტუაციასთან გვაქვს საქმე ქართულ ენასთან მიმართებითაც. გულუბრყვილო იქნებოდა ვინმეს ეფიქრა, რომ ენათა განვითარების ზოგადი კანონები თუ კანონზომიერებანი ქართულ ენაზე შეიძლება არ ვრცელდებოდეს.

ცალკე, ძალიან მოკლედ, გვინდა შევეხოთ „გაკეთებული მაქვს“, „ნამყოფი ვარ“, „უკვე დამთავრებული მქონდა“ ტიპის რთულ ანალიზურ კონსტრუქციებს, რომელთაც ტრადიციული ქართული ნორმატიული ენათმეცნიერება აგრეთვე არა სწყალობს. არადა ისინი ერთი მხრივ მყარად და საფუძვლიანადაა დამკვიდრებული თანამედროვე ქართულ ენაში, მეორე მხრივ კი მათ გარეშე ხშირად აზრის გასაგებად ჩამოყალიბება შეუძლებელი იქნება. საუბარი გვაქვს სიტუაციებზე, როდესაც საჭიროა ხაზის გასმა იმისათვის, რომ ერთი მოქმედება ან სიტუაცია მეორე მოქმედებაზე ან სიტუაციაზე უფრო ადრე შესრულდა და დასრულდა, ან მომავალში შესრულდება და დასრულდება სხვა მოქმედებაზე ადრე. „მე როდესაც მოვედი, შენ უკვე წასული იყავი“, „შენ რომ დაბრუნდები, მე სტატიის წერა უკვე დამთავრებული მექნება“, „საზღვარგარეთ ბევრჯერ ვარ ნამყოფი“. სხვა ენებში არსებული პერფექტის / სრული დროის ტიპის ამ კონსტრუქციებს სინთეზური, „არაანალიზური“ კონსტრუქციებით ადვილად ვერ შევცვლით.

პრინციპში, შეიძლება ვთქვათ „საზღვარგარეთ ბევრჯერ ვყოფილვარ“ მაგრამ თურმეობითს „თურმეს“ კონოტაცია აქვს და, რეალურად, გაუგებრობის თავიდან ასაცილებლად უპირატესობას, როგორც წესი, „ნამყოფი ვარ“ ტიპის კონსტრუქციებს ანიჭებენ ხოლმე. ისიც უნდა აღინიშნოს, რომ მსგავს კონტექსტებში თურმეობითი რამდენადმე არქაულად და მოძველებულად ჟღერს.

ვერ ვიტყვით, „მე როდესაც მოვედი, შენ უკვე წახვედი“, „შენ რომ დაბრუნდები, მე სტატიის წერას უკვე დავამთავრებ“ – ასეთი ქართულით ძირითადად ენის ცუდად მცოდნეები (ვისთვისაც ქართული მშობლიური ენა არაა) უფრო მეტყველებენ.

„გაკეთებული მაქვს“, „ნამყოფი ვარ“, „დაწერილი მექნება“ და მისთ., პრაქტიკულად ქართული პერფექტია და იგი ქართულში ისევე გაჩნდა, როგორც თავის დროზე ინგლისურში, გერმანულში და იტალიურში. ამ ანალიზური დროითი ფორმის „შერისხვა“ და „კანონგარეშედ გამოცხადება“ მიამიტობა და რეალობისაგან მოწყდომა იქნება.

4 თანამედროვე ქართული ენა ლექსიკოგრაფიისა და ბუნებრივი ენის დამუშავების (NLP) მეთოდების თვალსაზრისით

როგორც ზემოთაც აღვნიშნეთ, ენის (კონკრეტულად ქართულის) დიაქრონიული განვითარება და მასში მომხდარი ცვლილებები, ჩვენი საქმიანობის ხასიათიდან გამომდინარე, გარდა წმინდა ენათმეცნიერულისა, აგრეთვე (და უპირატესად) ლექსიკოგრაფიული და ლექსიკოლოგიური (და არა მარტო) თვალსაზრისითაც გვაინტერესებს.

ვერავის გავაკვირვებთ თუ ვიტყვით, რომ კომპიუტერული და ციფრული რევოლუციის ეპოქაში ვცხოვრობთ. ამას ყველანი აშკარად ვხედავთ და ვგრძნობთ. არსებულ ვითარებაში ნებისმიერი ენისათვის, თუკი მისი სრულფასოვნად განვითარებისა და თვით სამომავლო არსებობის უზრუნველყოფა გვსურს, სასიცოცხლოდ აუცილებელია ციფრულ სამყაროში ადგილის დამკვიდრება, მისი სრული თუ არა, შეძლებისდაგვარად მაქსიმალური გაციფრება მაინც.

ეს აგრეთვე გულისხმობს სრულფასოვანი ორენოვანი ქართულ-უცხოური და უცხოურ-ქართული ლექსიკონების შექმნას, რომლებშიც ქართულ ენაში დღეისათვის რეალურად არსებული სიტუაცია, დღევანდელი ქართულის რეალური ლექსიკური მარაგი და მორფოლოგია იქნება ასახული. ეს განსაკუთრებით ქართულ-უცხოენოვანი ლექსიკონებისთვისაა მნიშვნელოვანი, ვინაიდან თუ ამა თუ იმ ქართულ-გერმანულ, ქართულ-რუსულ ან ქართულ-ფრანგულ ლექსიკონში ესა თუ ის აქტუალური და ხშირად ხმარებული ქართული სიტყვა შეტანილი არ იქნა, ეს ლექსიკონის ხარისხსაც დააქვეითებს და იმ უცხოელსაც გაურთულებს საქმეს, ვინც ამ ლექსიკონის დახმარებით ქართული ენის შესწავლას ან ქართული ტექსტის გაგებას შეეცდება.

ორენოვან ლექსიკონებთან მჭიდროდაა დაკავშირებული ბუნებრივი ენის დამუშავებისა (natural language processing, NLP) და ავტომატური თარგმნის პროგრამული ინსტრუმენტების შემუშავების თემა. ამ დარგებს კი, თავის მხრივ, დიდი მნიშვნელობა აქვთ ენის დიგიტალიზაციის და ციფრულ განზომილებაში დამკვიდრების საქმეში.

ყველა ეს დარგი, თანამედროვე ციფრულ ეპოქაში, დაკავშირებულია ვრცელი ტექსტური კორპუსების შექმნასა და განვითარებასთან. ჩვენს შემთხვევაში უაღრესად მნიშვნელოვანია ისეთი ქართული ტექსტური კორპუსის ან კორპუსების არსებობა, რომლებიც ოპტიმალური იქნება თავისი **რეპრეზენტაციულობის** თვალსაზრისით. თუ ქართულ ენაში ხშირად ხმარებული და აქტუალური ესა-თუ-ის სიტყვა,

გამოთქმა ან კონსტრუქცია „კალკად“, „სტილისტურ ხარვეზად“ იქნება მიჩნეული გასული საუკუნის დასაწყისის დროინდელი, ან სულაც 2-3 საუკუნის წინანდელი ენობრივი ნორმებიდან გამომდინარე და შედეგად ამისა ამგვარი სიტყვებისა და გამოთქმების შემცველი ფრაზები და ტექსტები ქართულ კორპუსებში ვერ მოხვდება, ეს ამ კორპუსების რეპრეზენტაციულობას მნიშვნელოვნად დააქვეითებს. ტექსტური კორპუსების არარეპრეზენტაციულობა კი არსებითად დაამახინჯებს ქართული ენის რეალურად არსებულ სურათს და გამოუსადეგრად და ნაკლებად სასარგებლოდ გახდის თავად ამ კორპუსებსაც და იმ ლექსიკონებსა თუ ავტომატური თარგმნის პროგრამებსაც, რომლებიც ჩვენს ციფრულ ეპოქაში ძირითადად ერთენოვან ან პარალელურ ტექსტურ კორპუსებს ეფუძნება.

ამდენად, აუცილებელია არა მარტო ზემოთ აღწერილი ანალიზური კონსტრუქციების „ლეგალიზება“ (ამაში უპირატესად რეალობისთვის თვალის გასწორება და რეალური ლინგვისტური ფაქტის არსებობის აღიარება იგულისხმება), არამედ აგრეთვე რეალური, ცოცხალი თანამედროვე ქართული ენის ლექსიკური მარაგის ადეკვატურად დაფიქსირება.

საინტერესო იქნებოდა შეგვეფასებინა, თუ რა სიტუაციაა ქართულ ფილოლოგიაში ამ თვალსაზრისით. როგორც ლექსიკოგრაფი, ცალკე გამოვყოფთ ქართული განმარტებითი და ქართულ-უცხოენოვანი ლექსიკონების თემას.

თითქმის ოცდაათი წელია, ლექსიკოგრაფიის დარგში ვმუშაობთ, ამდენად კოლეგების მიერ შექმნილი ლექსიკოგრაფიული პროდუქტების ოდნავი კრიტიკაც კი ჩვენთვის საკმაოდ საჩოთირო და უხერხულია. მიუხედავად ამისა, ქართული ენის ინტერესებიდან გამომდინარე, სათქმელი პირდაპირ უნდა ითქვას (sed magis amica veritas, ასე ვთქვათ).

მაგალითად, ბოლო წლებში საკმაოდ აქტუალური სიტყვა „გასაჯაროება“ (არც პირიანი, არც უპირო ფორმებით) არ არის ასახული არც დონალდ რეიფილდის დიდ ქართულ-ინგლისურ ლექსიკონში, არც „ქართულ ლექსიკონში“ (*ქართული ლექსიკონი*, n.d.), არც ქართულ-ოსურ და ოსურ-ქართულ ლექსიკონში (*ქართულ-ოსური და ოსურ-ქართული ლექსიკონი*, n.d.), არც ქართულ-თურქულ და თურქულ-ქართულ ლექსიკონში (*ქართულ-თურქული და თურქულ-ქართული ლექსიკონი*, n.d.), არც ქართულ-აზერბაიჯანულში (*ქართულ-აზერბაიჯანული ლექსიკონი*, n.d.) და არც ბევრ სხვა ლექსიკონში, რომელთა ჩამოთვლასაც აქ აღარ მოვყვებით.

სავარაუდოდ, საქმე ისაა, რომ ყველა ეს ლექსიკონი, როგორც ჩანს, ამა თუ იმ ფორმით, თავისი სიტყვანის განსასაზღვრავად ქართული ენის განმარტებით ლექსიკონს (ქ.ე.გ.ლ.) იყენებს. ამგვარი დასკვნის საფუძველს ამ ლექსიკონების სტრუქტურისა და მასალის მოწოდების ხასიათის არაერთი თავისებურება იძლევა.

მართლაც, თავად ქ.ე.გ.ლ.-ში ვერ ვნახავთ სიტყვებს **„ასაჯაროებს“**, **„გაასაჯაროებს“**, **„გასაჯაროება“**, **„გასაჯაროვდება“** და ა.შ., მიუხედავად იმისა, რომ ინტერნეტში ქ.ე.გ.ლ.-ის მინიმუმ ორი ონლაინვერსია დევს (რომელთა უახლესი სიტყვებით შევსებაც ძნელი არ უნდა იყოს), **ა-ბ** და **გ** ასოები კი რედაქტირებული და განახლებული ფორმით წიგნებადაცაა დაბეჭდილი.

ქ.ე.გ.ლ.-ში სათანადოდ არ არის წარმოდგენილი სიტყვა „გადამისამართება“, „გადაამისამართებს“ და *ა.შ.*, რომელიც ჩვეულებრივი და საკმაოდ ხშირად ხმარებული

სიტყვაა როგორც სატელეკომუნიკაციო („ზარის გადამისამართება“), ასევე ზოგადი („ენდოკრინოლოგთან გადამამისამართეს“) მნიშვნელობით. ერთადერთი, რაც ამ ლექსიკონის განახლებულ (ნაბეჭდ და ონლაინ-) ვარიანტებში იძებნება, არის შემდეგი ლაკონიური სტრიქონები:

გადამისამართება (გადამისამართებისა) ხელოვნ. მისამართის შეცვლა. შეიძლება ითქვას, რომ ამ გამოთავისუფლებული თანხების გადამისამართება მოხდა (**«პარლ.»**). (*ქართული ენის განმარტებითი ლექსიკონი*, 2010). აბრევიატურა „ხელოვნ.“ ამავე ლექსიკონის შემოკლებათა სიის მიხედვით „ხელოვნურს“ ნიშნავს. ალბათ იგულისხმება, რომ იმ დროს, როდესაც ქ.ე.გ.ლ.-*ის* ახალი რედაქციის ასო გ-ზე მიმდინარეობდა მუშაობა, ქართველი პარლამენტარები გაუმართავი და ***ხელოვნური*** (რასაც არ უნდა ნიშნავდეს ეს) ენით მეტყველებდნენ.

დოკუმენტის **დახარვეზების** მნიშვნელობა (რაც დღესდღეობით საკმაოდ აქტუალური ცნების აღმნიშვნელი, ხშირად ხმარებული, ძირითადად საკანცელარიო ხასიათის ტერმინია) ქ.ე.გ.ლ.-*ში* საერთოდ არ არის ასახული. „დახარვეზებაზე“ წერია მხოლოდ შემდეგი:

დახარვეზება (დახარვეზებისა) **1.** კუთხ. (ქართლ.) სახელი დაახარვეზებს ზმნის მოქმედებისა, ხარვეზების დატოვება ხვნის დროს, ხარვეზებით მოხვნა. სახნავის დახარვეზება. **2.** სტამბ. სტრიქონების დაცილება ერთმანეთისაგან შპონით (*ქართული ენის განმარტებითი ლექსიკონი*, n.d.).

შესაბამისად, ეს სიტყვები („გასაჯაროება“, „გადამისამართება“, „დახარვეზება“) არც ზემოთ ნახსენებ ლექსიკონებში არაა შეტანილი.

ეს იმის მიუხედავად, რომ „გუგლის“ ძიება „გასაჯაროებაზე“, „გადამისამართებაზე“ და „დახარვეზებაზე“ ასობით და ათასობით „ჰიტს“ იძლევა.

გამოდის, რომ ქართული ენით დაინტერესებულმა თურქმა, აზერბაიჯანელმა ან ოსმა უკანასკნელი ცნობების გამოშვებაში სიტყვა „გასაჯაროება“ რომ გაიგონოს და მისი მნიშვნელობის გაგება სცადოს, არსებული ქართულ-თურქული, ქართულ-აზერბაიჯანული ან ქართულ-ოსური ლექსიკონების მეშვეობით ვერაფერს გაარკვევს. ანალოგიურად, ვინმეს პროექტი ან დოკუმენტი, ან სულაც საბაჟოზე მანქანა რომ **დაუხარვეზონ**, მარტოოდენ ამ ლექსიკონების მეშვეობით ვერც კი მიხვდება, რაშია საქმე (თუკი ეს სიტყვა რაიმე სხვა გზით მისთვის უკვე ცნობილი არ არის).

ცხადია, რომ რა პრინციპებითაც არ უნდა ეხელმძღვანელათ ამ ლექსიკონთა შემდგენლებს სიტყვანის შერჩევისას, მათი მეთოდიკის შედეგად ლექსიკონში შესატანი სიტყვების სიის მიღმა საკმაოდ აქტუალური და ხშირად ხმარებული სიტყვები დარჩენილა.

ამავე დროს ქ.ე.გ.ლ.-ში ბევრია ისეთი შედარებით ნაკლებად ხშირად ხმარებული სიტყვები, როგორიცაა „ააბაბანებს“, „ააბლატუნებს“, „ააგურგურებს“ და სხვ. (*ქართული ენის განმარტებითი ლექსიკონი*, 2008). დიდ და ყოვლისმომცველ ლექსიკონში, როგორიც ქ.ე.გ.ლ.-ია, ალბათ „აბლატუნება“ და „აგურგურებაც“ უნდა იყოს, მაგრამ ალბათ „გასაჯაროების“ და „გადამისამართების“ ტიპის გაცილებით უფრო ხშირად ხმარებული და თანამედროვე სიტყვებიც უნდა იყოს ასახული. მით უმეტეს,

რომ „ქართული ენის განმარტებითი ლექსიკონი“, როგორც ჩანს, სხვა ლექსიკონების სიტყვანების საფუძვლადაც გამოიყენება.

საკმაოდ კარგად არის წარმოდგენილი „გასაჯაროება“ და „გადამისამართება“ ქართულ ეროვნულ კორპუსში (GNC), თანაც ტექსტები, რომლებშიც ეს სიტყვები ჩნდება, ძირითადად ოფიციალურია (კანონები, პოლიტიკური ტექსტები, მობილურ კავშირთან დაკავშირებული ტექნიკური დოკუმენტაცია და *ა.შ.*) (*ქართული ენის ეროვნული კორპუსი*, 2012).

„აბაბანება“ და „აგურგურება“, მეორე მხრივ, GNC-ში, ამ ეტაპზე მაინც, არ იძებნება.

ეს ყოველივე ალბათ გვაჩვენებს, თუ რა მიმართულებითაა საჭირო მოძრაობა. ლექსიკონები ტექსტურ კორპუსებს უნდა ეფუძნებოდეს, თავად კორპუსები კი მაქსიმალურად რეპრეზენტაციული უნდა იყოს. იგივე „დახარვეზებაზე“ უამრავი ჰიტი იძებნება „გუგლის“ საძიებო სისტემით, ქართულ ეროვნულ კორპუსში კი ეს სიტყვა ვერ აღმოვაჩინეთ. სიტყვები „ბორდერიზაცია“ და „დაანონსება“ პირადად ჩვენ გვაღიზიანებს, მაგრამ მედიაში იმდენად ხშირად იხმარება, რომ მათი ამა თუ იმ ფორმით დოკუმენტირება და ასახვა ალბათ სასურველი იქნებოდა, თუმცა მათი მიგნება ქართულ ეროვნულ კორპუსში არ მოხერხდა. სავარაუდოდ, ზედმეტი არ იქნება ქართული ეროვნული კორპუსის მოცულობისა და რეპრეზენტაციულობის გაზრდა, რათა მასში თანამედროვე ქართული ლექსიკა კიდევ უფრო სრულად იქნეს ასახული.

5 დასკვნა

როგორც ვხედავთ, თანამედროვე ქართული ლექსიკა სრულად არ არის ასახული ქართულ ლექსიკონებში და, პრინციპში, არც ქართულ ეროვნულ კორპუსში. მრავალი ხშირად ხმარებული სიტყვა შეტანილი არაა თანამედროვე ლექსიკოგრაფიულ წყაროებში და არც ტექსტურ კორპუსებში.

ლექსიკონებისა და ტექსტური კორპუსების მიღმა ჩვენ მიერ ზემოთ საკმაოდ დეტალურად განხილული ანალიზური კონსტრუქციებიც თუ დარჩა, მაშინ განსხვავება ლექსიკონებში და კორპუსებში ასახულსა და რეალურ ქართულ ენას შორის უკვე სახიფათოდ გაიზრდება. ეს კი თავის მხრივ უარყოფითად აისახება ქართულ-უცხოური ლექსიკონებისა და ქართული ტექსტური კორპუსების ხარისხზე. მანქანური თარგმნისთვისაც ასეთი ნაკლოვანი წყაროები ადეკვატურ მონაცემთა ბაზად არ გამოდგება. ამგვარი ხარვეზები ქართული ენის დიგიტალიზაციაში კი ჩვენს ეპოქაში საკმაოდ არასასურველია და არასახარბიელო შედეგები შეიძლება მოუტანოს ენის განვითარებას, რაზეც ზემოთაც ითქვა.

ზოგადად, ქართული ენა უნდა აღვიქვათ ისეთად, როგორიც ის არის, და არა ისეთად, როგორიც იგი ჩვენი აზრით უნდა იყოს. განსაკუთრებით, როდესაც საქმე თანამედროვე ენის ციფრულად დაფიქსირებას ეხება. გარდა იმისა, რომ ბოლო წლებში სამყარო მნიშვნელოვნად შეიცვალა და უამრავი ახალი რეალია და მასთან დაკავშირებული სრულიად ახალი ტერმინები გაჩნდა, ქართულ ენასთან დაკავშირებით ამ ენაზე მოლაპარაკენი დამატებითი გამოწვევების წინაშეც ვდგავართ. საბჭოთა პერიოდში ქართული ენის გამოყენების სფერო შეზღუდული იყო და ჩვენი ენა, პრაქტიკულად, ეროვნული უმცირესობის ენის სტატუსამდე იყო ჩამოქვეითებული.

ქართულს დიდი საბჭოთა იმპერიის ფარგლებში ისეთივე შეზღუდული ფუნქცია და გამოყენების სფერო ჰქონდა, როგორიც, ვთქვათ, უელსურს შეიძლება ჰქონდეს დიდ ბრიტანეთში. არ არსებობდა ქართული ჯარი, არ არსებობდა ქართული საელჩოები საზღვარგარეთ და არც ქართული დიპლომატია. საზოგადოებრივი ცხოვრების მრავალ სფეროში რუსული ენა დომინირებდა. სახელმწიფო დაწესებულებებში სულ უფრო მეტად გადადიოდნენ რუსულად საქმეთა წარმოებაზე. სამწუხაროდ, ეს ბუნებრივიც იყო – რუსულად შედგენილ ოფიციალურ დოკუმენტს, ვთქვათ, ე.წ. შრომის წიგნაკს, გასავალი მთელს უზარმაზარ საბჭოთა კავშირში ჰქონდა, ქართულ საბუთს კი საქართველოს ფარგლებს გარეთ ვერაფერში გამოვიყენებდით.

საბედნიეროდ, ჩვენი ქვეყნის მიერ დამოუკიდებლობის აღდგენის შემდეგ ქართულმა ენამ დაიბრუნა სახელმწიფო ენის სტატუსი და რეალური ფუნქცია, რამაც მის განვითარებას მძლავრი სტიმული მისცა. ჩნდება და აქტიურ ხმარებაშია თავისუფალ მეწარმეობასთან, საბაზრო ეკონომიკასთან, საერთაშორისო ურთიერთობებთან და სხვა ისეთ დარგებთან დაკავშირებული ტერმინები, რომელ დარგებშიც ქართული ენის გამოყენება ადრე მნიშვნელოვნად იყო შეზღუდული (იგივე სამხედრო სფერო).

ამას ემატება კომპიუტერულ და ციფრულ რევოლუციასთან დაკავშირებული ახალი სიტყვები, გამოთქმები თუ წმინდა ტექნიკური ტერმინები, რომლებიც, თანამედროვე ტექნოლოგიზებული ცხოვრების ყაიდიდან გამომდინარე, ჩვენს ყოველდღიურობაში აქტიურად შემოდის.

ასეთ ვითარებაში ლექსიკოგრაფია და კორპუსული ლინგვისტიკა მაქსიმალურად ინკლუზიური უნდა იყოს და ფართოდ უნდა უღებდეს კარს ახალი რეალიების აღმნიშვნელ ახალ სიტყვებსა და გამოთქმებს. იგივე უნდა ითქვას ზემოთ განხილულ ანალიზურ კონსტრუქციებზეც, რომლებიც, საბოლოო ჯამში, საზოგადოებრივი თუ ეკონომიკური ცხოვრების გამრავალფეროვნებისა და განვითარების შედეგიცაა.

თუკი ვინმე საუკუნეების წინანდელი ენობრივი ნორმებით თანამედროვე ქართული ენის ლექსიკისა და მორფოლოგიის შეზღუდვას შეეცდება და, ვთქვათ, იმავე „განცხადების გაკეთების“ ან „ცვლილების განხორციელების“ ტიპის გამოთქმებს კალკებად და სტილისტურ შეცდომებად ან ხარვეზებად მოგვინათლავს, მისმა ასეთმა ქმედებებმა ლიტერატურული ქართული ენის განვითარება მნიშვნელოვნად შეიძლება შეაფერხოს.

ლიტერატურული თუ ოფიციალური ქართულის „დროში გაყინვა“ კი ზემოთ მოხსენიებული ნეგატიური შედეგების გარდა, პოტენციურად, სხვა, უფრო სერიოზული საფრთხის მომტანიც შეიძლება გახდეს. ლიტერატურულ / ოფიციალურ ენასა და სასაუბრო, არაოფიციალურ ენას შორის სხვაობის გაღრმავებამ შეიძლება საბოლოო ჯამში ე.წ. დიგლოსია, ანუ ორენოვნება გამოიწვიოს. დიგლოსია ეს არის სიტუაცია, როდესაც ოფიციალურ, ნორმატიულ ენაზე მოლაპარაკეთა და ენის სასაუბრო, ხალხში გავრცელებულ ვარიანტზე მოლაპარაკეებს ერთმანეთის არ ესმით, ან ურთიერთგაგება მნიშვნელოვნად აქვთ გართულებული. ასეთი სიტუაცია იყო ცოტა ხნის წინ საბერძნეთში, სადაც განსხვავება წიგნურ, ძველბერძნულ ენასთან ახლოს მდგომ, **კათარევუსას** (Καθαρεύουσα) და სალაპარაკო **დიმოტიკის** (Δημοτική) შორის სერიოზულ სირთულეებს ქმნიდა (“Greek language question”, n.d.).

მსგავსი მდგომარეობა იყო ცოტა ხნის წინ შვედეთშიც, სადაც სალიტერატურო შვედურს, **სკრიფტსპროკს** (skriftspråk) და სასაუბრო შვედურს, **ტალსპროკს** (talspråk) შორის არსებითი გრამატიკულ-მორფოლოგიური და ფონეტიკური განსხვავება აღინიშნებოდა.

ამ დიგლოსიისა და მის მიერ გამოწვეული პრობლემების დაძლევა კი ორივე ამ ქვეყანაში საკმაოდ მტკივნეული პროცესი გამოდგა. საბერძნეთში ხანგრძლივი დავის, კამათისა და დაპირისპირების შემდეგ ე.წ. **სტანდარტული თანამედროვე ბერძნული ენა** შეიმუშავეს, რომელიც როგორც *კათარევუსას*, ასევე *დიმოტიკის* ელემენტებს შეიცავს. შვედურში დამკვიდრდა ენის ვარიანტი, რომელსაც **ახლანდელი შვედური** (nusvenska) ეწოდება. უნდა ითქვას, რომ მთლიანობაში შვედურის ეს ვარიანტი ტალსპროკთან უფრო ახლოსაა, ვიდრე სკრიფტსპროკთან (“Swedish language”, n.d.).

დიგლოსიის ზემწვავე შემთხვევად შეიძლება ჩაითვალოს პერიოდი საფრანგეთის ისტორიიდან, სადაც XVI საუკუნემდე (!) ფრანგული ენა ლათინურის დამახინჯებულ თუ ბარბაროსულ ფორმად ითვლებოდა და ოფიციალური ენა ლათინური იყო. ამ რამდენადმე კურიოზულ ვითარებას, როგორც ცნობილია, ბოლო მოუღო საფრანგეთის მეფე ფრანსუა I-*მა* (1494 — 1547), რომელმაც 1530 წელს ფრანგული ქვეყნის ოფიციალურ ენად გამოაცხადა და სახელმწიფო დაწესებულებებში მისი გამოყენება სავალდებულოდ გახადა (“Francis I of France, n.d”).

სასურველია, ამგვარი პრობლემების წინაშე ქართველებიც არ აღმოვჩნდეთ და „სწორი“, არქაულ ენობრივ ნორმებზე დაფუძნებული წიგნური ქართულისაგან არავინ შეგვიქმნას ახალი ლათინური ან კათარევუსა, რომელიც საზოგადოების ცოცხალ, სასაუბრო ენაზე მეტყველი ნაწილისთვის გაუგებარი იქნება. აღსანიშნავია, რომ ილია ჭავჭავაძის მოღვაწეობის მნიშვნელოვან ნაწილს სწორედ ქართული სალიტერატურო ენის ხალხისათვის ხელმისაწვდომად და ადვილად გასაგებად ქცევა იყო.

იმედია, ენის დიაქრონიული განვითარებისადმი და თანამედროვე ქართულში მიმდინარე ლექსიკური, სინტაქსური თუ სხვა პროცესებისადმი სწორი და ადეკვატური მიდგომა ჩვენს მშობლიურ ენაში ძველისა და ტრადიციულის, ერთი მხრივ, და ახლისა და აქტუალურის, მეორე მხრივ, ისეთ ჰარმონიულ სინთეზამდე მიგვიყვანს, რისი შედეგიც იქნება რეალური ქართული ენა, რომელიც არც თავის ძირძველ ფესვებს არ მოსწყდება და არც თანამედროვე ენობრივ თუ სოციალურ რეალობას. ასეთი „დაბალანსებული“ ენა კი უფრო ადვილად და უპრობლემოდ გადავა ციფრულ ფორმატში, რაც ქართულისთვის მდგრადი და ჰარმონიული განვითარების საწინდარი იქნება.

გამოყენებული ლიტერატურა

Bede the Venerable (1907). *Ecclesiastical History of England*. London: George Bell and Sons.

https://www.ccel.org/ccel/bede/history.html

Francis I of France. (n.d.). *In Wikipedia. The Free Encyclopedia*.

https://en.wikipedia.org/wiki/Francis_I_of_France (ბოლო ნახვა 18.11.2022)

Greek language. (n.d.). *In Wikipedia. The Free Encyclopedia*.

https://en.wikipedia.org/wiki/Greek_language (ბოლო ნახვა 07.10.2022)

Greek language question. (n.d.). *In Wikipedia. The Free Encyclopedia*.

https://en.wikipedia.org/wiki/Greek_language_question (ბოლო ნახვა 07.10.2022)

Grimm, J. (1819-1837). *Deutsche Grammatik*. Göttingen: Dieterichsche Buchhandlung.

Grimm, W. (1812–1815). *Kinder- und Hausmärchen* (with Jacob Grimm). Berlin: Realschulbuchhandlung.

Herder, J. G. (1772). *Abhandlung über den Ursprung der Sprache*. In M. N. Forster (Ed.), Philosophical Writings. Cambridge: Cambridge University Press (English translation available).

Jespersen, O. (1922) Language, its Nature, Development and Origin (LONDON: GEORGE ALLEN & UNWIN LTD.)

Jespersen, O. (1894). *Progress in Language: With Special Reference to English*. London: S. Sonnenschein.

Jespersen, O. (2022). Editors of Encyclopaedia, *Encyclopedia Britannica.*

https://www.britannica.com/biography/Otto-Jespersen

McElvenny, J. (2013). *Otto Jespersen and progress in international language.* In "History and Philosophy of the Language Sciences" Otto Jespersen and progress in international language | History and Philosophy of the Language Sciences (hiphilangsci.net).

Rastorguyeva, T. (1983). *A History of English.* Moscow «Vyssaya Škola».

STEPBible (*STEP - Scripture Tools for Every Person*). (2022).

https://www.stepbible.org/?q=version=THGNT|reference=John.18&options=VGNHU

და

https://www.stepbible.org/?q=version=THGNT|reference=Mark.14&options=VGNHU

Swedish language. (n.d.). *In Wikipedia. The Free Encyclopedia*.

https://en.wikipedia.org/wiki/Swedish_language (ბოლო ნახვა 07.10.2022)

Wulfila Project. (1997). University of Antwerp

http://www.wulfila.be/gothic/browse/ (ბოლო ნახვა 06.11.2022)

რეიფილდი, დ. (2006). *„დიდი ქართულ-ინგლისური ლექსიკონი"* (დესკტოპ-აპლიკაცია).

ქართულ-აზერბაიჯანული ლექსიკონი. (n.d.).

http://www.nplg.gov.ge/gwdict/index.php?a=index&d=49 (ბოლო ნახვა 07.10.2022)

ქართულ-თურქული და თურქულ-ქართული ლექსიკონი. (n.d.).

http://www.nplg.gov.ge/gwdict/index.php?a=index&d=32 (ბოლო ნახვა 07.10.2022)

ქართულ ენის განმარტებითი ლექსიკონი. (n.d.). ენის მოდელირებს ასოციაცია.

http://www.ena.ge/explanatory-online

ქართული ენის განმარტებითი ლექსიკონი. (2008). ახალი რედაქცია. ტომი I (ა-ბ). რედაქტორი ავ. არაბული. თბილისი: საქართველოს სსრ მეცნიერებათა აკადემიის გამომცემლობა.

ქართული ენის განმარტებითი ლექსიკონი. (2010). ახალი რედაქცია. ტომი II (გ). მთავარი რედაქტორი ავ. არაბული. თბილისი: საქართველოს სსრ მეცნიერებათა აკადემიის გამომცემლობა.

ქართული ენის ეროვნული კორპუსი. (2012).

http://www.gnc.gov.ge (ბოლო ნახვა 07.10.2022)

ქართული ლექსიკონი. (n.d.)

https://www.ganmarteba.ge (ბოლო ნახვა 07.10.2022)

ქართულ-ოსური და ოსურ-ქართული ლექსიკონი. (n.d.)

http://www.nplg.gov.ge/gwdict/index.php?a=index&d=54 (ბოლო ნახვა 07.10.2022)

Frame-Semantic Perspectives on Bilingual Legal Dictionaries for Cross-Systemic Translation

Waldemar Nazarov
Romanisches Seminar, Johannes Gutenberg-Universität Mainz;
Centre Interlangues TIL, Université Bourgogne Europe
E-mail: wanazaro@uni-mainz.de, waldemar.nazarov@ube.fr

Abstract

Legal translators face the formidable task of making legal texts, embedded in a foreign legal system of reference, accessible to an expert audience in the target language. Unlike most specialized fields such as medicine or technology, this undertaking requires constant legal comparison due to the uniqueness of legal systems and the extreme system-dependence of legal terminology. Consequently, bilingual dictionaries must meet specific requirements when used as tools for legal translation, as several interlingual transfer strategies have been introduced that now coexist. This paper examines the various techniques, from identifying fully established functional equivalents to creating neologisms in the target legal language, from a frame-semantic perspective. Such an approach enables translators to examine the complex knowledge uniquely associated with linguistic signs proposed by dictionaries within specific legal systems and to make informed decisions when selecting and combining legal translation techniques to convey system-dependent legal knowledge across nations.

Keywords: legal dictionary; legal terminology; legal translation; frame semantics; equivalence

1 Introduction

The creation of bilingual legal dictionaries presents challenges in both scholarship and practice, as the legal field exhibits unique features – most notably, extreme system dependence – that are largely irrelevant in other domains. Consequently, legal translation models inherently involve a comparative analysis of concepts embedded in different national legal systems, making it impossible to create bilingual legal glossaries with one-to-one correspondence. To address this, translation and legal scholars have developed various transfer strategies aimed at rendering legal texts accessible to jurists within the target legal system.

This paper summarizes key observations and requirements for contrastive legal lexicography and assesses the various translation techniques specifically developed for this field from the perspective of frame semantics. This approach is grounded in a poststructuralist linguistic theory stemming from the works of Fillmore (1976), which focuses on holistic and

complex knowledge segments cognitively evoked by linguistic signs and consequently rejects the strict distinction between linguistic knowledge on the one hand and encyclopedic knowledge on the other (Busse, 2012). Such fruitful focus on the epistemic dimension has led to the application of frame semantics to specialized languages, particularly through the Frame-Based Terminology approach developed by Faber (2022), and specifically to legal translation, which Engberg (2021) conceptualizes as an act of knowledge communication. Against the backdrop of the difficulties in establishing equivalence and ensuring translatability, which are issues extensively discussed in legal linguistics and translation studies, such cognitive and epistemic approach enables the reconstruction of knowledge segments associated with linguistic signs in law. This, in turn, facilitates decision-making when selecting or combining target-language options presented by dictionaries for the purpose of inter-systemic legal communication.

2 System-Dependence and Directionality in Bilingual Dictionaries

One of the main features of legal terminology, often cited as a distinctive criterion of law within LSP research and specialized translation studies, is its extreme dependence on a specific legal system. This creates significant challenges in cross-systemic translation scenarios, where source and target languages are inevitably embedded in different legal cultures. While the linkage between a specialized language and its *referential system* is a well-established concept in terminology (Sager, Dungworth, & McDonald, 1980, pp. 70–80), it plays a uniquely prominent and practical role in the legal field. In his influential work on law as a special case in translation, Sandrini (1999, p. 30) observes that dictionaries often present entries that fail to account for the system-dependence of legal terms, rendering them ineffective tools for legal translators. This is primarily because cross-systemic legal translation is conceptualized not merely as a transfer from a source to a target language, but also from a source to a target legal system. Unlike other categories of specialized translation, this dual transfer makes legal comparison an intrinsic step in any viable translation model (Pommer, 2006).

This imposes new requirements for classifying legal language in cases of pluricentricity: An *English legal language* only exists when embedded in a specific Anglophone legal system, such as those of the U.S., England, or Australia (Nazarov, 2025). Law is considered a metaphysical phenomenon (Mattila, 2006, p. 105), constructed within a nation to objectively reflect the values and norms of its society. Designations and concepts therefore differ significantly, even across different English legal languages. For instance, the terms *solicitor* and *barrister*, which are two distinct types of lawyers in the UK, have no standing as legal terms in the U.S., where no such distinction exists. From a frame-semantic perspective, no meaning can be associated with these two linguistic signs; that is, no legally established knowledge is evoked when they are perceived by U.S. jurists. Similarly, the designation *Basic Law* evokes institutionalized legal knowledge only within the Hong Kong legal system, where it refers to a crucial constitutional concept. In other cases, the same designations are found in different legal languages but activate different semantic fields. The term bailiff, for example, exists in both England and the U.S., yet it refers to what in France is understood as a *huissier de justice* in the former system and to a law-enforcement officer present during a court

proceeding in the latter. These differences preclude the translation of any text containing such a term without precise contextualization within its source system, which means that bilingual dictionary entries cannot be proposed without these specific differentiations.

De Groot (1999, p. 214) argues that bilingual dictionaries, unless they consist of two official languages from a single multilingual system, should be restricted to two legal systems due to the necessity of constant legal comparison. In this context, a dictionary would be specifically tailored to a translation scenario involving a particular pair of source and target languages. A generalized English metalanguage cannot serve as the target language for either a dictionary or a translation scenario. To illustrate this point, de Groot (2002, pp. 227–228) presents the English translation of the Dutch Civil Code by Haanappel and Mackaay, which systematically employs the legal English of Quebec as its target language. Although this approach might curtail accessibility and usability due to the limited prominence of this specific legal terminology in the world, the legal comparatist does not criticize this choice. In fact, he suggests enhancing usability by translating the code into the legal languages of other Anglophone nations as well.

This requirement necessarily raises the question of directionality, which is a less pivotal topic in translation and terminology work relating to other domains such as medicine or technology. In the legal field, however, unidirectionality – to use a term pertaining to contrastive linguistics (Tekin, 2012) – is considered indispensable and, for that reason, source and proposed target terms may not be reversed (de Groot, 1999, p. 214). This issue is of great relevance for cross-systemic translation scenarios: Harvey (2002, p. 44), for instance, proposes a set of established translation techniques for law and presents backtranslation (*rétrotraduction*) as an advantage of the formal translation strategy over functional equivalents (see chapter 4). This would, however, create a bidirectional link between the common-law *notary* and the French *notaire*, two roles that present significant differences. For inter-systemic contexts, this must be rejected as a criterion, as the goal in court-certified legal translation is not to create a parallel language version, as is the case in joint drafting (*corédaction*) within multilingual systems like those of Switzerland or the European Union (Dullion, 2014). Its function is rather defined as communicating knowledge using linguistic signs that provide a jurist from the target nation with access to a foreign, system-dependent legal text (Engberg, 2021). As tools specifically used for translation purposes, bilingual dictionaries must also fulfill this requirement. De Groot (2002, p. 232) specifically criticizes that some legal dictionaries do reverse this direction, a practice he legally classifies as defective performance (*Schlechtleistung*; *Leistungsstörung*), a deficiency so severe that it reportedly could justify unwinding the purchasing agreement for such a dictionary.

3 Equivalence Between Legal Languages

Given the observed system dependence of law, the question of equivalence is regularly presented as a key issue in bilingual legal dictionaries. De Groot insists that a dictionary must clearly state whether a proposed translation is an *approximate equivalent* to the source term or whether there is only *partial equivalence*; however, it must avoid creating the impression of presenting fixed *standard equivalents*. In the not uncommon case where no equivalence exists at all, the dictionary is likewise required to make this gap transparent

by employing alternative transfer strategies such as paraphrasing (de Groot, 1999, p. 213). Regarding contrastive legal terminology work aimed at creating specialized dictionaries and terminological databases, Sandrini (1999, p. 30) claims that no equation or absolute equivalence can be expected between source and target terms. This objective is categorically rejected for legal terminology, which must instead focus on comparing individual concepts and describing their similarities and differences.

In this applied scenario, the theoretical observations on inequivalence in legal linguistics become clear. Many scholars, such as Grass (1999, p. 34), Kjær (1995, p. 51), and Pommer (2006, p. 65), advocate for a general principle of inequivalence between terminologies embedded in different legal systems. These researchers, however, do not present this thesis as an obstacle but as a foundational assumption upon which any comparative and translational models must be built. Others, predominantly legal comparatists, claim that law is fundamentally untranslatable and that any attempt to render a system-dependent legal text into another legal language is a "sublime aim never to be truly achieved" (Kischel, 2009, p. 7). Both presumptions rely on the same core argument: Identicalness between two legal terms from different legal systems is impossible. Given the general sociocultural dependence of language and recent research on the culture-specificity of non-legal specialized languages (Gautier, 2019), exact semantic correspondence can be rejected outside of law as well and cannot be considered a universal requirement for contrastive terminology and translation.

It is clear, however, that a medical dictionary proposing the French *pneumonie* as a translation of the English *pneumonia* does not require in-depth analyses of the differences between French and Anglophone medical cultures, which have in fact historically shaped each concept. In law, the transfer of terms like *Basic Law* (Hong Kong) or *solicitor* (UK) presents as many challenges as more transnationally comparable terms such as *contract*, since their legal effects are palpable. As Lundmark (1999, p. 63) states, there can be no perfect bilingual legal dictionary when source and target languages refer to different systems; users must therefore accept that transfer will remain inexact. Instead, the legal scholar suggests establishing bilingual legal encyclopaedias containing exhaustive descriptions. From a frame-semantic perspective, this suggests making the many details of highly complex and interconnected knowledge segments in law transparent to translators and terminologists. This aligns with Bocquet (2008, pp. 15–16), who conclusively denies a clear separation between a dictionary and an encyclopaedia in the legal field, arguing that it makes little sense to exclude translators as non-jurist recipients from the detailed knowledge of encyclopaedias designed for legal experts.

4 Frames and Translation Strategies in Law

The relevance of the epistemic dimension becomes apparent in the lack of semantic correspondence between source and target terms. Unlike bilingual dictionaries for subject matters like medicine or physics, such reference works for the legal field cannot limit themselves to simple translation without further explanation of terms. Glossaries (or word lists), therefore, serve only as memory aids for those who have already acquired and distinguished the relevant knowledge segments (de Groot, 1999, p. 213); they do not provide a tool for

recipients to grasp the meaning of the proposed terms or their interrelations (Lönneker-Rodman & Ziem, 2018, p. 254). Similarly, Sandrini (1999, p. 30) criticizes dictionaries for frequently proposing entries that merely juxtapose designations without additional information, a practice he deems useless for legal translation. The fact that several options are often suggested in law stems from the asymmetry between legal languages, where comparable concepts are available, but the choice depends on the context of the source term's use. Grass (1999, p. 6) rightly considers it a weakness of bilingual dictionaries to propose target-language designations without defining the source term, a failure that complicates the selection of a suitable equivalent. According to Bowker (2011), term banks in general tend to present entries without specifying linguistic context, a tendency linked to the flawed assumption that terms are monosemic within a given domain. Wiesmann (2004, p. 151) similarly rejects the usability of dictionaries for legal translation when no domain-specific and linguistic context is provided, when translated elements remain uncontextualized, and when similarities and differences between source and target terms are not made explicit.

The uniqueness of legal systems and cultures gives rise to specific requirements for bilingual legal dictionaries. The choice of so-called *functional equivalents* is a preferred translation strategy, as it proposes target-language concepts that are already established within the target legal system. From a frame-semantic perspective, this involves selecting a linguistic sign capable of evoking well-established knowledge among jurists who are experts of that system. At first glance, this appears to be an ideal tool for facilitating understanding. However, the risks associated with this strategy are crucial, as a culturally created concept carries significant target-system dependence, which can distort the source term to the point of being inadequately understood. This occurs, for example, when the French supreme court, the *Cour de Cassation*, is rendered in German as *Bundesgerichtshof*. This translation linguistically implies a hierarchy within a federal system that also includes state courts, a structure that does not exist in France. Consequently, the knowledge evoked by this linguistic sign by a German lawyer would be unsuitable for depicting the frame activated by the French term, which is embedded in a non-federal, centralized nation. Conversely, the functional comparison of the French *rupture conventionnelle* and the German *Aufhebung*, which share no formal correspondence, seems to face few obstacles, as both represent an alternative agreement for terminating employment, albeit linked to different requirements and processes. The decision to apply a functional approach therefore depends heavily on the term's context within the legal text, as the importance of various knowledge elements must be carefully evaluated. Nevertheless, differences in epistemic elements are inevitable.

Among the various transfer strategies often classified as alternative solutions, which implies the preponderance of the functional approach, *formal equivalence* is a frequent technique. This strategy involves translating as literally as possible, which would mean rendering the French *Conseil constitutionnel* as *Constitutional Council* in English. According to Harvey (2002, p. 44), the advantage of this strategy lies in its strong reflection of and respect for the source legal culture. Given the court-centric nature of legal translation, adequately mirroring the source system contributes to a successful outcome. However, the unfamiliarity of the resulting designations can complicate comprehension for target-system jurists, and the strategy can sometimes lead to the creation of false friends. The scholar exemplifies this based on the erroneous translation of the French *droit commun* as *common law*, and *Haute*

Cour de Justice, a court competent for criminal charges against the President, as *High Court of Justice*. In both cases, the English terms evoke frames that do not represent the knowledge elements pertaining to the respective source terms; instead, they convey knowledge entirely irrelevant to the target text.

Descriptive translation is a technique in which cultural specificities are explained using general-language terms. This is exemplified, by the same author, by the transfer of the French criminal category *délit*, a type of crime distinct from the less serious *contravention* and the more serious *crime*, without selecting an established equivalent (a functional approach) such as *misdemeanor*. Instead, it might be described as an *intermediate offence* (Harvey, 2002, p. 46). The scholar argues that the impossibility of backtranslation presents an obstacle to this method; however, this claim must be called into question given the necessary unidirectionality of legal translation (see chapter 2). From a frame-semantic perspective, descriptive equivalents facilitate the cross-systemic communication of system-dependent knowledge in situations where no unique linguistic sign in the target language is sufficient, or where the functional equivalent presents issues due to its intrinsic linkage to a specific legal concept within the target system. Such a conclusion, however, is always dependent on the term's context of use, since, as previously mentioned, it is impossible to evoke knowledge segments that are identical between the systems in question. The technique of *lexical expansion* can also be regarded as a descriptive one as words are willingly added to the linguistic sign in the target language "as a means of compensating for terminological incongruency," which is illustrated by Šarčević (1997, p. 250) on the expanding of the term *délit* towards *délit civil* in Canadian French in order to reflect the Canadian English *tort*.

This example, however, can also be aligned with the strategy of creating *neologisms*, which is regularly proposed for the legal field. As Grass (1999, p. 52) notes, such neologisms often emerge from literal translation. For instance, the frequent German rendering of the French *Cour de Cassation* as *Kassationshof* (see, for example, Griebel, 2013) relies on a formal technique but also introduces a new term into the German legal text, as no institution known as a *Kassationshof* exists in the German legal system. In this specific case, sufficient knowledge can, in fact, be evoked within a German legal audience: While the term *Kassation* is no longer part of contemporary German legal terminology, it historically existed as a concept, allowing jurists to associate epistemic elements with this linguistic sign. Neologisms, therefore, can be justified for legal translation from a frame-semantic perspective as well.

An even more controversial technique is *non-translation*, where a source term is transferred verbatim into the target text. If a legal translator, following Geeroms' (2002) recommendation not to functionally translate the French *cassation* due to its uniqueness, were to render the same institution for an American audience with a phrase such as *The* Cour de Cassation *affirms the lower court's decision*, this would implicitly expect the legal recipient to understand French. As Stolze (1999, p. 49) rightly points out, there is little justification for choosing this technique. It is unreasonable to expect a legal scholar in the U.S. system to cognitively activate any knowledge segment associated with this unfamiliar linguistic sign. For this reason, many researchers propose combining non-translation with a paraphrase or a footnote that provides an explanation in general language.

A fruitful example can be found in the largest English-to-German legal dictionary *Wörterbuch Recht, Wirtschaft & Politik*, which proposes the following entry:

> **Supreme C[ourt]** *n* **(of the United States) (the)** (*Abk* SC) *Am* Oberstes Bundesgericht *n* (das) *(in Washington)*. ⓘ Besetzt mit 9 Bundesrichtern (8 *Associate Judges* unter dem *Chief* → *Justice*). Höchstes Rechtsmittelgericht gegenüber Entscheidungen der Bundesgerichte und (in bestimmten Fällen) der einzelstaatlichen Gerichte. Der *Supreme Court* erfüllt Funktionen, die in der Bundesrepublik dem Bundesgerichtshof (und den anderen oberen Bundesgerichten) zufallen. Erstinstanzliches Gericht: a) ausschließliche Zuständigkeit in Streitsachen zwischen 2 oder mehr Einzelstaaten sowie in Rechtsstreitigkeiten gegen ausländische Botschafter oder Gesandte; b) (mit den Einzelstaaten) konkurrierende Zuständigkeit für Klagen ausländischer Botschafter oder Gesandter; ferner für Streitsachen zwischen den Vereinigten Staaten und einem Einzelstaat und für solche eines Einzelstaates gegen Angehörige eines anderen Einzelstaates oder gegen Ausländer. (Dietl & Lorenz, 2016, pp. 197–198)

Many relevant frame elements are provided in this entry to establish a foundation for decision-making within the knowledge communication process. The German translation suggestion *Oberstes Bundesgericht* constitutes a descriptive equivalent, as it translates to *Highest Federal Court*. The frame elements evoked by these linguistic signs within the German legal system are the hierarchy of the court and its embeddedness in a federal system. If the American legal term is used in a source text where these two (main) knowledge elements are central and sufficient, this descriptive equivalent proves suitable for conveying adequate knowledge without causing semantic or legal confusion. At the same time, the dictionary provides extensive details that constitute an elaborate basis for frame extraction. It primarily defines the court as a *höchstes Rechtsmittelgericht*, which is another descriptive equivalent focusing on its appellate function as a court of last resort – a function that may be the main focus in a specific source text. The entry then compares SCOTUS to a German institution, providing a functional equivalent: the *Bundesgerichtshof*. It also emphasizes that this comparison must remain applicable to the other appropriate jurisdictions, as the German system divides its court structure into several branches, resulting in separate supreme employment court (*Bundesarbeitsgericht*) and a supreme administrative court (*Bundesverwaltungsgericht*), for example. When all or some of these knowledge elements are relevant in a source text, a comparison to a functional equivalent, either in the main text or in a footnote, can provide the necessary knowledge to a German jurist. The entry further outlines scenarios where SCOTUS acts as a court of original jurisdiction, a function linked, for example, to specific diplomatic contexts. These frame elements cannot be evoked by the German linguistic signs proposed as functional or descriptive equivalents, which is why the inclusion of this information in the entry provides a basis for a footnote that can be added to the source text and combined with one of these equivalents, along with non-translation of the original term in brackets, if the source text uses the American concept in this specific context.

5 Conclusion

Various translation techniques are specifically adapted for the legal field to facilitate cross-systemic communication and the application of foreign legal texts in court proceedings. As tools and terminological databases for such undertakings, bilingual dictionaries face domain-specific challenges. These are primarily linked to the extreme system dependence of legal language, a characteristic often said to preclude the establishment of equivalence between terminologies embedded in different legal systems.

A frame-semantic approach allows for the prioritization and strategic application of different transfer options by focusing on the cognitive and epistemic dimensions of legal concepts. The functional approach, which holds a certain preponderance in inter-systemic legal translation and legal comparison, involves selecting a legal term that exists in the target system and thereby evokes a firmly established, complex knowledge segment. If its frame elements distort the source text and prove unsuitable for the specific context, alternative translation techniques can be selected. The strategic choice among non-translation, paraphrasing, formal translation, descriptive equivalents, lexical expansion, footnotes, or even neologisms – often in combination – makes it possible to systematically convey knowledge to jurists of the target system by creating linguistic signs that adequately reflect the relevant frame elements of the source text. When it is acknowledged that identical knowledge segments cannot be evoked in both systems, and that this is not the goal of any cross-systemic legal translation, frame semantics serves as a valuable tool for evaluating the options presented in bilingual terminological databases. It facilitates a comparative decision-making process aligned with the source-text context. Ideally, bilingual dictionaries should present the various linguistic signs generated by suitable established translation techniques in order to provide a helpful and effective tool for legal translators to extract frame elements relevant to the context, which allows them to convey complex legal knowledge across nations.

References

Bocquet, C. (2008). *La traduction juridique: Fondement et méthode*. Brussels: De Boeck.

Bowker, L. (2011). Off the record and on the Fly: Examining the impact of corpora on terminographic practice in the context of translation. In A. Kruger, K. Wallmach, & J. Munday (Eds.), *Corpus-based translation studies: Research and applications* (pp. 211–236). London: Continuum.

Busse, D. (2012). *Frame-Semantik: Ein Kompendium*. Berlin, Boston: de Gruyter.

De Groot, G.-R. (1999). Zweisprachige juristische Wörterbücher. In P. Sandrini (Ed.), Ü*bersetzen von Rechtstexten: Fachkommunikation im Spannungsfeld zwischen Rechtsordnung und Sprache* (pp. 203–227). Tübingen: Narr.

De Groot, G.-R. (2002). Rechtsvergleichung als Kerntätigkeit bei der Übersetzung juristischer Terminologie. In U. Haß-Zumkehr (Ed.), *Sprache und Recht* (pp. 222–239). Berlin, New York: de Gruyter.

Dietl, C. E. & Lorenz, E. (2016). *Wörterbuch Recht, Wirtschaft & Politik: Englisch-Deutsch*. Munich: Beck.

Dullion, V. (2014). Traduire les textes juridiques dans un contexte de plurilinguisme officiel : quelle formation pour quelles compétences spécifiques ? *Meta*, *59*(3), 636–653.

Engberg, J. (2021). Legal Translation as Communication of Knowledge: On the Creation of Bridges. *Parallèles*, *33*(1), 6–17.

Faber, P. (2022). Frame-Based Terminology. In P. Faber & M.-C. L'Homme (Eds), *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge* (pp. 353–375). Amsterdam, Philadelphia: Benjamins.

Fillmore, C. J. (1976). Frame Semantics and the Nature of Language. In S. R. Harnad, H. D. Steklis & J. Lancaster (Eds.), *Origins and Evolution of Language and Speech* (pp. 20–32). New York: New York Academy of Sciences.

Gautier, L. (2019). La recherche en « langues-cultures-milieux » de spécialité au prisme de l'épaisseur socio-discursive. In M. Calderón & C. Konzett-Firth (Eds.), *Dynamische Approximationen: Festschriftliches pünktlichst zu Eva Lavrics 62,5. Geburtstag* (pp. 369–387). Berlin: Lang.

Geeroms, S. M. F. (2002). Comparative Law and Legal Translation: Why the Terms Cassation, Revision and Appeal Should Not Be Translated. *The American Journal of Comparative Law*, *50*(1), 201–228.

Grass, T. (1999). *La traduction juridique bilingue français-allemand: Problématique et résolution des ambiguïtés terminologiques*. Bonn: Romanistischer Verlag.

Griebel, C. (2013). *Rechtsübersetzung und Rechtswissen: Kognitionstranslatologische Überlegungen und empirische Untersuchung des Übersetzungsprozesses*. Berlin: Frank & Timme.

Harvey, M. (2002). Traduire l'intraduisible: Stratégies d'équivalence dans la traduction juridique. *ILCEA*, 3, 39–49.

Kischel, U. (2009). Legal Cultures – Legal Languages. In F. Olsen, A. Lorz, & D. Stein (Eds.), *Translation Issues in Language and Law* (pp. 7–17). Basingstoke: Palgrave Macmillan.

Kjær, A. L. (1995). Vergleich von Unvergleichbarem. Zur kontrastiven Analyse unbestimmter Rechtsbegriffe. In H.-P. Kromann & A. L. Kjær (Eds.), *Von der Allgegenwart der Lexikologie: Kontrastive Lexikologie als Vorstufe zur zweisprachigen Lexikographie* (pp. 39–56). Tübingen: Niemeyer.

Lönneker-Rodman, B., & Ziem, A. (2018). Frames als Repräsentationsformat in modernen Terminologiesystemen. In A. Ziem, L. Inderelst, & D. Wulf (Eds.), *Frames interdisziplinär: Modelle, Anwendungsfelder, Methoden* (pp. 251–288). Düsseldorf: düsseldorf university press.

Lundmark, T. (1999). Über die grundlegende Unmöglichkeit, ein juristisches Wörterbuch mit der Zielsprache Englisch zu erstellen: Plädoyer für eine Rechtsenzyklopädie. In G.-R. de Groot & R. Schulze (Eds.), *Recht und Übersetzen* (pp. 59–65). Baden-Baden: Nomos.

Mattila, H. E. S. (2006). *Comparative Legal Linguistics*. Aldershot: Ashgate.

Nazarov, W. (2025). Zur Plurizentrik der deutschen Rechtssprache: Eine frame-ontologische Betrachtung. In C. Di Meola, J. Gerdes, & L. Tonelli (Eds.), *Forum für Fachsprachen-Forschung: Band 175. Sprachvariation im Deutschen zwischen Theorie und Praxis: Fachsprachlichkeit, Inklusion, Didaktik, Übersetzung, Kontrastivität*. Berlin: Frank & Timme Verlag für wissenschaftliche Literatur.

Pommer, S. (2006). *Rechtsübersetzung und Rechtsvergleichung: Translatologische Fragen zur Interdisziplinarität*. Frankfurt am Main: Lang.

Sager, J. C., Dungworth, D., & McDonald, P. F. (1980). *English Special Languages: Principles and Practice in Science and Technology*. Wiesbaden: Brandstetter.

Sandrini, P. (1999). Translation zwischen Kultur und Kommunikation: Der Sonderfall Recht. In P. Sandrini (Ed.), Übersetzen von Rechtstexten: Fachkommunikation im Spannungsfeld zwischen Rechtsordnung und Sprache (pp. 9-43). Tübingen: Narr.

Šarčević, S. (1997). *New Approach to Legal Translation*. The Hague: Kluwer Law.

Stolze, R. (1999). Expertenwissen des juristischen Fachübersetzers. In P. Sandrini (Ed.), Übersetzen von Rechtstexten: Fachkommunikation im Spannungsfeld zwischen Rechtsordnung und Sprache (pp. 54–62). Tübingen: Narr.

Tekin, Ö. (2012). *Grundlagen der Kontrastiven Linguistik in Theorie und Praxis*. Tübingen: Stauffenburg Verlag Brigitte Narr.

Wiesmann, E. (2004). *Rechtsübersetzung und Hilfsmittel zur Translation: Wissenschaftliche Grundlagen und computergestützte Umsetzung eines lexikographischen Konzepts*. Tübingen: Narr.

ქართული საფინანსო ტერმინოლოგიის პრობლემები

გიორგი ოქროპირიძე
ილიას სახელმწიფო უნივერსიტეტი
giorgi.okropiridze.4@iliauni.edu.ge

აბსტრაქტი

ენა დინამიკური სისტემაა, რომელიც საზოგადოების ცვლილებებს პირდაპირ ასახავს. ეს განსაკუთრებით თვალსაჩინოა დარგობრივ ტერმინოლოგიაში, სადაც გლობალიზაციისა და ტექნოლოგიური განვითარების ფონზე მუდმივად ჩნდება ახალი ცნებები. ქართული საფინანსო ტერმინოლოგია ამ თვალსაზრისით განსაკუთრებულ ყურადღებას იმსახურებს: საბჭოთა გეგმიური ეკონომიკიდან საბაზრო სისტემაზე გადასვლამ, შემდეგ კი დასავლურ სამყაროსთან ინტეგრაციამ ტერმინოლოგიაზე ღრმა კვალი დატოვა. ამდენად, ჩვენთვის საინტერესო აღმოჩნდა შესაძლებლობა, რომ შეგვემოწმებინა უცხო ენების გავლენა და ასევე მათგან დამოუკიდებელი უარყოფითი პროცესები დარგობრივ ტერმინოლოგიაზე.

წინამდებარე ნაშრომი იკვლევს თანამედროვე ქართული საფინანსო ტერმინოლოგიის ძირითად პრობლემებს. კვლევა აჩვენებს, რომ დარგი დღეს ტერმინოლოგიური სიჭრელის, შეუთანხმებლობისა და სტანდარტიზაციის სერიოზული ხარვეზების წინაშე დგას: ერთი და იგივე ცნებისთვის ხშირად რამდენიმე პარალელური ფორმა გვხვდება; ხშირად არ არის დაცული ტრანსლიტერაციის დადგენილი ნორმები; ზოგჯერ უგულებელყოფილია ქართული ენის მორფოლოგია; უცხოური ტერმინები კი ნასესხებია მაშინაც კი, როდესაც ქართული შესატყვისი სრულიად ადეკვატური იქნებოდა.

კვლევაში გამოყენებულია პარალელური კორპუსის პლატფორმა და თანამედროვე ტერმინოლოგიური ინსტრუმენტი SynchroTerm, რომელიც ორენოვანი კორპუსის ბაზაზე ტერმინთა ავტომატურ ამოღებას უზრუნველყოფს. მიღებული შედეგები ადასტურებს სისტემური ტერმინოლოგიური მუშაობის გადაუდებელ აუცილებლობას საქართველოში.

საკვანძო სიტყვები: ფინანსები, ტერმინოლოგია, საფინანსო ტერმინები, ლექსიკოგრაფია, ნეოლოგიზმები, სესხება, ტრანსლიტერაცია, ტერმინთშემოქმედება

Problems of Georgian Financial Terminology

Giorgi Okropiridze
Ilia State University
giorgi.okropiridze.4@iliauni.edu.ge

Abstract

Language is a dynamic system that directly reflects societal change. This is particularly evident in specialized terminology, where new concepts continuously emerge against the backdrop of globalization and technological advancement. Georgian financial terminology deserves special attention in this regard: the transition from the Soviet planned economy to a market system, followed by integration with the Western world, left a profound mark on the terminological landscape. It therefore seemed worthwhile to examine both the influence of foreign languages on specialized terminology and the negative processes that occur independently of such influence.

The present paper investigates the key problems of contemporary Georgian financial terminology. The research demonstrates that the field today faces serious challenges of terminological inconsistency, lack of standardization, and uncoordinated practice: multiple parallel forms are frequently found for the same concept; established transliteration norms are often disregarded; the morphological rules of the Georgian language are sometimes ignored; and foreign terms are borrowed even in cases where a Georgian equivalent would be perfectly adequate.

The study employs the parallel corpus platform and modern terminological tool SynchroTerm, which enables the automatic extraction of terms from a bilingual corpus. The findings confirm the urgent need for systematic and coordinated terminological work in Georgia.

Keywords: finances, terminology, financial terms, lexicography, neologisms, borrowing, transliteration, term formation

1 შესავალი

ენა დინამიკური სისტემაა, რომელიც მუდმივად ვითარდება საზოგადოების სოციალურ-ეკონომიკური ცვლილებების შესაბამისად. განსაკუთრებით აქტიური ცვლილებები / სიახლეები გვხვდება დარგობრივ ტერმინოლოგიაში, მათ შორის საფინანსო სფეროში, სადაც გლობალიზაციის, ტექნოლოგიური პროგრესისა და ეკონომიკური ურთიერთობების გაღრმავების პირობებში ჩნდება ახალი ცნებები და მათი ენობრივი ასახვის საჭიროება, რაც ქართული საფინანსო ტერმინოლოგიის კვლევას კომპლექსურ პროცესად აქცევს.

ბოლო ათწლეულების განმავლობაში ქართულმა ეკონომიკამ განიცადა რადიკალური ტრანსფორმაცია: საბჭოთა გეგმიური სისტემიდან საბაზრო ეკონომიკაზე გადავიდა, ჩამოყალიბდა ახალი ეკონომიკური ინსტიტუციები, მიმდინარეობს საერთაშორისო საფინანსო სისტემებთან ინტეგრაცია. ყოველივე ეს უშუალო გავლენას ახდენს ენის ტერმინოლოგიურ მხარეზე.

მეცნიერული კვლევის აუცილებლობა განპირობებულია რამდენიმე მნიშვნელოვანი ფაქტორით. პირველ რიგში, სახეზეა საფინანსო ტერმინოლოგიის ქაოსური განვითარება, რაც გამოიხატება ტერმინთა არაერთგვაროვნებაში, უცხოური სიტყვების უსისტემო შემოჭრაში და ხშირად ქართული ენისათვის არაბუნებრივი საკომუნიკაციო მოდელების დამკვიდრებაში. კერძოდ, თანამედროვე ქართულ საბანკო, საფინანსო და ეკონომიკურ ტექსტებში (არა მხოლოდ დოკუმენტებში, არამედ დარგობრივ საინფორმაციო საშუალებებში, ვიდეორგოლებში, სტატიებში და ა. შ.) ხშირად გვხვდება ტერმინოლოგიური სიჭრელე: ერთი და იგივე ცნება შეიძლება რამდენიმენაირად იქნეს გადმოცემული, ახალი საფინანსო ცნებების აღსანიშნავად იყენებენ როგორც უშუალო ნასესხობებს (დამახინჯებული ტრანსლიტერაციის გზით), ისე მექანიკურად თარგმნილ კონსტრუქციებს, რაც ზოგჯერ შესაძლოა არ იყოს ბუნებრივი და ადვილად აღსაქმელი.

საქართველოში ტერმინოლოგიური პრობლემების შესახებ ბევრი მეცნიერი წერს, განსაკუთრებით კი ბოლო პერიოდში (2014 წლიდან ენათმეცნიერების ინსტიტუტის მიერ გამოიცემა სტატიათა კრებული „ტერმინოლოგიის საკითხები“). ქვეყნის დამოუკიდებლობის აღდგენის შემდეგ, ომების, არეულობისა და ანარქიის პირობებში მოიშალა ან შეფერხდა სხვადასხვა სახელმწიფოებრივი და მათ შორის ენობრივ პოლიტიკასთან დაკავშირებული უწყების საქმიანობა. ენობრივ პოლიტიკაზე, ტერმინოლოგიასა და ენის ფუნქციონირებაზე ეფექტური ზემოქმედების მომხდენი ინსტიტუციის არარსებობის შედეგად წარმოიშვა უამრავი პრობლემა: ენაში ქაოსურად შემოვიდა უცხოური ტერმინოლოგია, სხვადასხვა დარგში ერთი და იგივე ცნებების აღსანიშნავად გაჩნდა განსხვავებული ტერმინები, ხშირად უგულებელყოფილი იყო ქართული ენის ნორმები და ტრადიციები.

2 ლიტერატურის მიმოხილვა

პოსტსაბჭოთა ეპოქაში მრავალმა პრობლემამ იჩინა თავი, გაუქმდა ენის სახელმწიფო კომისია, დაირღვა ცენტრალიზებული სტრუქტურა, არ მოხდა ტერმინოლოგიის ევროპულ სტანდარტებზე გადასვლა, ძველი რუსულ-ქართული და ახალი ინგლისურ-ქართული ტერმინების ჰარმონიზაცია, ლექსიკონები ხშირად იქმნე-

ბოდა შეუთანხმებლად. ამის შედეგად მივიღეთ ტერმინოლოგიური სიჭრელე (მ.შ. უთარგმნელად, ქაოსურად შემოსული ტერმინები), საერთაშორისო სტანდარტებისადმი შეუსაბამობა და უნდობლობა ტერმინოლოგიური ლექსიკონების მიმართ (ქაროსანიძე 2022–2023: 399–400). ზოგადი ტერმინოლოგიური პრობლემატიკა თავს ავლენს მეცნიერების პრაქტიკულად ყველა დარგში. თანამედროვე განვითარების პირობებში ჩნდება უამრავი ახალი მიმართულება, ახალი ცნებები, რომელთა აღსანიშნი ტერმინები ქაოსურად და უკონტროლოდ იჭრება ქართულ ენაში (კალკირება, უთარგმნელად გადმოღება, ზოგჯერ ორი ან მეტი პარალელური ტერმინის გაჩენა), რაც იმას მიუთითებს, რომ არ არის შემუშავებული ტერმინოლოგიური მუშაობის ეფექტური მექანიზმი (საღინაძე 2022–2023: 339).

მიუხედავად მრავალსაუკუნოვანი ტრადიციებისა, საქართველოში დღეს ტერმინოლოგიის მართვის ეფექტური სისტემა არ არსებობს. საბჭოთა პერიოდის ტერმინთშემოქმედების დადებითი მხარე იყო დარგობრივ ტერმინოლოგიაზე სისტემური მუშაობა (ქაროსანიძე 2022–2023: 396–397), თუმცა მის სუსტ მხარედ ითვლება ორენოვანი (რუსულ-ქართულ) სტრუქტურა და რუსულის ძალიან მომძლავრებული გავლენა. ტერმინთა ანალიზისას ნათლად ჩანს, რომ ტერმინოლოგია დიდწილად რუსულს მიჰყვება პირდაპირ. მაგალითად, თუკი კონკრეტული ცნების აღსანიშნავად რუსულში ტერმინი შექმნილია საკუთარი ენის ლექსიკური ფონდით, ქართულში მას უქმნიდნენ ეკვივალენტს ან თარგმნიდნენ (ანუ ადგილი ჰქონდა სემანტიკურ სესხებას ან კალკირებას): самолет – თვითმფრინავი, гранатомет – ყუმბარსატყორცნი, предложение – წინადადება და ა.შ. მაგრამ თუ ტერმინი რუსულში სხვა ენიდან იყო ნასესხები, ქართულში იგი რუსულის გავლით გადმოდიოდა ასევე ტრანსლიტერაციის გზით და რუსული ორთოგრაფიის გათვალისწინებით; მაგალითად, акция – აქცია (< ფრანგ. action), курс – კურსი (< ლათ. cursus); მაგრამ არის შემთხვევები, როდესაც სიტყვა მთლიანად რუსულ ორთოგრაფიას არ მისდევს: вексель – ვექსილი (< გერმ. Wechsel).

ამას გარდა რუსულის გავლენით ქართულში სემანტიკური სესხებით გზით შემოსული (რაც უარყოფითად ნამდვილად არ უნდა შეფასდეს) ან ქართული ენის ბაზაზე სპეციალურად ქართული ენისათვის შექმნილი ტერმინების გვერდით იმავე რუსულის გავლენით იწერებოდა ტრანსლიტერირებული სინონიმური ფორმები, რომლებთაგანაც ზოგიერთმა შემდეგში გამოდევნა ქართული სინონიმები. „საბჭოთა პერიოდამდელ ლექსიკონებში ქართული ტერმინები გამოიყენებოდა, შემდეგ კი ისინი უცხო სიტყვებმა ჩაანაცვლა“. 1920 წელს გამოცემულ რუსულ-ქართულ ტექნიკურ ლექსიკონში გვხვდება ტერმინები: სხივთაბნევა, კუთხზომი, მათრევი, ფრენოსანი, რომლებიც შემდგომ საბჭოთა პერიოდში თანდათან ჩანაცვლდა ინტერნაციონალიზმებით: აბერაცია, გონიომეტრი, ბუქსირი, ავიატორი. სინონიმია, როგორც ტერმინოლოგიის ერთ-ერთი სირთულე, „გარკვეულწილად საბჭოთა პერიოდის ენობრივი პოლიტიკით იყო გამოწვეული. სწორედ ამ ეპოქის მემკვიდრეობაა ორი ტერმინი ერთი ცნებისათვის: ერთი ქართული, ერთიც – უცხოური“ (ქაროსანიძე & ხურცილავა 2018: 28–30).

სინონიმიის პრობლემას გვიან საბჭოთა პერიოდში მკაცრი სტანდარტიზაციით უპასუხეს. ლექსიკონებში, რომლებსაც ოფიციალური ორგანოები ამტკიცებდნენ, ერთი ცნებისათვის შეჰქონდათ ერთი ტერმინი, მიუხედავად იმისა, რომ დარგობრივ ლიტერატურაში შესაძლოა მისი რამდენიმე ფორმა ყოფილიყო გავრცელებული. ქარ-

თულ ტერმინოლოგიას საბჭოთა სტანდარტის შესაბამისად დაევალა, რომ რუსულ ტერმინებზე ჰქონოდა სწორება. ეს მეთოდი ტერმინოლოგიაში სინონიმიის თავიდან აცილების მიზნით შესაძლოა გამართლებულად ჩაითვალა, მაგრამ ამავდროულად ის გზას უხსნიდა კალკირებულ ტერმინებს და მთლიანად რუსულის თარგზე „ჭრიდა“ ქართულ ტერმინოლოგიურ საქმეს (საღინაძე 2022–2023: 347–348).

ერთი შეხედვით სინონიმურ წყვილებში უარყოფითი ბუნება არ ჩანს, თუმცა თუკი დავაკვირდებით სამეცნიერო თუ არასამეცნიერო ტექსტებს, მარტივი შესამჩნევია, რომ ინტერნეციონალიზმებმა თანდათან გამოდევნა ენიდან მათი ქართული შესატყვისები (უფრო მეტი ასეთი სინონიმური წყვილისათვის იხ. უცხო სიტყვებისა და მათი ქართული შესატყვისების გლოსარიუმი – https://glosarium.iliauni.edu.ge/).

ვლადიმერ პაპავა სტატიაში „თანამედროვე ქართული ეკონომიკური ტერმინოლოგია: ახალი პრობლემები და ძველი შეცდომები“ თანამედროვე ქართული ეკონომიკური ტერმინოლოგიის განვითარების ორ მთავარ პერიოდს გამოჰყოფს: 1) XX საუკუნის 90-იანი წლები (დამოუკიდებლობის მოპოვებისა და საბაზრო ეკონომიკაზე გადასვლის პერიოდი) და 2) 2000 წლიდან მოყოლებული. პირველი პერიოდის ტერმინოლოგიურ მუშაობაში ის დიდწილად ისევ რუსული ენის გავლენას ხედავს (მაგალითად, ამ დროს დამკვიდრდა ტერმინი „მთლიანი შიდა პროდუქტი“ – валовой внутренний продукт და არა „მთლიანი სამამულო პროდუქტი“ – gross domestic product). მეორე, ანუ ახლანდელ პერიოდს კი უფრო მეტად ინგლისური ენის გავლენა ახასიათებს, რასაც ხშირად თან ახლავს ადრე ინგლისური ენიდან შემოსული სიტყვების მიერ ქართული ტერმინების განდევნა (მაგალითად, „მართვა“ თითქმის მთლიანად გამოდევნა „მენეჯმენტმა“) (პაპავა 2014: 141–142). ინგლისურის მომძლავრებული გავლენის მიუხედავად, ჯერ კიდევ მრავალი დარგი რჩება რუსულ ტერმინოლოგიასთან მიბმული. ინგლისურის გავლით შემოსული „სილაბუსის“, „პორტფოლიოს“ (არსებობს მანამდე შემოსული „პორტფელი“), „თინეიჯერის“ გვერდით აქტიურ ხმარებაში რჩება „ბოლტი“, „ნასოსი“ და სხვა. მიუხედავად იმისა, რომ რუსულიდან შემოსულ ბარბარიზმებს ტერმინოლოგიურ დონეზე დიდი ხანია ებრძვიან, დღემდე ვერ მოხდა ენიდან მათი განდევნა. ლიტერატურაში მათ ნაცვლად გამოყენებული „ქანჩი“, „საკისარი“, „ჭანჭიკი“ ენაში სრულყოფილად ჯერ ვერ დამკვიდრდა.

ქართული ტერმინოლოგიის დანერგვის სირთულის ერთ-ერთი მიზეზი ისიცაა, რომ ლექსიკონებში შეტანილი ზოგიერთი ქართული ტერმინი „თითქოსდა მიესადაგება მნიშვნელობას, მაგრამ იმდენად უხეიროდ, რომ ისინი საერთოდ არ გამოიყენება პრაქტიკაში“: იგრიხებადი, იწვებადი, მხუთრიანი, მეზარდი; არადა ბევრად უფრო ბუნებრივია (და თანაც ენაში გავრცელებული) გრეხადი, წვადი, მხუთავი, მზარდი (სუთიძე & იაკობაშვილი 2016: 120). თუკი ძველ ლექსიკონებში შეტანილი ტერმინები ენამ არ მიიღო და მათ ნაცვლად არსებობს ქართულისთვის უფრო ბუნებრივი და მოქნილი სიტყვები, ტერმინოლოგიური ბაზები (ლექსიკონი იქნება ეს თუ ონლაინბაზა) უფრო აქტიურად უნდა დაექვემდებაროს განახლებას.

ტერმინოლოგიის დანერგვის ახლანდელი მდგომარეობა იმდენად ქაოსურია, რომ ზოგჯერ ერთი სასწავლებლის სხვადასხვა კათედრაზეც კი არაუნიფიცირებულ ტერმინებს იყენებენ, ანუ ერთი ცნების აღსანიშნად შესაძლოა სხვადასხვა სიტყვას ასწავლიდნენ (მელაძე 2024: 124–139). ფინანსების და საბანკო საქმის სხვადასხვა

სასწავლო სახელმძღვანელოში გვხვდება თარგეთირება და ტარგეტირება, ოპციონი და ოფციონი, საბანკო საქმე და ბანკინგი, ნდობითი ოპერაციები და სატრასტო ოპერაციები. ერთ შემთხვევაში ადგილი აქვს ტრანსლიტერაციის ნორმების დაუცველობას, მეორეგან კი ქართული და უცხოური ვარიანტები გვერდიგვერდ იხმარება. ამგვარი მიდგომა ენის საფრთხედ ითვლება (ხურცილავა 2016: 158). თუკი ზემოაღნიშნულ პარალელური სინონიმური წყვილების მაგალითს გავიხსენებთ, ეს შიში უსაფუძვლო არ არის. ქართული საფინანსო (და არა მარტო) ტერმინოლოგიის ერთ-ერთ დიდ პრობლემად ნამდვილად დასტურდება ტერმინოლოგიის შეუთანხმებლობა და ტერმინების თვითნებურად ხმარება.

სასწავლო ლიტერატურაში დარგობრივი ტერმინების ქართული ეკვივალენტის არარსებობის დროს უცხო ენიდან შემოტანილი სიტყვა აუცილებლად მოითხოვს განმარტებას (ხურცილავა 2016: 160). სასწავლო ლიტერატურის შედგენისას გამოყენებული ეს პრაქტიკა უეჭველად მიუთითებს იმაზე, რომ უცხოური ტერმინების მასობრივი გამოყენება მნიშვნელოვნად ართულებს შინაარსის აღქმადობას. მაშინ რა საჭიროა ტექსტის გადავსება ამდენი უცხო სიტყვით? პრაქტიკულად ყველა მეცნიერი, რომელიც ტერმინოლოგიის პრობლემაზე წერს, ამ კითხვას სვამს.

ხშირად ერთმანეთისგან განსხვავდება ლექსიკონში შეტანილი და ყოველდღიურ დარგობრივ საქმიანობაში გამოყენებული ტერმინები. მაგალითად, სამხედრო და ტექნიკურ ლექსიკონებში შეტანილი „კუთხვილის“ და „კუთხვილიანის“ ნაცვლად ქართულ ჯარში გამოიყენება „ნაჭდევი“ და „ნაჭდევიანი“, „ცეცხლსაქრობის“ ნაცვლად ენაში დამკვიდრდა „ცეცხლმაქრი“, „საყურისის“ ნაცვლად – „ყურსასმენი“ (მელაძე 2024: 124–139), „ქარსარიდის“ ნაცვლად – „საქარე“ (მინა/შუშა) და ა. შ. ტერმინის შერჩევის, ლექსიკონში შეტანის, მის დამკვიდრებასთან დაკავშირებული საქმიანობის დროს ეს ფაქტორი აუცილებლად გასათვალისწინებელია – რა არის გავრცელებული ენაში, რომელი ფორმაა უფრო ბუნებრივი და გამჭვირვალე.

ტერმინოლოგიის პრობლემებს შორისაა ტერმინთა უმართებულოდ გამოყენება. მაგალითად, ხშირად governor-ისა და manager-ის შესატყვისად ორივე შემთხვევაში იყენებენ მენეჯერს (პირველს შესაძლოა შეესაბამოს მმართველი); tax-ისა და payment-ისათვის – გადასახადს (თუმცა არსებობს გადასახდელიც). მნიშვნელოვანი პრობლემაა ისიც, რომ ეკონომიკურ ტერმინთა მცდარად გამოყენება ხდება არამარტო პოლიტიკოსებისა და ჟურნალისტების, არამედ დარგის სპეციალისტების მიერაც (პაპავა 2014: 141–142).

როგორც ზემოთ აღინიშნა, რუსულის შემდეგ ქართულ ტერმინოლოგიას ზემოქმედების მთავარ ფაქტორად ინგლისური ჩაენაცვლა (თუმცა არა სრულად). როდესაც უცხო ტერმინების ქაოსურად შემოჭრაზე ვსაუბრობთ, აქ პირველ რიგში ინგლისურ ენას მოვიაზრებთ. ქართულ საფინანსო ტერმინოლოგიაში შემოსული (განსაკუთრებით ბოლო პერიოდში) უცხო სიტყვების დიდი უმრავლესობა ინგლისურია ან ინგლისურის გავლით მოხვდა ქართულში: კორელაცია, პროფიციტი, ლიზინგი, დისკონტი, სვოპი, სპოტი, ჰეჯირება და ა. შ. რუსულის მსგავსად ასევე ხშირად გვაქვს ქართულ-ინგლისური სინონიმური წყვილები (ხურცილავა 2016: 157).

სხვადასხვა დარგში გვაქვს ტერმინთა არასაჭიროდ ან ხელმეორედ სესხების შემთხვევებიც. ადრე რუსული ენის გავლით შემოსულ და გარკვეული მნიშვნელობით დამკვიდრებულ ლათინურძირიან სიტყვებს დღეს ხელმეორედ ვსესხულობთ

ინგლისურიდან – უკვე ახალი მნიშვნელობით: რეზერვაცია (როგორც იძულებით გადასახლების ადგილი) // რეზერვაცია (როგორც დაჯავშნა), არტისტი (როგორც მსახიობი) // არტისტი (როგორც მხატვარი), კოლაბორაცია (როგორც მტერთან თანამშრომლობა) // კოლაბორაცია (როგორც უბრალოდ თანამშრომლობა); ასეთი შემთხვევები ძლიერ მომრავლდა (მარგალიტაძე 2023: 16–18). სატყეო ტერმინოლოგიაში „ტყის მცველი“ ბოლო პერიოდში მასობრივად ჩაანაცვლა ინგლისურიდან შემოსულმა „რეინჯერმა“. რეინჯერი მანამდე ქართულში სამხედრო ტერმინოლოგიის ნაწილი იყო. ქართულ სამეცნიერო სივრცეში გავრცელებული აზრით, ეს არღვევს ტერმინის მონოსემიურობის პრინციპს (სუხიშვილი 2020: 317); თუმცა ტერმინთშემოქმედების ეს პრინციპები (მონოსემიურობა, ერთსიტყვიანობა, არასინონიმურობა...) დიდი ხანია მოძველდა საერთაშორისო პრაქტიკაში და ქართულიც ვერ იქნება გამონაკლისი. ერთი ტერმინი მხოლოდ ერთ დარგში და მხოლოდ ერთი ცნებისთვის ყოველთვის ვერ იქნება გამოყენებული. ამავდროულად, ვერ უარვყოფთ ტერმინების სინონიმიას, ესეც თანამედროვე რეალობაა, მაგრამ აქ მეორე მომენტია გასათვალისწინებელი: რა საჭიროა უცხო ენებიდან ისეთი ცნებებისათვის შემოვიტანოთ ტერმინები, რომელთათვისაც უკვე არსებობს მიღებული შესატყვისები? განსაკუთრებით სამწუხაროა, როდესაც ამას ვხვდებით ოფიციალურ დოკუმენტებში, უმაღლესი სასწავლებლების გამოცემებში, სასწავლო მასალებში. შევეცადეთ შეგვესწავლა, რამდენად ხშირია ეს ტენდენციები უშუალოდ ფინანსების დარგში.

თანამედროვე ტერმინოლოგიისათვის დამახასიათებელი კიდევ ერთ-ერთი ტენდენციაა საერთო ენის სიტყვების ტერმინებად ქცევა. ეს უკვე სიახლე აღარ არის, მაგრამ დღეს პროცესი ბევრად უფრო გააქტიურებულია, ვიდრე უწინ. ამავდროულად გახშირდა ტერმინების გადასვლა ერთი დარგიდან მეორეში. ყოველივე ეს ენაში აჩენს სიტყვების ახალ მნიშვნელობებს, ანუ პოლისემიას. დიდი რაოდენობით ახალი ცნებების გაჩენა ენას უბიძგებს გამოიყენოს ე. წ. ეკონომიის პრინციპი და სახელდებისას ისარგებლოს არსებული ლექსიკური ფონდით (მარგალიტაძე 2018). საინტერესოა საფინანსო ტერმინოლოგიის გაანალიზება ამ კუთხითაც.

ტერმინთშემოქმედების მეთოდების ეს მრავალფეროვნება ხელს უწყობს ტერმინოლოგიის გაჯანსაღებას; კონკრეტული შემთხვევებიდან გამომდინარე შესაძლებელია ტრანსლიტერაციის, სემანტიკური სესხების, დარგთაშორისი სესხების, საერთო ლექსიკის სიტყვების გამოყენება და ერთგვარი ოქროს შუალედის პოვნა პურიზმსა და ნასესხობას შორის. თუმცა ეს პროცესი თავისთავადი ვერ იქნება. დღეს ვიმკით საკუთარ დინებაზე მიშვებული, შეუთანხმებლად წარმოებული ტერმინოლოგიის უარყოფით შედეგებს. დინებაზე მიშვების არდაშვება არ უნდა წარმოვიდგინოთ როგორც „ავტორიტარული“ პროცესი (რაც პრაქტიკულ შედეგებს ნაკლებად მოიტანს), არამედ როგორც შეთანხმებული, დაკვირვებასა და ანალიზზე დაფუძნებული საერთო საქმე.

ამ თავში მოყვანილი ავტორები გამოსავლად ხედავენ ერთიანი ტერმინოლოგიური საცავის სისტემურ განვითარებას, პროცესში დარგის სპეციალისტების, ტერმინოლოგების, ლექსიკოგრაფების ჩართულობას, სტანდარტიზაციის პროცესის გაძლიერებას, საერთაშორისო სტანდარტებთან ინტეგრაციას, ენობრივ პროცესებზე დაკვირვებას, ენის და ამ ენაზე მოსაუბრე კოლექტივის ინტერესების ორმხრივ გათვალისწინებას.

3 კვლევის მონაცემები და მეთოდოლოგია

საანალიზო ტერმინების შესაკრებად გამოვიყენეთ კორპუსული მეთოდოლოგია, კერძოდ პარალელური კორპუსის პლატფორმა და ტერმინების ავტომატურად ამოღებას სპეციალური პროგრამა - SynchroTerm. ეს არის ტერმინოლოგიის ორენოვანი პროგრამული უზრუნველყოფა, რომელიც ბევრად აადვილებს ტერმინების ამოღებისა და დამუშავების პროცესს სტატისტიკური ალგორითმების, სინტაქსური და მორფოლოგიური ანალიზის გამოყენებით. პროგრამა ავტომატურად პოულობს ტერმინებს და მათ სხვაენოვან ეკვივალენტებს; თავსებადია სხვადასხვა ფორმატის დოკუმენტებთან; შესაძლებელია მონაცემების ექსპორტი ხუთ სხვადასხვა ფორმატში; აქვს 30 ენის მხარდაჭერა (რაც ამ ენების ტერმინების ამოცნობას ამარტივებს; ქართული ენა χერ-χერობით არ აქვს).[1]

პარალელური კორპუსის პლატფორმად გამოვიყენეთ ილიას სახელმწიფო უნივერსიტეტის ლექსიკოგრაფიისა და ენობრივი ტექნოლოგიების ცენტრის მიერ შექმნილი *ინგლისურ-ქართული პარალელური კორპუსი* (https://enkacorpus.iliauni.edu.ge/).[2] კორპუსი მოიცავს როგორც მხატვრულ ლიტერატურას, ისე სამეცნიერო და დარგობრივ ტექსტებს. კორპუსიდან ჩამოიტვირთა საფინანსო სფეროსთან დაკავშირებული ტექსტების ქართული და ინგლისური ვერსიები, რომლებიც შემდეგ აიტვირთა SynchroTerm-ში საფინანსო ტერმინების ამოსაკრებად. პროგრამა გაანალიზების შემდეგ გვთავაზობს ტერმინების კანდიდატებს, მის კონტექსტებს და შემდეგ უკვე მომხმარებელი იღებს გადაწყვეტილებას, დატოვოს თუ არა ეს ტერმინოლოგიური წყვილი ტერმინთბაზაში. შესაძლებელია როგორც წყარო, ისე სამიზნე ენის ტერმინების ჩასწორება, პროგრამის მიერ გამორჩენილი ტერმინების მონიშვნა და ბაზაში ჩამატება. არჩეული ტერმინები თავს იყრის ტერმინთბაზაში, სადაც ნებისმიერ დროს არის შესაძლებელი გადასვლა და გლოსარიუმის დათვალიერება, კონტექსტების გადახედვა, დამთხვევების რაოდენობის შემოწმება, ტერმინების ამოშლა და სხვ.

მუშაობის დასრულების შემდეგ მომხმარებელს აქვს მონაცემთა ექსპორტის შესაძლებლობა. შესაძლებელია შეგროვებული ტერმინებისა და მათი კონტექსტების კონვერტირება სხვადასხვა ფორმატში, მაგალითად Excel-ში, რის შემდეგაც მარტივდება ტერმინების ანალიზი. ამ ეტაპის დასრულების შემდეგ პროგრამის დახმარებით ამოვიღეთ 200-მდე ინგლისური ტერმინი ქართული შესატყვისებით.

საანალიზო ტერმინოლოგიური მასალის მოსაგროვებლად ასევე გამოვიყენეთ არსებული ლექსიკონების სიტყვანი და განმარტებები. ერთ-ერთი ასეთი ახალი დარგობრივი ლექსიკონია „ფინანსურ ტერმინთა ლექსიკონი“ (რედ. ვ. ლეჟავა, 2022, თავისუფალი და აგრარული უნივერსიტეტების გამომცემლობა; https://finterms.ge/),

1 *SynchroTerm – bilingual term extraction software*. Terminotix. https://terminotix.com/docs/factsheet_synchroterm_en.pdf [14/04/2025].

2 *ინგლისურ-ქართული პარალელური კორპუსი*. ლექსიკოგრაფიისა და ენობრივი ტექნოლოგიების ცენტრი. https://lexicography.iliauni.edu.ge/inglisur-qarthuli-paraleluri-korpusi/ [14/04/2025].

რომელიც 300-მდე ტერმინს შეიცავს. მასში მოცემულია ინგლისური ტერმინი, მისი ქართული ეკვივალენტი და ცნების განმარტება ქართულ ენაზე.

4 თანამედროვე ქართული საფინანსო ტერმინოლოგიის პრობლემები

4.1 არასწორად შედგენილი ტერმინები

ტერმინთა მოძიებისა და ანალიზის დროს გამოვლინდა ლექსიკონებში შეტანილი, ოფიციალურ დოკუმენტაციებსა თუ საიტებზე დაფიქსირებული მრავალი ისეთი ტერმინი, რომელიც თავისი შედგენის წესით არღვევს ტერმინშემოქმედების პრინციპებს, ზოგჯერ თვით ქართული ენის მორფოლოგიასაც კი.

ერთ-ერთი ყველაზე გავრცელებული (არამარტო ფინანსების დარგში) ტერმინოლოგიური ხარვეზია რამდენიმენაწილიანი კომპოზიტების ცალ-ცალკე წერა, რასაც მეორენაირად, რიგ შემთხვევებში, შეგვიძლია ვუწოდოთ ნათესაობითი ბრუნვის დაკარგვა. ერთი შეხედვით ასეთი კომპოზიტების პირველი ნაწილი არსებითი სახელის (რთული სიტყვის კომპონენტის) ატრიბუტულ ხმარებადაც შეიძლება ჩავთვალოთ, მაგრამ ქართულში ატრიბუტულად გამოყენებული არსებითი სახელი სახელობით ბრუნვაში უნდა იყოს; ამ შემთხვევაში კი ფუძის ფორმით გვაქვს წარმოდგენილი (ე.წ. წრფელობით ბრუნვაში, ბრუნვის ნიშნის გარეშე). აქ ჩამოვთვლით რამდენიმე ასეთ ტერმინს:

[1] ბლოკჩეინ ტექნოლოგია – blockchain technology: ერთ-ერთ სამეცნიერო სტატიაში[3] ტერმინის გვერდით გამოტანილია ქართული ვარიანტი „ბლოკჯაჭვი“, თუმცა ამავე სტატიაში უგულებელყოფილია ამგვარ კომპოზიტთა ჩაწერის წესი, რომელიც გულისხმობს სიტყვების შეერთებით მიღებული კომპოზიტის ერთ სიტყვად დაწერას (შდრ. ინტერნეტბლოგი, ვიდეოგადაღება, ავტომომსახურება და ა. შ.). ქართულად მართებული ფორმა იქნება ბლოკჩეინტექნოლოგია ან ბლოკჩეინის ტექნოლოგია.

[2] მეზანინ დაფინანსება – mezzanine financing: გულისხმობს ფინანსური საშუალებების მოძიების ისეთ გზას, როდესაც ერთმანეთთან შერწყმულია სესხი (კრედიტი) და საკუთარი კაპიტალი; ანუ მათი შუალედური ვარიანტია. ფინანსური ტერმინი წარმოსდგება არქიტექტურული ტერმინიდან „მეზონინი“ (ინგლ. mezzanine, რუს. мезонин), რომელიც შენობის მთავარ სართულებს შორის არსებულ შუალედურ სართულს ნიშნავს (ინგლ. *mezzanine* < ფრანგ. *mezzanine* < იტალ. *mezzanino* კნ. < იტალ. *mezzano* < ლათ. *medianus* „შუა“).[4] ასეთი კომპოზიტი უნდა დაიწეროს

3 კანდელაკი, ნ. (2019). ბლოკჩეინ ტექნოლოგიის გამოყენების პერსპექტივები ელექტრონულ მმართველობაში. კრებულში: საქართველოს ტექნიკური უნივერსიტეტის შრომები. N 3(513). გვ. 11–18.

4 Harper D. ‘Etymology of mezzanine’. *Online Etymology Dictionary*. https://www.etymonline.com/word/mezzanine [02/05/2025].

ერთად. ცალკე დაწერილი „მეზანინ“, რომელსაც არ აქვს საზღვრული სახელისათვის დამახასიათებელი რაიმე აფიქსი, შეცდომაა. ტერმინი უნდა დაიწეროს შეერთებით – მეზანინდაფინანსება, ან დაერთოს ზედსართავი სახელის წარმოებისათვის საჭირო რომელიმე აფიქსი – მაგალითად -ური (მეზანინური დაფინანსება). ასევე შესაძლებელია მოიძებნოს ქართული შესატყვისიც: შუალედური დაფინანსება, შერეული დაფინანსება ან სხვა.

[3] წარმოებული კომპოზიტები, რომლებიც ერთად უნდა დაიწეროს: **მიკრო სესხი** – უნდა იყოს მიკროსესხი; **ბიზნეს სესხი** – უნდა იყოს ბიზნესსესხი; **ბიზნეს მოდელი** – უნდა იყოს ბიზნესმოდელი ან ბიზნესის მოდელი; **ქოლ ცენტრი** – უნდა იყოს ქოლცენტრი.

სხვა ენებიდან ნასესხები და ტრანსლიტერაციით გადმოტანილი ზოგიერთი ტერმინი არღვევს ტრანსლიტერაციის წესებს:

[4] აუთსორსინგი – outsourcing: უნდა იყოს აუ**ტ**სორსინგი

[5] დემპინგი – dumping: უნდა იყოს დამპინგი. ამ ფორმით ფიქსირდება ზოგიერთ შედარებით ახალ ლიტერატურაში. რამდენიმე ადრინდელ ლექსიკონში ფიქსირდება ქართული სინონიმი „გადასაყარი ექსპორტი“, თუმცა პრაქტიკაში მას არსად არ იყენებენ.

[6] დოუ ჯონსის ინდექსი – Dow Jones Average: უნდა იყოს დ**ა**უ ჯო**უ**ნ**ზ**ის ინდექსი.

[7] ფილიპსის მრუდი – The Phillips curve: უნდა იყოს ფილი**ფ**სის მრუდი.

[8] ტრესტი – trust: უნდა იყოს ტრასტი. ამ ფორმით გვხვდება ზოგიერთ შედარებით ახალ გამოცემაში.

ზოგჯერ ტერმინი ქართულ ენაში პირდაპირ ინგლისური სიტყვით აქვთ გადმოტანილი მაშინ, როდესაც არსებობს ქართული სიტყვებით უფრო გასაგებად გადმოცემის შესაძლებლობა. მაგალითად:

[9] ლექსიკონებში, ინტერნეტგამოცემებში და ლიტერატურაში გვხვდება ტერმინი **agile ტრანსფორმაცია** (ლათინური ასოებით[5]), რომელიც ორგანიზაციის საქმიანობის მოქნილ პრინციპებზე გადაყვანას გულისხმობს (agile *adj.* „მარჯვე, მოქნილი“). ქართულში agile-ს ან ტრანსლიტერირებულ ეჯაილს პირდაპირ შეესატყვისება „მოქნილი“, ანუ გვექნება ტერმინი „მოქნილი ტრანსფორმაცია“, მაგრამ მეორე ნასესხები სიტყვის ჩანაცვლებაც შეიძლება ბუნებრივი ვატიანტით – „მოქნილი გარდაქმნა“.

[10] რისკების მიტიგაცია – risk mitigation: დიდი ინგლისურ-ქართული ლექსიკონის მიხედვით, mitigation-ის ქართული თარგმანი არის შემსუბუქება, შემცირება. ტერმინი risk mitigation გულისხმობს რისკებისა და საფრთხეების გააზრებას, მათი არსებობის აღიარებას და შესაბამისი ზომების მიღებას შესაძლო შედეგების შემცირების მიზნით. ჩვენთვის ნათელია, რომ ამ ცნებას პირდაპირ შეესატყვისება „შემცირება“. ჩვენი აზრით ქართული ტერმინი უნდა იყოს „რისკების შემცირება“.

5 იგივე შემთხვევაა SWOT-ანალიზი. ჩვენი აზრით, ტერმინების ჩაწერა სხვა დამწერლობების გრაფემებით ქართულში არ შეიძლება.

[11] რითეილ (კომპოზიტებში) – retail: ქართულში დიდი ხანია მის შესატყვისად შემოღებულია „საცალო“, მაგრამ მაინც ხშირად გვხვდება არასწორი ტერმინი.

[12] დერივატივი – derivative: აღნიშნავს სხვადასხვა საბაზისო აქტივიდან წარმოებულ ფინანსურ ინსტრუმენტს და სასწავლო ლიტერატურაში სწორედ ამ სახელით გვხვდება: „წარმოებული ფინანსური ინსტრუმენტი“ (შდრ. derivative – „წარმოებული“ მათემატიკაში). ამის მიუხედავად, ლექსიკონებსა და ონლაინგამოცემებში მაინც დავაფიქსირეთ დერივატივი.

[13] ტერმინალური ღირებულება – terminal value: გულისხმობს რისამე ახლანდელ დროში დისკონტირებულ სამომავლო ღირებულებას. *terminal*-ს ქართულში შეესატყვისება „ბოლო, საბოლოო, დასკვნითი“, ამიტომ ჩვენი აზრით უმჯობესია გამოვიყენოთ „საბოლოო ღირებულება“ ნაცვლად „ტერმინალური ღირებულებისა“.

[14] სუბორდინირებული სესხი – subordinated loan: ეწოდება სხვებთან შედარებით ნაკლებად პრიორიტეტულ სესხს. ქართულში „სუბორდინირებული“ დაქვემდებარებულის მნიშვნელობითაა შემოსული გაცილებით ადრე. ინგლისურ subordinated-ის აქ გამოყენებული მნიშვნელობის ქართული თარგმანია „მეორეხარისხოვანი, ნაკლებად მნიშვნელოვანი“. შესაბამისად, მართებული იქნება ტერმინის „მეორეხარისხოვანი სესხის“ გამოყენება.

სამწუხაროდ გვხვდება ისეთი ტერმინებიც, რომლებიც თავისი შემადგენლობით აცდენილია აღსანიშნი ცნებისაგან. ხშირ შემთხვევაში ამის მიზეზია ინგლისური სიტყვების პირდაპირ, ან ისეთი მნიშვნელობით სესხება, რაც არანაირ კავშირში არ არის ტერმინის მნიშვნელობასთან. ჩამოვთვლით ზოგიერთ მათგანს:

[15] რისკების აპეტიტი // მადა – risk appetite: ტერმინი ქართულში ამ ფორმით გავრცელებულია მათ შორის ოფიციალურ დოკუმენტებში. აღნიშნავს რისკის ისეთ დონეს, რომელზეც წამსვლელია კონკრეტული ორგანიზაცია, რომ მიიღოს სასურველი უკუგება ან სხვა სარგებელი. ქართულ ტერმინში „მადა“ ან „აპეტიტი“ არ არის დაკავშირებული აღსანიშნ ცნებასთან. ჩვენი აზრით, „რისკების მისაღები დონე“ ან „რისკების ზღვარი“ უფრო მეტად შესაბამისი იქნებოდა ამ ინგლისური ტერმინისათვის.

[16] არაფულადი კონტრიბუცია – Payment-in-Kind: „კონტრიბუცია“ ქართულში მიღებულია სამხედრო-პოლიტიკური მნიშვნელობით: ის აღნიშნავს იმ ფულად გადასახადს, რომელსაც ომში დამარცხებული ქვეყანა უხდის გამარჯვებულ მხარეს“. თუმცა ინგლისური ენის გავლენით ქართულ ენაში „კონტრიბუციამ“ დღეს ჩაანაცვლა „წვლილი“[6] და contribution-ის ისეთი მნიშვნელობები, რომლებსაც ქართულში სხვა შესატყვისი ჰქონდათ (შესატანი, შეწირვა და სხვ.). Payment-in-Kind გულისხმობს გადახდას არა ფულით, არამედ საქონლით ან მომსახურებით. საფინანსო ტერმინის ქართული სწორი შესატყვისი ჩვენი აზრით იქნება „არაფულადი ანგარიშსწორება“, ნაცვლად კონტრიბუციისა. სხვა შემთხვევაში ასევე შესაძლებელია გამოყენებული იყოს „არაფულადი წვლილი“ (როდესაც საქმე არ უკავშირდება საფინანსო გარიგებას).

6 მარგალიტაძე, თ. (2023). *ფიქრები მშობლიურ ენაზე*, თბილისი. გვ. 17.

იშვიათად გვხვდება ისეთი შემთხვევები, როდესაც ახლოს მდომი ტერმინებიდან არასწორადაა შერჩეული ერთ-ერთი:

[17] საოპერაციო შემოსავალი // საოპერაციო მოგება – operating income: ამ შემთხვევაში თარგმნის პრობლემას იწვევს ის ფაქტი, რომ ინგლისურ ტერმინ income-ს საფინანსო სფეროში ორი ცნება შეესაბამება. ქართულად მათ ვთარგმნით, როგორც შემოსავალს და მოგებას. operating income გულისხმობს სხვაობას მთლიან მოგებასა და ამ მოგებისათვის გაწეულ საოპერაციო ხარჯებს შორის. შესაბამისად, ამ შემთხვევაში „საოპერაციო შემოსავლის“ გამოყენება არასწორია.

4.2 პარალელურად ხმარებული ფორმები

ზოგჯერ გვხვდება ერთი და იგივე ცნების აღსანიშნად ორი ან მეტი ტერმინის პარალელურად გამოყენება, რაც რიგ შემთხვევებში იწვევს ბუნდოვანებას, ერთი ცნებისათვის ადამიანთა სხვადასხვა წრეში განსხვავებული ტერმინების გაჩენა-დამკვიდრებას. ამიტომ ვფიქრობთ, რომ ასეთი ტერმინებიდან უნდა დარჩეს ერთ-ერთი.

[1] პერპეტუიტი[7] **// პერპეტუიტეტი**[8] – perpetuity: როგორც ჩანს, ავტორებს ტერმინები გადმოაქვთ იმის მიხედვით, თუ რომელი ენიდან აქვთ ნათარგმნი წიგნის ან სტატიის კონკრეტული ნაწილი (რუს. перпетуитет vs ინგლ. perpetuity). არსებობს ამავე პრინციპით მიღებული ტერმინი **ანუიტეტი** (annuity). ამიტომ ეს ორი ტერმინი ერთმანეთთან ჰარმონიაში უნდა იყოს მოსული: 1) ანუიტეტი – პერპეტუიტეტი ან 2) ანუიტი – პერპეტუიტი.

[2] ამორტიზირებული // ამორტიზებული ღირებულება – amortised cost: ტერმინის ორივე ფორმა ერთსა და იმავე დოკუმენტში დავაფიქსირეთ. დარგობრივ და სასწავლო ლიტერატურაში ასევე ორივე ფორმით გვხვდება. -ირება სუფიქსი ამ ტერმინში რუსულის გავლენას უნდა მივაწეროთ (შდრ. амортизирование). ქართულად ფუძე ტერმინის სწორი ფორმა იქნება ამორტიზება ან ამირტიზაცია[9], ამიტომ amortised cost-ის ქართული შესატყვისი უნდა იყოს ამორტიზებული ღირებულება.

[3] დისკონტირებული // დისკონტის // დისკონტირების განაკვეთი – discount rate: ქართულ წყაროებში სამივე ფორმით გვხვდება. დისკონტირება ტერმინია და შეესაბამება ინგლისურ discounting-ს, ამიტომ discount rate უმჯობესია ქართულად გადმოიცეს, როგორც დისკონტის განაკვეთი.

[4] ცვლადი // მცურავი საპროცენტო განაკვეთი – floating interest rate: ეს ისეთი საპროცენტო განაკვეთია, რომლის მოცულობა საპროცენტო პერიოდის განმავლობაში შეიძლება გადაიხედოს და შეიცვალოს. „მცურავი“ ინგლისური floating-ის ძი-

7 ასლანიშვილი, დ., ომაძე, ქ. რიკარდოს ეკვივალენტი ანუ სახელმწიფო ვალის წარმოშობისა და პრაქტიკული დანიშნულების აზრი. http://www.conferenceconomics.tsu.ge/?mcat=0&cat=arq&leng=ge&adgi=465 [09/05/2025]

8 ჭელიძე, ი. (2021). *მმართველობითი აღრიცხვა 2*, თბილისი: თსუ. გვ. 159.

9 ‘ამორტიზება’. *ganmarteba.ge*. https://www.ganmarteba.ge/word/ამორტიზება [18/05/2025]

რითადი მნიშვნელობის კალკია. ქართულად ტერმინის მიერ აღნიშნული ცნების არსს უკეთესად გადმოსცემს „ცვლადი“, რომელიც floating-ის ერთ-ერთი სხვა მნიშვნელობაც არის.

[5] ლიკვიდობა // ლიკვიდურობა – liquidity: ლიკვიდური ეწოდება ისეთ აქტივს, რომლის ფულად გადაქცევა მარტივადაა შესაძლებელი. ტერმინოლოგიური თვალსაზრისით ვხედავთ სიჭრელეს, ზოგჯერ ერთსა და იმავე წყაროში დაფიქსირებულია ლიკვიდურობის კოეფიციენტი და ლიკვიდობის გადაფარვის კოეფიციენტი, ლიკვიდურობის რისკი და ლიკვიდობის რისკი. ამგვარი ტერმინოლოგიური ვარიანტულობა დაუშვებელია. სწორ ფორმებად უნდა დარჩეს „ლიკვიდური“, „ლიკვიდურობა“ და აქედან წარმოებული მრავალსიტყვიანი შესიტყვებები – ლიკვიდურობის კოეფიციენტი, ლიკვიდურობის რისკი და სხვ.

[6] საანგარიშგებო // საანგარიშო პერიოდი – reporting period

[7] ბაზრის რისკი // საბაზრო რისკი – market risk: ერთი ფუძით და სხვადასხვა აფიქსოიდებით წარმოებული პარალელური ფორმები აგრეთვე ართულებს ტერმინოლოგიის ნორმირების საკითხს.

[8] საწარმოს ფუნქციონირების დაშვება // ფუნქციონირებადი საწარმოს პრინციპი – going concern

[9] ბანკინგი // საბანკო საქმე, საბანკო მომსახურება – banking: უცხოურ-ქართული ტერმინოლოგიური წყვილები, რომელსაც ენაში ხშირად ვხვდებით და არა მარტო საფინანსო სფეროში, ერთი მხრივ მიღებული ტერმინოლოგიური პრაქტიკაა, თუმცა მეორე მხრივ ხშირ შემთხვევაში ხელს უწყობს ქართული ან ანალიტიკური ტერმინების ენიდან გამოდევნას და ნასესხები ფორმებით ჩანაცვლებას (იხ. აგრ. თავი „ლიტერატურის მიმოხილვა“). იქ, სადაც ორი ან რამდენიმე სიტყვით უფრო გასაგებადაა შესაძლებელი აზრის ჩამოყალიბება, უმჯობესია გამოყენებულ იქნეს პრაქტიკაში მიღებული ანალიტიკური ტერმინები:

> ღია <u>ბანკინგი</u> ბევრ ახალ შესაძლებლობას ქმნის ბიზნესებისთვის; თუ
>
> ღია <u>საბანკო მომსახურება</u> ბევრ ახალ შესაძლებლობას ქმნის ბიზნესებისთვის?

[10] ფრანჩიზი // ფრანშიზი // ფრანშიზა // ფრენჩაიზი // ფრენშაიზი – franchise: პარალელური ფორმების ამგვარი სიჭრელე გამოიწვია ტერმინის სხვადასხვა ენიდან სხვადასხვა დროს შემოტანამ და ტრანსლიტერაციის ნორმების (და წარმოთქმის) გაუთვალისწინებლობამ. თავდაპირველად შემოსულია ფრანგულიდან რუსული ენის გავლით და ორი ფორმით გვხვდებოდა: ფრანშიზი და ფრანშიზა (ამ უკანასკნელში -ა რუსული ენის გავლენაა). ბოლო დროს იგივე ტერმინი ხელახლა გადმოაქვთ, მაგრამ უკვე ინგლისურიდან: ფრენჩაი ან დამახინჯებული ფრენშაიზი (ინგლისურშიც ფრანგულიდანაა შესული)

4.3 აბრევიატურები

ინგლისურ მრავალსიტყვიან ტერმინებს ხშირად აქვს შემოკლებები ან აბრევიატურები. ჩვენ მიერ მოძიებულ ტერმინოლოგიაში გვხვდება შემდეგი აბრევიატურები:

- Accounts Receivable (AR)
- Annual Percentage Rate (APR)
- Break-Even Point (BEP) Analysis
- Cash Conversion Cycle (CCC)
- Cost of Goods Sold (COGS)
- Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA)
- Gross National Product (GNP)
- Liquidity Coverage Ratio (LCR)
- Over-the-counter (OTC)
- Payment-in-Kind (PIK)
- Purchasing Power Parity (PPP)
- Shareholder Equity (SE)

ქართულ საფინანსო ტერმინებში სულ რამდენიმე დავაფიქსირეთ ისეთი, რომელსაც აბრევიატურა აქვს:

- მთლიანი შიდა პროდუქტი (მშპ)
- სააქციო საზოგადოება (სს)
- შეზღუდული პასუხისმგებლობის საზოგადოება (შპს)
- დამატებული ღირებულების გადასახადი (დღგ)
- მიკროსაფინანსო ორგანიზაცია (მისო)

აბრევიატურების არქონის გამო ქართულენოვან ტექსტებში ხშირად იყენებენ ინგლისურ აბრევიატურებს, ზოგჯერ ლათინური ასოებით დაწერილს. მაგალითად, შერწყმა და შეძენის (Mergers and Acquisition) შემოკლებად ქართულში გვხვდება M&A. First in, First Out ინგლისურ ტერმინს ქართულად შეესაბამება „პირველი შემოსავალში – პირველი გასავალში“, ტერმინის სიდიდის გამო ინგლისურში იყენებენ FIFO-ს, რაც ქართულში პირდაპირი სესხებით გადმოაქვთ – ფიფო (ეს ფორმა გვხვდება მათ შორის სასწავლო ლიტერატურაში). Foreign Exchange ანუ სავალუტო ბაზარი ინგლისურში შემოკლებულია როგორც Forex; ქართულში შემოკლებული ფორმა ასევე პირდაპირაა ნასესხები – ფორექსი. ქართულ ტერმინოლოგიაში აბრევიატურებისა და შემოკლებების სიმცირეს (ნაკლებობას) ერთ-ერთ პრობლემად გამოვყოფთ.

5 დასკვნა

ქართულ ენაში დარგობრივი ტერმინოლოგიის გამართულობა და სტანდარტიზაცია ერთ-ერთი აქტუალური საკითხია, განსაკუთრებით კი იმ დარგებში, სადაც ტექნოლოგიური აქტივობა და უცხოეთთან ინტეგრაცია გავლენას ახდენს ენობრივ საკითხებზე, იწვევს ცნებათა და ტერმინთა სწრაფ ცვლილებებს. ერთ-ერთი ასეთი დარგია ფინანსები, რომლის ტერმინოლოგია საკმაოდ დიდი ხანია განიცდის უცხო ენების გავლენას და ბოლო წლებში იგივე ტენდენცია შეინიშნება.

სტატიაში გამოყენებული იყო შერეული მეთოდოლოგია, რომელიც მოიცავდა როგორც ტრადიციულ (არსებული ლექსიკონებისა და სხვა დარგობრივი ლიტერატურის ანალიზი, ტერმინთა სტრუქტურულ-სემანტიკური კლასიფიკაცია, ტერმინთშემოქმედებითი მოდელების იდენტიფიცირება), ისე თანამედროვე ტექნოლოგიებზე დაფუძნებულ მიდგომებს, მაგალითად, SynchroTerm-ის გამოყენებას ორენოვანი კორპუსიდან ტერმინთა ამოსაღებად.

ტერმინთა ანალიზის შედეგად დადასტურდა უცხო ენათა დიდი გავლენა ქართულ საფინანსო ტერმინოლოგიაზე, რაც თავს იჩენს როგორც სტრუქტურულ, ასევე ლექსიკურ-სემანტიკურ დონეზე. ცხადია, რომ ქართულ საფინანსო ლექსიკაში მნიშვნელოვანი ადგილი უკავია რუსული და ინგლისური ენებიდან შემოსულ ტერმინებს. ერთი მხრივ, ეს განპირობებულია რუსული ენის მომძლავრებული გავლენით თითქმის ორი საუკუნის განმავლობაში (და რაც ძალიან საინტერესოა, ეს გავლენა ახლანდელ პროცესებშიც იგრძნობა), მეორე მხრივ, დასავლურ სამყაროსთან ეკონომიკური ინტეგრაციის შედეგად უკვე ინგლისურენოვანი გავლენის გაძლიერებით (თუმცა აუცილებლად უნდა აღინიშნოს, რომ რუსულის გავლით ნასესხები ტერმინებიდან პრაქტიკულად ყველა ევროპულია).

რუსული ენის გავლენა განსაკუთრებით თვალსაჩინოა ისეთი ტერმინების სტრუქტურაში, რომლებიც კალკირების გზით მივიღეთ (მაგ., ძირითადი საშუალებები, საბალანსო ღირებულება, პრივილეგირებული აქცია და სხვ.), ინგლისურის გავლენა კი უფრო მეტად გამოიხატება ბოლო წლებში ნასესხებ ისეთ ტერმინებში, როგორიცაა ჰეჯირება, ლევერიჯი, დერივატივი, ფიუჩერსი, ფრენჩაიზი და სხვ.

კვლევის უშუალო შედეგებმა გამოავლინა რამდენიმე ტიპის მნიშვნელოვანი პრობლემური საკითხი უშუალოდ ტერმინთა დონეზე:

- ტერმინოლოგიური სიჭრელე, კერძოდ ერთსა და იმავე ცნებას ხშირად შეესაბამება რამდენიმე განსხვავებული ფორმა, რაც არღვევს ტერმინოლოგიის უნიფიცირების პრინციპს;
- ტერმინთშემოქმედებისას ქართული ენის მორფოლოგიური ნორმების უგულებელყოფა, მათ შორის მრავალწევრიანი ტერმინების არასწორად ჩაწერა, უცხო ენების ტრანსლიტერაციის ნორმების დაუცველობა;
- უცხოური ტერმინების სესხება ისეთ შემთხვევებშიც კი, სადაც ამის საჭიროება არ დგას;
- იმგვარი სესხება, რომელიც რიგ შემთხვევებში ხელს უშლის ტერმინის შინაარსობრივ იდენტიფიკაციას.

გამოყენებული ლიტერატურა

ლეჟავა, ვ. (2022). *ფინანსურ ტერმინთა ლექსიკონი*. თბილისი: თავისუფალი და აგრარული უნივერსიტეტების გამომცემლობა.

მარგალიტაძე, თ. (2018). ტერმინოლოგიური პოლიტიკის საკითხისათვის საქართველოში (პლენარული მოხსენება). კრებულში: *საერთაშორისო კონფერენცია ჰუმანიტარული მეცნიერებები ინფორმაციულ საზოგადოებაში - III*. ბათუმი: ბათუმის შოთა რუსთაველის სახელობის სახელმწიფო უნი*ვერსიტეტი.*

მარგალიტაძე, თ. (2023). ფიქრები მშობლიურ ენაზე. თბილისი.

მელაძე, გ. (2024). ენობრივი ცვლილების ბუნებისათვის (ქართული ნეოლოგიზმების მაგალითზე). საერთაშორისო კონფერენციის კრებულში: *ლექსიკოგრაფია 21-ე საუკუნეში,* თბილისი: ილიას სახელმწიფო უნივერსიტეტი.

პაპავა, ვ. (2014). თანამედროვე ქართული ეკონომიკური ტერმინოლოგია: ახალი პრობლემები და ძველი შეცდომები. კრებულში: *ტერმინოლოგიის საკითხები I.* თბილისი: თსუ არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი.

საღინაძე, რ. (2022–2023). ქართული ტერმინოლოგიური მუშაობის ასპექტები. კრებულში: *ტერმინოლოგიის საკითხები V*. თბილისი: თსუ არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი.

სუთიძე, ლ., იაკობაშვილი, გ. (2016). ტერმინოლოგიის პრობლემები ტექნიკურ სასწავლო ლიტერატურაში. კრებულში: *ტერმინოლოგიის საკითხები II*. თბილისი: თსუ არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი.

სუხიშვილი, თ. (2020). ტერმინოლოგიური პრობლემები ქართულ თარგმანში. კრებულში: *ტერმინოლოგიის საკითხები IV*. თბილისი: თსუ არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი.

ქაროსანიძე, ლ. (2022–2023). ქართული ტერმინთბანკი და ტერმინოლოგიური პოლიტიკა საქართველოში. კრებულში: *ტერმინოლოგიის საკითხები V*. თბილისი: თსუ არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი.

ქაროსანიძე, ლ., ხურცილავა, ა. (2018). *ქართული ტერმინთშემოქმედების პრობლემები: ისტორია და თანამედროვეობა,* თბილისი: თსუ არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი.

ხურცილავა, ა. (2016). თანამედროვე ქართული საბანკო ტერმინოლოგიის პრობლემები. კრებულში: *ტერმინოლოგიის საკითხები II*. თბილისი: თსუ არნ. ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი.

Developments in the Structure of Dictionaries of Foreign Words in Slovak

Renáta Panocová
Pavol Jozef Šafárik University in Košice
E-mail: renata.panocova@upjs.sk

Abstract

Slovak dictionaries developed significantly over the course of the past century. There is a long tradition of dictionaries of foreign words, inspired in part by German and Slavic examples. Since 1922, when the first dictionary of foreign words was published, more than forty dictionaries of this type have appeared. Such a quantity is impressive as it implies that on average, a new dictionary or revised edition was published nearly every two years. The main aim of this paper is to investigate the development in this genre of dictionaries. In order to do so, I compared five dictionaries of foreign words in Slovak SCSVS (1939), SCS (1953), SCS (1966), VSCS (1997) and SCSA (2005) with a central focus on their properties such as macrostructure, microstructure, intended users, and medium. A key selection criterion was that each of the five dictionaries represents a different historical period, which can be assumed to influence these properties. The results of the analysis are compared with the situation in German and English in the respective periods. This comparison determines the position of the Slovak tradition of dictionaries of foreign words in the broader European context.

Keywords: loanwords; dictionary definitions; native and non-native suffixes

1 Understanding of foreign words

Dictionaries of foreign words are monolingual dictionaries that are primarily designed for adult native speakers. In some cases, the title of a dictionary of foreign words implies that it is aimed at school pupils or university students. The most frequent reasons to use monolingual dictionaries are for the quick verification of orthography and approximate primary meaning (Nesi 2013). Although these findings concern monolingual dictionaries of general language, it can be assumed that understanding of foreign words is often a reason to use dictionaries of foreign words.

The origin of dictionaries in Europe was connected with the need to understand foreign words. Starting from the Middle Ages glosses of difficult words were included in Latin manuscripts. The glosses appeared in simple Latin, in the vernacular language or in both (Béjoint 2016: 9-10). Later, they were collected in separate, alphabetically ordered word

lists called *glossaries*. The glossaries were not dependent on any particular text. Then, glossaries were gradually expanded and included more information. In this way, glossaries developed into dictionaries. In Europe, bilingual dictionaries preceded monolingual dictionaries. Van Sterkenburg (2003: 9) further explains that the main reason was that people in Europe had edited more basic texts in foreign languages (in particular Latin) than in their own. This situation started changing from the sixteenth century when vernacular languages were used more frequently and in all contexts. This resulted in a more urgent need for a monolingual dictionary to be used by people in their professional lives and cultural activities (Béjoint 2016: 10).

The situation for English is characterized as "the hard word tradition" by Osselton (1983: 15). The first monolingual English dictionary *A Table Alphabeticall, conteyning and teaching the true writing and understanding of hard vsvall English wordes* by Robert Cawdrey contains 2,500 entries for hard words. It was published in 1604. Entries in Cawdrey's dictionary such as *obdurate*, or *obeisance* represent words that were hard to understand. In contrast, common words such as *oak tree*, or *obey* were not included. Osselton (1983: 16) explains that understanding the meaning of hard words was crucial at that time in order to "enable a wider, unlatined, reading public to understand and to learn to use the new technical and abstract vocabulary of learned words, which in many cases thus became less 'hard' and were assimilated into the language". The main functions of the earliest English dictionaries up to 1750 were cultural and educational. Béjoint (2000: 94) explains that especially many children, women and foreigners did not have access to education and dictionaries of hard words were often used as self-teaching tools. This changed in the eighteenth century when dictionaries "came to be seen as a scholarly record of the whole language" (Osselton 1983: 17). Here English followed languages such as Italian and French. It is understood that such a change needs some time to develop. Durkin (2016: 3) observes that "it was only very slowly that the general dictionary of a language emerged as a fundamental type, alongside the specialist lexicon of hard words, or the technical vocabulary of a particular field, and alongside the bilingual dictionary".

Dictionaries of foreign words in German have a deep-rooted tradition of more than two hundred years. Heier (2012) gives a detailed account of the development of the lexicography of foreign words from 1800 to 2007. Hausmann (1985: 391) emphasizes two main functions of dictionaries of foreign words in German. On one hand, in line with Osselton (1983), it is the explanation of difficult words. On the other hand, it is a purist attempt to replace these words by words of native origin. Wiegand (2001: 59) prefers to view dictionaries of foreign words in German from a perspective of language contact dictionaries.

Dictionaries of foreign words in Slovak have a well-established tradition of more than one hundred years. The Slovak linguist, lexicographer and translator Peter Tvrdý (1850-1935) compiled the first dictionary of foreign words in Slovak in 1921. The first edition was published in 1922 and a revised and expanded edition appeared in 1932. The 1922 edition has 126 pages. Jóna (1961: 25) describes this dictionary as purely practical. In compiling it, Tvrdý was inspired by Czech, German and Hungarian encyclopaedic dictionaries (Jóna 1961: 25). Jarošová (2001: 28) emphasizes that Tvrdý provided a number of Slovak equivalents of foreign words that were missing in the Slovak vocabulary at that time. In mapping the history of Slovak dictionaries Hayeková (1992: 25) explains that the need for dictiona-

ries of foreign words increased after the Second World War, when borrowed words were often misused. In the Slovak context, an interest in language cultivation, which includes knowledge of words of non-native origin, was one of the key factors that contributed to the popularity of dictionaries of foreign words especially starting from the 1970s (Jarošová 2006: 423). On the basis of the intended user, Mistrík (1993: 406) distinguishes dictionaries of foreign words for the general public, academic users, and school users. The main functions of these types of dictionaries of foreign words are, in line with the English tradition described above, educational and cultural.

Against this background, a central aim of this paper is to investigate the developments in the structure of dictionaries of foreign words in Slovak over the past one hundred years. These developments are analysed in a sample of loanwords with clearly recognizable suffixes. Such loanwords with the same suffix tend to occur in larger sets. This means that speakers recognize their similarity and regularity, which may lead to a generalization and increase the degree of acceptability. The rest of the paper is organized in three sections. Section 2 gives a detailed account of dictionaries of foreign words in Slovak in the past one hundred years. Section 3 focuses on the representation of suffixed loanwords in the selection of Slovak dictionaries of foreign words. The final section summarizes the main findings and formulates some generalizations.

2 Slovak dictionaries of foreign words in the past one hundred years

In the Slovak lexicographic tradition, Mistrík (1993: 406) considers the dictionary of foreign words the most commonly used type of dictionary. It is a monolingual dictionary which explains the meaning of words and terms that were borrowed in Slovak. The main criteria for the explanation of meanings of loanwords are clarity, intelligibility, and accessibility to general users. In the case of specialized vocabulary items, explanations tend to be closer to regular dictionary definitions. Dictionary entries usually include information about the source language of a loanword, basic grammatical information, pronunciation especially if it differs considerably from the one predictable from the orthography, and style or register in which a loanword is used. Some dictionaries also give orthographic variants of a loanword reflecting the degree of adaptation in Slovak at the time of publication. The size of the dictionaries of foreign words in Slovak typically ranges from 10,000 to 60,000 words, but dictionaries of smaller (pocket) size have also been published.

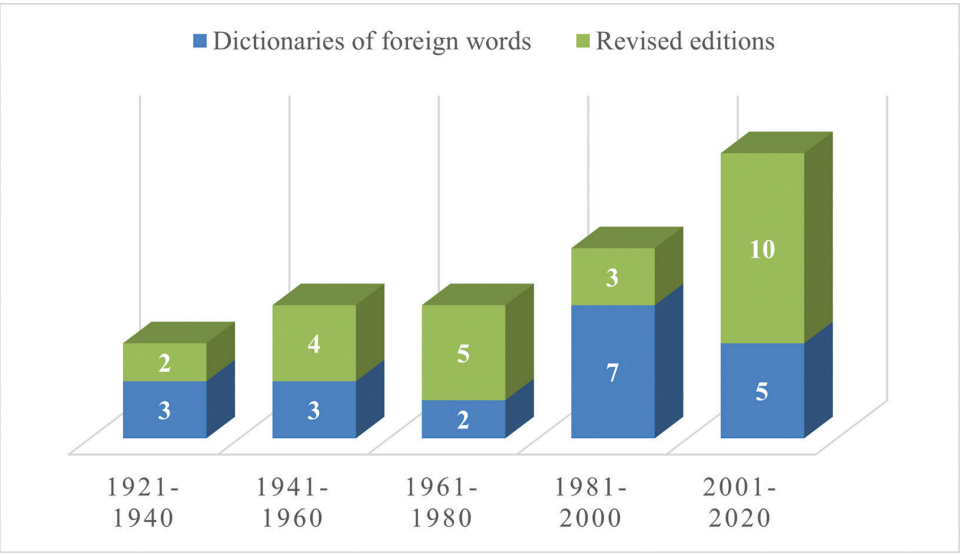


Figure 1: Dictionaries of foreign words in Slovak published by period

Figure 1 gives an overview of the number of dictionaries of foreign words in Slovak that were published during five twenty-year periods starting from 1921. Since 1922, when the first dictionary of foreign words was published, more than forty dictionaries of this type have appeared. This number includes revised editions of previously published dictionaries. It is obvious that the number of dictionaries of foreign words increased considerably in the most recent periods. While in 1921-1940 one dictionary appeared every 4 years on average, in 2001-2020 one dictionary was published every 16 months. The number of such dictionaries is striking when we take into account the number of speakers of Slovak, amounting to slightly over 5 million potential dictionary users (Statistical Office of the Slovak Republic 2025).

For the analysis, I selected one dictionary of foreign words from each of the five periods, SCSVS (1939), SCS (1953), SCS (1966), VSCS (1997) and SCSA (2005). A key criterion was that each of the five dictionaries represents a different historical period. This was assumed to have an impact on the inclusion of dictionary entries and their description. Before proceeding to the analysis, I will briefly describe each dictionary in the selection.

SCSVS (1939) was compiled by Jozef M. Prídavok. It was the first large dictionary of foreign words in Slovak. The dictionary has 704 pages and around 35,000 entries. The difference is considerable when we compare SCSVS (1939) with the very first dictionary of foreign words by Tvrdý (1922), which has 126 pages. Despite quite significant expansion in the second revised edition, Tvrdý (1932) has 216 pages, so that SCSVS (1939) stands out as a large dictionary. The foreword to SCSVS (1939) was written by Henrich Bartek (1907-1986), a prominent Slovak linguist, who helped Prídavok in finalizing the dictionary. Bartek views SCSVS (1939) as a continuation of the work started by Tvrdý. According to Bartek, the main advantage is the large number of appropriate Slovak equivalents offered for foreign words. At that time, a full-scale monolingual dictionary of contemporary Slovak was not available yet. This

means that SCSVS (1939) to some extent compensated for a missing monolingual dictionary. SCSVS (1939) contains a section with basic pronunciation rules in different languages from which loanwords are included in the dictionary. Each entry includes information about the origin of the loanword, sometimes also from which language it was borrowed into Slovak. For some loanwords, the pronunciation is provided in brackets as part of the dictionary entry. This can be illustrated by the English loanword *out-fighting* (aut fajtin) used in boxing or the French word *nouveau* (nuvó). SCSVS (1939) reflects a degree of adaptation of some loanwords, e.g. there is a separate entry for *orangeade* (oranžád) from French with a cross-reference to the orthographically adapted entry *oranžáda*.

SCS (1953) was edited by Anton Hollý and Ľudovít Zúbek. The editors collaborated with experts from different fields. SCS (1953) has 1052 pages and includes approximately 16,000 commonly used loanwords. The editors were inspired by Russian and Hungarian dictionaries of foreign words. The dictionary entries give information about the pronunciation where necessary as well as about the origin of the loanwords. In her review, Šalingová (1954: 95-96) gives a critical account of the unsystematic and therefore sometimes misleading information about pronunciation. For example, *chevrette* borrowed from French does not give the pronunciation. For another French loanword *ateliér* there is a note (vyslov 'pronounce' atelié) despite the meanwhile established adapted pronunciation *ateliér*. The dictionary is not only linguistic in nature but also encyclopaedic (Mistrík, 1993: 406). This explains that despite the larger number of pages, it has fewer entries than SCSVS (1939). Šalingová (1954: 97) observes that the structure of individual entries does not follow any systematic pattern. Considerable differences occur in the length and quality of explanations. For Šalingová (1954: 97) it is obvious that the number of contributors was large, which may be seen as an advantage, but the final outcome is an inconsistent representation of entries. The editors applied a new approach to processing of word formation constituents used productively in other formations such as *chrono-*, *mega-*, or *chryzo-*, which Šalingová (1954: 98) sees positively. In Šalingová's overall evaluation of SCS (1953), the advantages outnumber the deficiencies.

SCS (1966) was edited by Samo Šaling, Mária Šalingová (aka Ivanová-Šalingová) and Ondrej Peter. This was the second, revised edition of a dictionary of foreign words first published in 1965. The third edition appeared in 1970, which shows the intensity of the effort of the editorial team. At the same time, the influx of new loanwords at these times is evident as significant revisions were introduced in a relatively short time. The 1966 dictionary includes around 27,000 entries and has 1139 pages. At the time of publication, it was the largest dictionary of foreign words (Michalus, 1966: 236) and one of the few to apply modern lexicographic methods. In his review, Michalus (1966: 236) emphasizes the careful selection of entries and the exclusion of loanwords that Slovak speakers do not perceive as foreign, e.g. *cisár* ('emperor'), *koruna* ('crown'), *kráľ* ('king'). Among its strengths is the precise and succinct explanations of meaning. This contrasts with the approach in SCS (1953). Michalus (1966: 236) observes numerous loanwords that were not included in previously published dictionaries of foreign words, e.g. *beatnik*, *bestseller*, or *laser*. He makes only very few minor critical remarks. One of them concerns the imprecise information about currencies. The information about *libra* ('pound') as a currency in Australia was no longer valid due to the change to the Australian dollar in 1966, when SCS was published. SCS (1966) gives separate

entries for homonyms. Nesting and grammatical information are reduced to a minimum, which is in line with the purpose of this type of dictionary.

In VSCS (1997) the editors are again Samo Šaling and Mária Ivanová-Šalingová, now in collaboration with Zuzana Maníková. The dictionary is an expanded and thoroughly revised edition of the *Slovník cudzích slov* A/Z (1979, 1983 2nd edition, 1990 3rd edition) ('Dictionary of foreign words A/Z') edited by the same editorial team. VSCS (1997) has approximately 60,000 entries and 1310 pages. Four substantially revised editions were published in 2000, 2003, 2006 and 2008. The 2008 edition has more than 70,000 entries, which makes it a much larger dictionary of foreign words than earlier ones.

SCSA (2005) has a special status among the dictionaries in the selection, because it is a revised and expanded translation of a Czech dictionary, the *Akademický slovník cizích slov* ('Academic dictionary of foreign words'), published in 1995. Jiří Kraus was the main editor of the original Czech dictionary. The editors of the Slovak translation, revision and expansion were Ľubica Balážová and Ján Bosák. The first Slovak edition was published in 1997, the second appeared in 2005. It includes approximately 100,000 meanings of commonly used words, abbreviations and quotes of foreign origin. Loanwords typically used in Czech but not in Slovak were omitted and loanwords used currently in Slovak were added. Balážová and Bosák (1998: 344) describe that the use of a Czech dictionary as a basis was motivated on one hand by the lack of capacity for a completely new Slovak dictionary, on the other hand by the quality of the methodology adopted by Kraus. Kraus's team collaborated with a large database of consultants, around one hundreds experts from various fields, and designed a concise lexicographic structure of dictionary entries including frequent examples of their use and collocations. In her review of the first edition, Koncová (1997) observes a considerable increase of loanwords formed with neoclassical formatives in the first 1997 edition. In comparison with the Czech dictionary that includes 35 loanwords with *eko-* ('eco-'), the 1997 Slovak edition has approximately 60.

3 Representation of suffixed loanwords

For the analysis, I selected a section of 25 pages from SCSVS (1939), starting at Ni and ending with the first two pages covering the letter P. For the other 4 dictionaries, the corresponding section was selected. All omissions and new inclusions were recorded. Table 1 lists the number of loanwords in each dictionary of foreign words in the selection.

Dictionary of foreign words	Number of loanwords in the selection
SCSA (2005)	1,689 words
VSCS (1997)	1,206 words
SCSVS (1939)	1,094 words
SCS (1966)	928 words
SCS (1953)	633 words

Table 1: Number of loanwords in the selection of dictionaries of foreign words in Slovak.

In Table 1, the number of words in the sample in SCSVS (1939) is in the middle, with the two dictionaries published during communist times having fewer and the two latest dictionaries more entries. I analysed the entries from the perspective of their origin. For a more detailed analysis, I focused on loanwords that have a clearly recognizable suffix. The main reason was that in the case of larger sets of loanwords with the same suffix, speakers tend to recognize their similarity and regularity, which may lead to a generalization and increase the degree of acceptability. It may also lead to the emergence of competing, more native-like variants. The most frequent suffixes are *-ácia* ('-ation'), for instance, in *notifikácia* ('notification'), *-izmus* ('-ism') as in *noktambulizmus* ('noctambulism'), and *-ita* ('ity') as in *nudita* ('nudity'). These examples show that Slovak and English equivalents are very similar and easily recognisable. This influence of Latin and Greek can be observed not only in Slovak, but also in other European languages. A large number of words have been adopted in several languages with little modification. Such words usually formed on the basis of Greek or Latin roots are called *internationalisms*. In the sample, loanwords with the elements from Latin represent the largest class in all dictionaries except for SCSA (2005) and loanwords with the elements of Greek origin are the second largest class.

In the next step, I reduced my selection to suffixed loanwords of non-native origin for which a competing word with a native suffix is available. This is illustrated in Table 2.

Non-native suffix	Native counterpart	Examples	
-ácia	-nie	Oscilácia	oscilovanie
-izmus	-stvo	nomádizmus	nomádstvo
-ita	-osť	objektivita	objektívnosť

Table 2: Loanwords with native and non-native suffixes

Table 2 displays three pairs of suffixed loanwords, one with a non-native suffix and the other one with a suffix, which is a native counterpart to the former. The two nouns are more or less synonymous. The suffixes *-ácia*, *-izmus*, and *-ita* tend to have equivalents in a number of other languages and are part of what is called international vocabulary. Doublets of the type in Table 2 are quite frequent and the nouns with a native suffix reflect a greater degree of nativization in Slovak. The data set includes a total number of 494 such loanword pairs with these three suffixes.

In the next step, the explanations used by lexicographers for the meanings of loanwords with the competing suffixes we saw in Table 2 were investigated. Their analysis in the dictionaries in the selection shows that the explanations can be divided into four basic types. Panocová (2024) gives a more detailed discussion of individual types in relation to suffixed loanwords. Type 1 is illustrated in (1).

(1) SCSVS (1939) nikotinizmus ('nicotinism') otrava nikotínom ('poisoning by nicotine')

In (1), the definition gives a brief description explaining the meaning of the loanword. The dictionary definition in (1) is an example of a conventional definition. Here, SCSVS (1939) does not provide any additional information, e.g. about grammatical gender, inflection, origin. However, in most other entries SCSVS (1939) typically includes at least information about the source language of loanwords. Suffixed loanwords explained by a definition of Type 1 usually do not have corresponding one-word Slovak equivalents. This increased the chances of the use of such loanwords by speakers of Slovak. The use of loanwords is also supported by the principle of economy, because the alternative would involve using a longer descriptive phrase.

Type 2 includes definitions of suffixed loanwords with a direct native equivalent as presented in (2).

(2) SCS (1966)	observácia	-ie ž. <lat.> odb.	pozorovanie, dohľad
	(‘observation’)	gen.sg fem. <Latin.> spec.	(‘observation, surveillance’)

The example in (2) represents many cases in which also the base is a native equivalent. In (2), the noun *pozorovanie* (‘observation’) is based on the verb *pozorovať* (‘observe’). In such instances, the use of the foreign word is stylistically marked. The native words tend to be more frequent in everyday communication. The dictionary entry illustrated in (2) provides information about the Latin origin of the suffixed loanword. Inflectional forms (genitive singular) and grammatical gender are included as a standard in the lexicographic representation of SCS (1966) and SCSA (2005). From the entry in (2), we can also read that *observácia* (‘observation’) is an example of specialized vocabulary.

Type 3 covers definitions as in (3), with a variant formed by means of a competing native suffix attached to the same base as the foreign word.

(3) SCSA (2005)	pacifikácia	-ie ž. <l>	obnova mieru, pacifikovanie
	(‘pacification’)	gen.sg fem. <Latin.>	(‘restoration of peace, pacifying’)

Entries of this type have the word with the native suffix as the last word in the definition, as an alternative to the preceding description. Dictionaries of foreign words published later, 1997 and 2005, tend to present such a variant with a native suffix more frequently in their definitions.

Type 4 includes cases when a dictionary of foreign words has both an entry for the loanwords and a separate entry for the variant with the competing native suffix. This is illustrated in (4) and (5).

(4) SCSA (2005)

nitrocementácia	-ie ž. <g+l> hut.	nitrocementovanie, kyanovanie
(‘nitrocementation’)	gen.sg fem. <G+L> metallurgy	(‘nitrocementing, cyanidation’)

In (4), the loanword with the native suffix appears in the definition, but it also has a separate entry in SCSA (2005). The loanword *nitrocementácia* (‘nitrocementation’) is of mixed Greek and Latin origin, the first component *nitro-*, related to *nitrogen*, is of Greek origin and

the second component *cement-* is of Latin origin. In SCSA (2005), the loanword with the native suffix has the entry given in (5).

(5) SCSA (2005)

nitrocementovanie -ia s. <g+l> hut. sýtenie povrchu ocele súčasne dusíkom a uhlíkom, nitrocementácia, kyanovanie

('nitrocementating') gen.sg neut. <G+L> metallurgy ('saturation of the surface of steel simultaneously by nitrogen and carbon, nitrocementation, cyanidation')

Both variants, with *-ácia* in (4) and with *-nie* in (5), are terms used in the field of metallurgy, which motivates listing them as two independent entries. Sometimes two entries are necessary due to semantic specialization. However, (4) and (5) are synonyms. The fact that the entry for the word with the native suffix contains the explanation of the meaning indicates that it is the preferred form. The entry in (4) can be interpreted as a referring entry.

The historical development of the use of the different types of definition can be represented as in Figure 2. Here all instances of a particular type are taken as one hundred per cent and the proportion of them occurring in each dictionary gives the value on the vertical scale.

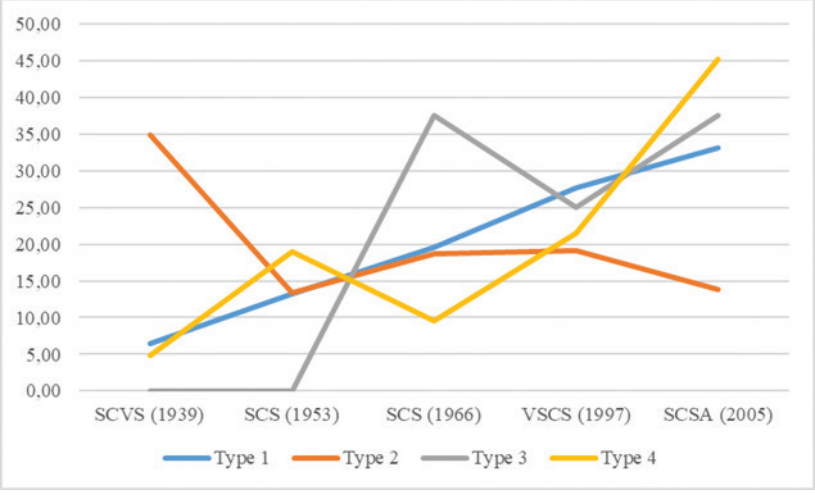


Figure 2: Development of types of definitions.

Dictionary definitions of Type 1 provide a concise descriptive phrase explaining the meaning of the loanword. Figure 2 shows that of all the occurrences of Type 1, nearly 7 per cent is in the oldest dictionary SCSV (1939) and about 33 per cent in the latest one SCSA (2005). The line shows a steady increase over time. This contrasts with the situation for Type 2. Type 2 includes definitions of suffixed loanwords with the help of a direct native equivalent. This type is very frequent in SCSV (1939), but then it declines sharply. More modern dictionaries tend to use it less frequently. This development reflects the emerging insight that synonyms are generally not ideal as definitions. Type 3 includes definitions with a variant formed with the same base but a competing native suffix. Figure 2 shows that Type 3 is not

used before 1966. The emergence of this type is important because it shows that the variant with a native suffix is perceived as less foreign and therefore suitable for explaining the meaning of the loanword with a non-native suffix. Alternatively, it might be a suggestion to replace the less adapted form by one with a native affix. The use of Type 3 can be linked to the more frequent use of words with a foreign base and a native suffix among Slovak speakers at the time of the compilation of this dictionary. The majority of examples are for nouns in *-ácia*. Type 4 includes cases where a dictionary of foreign words has entries both for the loanword and for the variant with a competing native suffix. The difference from Type 3 is that the word used as an explanation in Type 3 gets a separate entry in Type 4. The line in Figure 2 indicates that Type 4 is relatively rare in earlier dictionaries. Almost half of the occurrences are in the SCSA (2005) dictionary. A slight gradual increase in the use of Type 4 in more recent dictionaries indicates the relationship between these variants, which can be synonymous or semantically or stylistically differentiated. The resulting proportions of the definition types for each individual dictionary are not directly represented in Figure 2. Figure 3 shows them for the first and the last dictionary of my sample.

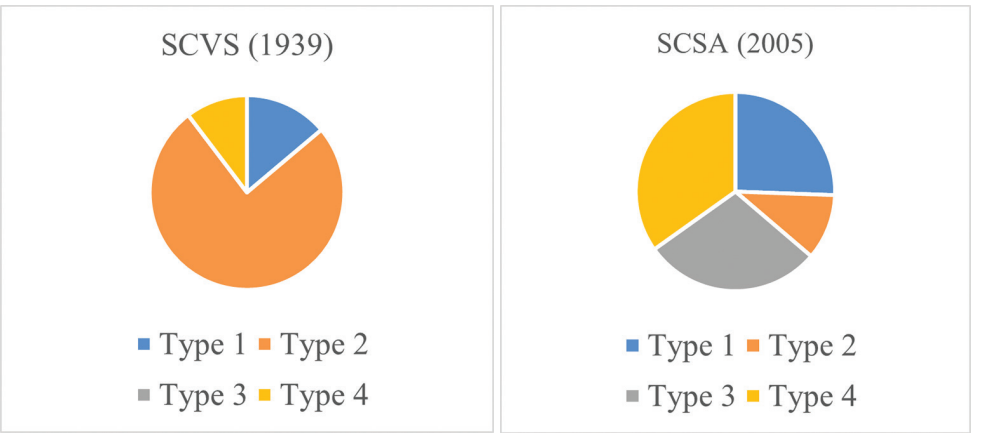


Figure 3: Comparison of definition types in SCVS (1939) and SCSA (2005).

On the left in Figure 3, we see how the different definition types were used in the oldest dictionary SCSV (1939) in the sample. This contrasts with the corresponding distribution for the latest dictionary SCSA (2005) on the right in Figure 3. The most striking difference between the two dictionaries concerns Type 2. Providing a synonym with a native base is by far the most common one in SCSV (1939), accounting for nearly 75 per cent of the cases. However, in SCSA (2005), Type 2 accounts for only a small minority of definitions, less than 14 per cent. A significant increase in the use of the Type 4 strategy, i.e. treating the words with foreign and with native suffixes in separate entries, can be observed in SCSA (2005). In addition, Type 3, which includes the word with the foreign base and the native suffix as part of the definition, did not occur in the older dictionary. On the other hand, this type is quite frequent in SCSA (2005). This development in Figure 3 reflects increasing insight in lexicographic theory. Synonyms are problematic as definitions. However, giving a word with the same base and a native suffix creates awareness of the word formation processes.

Treating these cognate forms in separate entries can be considered the most systematic presentation.

4 Conclusion

The main aim of this paper was to explore the developments in the structure of dictionaries of foreign words in Slovak over the past one hundred years. In this period, more than forty dictionaries were published. For the analysis, the selection included five dictionaries: SCSVS (1939), SCS (1953), SCS (1966), VSCS (1997) and SCSA (2005). Each of the five dictionaries represents a different historical period. The main function of the oldest dictionary was to explain difficult words. At this time, there was no monolingual dictionary of Slovak. This meant that SCSVS (1939) played an important role in providing native Slovak equivalents to foreign words. This corresponds to the observation that in many European languages, early dictionaries were not intended as full accounts of the vocabulary, but rather as dictionaries of hard words. SCS (1953) combined understanding of foreign words with expert knowledge mapped into encyclopaedic explanations. From the perspective of lexicographic methodology, early dictionaries were less consistent and less systematic. This changed with SCS (1966), which provides a good model of modern lexicographic practice. Distinctions between foreign words and adapted words in terms of orthography, pronunciation and perception of a degree of foreignness by Slovak speakers were elaborated in more detail. VSCS (1997) and SCSA (2005) represent a further stage in the qualitative development of this type of specialized dictionaries.

The analysis of the sample of loanwords with the clearly recognizable suffixes *-ácia* ('-ation'), *-izmus* ('-ism'), and *-ita* ('ity') illustrates this development. Four types of dictionary definitions were recognized. The use of these types in Slovak confirms the gradual integration of foreign words into present-day Slovak. This tendency is obvious in the emergence of competing equivalents combining a non-native base with a native suffix, synonymous with a non-native one. This indicates that large sets of suffixed loanwords are more likely to be perceived as analysable by Slovak speakers. It can also lead to the emergence of competing, more native-like variants. Therefore their lexicographic treatment in separate entries seems an appropriate and systematic solution.

References

Béjoint, H. (2000). *Modern Lexicography: An Introduction*. Oxford: Oxford University Press.

Béjoint, H. (2016). Dictionaries for General Users: History and Development; Current Issues. In Durkin, P. (ed), *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, pp. 7-24.

Durkin, P. (2016). 'Introduction'. In Durkin, P. (ed), *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, pp. 1-4.

Hausmann, F. J. (1985). Lexicographie. In Schwarze, Ch. and D., Wunderlich (eds), *Handbuch der Lexikologie*. Königstein/Ts.: Athenäum, pp. 367-411.

Hayeková, M. (1992). *Dejiny slovenských slovníkov II 1946-1975*. [History of Slovak dictionaries II 1946- 1975]. Bratislava: Filozofická fakulta Univerzity Komenského.

Heier, A. (2012). Deutsche Fremdwortlexikografie zwischen 1800 und 2007: Zur metasprachlichen und lexikografischen Behandlung äußeren Lehnguts in Sprachkontaktwörterbüchern des Deutschen. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110282672

Jarošová, A. (2001). Poznámky k lexikografickej koncepcii Petra Tvrdého. [Remarks on lexicographic conception of Peter Tvrdý]. In: S. Ondrejovič, K. Fircáková, D. Lechner (eds.) *Peter Tvrdý. Zborník zo seminára k 150. výročiu narodenia*. Bratislava: Univerzitná knižnica – Jazykovedný ústav Ľ. Štúra SAV 2001, pp. 27-34.

Jarošová, A. (2006). Slovak Lexicography. In Brown, K. (ed) *Encyclopedia of Language and Linguistics*. Volume 11, 2nd edition. Oxford: Elsevier, pp. 422-423.

Jóna, E. (1961). Slovenský slovnikár Peter Tvrdý (1850-1935). [Slovak dictionary-maker Peter Tvrdý (1850-1935)]. In: *Slovenská reč* 26(1), pp. 18-31.

Koncová, M. (1997). Nad novým Slovníkom cudzích slov. In *Slovenská reč* 62(3), pp. 357-361.

Michalus, Š. (1966). S. Šaling, M. Šalingová, O. Peter, Slovník cudzích slov. Slovenské pedagogické nakladateľstvo, Bratislava 1965, 1140 strán. [S. Šaling, M. Šalingová, O. Peter, Dictionary of foreign words. Slovenské pedagogické nakladateľstvo, Bratislava 1965, 1140 pages] *Slovenská reč* 31(4), pp. 236-238.

Mistrík, J. (1993). *Encyklopédia jazykovedy*. [Encyclopedia of Linguistics]. Bratislava: Obzor.

Nesi, H. (2013). Researching Users and Uses of Dictionaries. In Jackson, H. (ed), *The Bloomsbury Companion to Lexicography*. London: Continuum, pp. 62-74.

Panocová, R. (2024). Definitions of Suffixed Loanwords in Dictionaries of Foreign Words in Slovak. In *International Journal of Lexicography* 37(3), pp. 337-353.

Osselton, N. E. (1983). On the history of dictionaries: the history of English-language dictionaries. In Hartmann, R. R. K. (ed), *Lexicography: Principles and practice*. London: Academic Press, pp-13-21.

SCSVS (1939). *Slovník cudzích slov a výrazov v slovenčine*. [Dictionary of Foreign Words and Expressions in Slovak]. Prídavok, J. M. (ed). Praha: Československá grafická Unie úč. Spol.

SCS (1953). *Slovník cudzích slov* (1. vyd.) [Dictionary of Foreign Words, 1st edition]. Hollý, A. and Ľ. Zúbek (eds). Bratislava: Tatran.

SCS (1966). *Slovník cudzích slov* (2. zrev. vyd.) [Dictionary of Foreign Words, 2nd revised edition]. Šaling, S., Šalingová, M., and O. Peter (eds). Bratislava: Slovenské pedagogické nakladateľstvo.

VSCS (1997). *Veľký slovník cudzích slov*. (1. vyd.). [Large Dictionary of Foreign Words, 1st edition]. Šaling, S., Ivanová-Šalingová, M., and Z. Maníková (eds). Veľký Šariš: Vydavateľstvo SAMO.

SCSA (2005). *Slovník cudzích slov (akademický)*. (2. doplnené a upravené vydanie). [Dictionary of Foreign Words (Academic), 2nd revised edition]. Balážová, Ľ., and J. Bosák (eds). Bratislava: Slovenské pedagogické nakladateľstvo - Mladé letá.

Statistical Office of the Slovak Republic. Accessed at: https://slovak.statistics.sk/ [04/03/2026].

Šalingová, M. (1954). Z problematiky slovníka cudzích slov. In *Slovenská reč* 19(3-4), pp. 90-99.

Van Sterkenburg, P. (2003). The dictionary: Definition and history. In Van Sterkenburg, P. (ed), *A Practical Guide to Lexicography*. Amsterdam: John Benjamins, 3-17.

Tvrdý, P. (1922). *Slovník inojazyčný*. [Dictionary of Foreign Words]. Bratislava: Comenius.

Tvrdý, P. (1932). *Slovník inojazyčný*, druhé vydanie [Dictionary of Foreign Words, 2nd revised edition]. Žilina: O. Trávniček.

From Sociolinguistic Database to Classroom Tool: Documenting and Teaching Pluricentric Hungarian through the Termini Online Hungarian Dictionary

Katalin P. Márkus, Tibor M. Pintér
Károli Gáspár University of the Reformed Church in Hungary, Department of English Linguistics,
Károli Gáspár University of the Reformed Church in Hungary, Department of Hungarian Linguistics
p.markus.kata@kre.hu; m.pinter.tibor@kre.hu

Abstract

The Termini Online Hungarian Dictionary describes the lexicon of the Hungarian language as spoken in the countries neighboring Hungary. It is considered a general dictionary of present-day Hungarian. Each entry contains authentic example sentences to illustrate the use of the headword, allowing for the examination of a word or construction's special use in a grammatical and pragmatic context. The dictionary placed in a lexicographical database is edited online in eight countries. Online editing makes it possible for the dictionary to expand – even simultaneously – as a result of activity in eight countries. In the present study, the authors review the novelties and peculiarities of the dictionary in some detail, touching on the following topics: dictionary structure, IT support, database character, multimedia elements, and labelling system and provide some tips for practicing dictionary use or skills. Since dictionary use forms part of the Hungarian National Curriculum, this paper focuses not only on dictionary skills but also on pedagogical strategies that familiarise students with the Termini Online Hungarian Dictionary as a resource. Learning about the existence and function of this dictionary helps students recognise the pluricentric nature of Hungarian, become aware of its regional varieties, and develop a broader, more inclusive perspective on language use and language identity in the digital age.

Keywords: Hungarian language, online dictionary, lexicography, dictionary use, dictionary skills, dictionary didactics.

1 Introduction

The linguistic communities of minority Hungarians beyond the present-day borders of Hungary came into being in 1920, when significant parts of the Hungarian language area were annexed to newly formed non-Hungarian states as a consequence of the Treaty of Trianon (1920). The number and share of Hungarian-speaking people in some of the countries neighbouring Hungary are still significant this day: 1,248,623 in Romania (2011) (6.2% at the state level, 17.9% in Transylvania) (Péntek and Benő 2020: 59, 61), 458,467 (8.5%) in Slovakia

(2011) (Horony 2016: 47), 254,000 (3.9%) in Serbia (2011), about 141,000 in Ukraine (2015) (11.28% in Subcarpathia, data from the census of 2021 are still unavailable for this region) (Kapitány 2015: 234–236). For this reason, we use data from the previous censuses. The number of Hungarian speakers is around 10,000 in Austria (in Burgenland and in Vienna, the latter being home to Hungarian immigrants), 14,408 in Croatia, and 5,544 in Slovenia (Kapitány 2015: 236–237). As a consequence of the minority situation, most Hungarian speakers living in these seven countries are bilingual: they speak Hungarian as their mother tongue and also the official language of the state, which is the native language of the majority group (Romanian, Slovak, Ukrainian, Serbian, etc.). Their active bilingualism creates a language contact situation which makes the lexicons of these varieties special, affected by language contact phenomena, a greater number of regionalisms, and lexical conservatism as a tendency to preserve older administrative terms. (For an exhaustive – and still actual – sociolinguistic description of these Hungarian language varieties, see Fenyvesi 2005).

The Treaty of Trianon did not only divide the Hungarian language community with political borders, creating linguistic minority groups in their own historic homeland, but also led to divergent development in the use of the Hungarian language in Hungary and in the states neighbouring Hungary (Huber 2020: 17). The various *state varieties* of Hungarian formed different *standard varieties* and varying norms of speech of these minority language communities due to the close links to each respective majority culture, institutional and legal system. Therefore, Hungarian can be regarded as a *pluricentric language* (Lanstyák 1996, Csernicskó & Kontra 2018: 104). But this pluricentricity characteristic of minority varieties of Hungarian is ”lacking the appropriate status” (Muhr 2012: 33) since the codification of lexical items used in regional standard function is not fully recognized due to the dominant prescriptive and exonormative model of codification in Hungary (Huber 2020: 23–25).

In November 2001, at the initiative of the Hungarian Academy of Sciences, linguistic research centres were established in Hungarian-speaking areas neighbouring Hungary in Romania, Slovakia, Ukraine, and Serbia. In the same year, linguistic individual study bases came into being in Austria, Slovenia and Croatia. These language centres started operating in 2002: the Attila Szabó T. Language Institute in Cluj, Romania; the Gramma Language Office in Dunajská Streda, Slovakia; the Antal Hodinka Institute in Beregovo, Ukraine; and the Hungarian Scientific Research Society in Subotica, Serbia, whose role was later taken over by the Verbi Language Research Workshop. The other neighbouring countries (Slovenia, Austria, Croatia) also had at least one researcher based locally closely collaborating with them. This linguists’ network was later expanded to include Imre Samu Language Institute, which studies the language use of Hungarian communities in Slovenia and Austria, and, most recently, the Glotta Language Institute in Croatia. The tasks of these linguistic research centres include the sociolinguistic study of Hungarian spoken in the regions neighbouring Hungary, its varieties, lexicological and terminological issues, analysis and exploration of language problems in the context of language planning and language management.

Based on this network organisation, these linguistic offices established the Budapest-based Termini Association in 2013, creating a legal framework for the then already one decade long cooperation (more information on the association, about its work can be found at http://termini.nytud.hu) or by using the following QR-code.

Figure 1: The QR-code leading to the online dictionary.

From the very beginning, one of the main research topics of these research centres was the collection and study of the special vocabulary of the Hungarian minority varieties. As a result, the Termini Online Hungarian Dictionary (TOHD) was established in 2007, it has been continuously developing ever since.

The lexicographic database supplies lexical data concerning contact phenomena (widespread direct and indirect loanwords) and independent lexical creations present in the minority varieties of Hungarian. In addition to these items, the dictionary also contains common regional lexical elements characteristic of these regions and vernacular words as well. Since lexicalized forms are not limited to certain geographical and social dialects or registers only, the dictionary can be considered a general dictionary of today's Hungarian language. By documenting older linguistic forms used today in minority situations, the online dictionary reflects linguistic tradition, and by presenting neologisms, it registers current linguistic change and renewal.

The dictionary also functions as a digital database. The lexical data are placed in a relational database so that virtually any statistical analysis can be performed on them. Various glossaries and statistical statements ensure rich processing opportunities. In addition, the large number of example sentences forms a structured living language corpus that offers many additional research opportunities for digital word processing (Benő et al. 2020: 156–158).

The concept of the dictionary is based on treating the Hungarian language as unified, disregarding political borders (defragmenting the language, as it were). The aim of the linguistic defragmenting or de-bordering program is to collect and enter elements from the entire Hungarian language area into new dictionaries (Péntek 2009: 104–105), since until the political changes of 1989 the codification of standard Hungarian did not take into account the regional standard varieties of Hungarian spoken outside Hungary by more than 2.5 millions of Hungarian native speakers. The so-called state varieties were taken into linguistic account only decades after the political (and also scientific) transition.

In this respect, lexicographic antecedents of TOHD are those Hungarian dictionaries that display lexicographic data from the varieties of the Hungarian language spoken in regions neighbouring Hungary. The second edition of *Magyar értelmező kéziszótár* [The Concise Explanatory Dictionary of Hungarian] (Pusztay 2003) indicates a commitment in this direction since the lexical data of the dictionary had been expanded to include representative Hungarian words and meanings used in the Hungarian spoken in Romania, in Slovakia and Ukraine (approximately 250 Hungarian words and meanings were included from Slovakia, 91 from Romania, and 38 from Ukraine – see Kiss 2004). Since 2003 other Hungarian dic-

tionaries have been expanding their lexicographical data taking into account the lexicon of Hungarian speakers living in minority situations in the countries neighbouring Hungary. Such a lexicographical work is the Hungarian orthographic dictionary (Laczkó & Mártonfi 2004), which displays place- and institution names used in the minority varieties of Hungarian (while the orthographic dictionary of the HAS still lacks toponyms connected to these regions). The dictionary of recent lexical borrowings [*Idegen szavak szótára*] (Tolcsvai 2008) also makes use of the loans collected by linguists of the Termini Hungarian Linguistic Network in the regions outside Hungary. Thanks to further collaborations, new entries, meanings and synonyms used in bilingual contexts were proposed for *Értelmező szótár+* [Hungarian explanatory and etymological dictionary] (Előry 2013). According to M. Pintér (2017: 166), these dictionaries contain entries from TOHD in the following distribution: Hungarian explanatory and etymological dictionary: 55 entries, 115 meaning, 850 synonyms; the concise explanatory dictionary of Hungarian: 273 entries; Hungarian orthographic dictionary: 739 words; the dictionary of recent borrowings: 1.501 headwords.

The lexical database has *practical benefits.* It makes it possible for Hungarian speakers to learn the words and phrases of the regional varieties used in different countries neighbouring Hungary. This is envisaged also by an interface offering detailed searches using the elements of the microstructure.

The words included in the dictionary are *legitimized* in a sense, but "legitimizing" should not be confused with their *codification* (with their incorporation into the standard variety). Legitimatization here means the acknowledgment of the existence of these lexical items, but it does not mean that all such words would be used in high-level, formal-style writing – as it is indicated with the stylistic labelling.

In addition to lexicological and semantic research, the lexicographical database can also be used in *morphological, syntactic, pragmatic* and *translation* studies since the lexical items appear in context.

The *theoretical significance* of the research program cannot be neglected either. It is important to examine how a language adapts to the specific needs of speakers living in eight different countries and what types of contact phenomena are introduced in each community, what it is that can enter easily into the receiving language system, and what enters only with certain modifications (Lanstyák et al. 2010: 56–58).

2 Technology in the Editorial Process

Digital lexicography is considered to be a part of digital humanities (just like the online Bible readers, for example): lexicographic work is nowadays driven by computer assisted technology supporting the so called computer assisted lexicography (see Wooldridge 2004: 69). In case of TOHD, digital humanities are used to build the elements of microstructure or build connections between entries or producing glossaries from the data sets. With the help of database structure human lexicon is described as a structured text enhanced with several functions. For this reason, semi-structured mark-up languages (such as XML) and databases are used to store the envisaged elements. As described in Dziemianko (2012: 319), these

characteristics of the digitized texts were outlined in the beginning of the era of electronic dictionaries (see De Schryver 2003).

TOHD is technically driven by an SQL-database, which enables queries of selected data or searches in entries, providing user dictionaries based on selected data sets. This solution has proved to be useful as structured data and multiuser data access are used. The database handles the user rights of the clients (registered and unregistered users), enables simple browsing, creates lists, statistics, and edits the entries. A user interface built for editors is straightforward to use (however a new design would help in aesthetics), helping editors with dropdown lists and simple technical issues. The webpage is in Hungarian, no localisations have been done, as the dictionary covers only Hungarian lexical data (except for etymological information).

Because the database is distributed, editors can edit entries and add descriptive suggestions seen only by other editors (the database also enables sending emails to other editors – with the embedded HTML-version of the chosen entry). This feature is useful as the dictionary is edited from eight different countries. Modifications appear in the dictionary immediately after editing, which distinguishes digital lexicography from 'classical' one. The technical background used in TOHD is nowadays widely used in terminography and lexicography, but at the time of its beginnings (in the early 2010s) it was still unique in Hungary.

The database and the online interface are useful not only for quick editing and effective cooperation between editors, but they have also proved to be good for providing statistics and data mining – this is a feature that is not available in paper dictionaries. As TOHD is used online, multimedia data such as images and audio files can be easily linked (or later – deleted) to its entries. The database setup allows providers to gain information on lexical data, trying to catch the most exhaustive information on lexical and conceptual content.

The dictionary can be used by two types of user profiles: registered and non-registered users. Non-registered users have fewer options for queries than registered users (editors and linguists). Non-registered users can only use the *Search* function (*Keresés a szótárban*), allowing them to set the search in three ways. First, users can set restrictions to the position of character strings (characters typed into the edit box): they can be searched in the headword list, meaning field, or example sentences. Second, strings can be searched as a beginning of a word, as a set of characters in the middle of a word, or as a set of characters at the end of a word. Third, users can define the number of example sentences to be shown to the user: zero, all or only two sentences. In contrast, registered users can view a more detailed setting of the searching tool, making the dictionary much more interactive, but this setting does not contain all fields represented in the database either, only those which contain important linguistic data. Linguistic information to be used in lexicology, lexicography, sociolinguistics or data mining will only be visible after the query of represented data (e.g., the log data remain covered).

Technology has changed the way a dictionary is compiled: the process used for compiling a dictionary is conducted in a distributional database where all modifications can be viewed immediately in the dictionary, and modifications leave footprints in the statistics and the log files (entries can have their own "history", their versions can be viewed in their "life pe-

riod"). Editors can be informed about the changes through the database, and changes can be done anytime as there is no paper format consuming extra costs for printing.

3 Interactive Dictionary

As Lew states, the effectivity of e-dictionaries can be measured only according to the expectations towards them, paper dictionaries are as useful as the digital ones: the differences between them are mostly based on the conditions of the needs of a particular user in a particular situation (Lew 2012: 344, 360). However, access to data is quicker and more effective (in comparison to the paper dictionaries), and additional materials linked to entries (such as hyperlinks and multimedia material) are also enabled.

TOHD is not simply a dictionary, but rather a database of language use, varieties, morphological data, and lexicon – it can be used for data mining, and its content can be visualized according to several criteria. TOHD makes it possible to create glossaries and lists, and the classical dictionary view can also be generated.

The user interface determines the queries in various ways. Users can gain data (make according to specific criteria of the microstructure their own dictionaries) and information with the help of three views. Its speed depends on the local computer, browser and Internet connection, e.g., on circumstances independent of the dictionary and its system. Views (query interfaces and possible formats of the dictionary entries assigned to users) are connected to user rights and registration. Registered users can exploit the database more deeply and view more options in queries. The interface built for non-registered users offers only the basic options, enough for browsing data and reading the lexicographic information as described above (*Keresés a szótárban* / Search function).

The enhanced search tool *Szójegyzékek* (Glossaries) is built for registered users (especially with linguists in mind) and serves as a tool for data mining or helps dictionary editors in the editorial process. Queries facilitated through dropdown lists can be done in any part of an entry. *Glossaries* can provide four types of lists, focusing on specific data within the entries. Selected entries can be organised into glossaries according to the time period of the creation/modification (*Időszak* / Time period), entries listed connected with any editor (*Kutatási listák* / Editorial lists), parts of the microstructure rendered as lists for editorial work (*Fejlesztési listák* / Development lists), or glossaries to view the work of any editor (*Dokumentációs listák* /Documentation lists). The structure of entries or lists presented in the Glossaries section can be set according to data in microstructure or metadata in entries.

As has been mentioned above, editors can add comments to any entry (and send them to any other editor for further correction). Comments can be viewed only by the editors (as they are written for work purposes – this feature is hidden from the public, only editors can utilise it). Unfinished entries can be set as hidden entries viewed only by editors – in the glossaries function editors can generate a list showing all hidden entries and work on them.

Lists and glossaries are generated for lexicographic enhancement of the entries, and they also provide information on varieties of Hungarian spoken outside Hungary. As the database and its front end (user interface) are flexible, any glossaries and lists used for quantita-

tive research can be changed or new options for generating lists/glossaries can be added according to the needs of the editors.

With the help of lists and views, users and editors of the dictionary can select the information they want to see, while all the other, distracting information is excluded. Data in the entries are rendered with the help of colours, font types and tool tips, helping editors and readers access the most accurate linguistic information.

TOHD is a unique lexicographic tool in Hungarian lexicography. Its uniqueness concerns the coverage of varieties, its detailed microstructure and editorial process allowing simultaneous compilation from eight countries. The database-based dictionary is specially built for online use, although its entries were published in several monolingual Hungarian dictionaries published by major publishers. TOHD is the only dictionary focusing on the varieties of Hungarian spoken outside Hungary.

3.1 Multimedia Elements

Today, the growing popularity of digital dictionaries is undeniable. Their biggest advantage is that access to information is easy and fast. Furthermore, information can be accessed from anywhere at any time. Digital storage also enables a wide range of multimedia tools to be used, so that dictionary users can read not only the "traditional" dictionary entries found in paper dictionaries, but visual and acoustic elements are also incorporated. The use of the image as an element supporting the description of meaning is a traditional feature in the history of dictionaries (cf. picture dictionaries). If we look back at the history of Hungarian dictionaries, we find Comenius' pictorial "dictionary" as early as 1653, which is a great forerunner of the use of visual possibilities (P. Márkus 2019). Another good example is the first edition of *The Concise Explanatory Dictionary of Hungarian* (*Magyar értelmező kéziszótár*, Juhász et al. 1972), which explained certain headwords not only through definitions but also by means of pictures. Pictorial illustrations in dictionaries are incorporated in order to assist the user with the process of decoding meaning. They were considered typical in encyclopaedias and, therefore, encyclopaedic by nature, since illustrations are concerned with the world, not with linguistic signs (Rey-Debove 1971). However, modern dictionaries include not only linguistic but encyclopaedic information as well. The latter type of information can be provided in written and visual form.

In TOHD, pictures which can be inserted separately for each meaning in the dictionary entry are mostly used to represent nouns and concrete objects, especially for culture-specific ones. Captions and the sources of the illustrations can also be added. The captions do not necessarily repeat the lemma but are useful for descriptive and explanatory purposes. This can be seen in the entry for *blokk* 'block of flats; apartment block', where under the equivalent 'block of flats' the caption reads "Typical socialist-era block of flats in Szered".

Following Hupka (1989), who classifies visual elements into several types, TOHD uses three main types. The first type includes images which show just one specimen of a kind (*single illustration*). In these cases, it is recommended to choose the most typical elements. The illustration in the entry of *baklazsán* 'aubergine; eggplant' is a good example, where the shape, form and colour of the vegetable are sufficient for interpretation. The dictionary also

contains brand names, which can also be included in this group. Images of brand-related products can be used to present these products in a clear and concise way (e.g., *ticsinki* 'salty sticks', where the packaging and appearance of the product speak for themselves).

ticsinki (fn) ~k, ~t, ~je

(Gaszt) Fv *(ált) (közh) (biz)* ropi ♠ Fv *én viszek magammal mexicornt, ticsinkit,* sajtostekercset (ez olyan anyukám féle finomság). szerintem valamit kifelejtettem :-). (http://zene.sk/forum) ♠ Fv *Természetesen azok, akik nem szeretik, ha labdával futkároznak a férfiak, akik egy pillanatra sem ültek le férjük, vőlegényük, udvarlójuk, szeretőjük mellé, hogy "ticsinkit"* rágcsálva drukkoljanak a képernyő előtt ennek vagy annak a csapatnak. (http://mek.oszk.hu; Bodnár Gyula: Szappanopera a medencében)

Figure 2: A packet of *ticsinki* with the original Slovak spelling *tyčinky*.

In the next type, the object is presented in its normal surroundings (*structural illustration*). This type can provide the reader with a lot of extra information (e.g., *dodávka* 'van', where the vehicle can be seen in front of a warehouse entrance, in the driveway, illustrating its commercial vehicle character).

dodávka (fn) ~´k, ~´t, ~´ja

(Gépk) Fv *(ált) (közh) (biz)* furgon ♥▪ Fv *Dodávkával hozták az árut.* (ht-kut) ♥▫ Fv *Dodávkával hozatta haza a szőnyeget, a személyautójába nem fért bele.* (f.n.)

Figure 3: Perfect illustration of the commercial vehicle.

The third type of images shows the object in operation (*functional illustration*), since the meaning is best illustrated with the demonstration of the function (e.g., *bublifuk* 'soap bubble blower', where a child is blowing soap bubbles).

Because of the typological characteristics of the TOHD, it does not use other types of images, such as nomenclatural illustrations, which are particularly common in learner's dictionaries, as they support vocabulary development. There is also no need for the type illustrating actions and processes of various kinds (*sequential illustration*). Sequential illustrations are used to graphically explain different types of processes, such as the stages of insect development in encyclopaedias.

In addition to the visual content, the dictionary also includes audio material, as some of the example sentences in the entries illustrate spoken language usage. The audio recordings are selected from the editors' collections. They can be read as written examples and listened to as audio files in the entries. The dictionary contains relatively little audio material since it is a new area of development and is still in an experimental phase.

4 Editorial and Structural Details

4.1 Entry Structure

The microstructure of the dictionary comprises three parts, which are typographically distinct from each other in the dictionary: the head, the body, and the foot. There are two types of elements in the head: elements that are included in all the dictionary entries (obligatory elements: headword, part of speech specification, and inflected forms) and elements that are only included in certain types of entries (optional elements: spelling variants or pronunciation; information on the typology of loanwords is being added now).

The dictionary indicates the pronunciation or spelling variant(s) of the headwords that have been borrowed in their unchanged form during language contact. Spelling variants, listed in round brackets after the headword, are pronounced as the corresponding lemma, differing only in their written forms. The dictionary only indicates the pronunciation of the headword if it differs from the written form or if it does not follow the pronunciation rules of the Hungarian language. Pronunciation is presented after each headword in square brackets. A pronunciation variant is provided for headwords if there is a sound absent from the Hungarian sound system or the spelling of the sound is different from that used in Hungarian (e.g., the specific speech sounds of Serbian and Ukrainian, or the diacritical marks of Slavic languages). Because pronunciation is less regulated than spelling, there may be more than one spoken form of a headword, and these are indicated in the entry of the standardised headword.

The versatility of the microstructure is demonstrated by the richness of the information in the main body of the entry, which contains two separate sets of information: information needed to understand the entry word and example sentences to illustrate its use and meaning. The interpretation of the lemma is aided by the elaboration of the meanings and

senses as well as the 46 status labels (subject, regional, temporal and stylistic labels). Status labels are used to specify the scope of the use of a headword or a meaning.

TOHD being a monolingual, explanatory dictionary with etymological references, it has two main types of definitions: synonym definitions and lexicographic definitions (cf. Svensén 2009). The dictionary's example sentences are also a means of presenting meanings as accurately as possible since meaning can be clarified through example sentences. Furthermore, the stylistic and linguistic value of the headword can be demonstrated through the different types of example sentences. Example sentences are increasingly moving away from literary texts towards real-life examples. Earlier Hungarian dictionaries abound with examples from literature, while today's dictionaries try to draw not only from literary texts but also from real-life language use. Corpora can be a great source in this process because corpus-based dictionaries are no longer an unattainable dream but have become one of the hallmarks of modernity. As TOHD is a dictionary of language use, utilized primarily for sociolinguistic (and lexicological) research, and the primary classification of example sentences is based on their authenticity. The editors aim to use authentic, spoken or written example sentences, and avoid the use of lexicographer's made-up examples. In this respect, example sentences in the dictionary are of three types: (1) authentic written language example sentences, (2) authentic spoken language example sentences, and (3) non-authentic example sentences (whose sources cannot be verified). These categories are further subdivided into subcategories, specifying the extralinguistic circumstances in which the example sentence was created. Based on the above, the types of example sentences in the dictionary present the following typology (see Oktober 2025 statistics in brackets): authentic written language (65,907); authentic spoken language (4,724); undocumented and thus not authentic (242); meta-linguistic, authentic written, lay speaker (4,076); meta-linguistic, authentic written, linguist speaker (765); meta-linguistic, authentic spoken, lay speaker (892); meta-linguistic, authentic spoken, linguist speaker (104); constructed by a lay speaker (22,670); constructed by a linguist (8,795); translated text (143); and other (519).

An authentic example sentence is one that is not taken from an elicited speech event, i.e., it is a statement that has been published in writing (either in an online forum or in a literary work) or is a documented oral statement made by a speaker. Another interesting feature of the labelling of example sentences is that the editors indicate not only the regional affiliation of the word, but also the regional affiliation of the speaker, and in the case of authentic example sentences, the source.

The references in the foot of the entry include three types of information (each in a separate line): variants in the same region, analogous words in other regions, and other words of the dictionary from the same word family, regardless of their occurrence.

The dictionary is primarily intended for sociolinguists, but researchers from other fields of linguistics such as contact linguistics, descriptive linguistics, as well as translation theory and practice also use it. In addition, it is also important for a variety of practical uses: the editors of the dictionary are particularly interested in enabling lay language users to learn specific words and meanings of other varieties of Hungarian (Lanstyák et al. 2010: 44). The lexicological (temporal, regional, and stylistic) and lexicographic description of the vocabulary (e.g., the structure of the entry) provides "layers of use" for the dictionary in several

ways, as an explanatory dictionary of the Hungarian language, a general dictionary, a dictionary of foreign words, as well as a source of language statistics and data-mining.

The wordlist of the dictionary aims to reflect primarily synchronic language use – this is one of the reasons why it also includes neologisms. In addition to the present-day Hungarian vocabulary, there are also some typical words from the interwar period and from the socialist era. Headwords or meanings marked *Hu*, i.e. elements used as archaic in Hungary are used in various varieties in bordering regions – explaining the archaic nature of marginal dialects. Although reflecting a synchronic situation, the obsolete, old-fashioned or somewhat old-fashioned *Hu* headwords in the dictionary are now commonly used words or meanings in some regions. As the wordlist of the dictionary is not in any way linked to a particular language variety, its general dictionary character is also obvious. As far as the criteria headwords need to meet to be included in the dictionary is that the word, word combination or meaning has to have widespread use (it should not be limited to a small geographical area or to any stylistic variety or register). As the entry words and meanings included in the dictionary typically originate in a state language but have become an integral part of the Hungarian language usage of the given region, the dictionary can also be seen as a kind of dictionary of recent loans (called foreign words, *idegen szavak*, in the Hungarian linguistic tradition). Among the headwords there are some that are also used in Hungary, but with a meaning that is specific to the given minority language variety and different from the Hungarian one. The vitality of TOHD lies precisely in the fact that it functions as a database: users access the headword or meaning they are looking for through queries. The database feature of the dictionary has several advantages (which is perhaps why the editors do not plan to publish the dictionary as a book), providing easier editing and simpler, more customised presentation. This can guarantee the continued success of the dictionary.

4.2 Labelling System

The labelling system used in the dictionary is based on the linguistic labels of *The Concise Explanatory Dictionary of Hungarian* – we tried not to deviate too much from them, as most Hungarian dictionary users are familiar with the labels of this dictionary and encounter them in other dictionaries, so users can interpret them most easily. The stylistic value of each vocabulary item or meaning was defined along various sociolinguistically definable categories – these are the dimensions. The system currently consists of five main dimensions. In defining the dimensions, we have taken as our starting point the linguistic reality, but we have also been influenced by M. A. C. Halliday's classification (Halliday et al. 1964). The following dimensions are used:

The labels of the dialectal dimension refer to the stylistic value of lexemes as a result of their use being specific to certain geographical regions and social groups. Regional dialects are referred to as 'regional'. Examples of labels referring to social dialects are 'educated' and 'uneducated' – referring to the level of education of the language user – and 'child language' and 'youth language', which refer to the age group.

The register dimension is largely made up of the stylistic values under which headwords associated with certain specialised activities can be classified. This category includes, for

example, 'technical', 'literary', and 'slang'. The fact that a word or a meaning is not linked to any specialised register but can occur in discourse belonging to any register is referred to by the qualification 'commonly used'.

The stylistic variation dimension consists of registers defined by the formality of the speech situation; the stylistic value of individual headwords is determined by whether they are generally used in formal or informal speech situations. Examples include the style labels 'formal' and 'informal'. Words that are not associated with a style but can be used in any speech situation are labelled 'neutral'.

In the dimension of temporality there are labels such as 'old-fashioned', 'obsolete', and 'historical'. Since most of the headwords and meanings do not refer to temporality, for the sake of simplicity we do not refer to these with any special label.

The labelling system also includes, of course, 'real' stylistic classifications, which express the speaker's emotions and positive or negative attitudes towards the reality indicated by the headwords. This dimension includes labels such as 'euphemistic', 'humorous', 'ironic', 'taboo', 'pejorative', and 'rude'.

Finally, in addition to the dimensions mentioned above, in some specific cases the frequency of words is also indicated by the label 'rare', and the regional affiliation of the headwords is also marked.

4.3 Standardization

As the dictionary is edited by linguists residing in several countries, certain measures for standardization are necessary to maintain linguistic consistency, since editorial work is not a common daily task for all editors. Online meetings and shared documents are used to ensure the unity of editorial practices and lexical content. However, as the editors are not professional lexicographers and online communication can be cumbersome, standardization tasks must periodically be undertaken.

In 2024, a major effort was carried out to identify errors and inconsistencies in the microstructure and to integrate new knowledge into the editorial process. Following extensive unification work, several aspects of the dictionary were revised.

At the level of the entry headwords, grammatical forms were corrected and inconsistencies in their representation resolved (e.g., order of forms, representation in compound words, and use of the tilde). Within the body, example sentences were revised where the affected word was part of a compound rather than the headword itself, and bibliographic references were elaborated. In the footnotes, erroneous etymologies and malfunctioning cross-references were corrected. In the final stage of the process, new functionalities for the website were defined, including localization into English.

In 2025, the work was consolidated, resulting in a more coherent lexicon with fewer inconsistencies.

5 Educational Use

In contemporary language education, increasing emphasis is placed on developing learners' language awareness alongside their lexical and communicative competence. Digital lexicographic resources play a crucial role in this process, as they offer access to authentic language data and reflect current patterns of language use.

TOHD holds particular educational significance, as it provides learners with authentic examples that reflect the linguistic diversity of Hungarian across regions. Its inclusion in the Hungarian National Curriculum underscores its role as a key pedagogical tool for developing vocabulary, language awareness, and dictionary skills. By integrating its use into classroom practice, educators can design activities such as comparing entries from different regions, identifying loanwords influenced by majority languages, or exploring audio samples to practice pronunciation variations. Students might also engage in tasks like creating regional word maps, discussing why certain terms differ across borders, or using TOHD to check the appropriateness of expressions in formal versus informal contexts. These activities foster understanding of linguistic variation and promote a comprehensive, modern approach to language learning. Generally, the use of dictionaries constitutes an essential component of language learning and teaching (P. Márkus, 2023), as they enable learners to understand precise meanings of words and recognise their appropriate usage across various contexts. This aspect is particularly significant in Hungarian, where state varieties have developed distinct norms and differing patterns of usage within minority language communities. These differences have emerged due to close interaction with majority cultures and differing institutional and legal frameworks. Consequently, Hungarian can be regarded as a pluricentric language, characterised by several national varieties that coexist and influence one another. In a pluricentric linguistic environment, dictionaries play a crucial role in documenting contact varieties and guiding users in navigating between regional and standard norms, thereby promoting a context-sensitive understanding of the language. TOHD contributes significantly to vocabulary development by offering authentic example sentences, grammatical information, and pronunciation guidance (see Karmacsi et al. 2022). Its integration into education can follow diverse methodological approaches aligned with linguistic and pedagogical objectives. For instance, contrastive analysis tasks can be designed where students compare dictionary entries from different regions, identifying semantic shifts or register differences. This activity develops metalinguistic awareness and encourages discussion about sociolinguistic factors influencing variation. In grammar instruction, teachers can select example sentences from TOHD to illustrate syntactic structures in authentic contexts, such as the use of case endings or verb conjugations in minority varieties, reinforcing functional grammar learning. Furthermore, project-based learning offers opportunities for collaborative research: students can collect regional expressions from their communities, verify them in TOHD, and present findings through digital portfolios or classroom posters. This approach not only enhances lexical and grammatical competence but also cultivates digital literacy and critical thinking. Teachers may also implement task-based activities, such as role-playing scenarios where students choose regionally appropriate vocabulary for formal letters versus informal conversations, using TOHD as a reference tool. Additionally, the open-access nature of TOHD supports inquiry-based learning, where learners explore language contact phenomena, such as loanwords from Slovak or Romanian, through guided

research questions. These practices align with communicative and constructivist methodologies, promoting active engagement, learner autonomy, and a research-oriented perspective on language learning (cf. P. Márkus 2023; Richards & Rodgers 2014; Yamada 2014).

The following practices illustrate how TOHD can be effectively integrated into classroom instruction using established language teaching methodologies. The first activity employs discovery learning and contrastive analysis, encouraging students to explore regional semantic variation through the dictionary's advanced search functions. Since TOHD allows searches by headword, meaning, and example sentence, it is ideal for comparative lexical analysis. For example, when students look up the headword 'letér', they find that in Hungary it means "to leave the proper path," whereas in Slovakia it means "to turn into a street." This task aligns with constructivist pedagogy, as learners actively engage in meaning-making by comparing authentic data. Teachers can scaffold the activity using guided discussion and pair work, prompting students to hypothesise why meanings shift across varieties and to consider cultural and geographical influences. Such reflective tasks foster critical thinking, metalinguistic awareness, and intercultural competence, consistent with communicative and inquiry-based approaches.

Another effective exercise applies data-driven learning and interpretive skills development by focusing on the use of geographic labels in TOHD to explore the regional distribution of words and meanings. The dictionary uses geographic labels such as *Fv* for Felvidék (a Slovakian Hungarian-speaking region) and *Ka* for Kárpátalja (a Ukrainian Hungarian-speaking region) to indicate where a headword or sense occurs. For example, when students examine these labels for the headword 'letér', they can identify which meanings are region-specific and compare lexical patterns across areas. This activity aligns with task-based learning and scaffolding techniques, as learners work through authentic data to interpret lexicographical conventions. Teachers can incorporate pair or group work for collaborative analysis, followed by guided reflection to discuss why such regional distinctions exist and how they relate to sociolinguistic realities. These practices deepen students' understanding of linguistic variation, enhance their ability to critically interpret dictionary metadata, and introduce them to the complexity of lexicography, highlighting its context-sensitive and descriptive nature (cf. Leaney 2007; P. Márkus 2023).

Another useful practice employs semantic analysis and context-based learning by examining differences between general and specialised meanings. When students search for the headword 'abszolvál' in TOHD, they discover that in Slovakia the word is used broadly to mean "to finish something," whereas in Hungary it has a narrower, context-specific meaning, referring mainly to completing studies or passing an exam. This exercise aligns with communicative language teaching and contrastive analysis, as learners explore how lexical items develop distinct semantic ranges depending on sociolinguistic environments. Teachers can structure the activity using guided discovery and discussion tasks, prompting students to reflect on contextual appropriateness and the influence of social and institutional settings on word meanings. To extend the task, educators might incorporate role-play scenarios where students choose the correct meaning for different contexts (e.g., academic vs. everyday situations), reinforcing pragmatic competence and critical language awareness (cf. Leaney 2007; P. Márkus 2023).

A completely different aspect of dictionary use is highlighted in an exercise focusing on meaning-based searches rather than headword searches. Students are given a standard Hungarian word, such as 'autópálya' (motorway), and asked to find regional equivalents using TOHD's semantic search function. Through this activity, they discover that different words are used across regions, for example, 'autobahn' in Austria and 'autoceszta' in Slovenia. This task applies inquiry-based learning and data-driven learning, encouraging learners to explore lexical variation influenced by language contact and cultural integration. Teachers can scaffold the process with guided research questions (e.g., "*Why might these regional terms differ?*") and follow up with group discussions or concept-mapping activities to visualise patterns of linguistic diversity. By emphasising meaning-based searches, the exercise demonstrates how semantic exploration can uncover variation that headword-based searches might overlook, fostering critical language awareness, intercultural competence, and digital literacy (cf. Leaney 2007; P. Márkus 2023).

The final exercise focuses on exploring word families and semantic relations within TOHD, applying principles of lexical field analysis and autonomous learning. Students are asked to find the standard Hungarian equivalent of a given headword and then identify all related words and derived forms. For example, when searching for 'nanuk' (ice cream), learners locate the standard equivalent and explore related words listed as hyperlinks at the end of the entry, such as 'nanukszelet' (a slice of ice cream) and 'nanuktorta' (ice cream cake). By navigating these links, students engage in network-based vocabulary learning, which promotes awareness of morphological relationships and semantic fields. To deepen engagement, educators might integrate concept-mapping tasks or digital mind maps, where students visualise the relationships among related words (cf. Leaney 2007; P. Márkus 2023). This approach supports vocabulary expansion, morphological awareness, and digital literacy, while fostering learner autonomy through discovery-oriented strategies.

The outlined activities demonstrate how TOHD can be effectively integrated into language teaching to develop vocabulary, enhance linguistic awareness, and foster critical thinking. By engaging students in tasks such as exploring regional semantic variation, interpreting geographic labels, analysing general versus specialised meanings, conducting meaning-based searches, and investigating word families, educators can apply methodologies like inquiry-based learning, contrastive analysis, and project-based approaches. These practices not only strengthen lexical and grammatical competence but also promote digital literacy, intercultural understanding, and autonomous learning. Ultimately, incorporating TOHD into classroom instruction supports a modern, research-oriented perspective on language education, helping students appreciate the pluricentric nature of Hungarian and the sociocultural factors shaping its varieties.

6 Conclusions

TOHD is a unique lexicographical project work in Hungarian lexicography. It is unique in terms of the varieties covered, unique in the elaboration of its microstructure, and unique in the process of its production since it is a database dictionary edited by linguists, done as an electronic dictionary of special varieties of Hungarian. This dictionary is the only Hungarian dictionary to provide a lexical elements and meanings of the Hungarian language used

beyond the borders of Hungary, employing a highly sophisticated lexicographic and lexicological apparatus and information and communication technology.

Using the online interface dictionary, users can create their own dictionaries on the fly, and compile glossaries and statistics by using its sophisticated query system searching the database in the background. The material of the dictionary is used and exploited not only by 'lay' language users but also by linguists and researchers of Hungarian varieties providing the basis for numerous quantitative and qualitative analyses. Although not intended for linear reading and not published as a book, it is useful for those who wish to learn more about the meanings and usage rules of specific vocabulary of the state varieties of the Hungarian language in the Carpathian Basin. The dictionary aims to provide a detailed picture of the lexical and conceptual content behind headwords.

Work on the dictionary is supported by the Hungarian Academy of Sciences. The annual projects focus primarily on the development of micro- and macrostructure, the correction and improvement of the dictionary entries, and the implementation of documentation activities to support the development work.

The TOHD can be used in almost all areas of linguistic research into present-day Hungarian, as it is a systematic and easily searchable database of the living language, since its data come from actual language use and communication.

References

A. Dictionaries

Benő, A., Ćurković-Major, F., Csernicskó, I., Gaál, P., Juhász, T., Kitlei, I., Kolláth, A., Kovács, L., Lanstyák, I., Lehocki-Samardžic, A., M. Pintér, T., Márku, A., Molnár Csikós, L., Nádor, O., P. Márkus, K., Sági, Zs., Sófalvi, K., Szoták, Sz., Váradi, K., Varga, & T., Žagar Szentesi, O. (2025, eds.) *Termini Online Hungarian Dictionary*. Törökbálint: Termini Association.

Előry, V. (ed.) (2013). *Értelmező szótár+* [Hungarian explanatory and etymological dictionary]. Budapest: Tinta Könyvkiadó.

Juhász, J., Szőke, I., O. Nagy, G. & Kovalovszky, M. (1973, eds.). *Magyar értelmező kéziszótár* [The concise explanatory dictionary of Hungarian]. Budapest: Akadémiai Kiadó.

Laczkó, K. & Mártonfi, A. (2004, eds.). *Helyesírás* [Orthography]. Budapest: Osiris Könyvkiadó.

O. Nagy, G., Juhász, J., Szőke, I. & Kovalovszky, M. (1972). *Magyar értelmező kéziszótár* [A concise explanatory dictionary of Hungarian]. Budapest: Akadémiai Kiadó.

Pusztai, F. (2003, ed.). *Magyar értelmező kéziszótár* [A concise explanatory dictionary of Hungarian]. Budapest: Akadémiai Kiadó.

Tolcsvai Nagy, G. (2008, ed.). *Idegen szavak szótára* [Dictionary of foreign words]. Budapest: Osiris Könyvkiadó.

B. Other literature

Benő, A. (2020). Terminology and language planning for Hungarian spoken in Romania. In Vančo, I., Muhr, R., Kozmács, I. & Huber, M. (eds.) *Hungarian as a pluricentric language in language and literature,* Berlin – Wien: Peter Lang, pp. 113–126.

Benő, A., Juhász, T. & Lanstyák, I. (2020). A Termini „határtalan" szótára [The 'Borderless' Dictionary of the Termini Research Network]. *Magyar Tudomány*, 181(2), pp. 153–163.

Csernicskó, I. & Kontra, M. (2018). Határtalanítás a magyar nyelvészetben [Debordering in the Hungarian Linguistics]. *Fórum Társadalomtudományi Szemle*, 20(1), pp. 91–104.

De Schryver, G.-M. (2003). Lexicographer's dreams in the electronic-dictionary age. *International Journal of Lexicography*, 16(2), pp. 143–199.

Dziemianko, A. (2012). On the use(fulness) of paper and electronic dictionaries. In S., Granger & M., Paquot (eds.) *Electronic lexicography*. Oxford: Oxford University Press, pp. 319–342.

Fenyvesi, A. (2005, ed.). *The Hungarian language outside Hungary*. Amsterdam: John Benjamins.

Halliday, M. A. K., McIntosh, A. & Strevens, P. (1964). *The linguistic sciences and language teaching*. London: Longman.

Horony, Á. (2016). Az anyanyelvhasználat jogi keretei Szlovákiában [The legal framework of using Hungarian as a mother tongue in Slovakia]. *Acta Humana*, 4(3), pp. 45–55.

Huber, M. (2020). Hungarian as a pluricentric language. In Vančo, I., Muhr, R., Kozmács, I. & Huber, M. (eds.) *Hungarian as a pluricentric language in language and literature.* Berlin – Wien: Peter Lang, pp. 19–32.

Hupka, W (1989). *Wort und Bild: Die Illustrationen in Wörterbüchern und Enzyklopädien. Lexicographica, Series Maior 22*. Tübingen: Max Niemeyer.

Kapitány, B. (2015). Ethnic Hungarian in neighbouring countries. In Monostori, J., Őri, P. & Spéder, Zs. (eds.) *Demographic portrait of Hungary*. Budapest: Hungarian Demography Research Institute, pp. 225–239.

Kiss, J. (2004). Egy régi-új nyelvi sikerkiadvány: a Magyar értelmező kéziszótár [An old linguistic publication renewed: The Concise Explanatory Dictionary of Hungarian]. *Magyar Tudomány*, 49(5), pp. 670–673.

Lanstyák, I. (1996). Gondolatok a nyelvek többközpontúságáról (különös tekintettel a magyar nyelv Kárpát-medencei sorsára) [Thoughts on the pluricentricity of languages, with special attention to the fate of the Hungarian language in the Carpathian Basin]. *Új Forrás*, 28(6), pp. 25–38.

Lanstyák, I., Benő, A. & Juhász, T. (2010). A Termini magyar–magyar szótár és adatbázis. *Regio*, 21(3), pp. 37–58.

Leaney, C. (2007). *Dictionary Activities*. Cambridge: Cambridge University Press.

Lew, R. (2012). How can we make electronic dictionaries more effective? In S. Granger & M. Paquot (eds.) *Electronic lexicography*. Oxford: Oxford University Press, pp. 343–362.

M. Pintér, T. (2017). Adatbányászat a kétnyelvűség mögött [Data mining beyond bilingualism]. In Benő, A., Gúti, E., Juhász, D., Szoták, Sz., Terbe, E. & Trócsányi, A. (eds.) *Tudományköziség és magyarságtudomány a nyelvi dimenziók tükrében*. *A VIII. Nemzetközi Hungarológiai Kongresszus (Pécs, 2016) három szimpóziumának előadásai* [Linguistic dimensions of interdisciplinarity and Hungarian studies: Papers of the three symposiums of the 8th International Conference of Hungarian Studies, Pécs, 2016]. Budapest: Termini Egyesület, pp. 158–167.

Péntek, J. (2009). Termini: The network of Hungarian linguistic research centres in the Carpathian Basin. *Minorities Research,* 11, pp. 97–123.

Péntek, J. & Benő, A. (2020). *A magyar nyelv Romániában (Erdélyben)* [The Hungarian language in Transylvania, Romania]. Kolozsvár – Budapest: Erdélyi Múzeum-Egyesület – Gondolat Könyvkiadó.

P. Márkus, K. (2019). Szótárak: Kezdetektől napjainkig [Dictionaries: From the beginnings to the present]. In Furkó, P. & Szathmári, É. (eds.) *Népszerű tudomány, tudománynépszerűsítés: A Károli Gáspár Református Egyetem 2019-es évkönyve* [Popular science and popularizing science: The 2019 yearbook of the Károli Gáspár University of the Reformed Church]. Budapest: Károli Gáspár Református Egyetem, pp. 146–155.

P. Márkus, K. (2023). *Teaching Dictionary Skills*. Budapest: Károli Gáspár University of the Reformed Church – L'Harmattan.

Rey-Debove, J. (1971). *Etude linguistique et semiotique des dictionnaires frangais contemporains. Approaches to Semiotics*, 13. De Gruyter Mouton. https://doi.org/10.1515/9783111323459

Richards, J. C. & Rodgers, T. S. (2014). *Approaches and Methods in Language Teaching*. Cambridge: Cambridge University Press.

Svensén, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge: Cambridge University Press.

Tolcsvai Nagy, G. (2021). The indigenous status of the Hungarian language community in the Carpathian Basin: A historical and contemporary interpretation. *Foreign Policy Review*, 2, pp. 47–61.

Wooldridge, R. (2004). 'Lexicography'. In S. Schreibman, R. Siemens & J. Unsworth (eds) *A companion to digital humanities*. Oxford: Wiley-Blackwell, pp. 69–78.

Where Generative AI Fails: Explaining International Standards through an Explanatory Dictionary

Emmanuelle Pensec, Geoffrey Williams
ARIANE, France, Université Grenoble Alpes
E-mail: emmanuellepensec@outlook.fr, williamg@univ-grenoble-alpes.fr

Abstract

Lexicography has always been at the forefront of technology whether it be with the introduction of printing or the use of digital technology. It is thus not surprising that generative artificial intelligence (AI) along with its so-called Large Language Models (LLM) have been quickly put to lexicographical use. This paper sets out to demonstrate that the advantages brought by tools as ChatGPT in rapidly building entries and even setting them in a desired XML format has limitations. Here, we look at how AI can help in building an experimental explanatory dictionary on Corporate Social Responsibility in French. This entails explaining usage from a special language, that of ISO standards, to a non-specialised audience. What we show is that whilst AI can explain the terms, a number of difficulties arise in that the standard has been translated leading to cross-cultural misinterpretations and that it cannot adapt to a targeted audience which is a discourse community that must rely on mediation through an expert in the language norms of standards and business practice. The solution adopted is to use a corpus-driven approach alongside TLex Tools for management of both the source data and the overall dictionary.

Keywords: Explanatory dictionary; TLex; Corporate Social Responsibility; ISO standards; Sustainability reporting; Discourse communities;

1 Introduction

Lexicography has always been at the forefront of technological change. It was quick to adopt the new Roman characters, and also the printing press (Hanks, 2010). In so doing, it also developed a layout which is still used today. It was therefore obvious that it would adopt the advantages of computers and of corpora (Sinclair, 1987). The task of analysing corporate data was made far easier with the advent of Sketch Engine (Kilgarriff et al., 2004). Digital data, digital dictionaries, dictionary management systems, such as TLex, greatly facilitate the work of lexicographers (Joffe et al., 2003). Given such advances, it would seem obvious that artificial intelligence has a role to play and it is now fully integrated into Sketch Engine, but rather than being plug and play, the use of AI requires a lexicographer's knowledge to formulate requests (De Schryver, 2023). The advantages of professionally used AI are obvi-

ous, but there are also drawbacks inherent in the very nature of large language models as we shall see when discussing the concept of sustainability.

This paper concerns the compilation of a specialised dictionary aimed at French small businesses confronted with the need to apply Corporate Social Responsibility legislation. Such sustainability reports are compulsory for large enterprises and organisations, but they are now also being applied to small and medium-sized enterprises (SME). This creates numerous problems for small companies that do not have specialised management departments to handle such reporting. They are therefore obliged to build their own operational strategies. Apart from the complication of doing this, they are also faced with a problem of language as legislation uses specialised terminology, which have also been translated from English, thereby leading to certain problems of understanding. Here, we shall show some of the problems of using the AI tool ChatGPT to extract terminology so as to build a specialised dictionary that would be adapted to the needs of SME users. We show what AI can do, where its limits lie, and how we overcome these problems when building a specialised dictionary, using the T-Lex tools. Our aim is to build a specialised explanatory online dictionary that can be used by students following managerial courses in Sustainability management and also SMEs faced with implementing a sustainability strategy. We aim to make this dictionary available to managers and students but it remains an exploratory exercise and not a commercial one.

2 Sustainability Reporting

Corporate Social Responsibility, as enshrined by the Global Responsibility Initiative (GRI), has been in existence since 1999, and all major corporations have to file annual reports, which have often been accused of being a form of green-washing. These are often in English and have been studied by (Pensec, 2015, 2016, 2019) both in terms of their linguistic and managerial aspects. With the increasing importance of the United Nations Sustainable Development Goals and their implementation in both national and European policies, sustainability reporting has spread to a wider range of organisations, notably Small and Medium Enterprises. In France, such reporting is essentially enshrined in the application of different international standards, notably ISO 9001 (International Standards Organisation, 2015) for quality management and ISO 26000 (International Standards Organisation, 2010) for social responsibility.

Putting into place sustainability strategies brings notable benefits to companies, but also brings in numerous difficulties. One problem is that smaller companies do not have specialised personnel. They can call upon a consultant to set up a strategy, but they will still have to either recruit or train someone. In both cases, the person will not be a specialist of the standards, but a technician who is applying them. Such persons will require training. Another problem is that standards evolve requiring managers to read and understand updates. The latter brings us to the problem of language.

ISO standards are first published in English, everything else is in translation. Even if the dogma of exactitude in terminology is accepted, translation, inevitably, introduces difficulties of interpretation. If the terminology is understood by specialists, its interpretation by

users will not necessarily be so. This is precisely the problem we are faced with here. One solution that users now have is to have recourse to an AI tool such as ChatGPT, as will be demonstrated in the following section.

3 What is Sustainability?

If we consider the UN Sustainable Development Goals, this notion of sustainability becomes central in all economic activity. The problem lies in how to translate the concept.

We can illustrate this with the concept of sustainability, which ChatGPT defines as:

Sustainability refers to the ability to meet present needs without compromising the ability of future generations to meet their own needs. It involves maintaining a balance between environmental health, economic development, and social well-being.

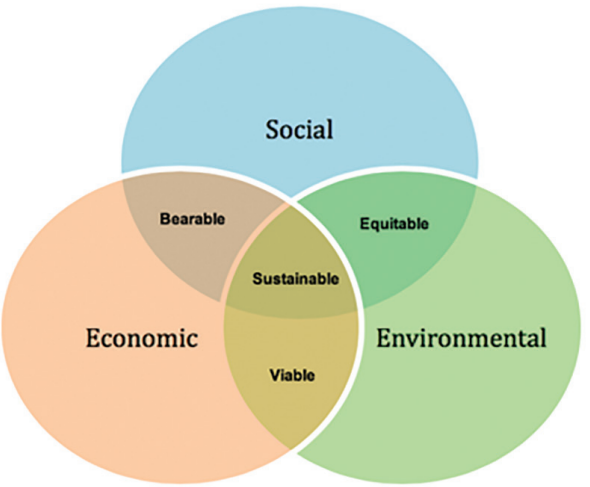


Figure 1: The three pillars of Sustainable Development.

Three pillars of Sustainability:

1. Environmental sustainability:
   - protecting natural resources (like water, air, soil, and biodiversity)
   - reducing pollution and waste
   - combatting climate change
2. Economic sustainability
   - supporting long-term economic growth without negatively impacting social or environmental health
   - promoting responsible consumption and production

3. Social sustainability

- ensuring human rights, equity, and access to basic services (healthcare, education, etc.)
- supporting communities and cultural diversity

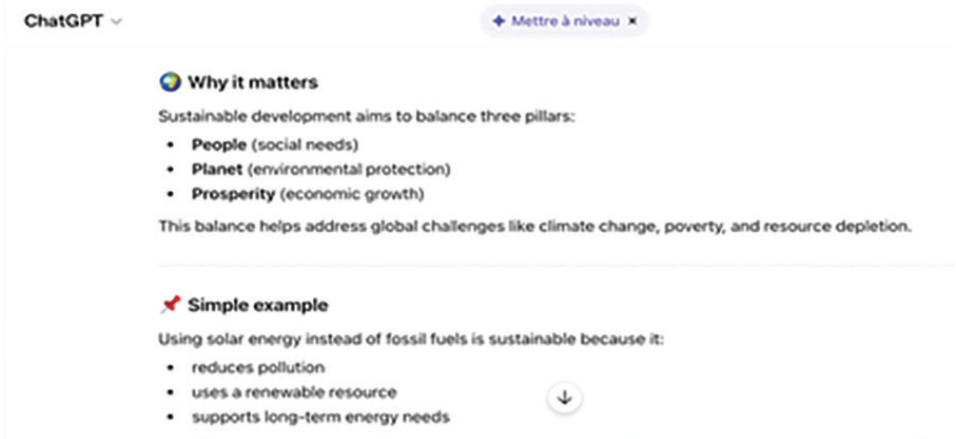


Figure 2: Extract from ChatGPT for 'sustainability'.

Such a definition is perfectly exact, but extremely general, and not really useful to someone who needs to apply sustainability in a business context. Somebody directly concerned with the problem, would probably say 'so what?' Now if we ask a similar question in French, we are faced with the problem of three potential keywords: *développement durable, durabilité*, *soutenabilité*.

According to ChatGPT, the term translates thus:

Le développement durable est un concept qui désigne une manière de répondre aux besoins du présent sans compromettre la capacité des générations futures à répondre aux leurs.

Cette définition vient du rapport publié en 1987 par la Commission mondiale sur l'environnement et le développement (souvent appelé rapport Brundtland).

Les 3 piliers du développement durable

Le développement durable repose sur un équilibre entre trois dimensions essentielles :

1. **Environnemental**

   Protéger la planète, préserver les ressources naturelles, limiter la pollution et lutter contre le changement climatique.

2. **Social**

   Assurer le bien-être de tous : accès à l'éducation, à la santé, respect des droits humains, réduction des inégalités.

3. **Économique**

   Développer l'économie de manière responsable, en créant des richesses et des emplois sans nuire à l'environnement ni à la société.

L'idée clé

Ces trois piliers doivent être **équilibrés**: un développement n'est durable que s'il est **écologiquement viable, socialement équitable et économiquement viable.**

In the same way as for English, this definition is correct and provides interesting information, but remains far too general compared to the precise and specific needs of the target of our dictionary. The first problem is that the term is generally understood in France as only concerning the environment, and anyway, the content is so general that it does not address the real needs of the SMEs managers who are not specialists on the subject of sustainability, and who are looking for a pragmatic approach to implement a sustainability initiative adapted to their specific situation. If we look at the alternative terms, we are not better off as 'durabilité' is seen as purely conceptual, although it does have a technical application in ways the SME reader would not immediately understand. 'Soutenabilité' is widely used in French Canadian usage, but not in France. Thus, the entry points to an issue that is still far from clear to a business user.

Figure 3: Sustainability in the *Grand Dictionnaire Terminologique.*

This observation highlights the central importance of reflecting on, and understanding, the typical user of the standard, his or her practices, and real needs. A definition as proposed by ChatGPT is conceptual and macro, whereas the dictionary user is looking for clear, simple definitions intended for non-specialists, equipped with examples of good practices. This means adapting definitions to a user community, which is something an AI system cannot do without heavily refining the query, something that only someone who is already an expert can do.

In theory, all dictionaries are designed to fit a target audience (Atkins & Rundell, 2008), which can be very general or very precise, although users do not always understand this targeting as the commercial claims of a publisher do not like to limit its appeal. A good example is the OED, an excellent reference work, but not adapted to everyday needs. Similarly, the pragmatic approach of advanced learner's dictionaries is often suitable to native users (McCreary, 2002), but this is because of their clarity. In both cases the user group is broad, but an AI tool could still not fulfil the role of the OED as it would be unable to access

and assess such historical data. This leaves AI as a tool for general usage. For more precise target groups, we need to look at the notion of discourse communities (Swales, 1990).

4 Adapting data to the user: discourse communities

The concept of a discourse community, as developed by Swales (1990), explains how language operates within specialised groups. A discourse community is not just any group of people who communicate, it is a group that shares common goals, and specific ways of using language to achieve those goals. So, what defines the community is not only who belongs, but how communication works inside it.

This definition highlights three key ideas:

- Language is socially situated, in the sense that meaning depends on the community who is using it, so, the same word or structure can mean different things in different discourse communities;
- Genres are not only forms, they are goal-oriented communicative actions shaped by the community itself;
- Being a member of a discourse community means learning its genres, mastering its vocabulary and understanding its norms.

It appears that a discourse community is not just a speech community (shares a language), nor just a social group (social interactions). A discourse community shares communicative practices and goals and is organised around communication and knowledge exchange.

To understand the needs of the dictionary user, it is appropriate to identify the characteristics of the target. In the case of small and medium-sized companies, this means an 'operational' user who is both:

- Pragmatic: they are not looking for a militant, ideological, or conceptual discourse but to address a concrete constraint. In sum, sustainability is a topic among others on a busy agenda
- Non-specialist in the subject: vague topic, full of jargon, cosmetic, but whose rational is to understand what sustainability will concretely bring them

From there, it is appropriate to define the user's needs in terms of sustainability when looking at sustainability. Thus, in our case, the SME manager is looking for:

- Immediate clarity: a simple explanation of sustainability topics without abstract concepts. A manager is seeking a direct translation of them into issues for his or her type of company.
- Educational support: an approach based on popularisation without infantilisation and a reassuring stance in which he or she does not feel judged.
- On-the-ground credibility: explanations adapted to his type of company and not inspired by large companies that do not correspond to his or her reality.

From an analytical point of view, the above highlights two things:

- The user of the dictionary constitutes a community defined by the characteristics we have just listed;
- The language used in the user's context constitutes a discursive genre characterised by a rejection of jargon, abstract concepts, and unnecessary complexity.

This is where a new complexity comes into play. French SME users form a discourse community with their own language and norms. Their usage requires understanding the language of the standards and relating it to their own practice within a French legal context. It also must take into account cultural aspects leading to semantic prosodies (Louw, 1993) that are not immediately apparent. Different languages can interpret what may seem a simply translation in culturally biased ways (Williams & Pensec, 2025). The added layer of complexity comes in through the fact that the ISO standard community is also a discourse community, so that the dictionary writer must mediate needs between the two communities.

5 The Dictionary

This is not an experimental explanatory dictionary in the Mel'čukian sense (Mel'čuk, 1996), but one that explains following Hank's call for a reversal of the Aristotelian paradigm of *definiendum-definiens* by explicandum-explicans (Hanks, 2013). The object here is to take a term and explain it in language that the target user can understand. This is about decoding, not encoding. The dictionary will also be corpus-driven in part, and certainly corpus-based.

Since the introduction of corpora in dictionary production during the COBUILD project (Sinclair, 1987), the use of corpora has become the norm. A so-called Large Language Model, the textual elements behind an AI tool, can be seen as a large corpus, but as we have seen they can only handle very general language use. Consequently, in this case we have adopted a mixed corpus approach.

Dictionaries in the 21st century are based on digital documentation, including corpora, in the sense of Sinclair:

- a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research (Sinclair et al., 1987)
- a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language (Sinclair, 1996).

The corpus is therefore not a collection of texts, as LLMs, but a set of texts selected in such a way as to be able to analyse the use of language in a specific context. Unlike LLMs, the corpus is built to answer a previously defined question and of which it is representative.

In the context at hand, the dictionary aims to provide clear information for a non-expert audience and address the problem they meet. The content of the selected textual data therefore plays an essential role and needs to reflect the needs of the discursive community concerned. The data selected to build the corpus must meet three criteria:

- be as neutral as possible
- be representative of the subject
- have proven its legitimacy
- correspond to the reality on the ground for future users

To build the most relevant corpus, two options have been considered:

The GRI corpus, a collection of reports in three languages composed of company reports in the context of the Global Reporting Initiative, is:

- built to analyse sustainability language as it is used by companies
- this corpus respects the constraints of corpus creation in the sense of Sinclair
- based on a sustainability reporting framework whose objective is to organise the sustainability reports
- corpus already created (Pensec, 2019)

The ISO26000 corpus, consisting of the current ISO standards :

- achieves the objective of ISO26000, to explain how to deploy a sustainability strategy in organisations
- does not meet the constraints for creating a corpus in the sense of Sinclair but the methodology is the same
- sustainability approach framework whose objective is to organise the process of deploying the sustainability approach
- corpus to be created.

The two standards are not intended to define what is right or wrong (there is no moral dimension in these two standards). They therefore do not indicate the concrete content of sustainability within an organisation. They provide a framework and methodology through which the process is organised, allowing relevant stakeholders to define their vision of the fair operation of the organisation. These two standards structure, guide, and organise the discursive space.

Among these two possibilities, the choice of the ISO26000 corpus clearly appeared as a more pragmatic and relevant choice in view of the needs of the future users of the dictionary.

- SMEs managers implement an approach before organising its reporting. The primary need of users is therefore to understand the sustainability approach. It is ISO26000 that meets this need.
- This international standard is the foundation on which the French sustainability labels used by organisations are based.
- This dictionary has a practical purpose and seeks to meet a concrete need of its non-expert users.

6 TLex as Dictionary Management System

TLex (Joffe et al., 2003) is a set of tools with dictionary production and management that were developed in South Africa. They are available for commercial use, but there is a price for academics who wish to use the tools for the creation of dictionaries within the teaching and research environment. The tools allow the building of structured dictionary entries, along with their export in different formats. Associated tools include Tl corpus, which is a corporate analysis system from which examples can be inserted directly into the dictionary. Another tool is tlTerm which allows the building of terminology data bases. We shall not be using this here, although the starting point is terminological, our aim is to build a special language dictionary.

The latest versions of the TLex tools allows the lexicographer to work directly with AI, contain the full dictionary structure, including the definitions (de Schryver & Joffe, 2023). This is not something we intend to do for the reasons given above. However, we do make use of TLex so as to have a consistent, coherent and well-structured dictionary. Although it can be integrated with Sketch Engine, we are not doing this but have preferred to extract a list of terms with Sketch Engine, but seek examples directly from the GRI French corpus using tlCorpus. As in all good terminological practice, the terms attracted the terms. The candidate terms extracted are checked by an expert before being included in the dictionary. In this case, we pay especial attention to acronyms as these are often the biggest problem for someone who is new to a field.

	B	C	D	E	F	G
1						
2						
3	method name: extract_keywords					
4	corpus: user/emmanuellepensec0/corpus_rse_iso26000					
5	subcorpus: -					
6	**Item**	**Frequency (focus)**	**Frequency (reference)**	**Relative frequency (focus)**	**Relative frequency (reference)**	**Score**
7	Responsabilité sociétale	531	8334	5976,36475	0,46807	4071,585
8	Question centrale	86	16117	967,92346	0,90519	508,571
9	Partie prenante	252	109007	2836,24097	6,12223	398,364
10	Intégration de la responsabilité sociétale	35	20	393,92233	0,00112	394,479
11	Norme internationale de comportement	23	61	258,86325	0,00343	258,976
12	Devoir de vigilance	22	2599	247,60832	0,14597	216,941
13	Contribution au développement local	19	114	213,84355	0,0064	213,477
14	Dialogue avec les parties prenantes	19	636	213,84355	0,03572	207,434
15	Développement local	46	27069	517,7265	1,52029	205,82
16	Groupe vulnérable	21	4613	236,35341	0,25908	188,513
17	Responsabilité sociétale dans une organisation	13	0	146,31401	0	147,314
18	Comportement éthique	14	1624	157,56894	0,09121	145,315
19	Intérêt des parties prenantes	13	301	146,31401	0,01691	144,865
20	Participation des parties prenantes	13	521	146,31401	0,02926	143,126
21	Chaîne de valeur	27	21051	303,88293	1,1823	139,707
22	Brève description objective	12	0	135,05908	0	136,059
23	Consommation durable	13	2039	146,31401	0,11452	132,177
24	Attente de la société	13	2151	146,31401	0,12081	131,436
25	Description objective	12	630	135,05908	0,03538	131,409
26	Initiative volontaire	12	665	135,05908	0,03735	131,16
27	Performance en matière de responsabilité sociétale	11	0	123,80416	0	124,804

Figure 4: SketchEngine list before cleaning.

Once imported into TLex, each entry is structured.

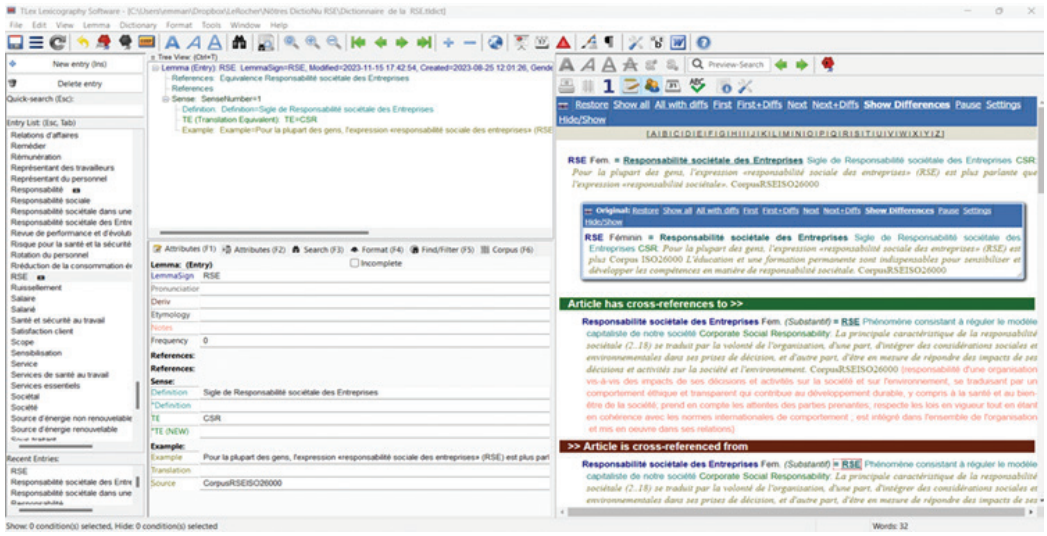


Figure 5: Structuring data in TLex.

This example concerns an acronym, RSE. It shows the acronym, its signification and the English term, CSR or Corporate Social Responsibility. The idea is not to create a bilingual dictionary as the user community is French speaking and using the tool for decoding only. However, we consider it important that the user have access to the language in which the term was initially created. In this case, the expert adds the signification manually, but the examples are drawn from the text of the current version of the standard. Acronyms and abbreviations are considered as entries in their own right so as to save the user's time, they are, however, rigorously cross-referenced in both directions.

Definitions are extracted from a number of sources and compiled and adapted using the knowledge of the expert who is both a consultant and trainer, thus capable of making the definition understandable to the target community. Using an ISO definition directly would defeat the purpose.

This can be illustrated by the entry for *Communautés locales (local communities)* which is defined as:

des personnes ou des groupes de personnes vivant ou travaillant dans des zones qui sont affectées ou qui pourraient être affectées par les activités de l'organisation. La communauté locale s'étend aux personnes vivant à proximité des opérations de l'organisation à celles vivant à distance

The first thing that we have done in standardising the entries is the use of the formula *person* or *group of persons* so as to avoid vague third person indications. Such practice is carried out throughout the dictionary so as to look language that is closer to that of the user. We are also faced with the problem that *community* does not have the same meaning between French and English and is yet another case of the lost in translation that we have explored with the notion of *work* (Williams & Pensec, 2025). In other cases, definitions require further precision such as when we speak of clients which is potentially ambiguous as

it can refer to other companies or to the final user, the consumer of a product or service. These are legal terms with reference to French and EU law, whereas the ISO standard is obviously neutral.

7 Conclusion

Without a doubt, AI models based on Large Language Models can bring a lot to lexicography. However, the limits of automatised systems must be taken into account. The inclusion of AI in corpus analysis tools such as SketchEngine and a Dictionary Management System such as TLex will greatly ease the work of lexicographers working on general language, but will be more limited on special language works. Here, we have illustrated the difficulties involved in building an experimental explanatory dictionary in French. We have highlighted a number of problems starting with the fact that the source language is English and therefore the usual document which is in French suffers from the dangers of translation with cultural interpretation of some language. The other problem is that building the resource from an ISO standard is fine if we are dealing with the terminology, provided the source is the most recent version of the standard. However, what we are finding in this case is that we have a clash of two different discourse communities, that of ISO standards and that of business users. The dictionary must then be mediated by an expert who understands both special languages and can render the text meaningful for a particular target community. The following stage is classic lexicographical procedure using the dictionary management system, in our case TLex, but here too, a great deal of human intervention is needed using an integrated corpus, but not necessarily the AI tools. In sum, AI tools can bring a lot to lexicography, but the human input remains central as dictionaries are essentially tools for humans compiled by humans with a real knowledge of social needs. If Artificial Intelligence is useful, it must remain a tool with real human intelligence dominating.

References

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

De Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, *36*(4), 355–387.

de Schryver, G.-M., & Joffe, D. (2023, February 27). *The end of lexicography, welcome to the machine: On how ChatGPT can already take over all of the dictionary maker's tasks*. 20th CODH Seminar. https://tshwanedje.com/articles/CODH-video-presentation-TLex-OpenAI-Integration/

Hanks, P. (2010). Lexicography, Printing Technology, and the Spread of Renaissance Culture. In *Proceedings of Euralex 2010*. Fryske Akademy. http://www.patrickhanks.com/uploads/5/1/4/9/5149363/renaissancelexicography.pdf [07/02/2026]

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations* (MIT Press).

International Standards Organisation. (2010). *ISO 26000: Guidance on Social Responsibility*.

International Standards Organisation. (2015). *ISO 9001: Quality management systems—Requirements*.

Joffe, D., de Schryver, G.-M., & Prinsloo, D. J. (2003). Introducing TshwaneLex – A New Computer Program for the Compilation of Dictionaries. *TAMA 2003 South Africa: Conference Proceedings*, 97–104.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress* (pp. 105–115). Université de Bretagne Sud. https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202004/011_2004_V1_Adam%20KILGARRIFF,%20Pavel%20RYCHLY,%20Pavel%20SMRZ,%20David%20TUGWELL_The%20%20Sketch%20Engine.pdf [07/02/2026]

Louw, W. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker (Ed.), *Text and Technology*. John Benjamins.

McCreary, D. R. (2002). American Freshmen and English Dictionaries: 'I Had Aspersions of Becoming an English Teacher'. *International Journal of Lexicography*, *15*(3), 181–205. https://doi.org/10.1093/ijl/15.3.181

Mel'čuk, I. (1996). Lexical functions: A tool for the description of lexical relations in the lexicon. In L. Wanner (Ed.), *Lexical functions in lexicography and natural language processing* (pp. 37–102). John Benjamins.

Pensec, E. (2015). La traducción de los informes de sostenibilidad de la Global Reporting Initiative: Qué papel para la contingencia cultural? In D. Gallego-Hernández (Ed.), *Enfoques Actuales en Traducción Económica e Institucional—Actas del Congreso Internacional de Traducción Económica, Comercial, Financiera e Institucional* (p. 29-39). Peter Lang.

Pensec, E. (2016). Building meaning in Context: Collocational analysis of the verb « work » in a specialised corpus of English non-financial reports. In D. Gallego-Hernández (Ed.), *New Insights into Corpora and Translation* (pp. 17–39). Cambridge Scholars Publishing.

Pensec, E. (2019). *L'impact de la GRI sur l'homogénéisation du discours RSE: Analyse collocationnelle trilingue d'un corpus de rapports RSE normalisés selon le référentiel de la GRI*. Université Bretagne Sud.

Sinclair, J. (Ed.). (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Collins.

Swales, J. (1990). *Genre analysis*. Cambridge University Press.

Williams, G. C., & Pensec, E. (2025). "All Work and No Play": The Collocational Resonance of 'Work'. *International Journal of Lexicography*. https://doi.org/https://doi.org/10.1093/ijl/ecaf006

A Model of a Georgian-English Learner's Online Dictionary of Collocations (based on the principles of collocations dictionaries and I. Mel'čuk's theory of Lexical Functions)

Salome Tchighladze
Ilia State University
salome.tchigladze.1@iliauni.edu.ge

Abstract

The present article aims to propose a model of a Georgian–English online learner's dictionary of collocations based on the principles of collocations dictionaries and Igor Mel'čuk's theory of lexical functions. It presents a model of dictionary entries in which collocations are organized according to parts of speech and enriched with relevant lexical functions and authentic usage examples. The proposed approach seeks to develop an active, text-synthesis-oriented dictionary that will contribute to Georgian learner lexicography and support the effective use of lexicographic resources in both teaching and research contexts.

In the process of foreign language acquisition, lexical competence extends beyond knowledge of the meanings of individual words. Effective communication largely depends on the ability to use correct and natural word combinations, enabling language learners not only to comprehend texts but also to produce well-formed sentences. From this perspective, collocations are of particular importance, as they constitute relatively stable and coherent combinations of lexemes. Knowledge of collocations contributes significantly to the naturalness and fluency of speech and distinguishes native-like language use from learners' language.

Although collocation dictionaries are well established in English lexicography, no comparable resource currently exists in Georgian-English learner lexicography. Most existing bilingual dictionaries focus primarily on the translation of individual words and devote little attention to their collocational potential. This creates a significant challenge for language learners, particularly in the context of text production and synthesis.

Keywords: dictionary of collocations; Mel'čuk's lexical functions; model of the dictionary entry; Georgian-English lexicography

1 Introduction

As early as the 1920s and 1930s, Harold Palmer and Albert Hornby emphasized the importance of word combinations in the process of language learning (Cowie 1999: 56). However,

this linguistic phenomenon as a term was first introduced into linguistics in the 1950s by the British linguist John Rupert Firth (1957). Following Firth, John Sinclair's research significantly expanded and deepened the understanding of the practical importance of collocations, which came to be regarded as one of the main principles of corpus-based research and lexical analysis.

In this regard, the COBUILD project (Collins Birmingham University International Language Database) deserves special mention, as it was the first electronic corpus. It brought together examples of the natural use of English, on the basis of which the first corpus-based dictionary – the *COBUILD English Dictionary* – was compiled. Through this project, Sinclair shifted the focus from grammatical rules to the issue of word combinability and collocational behavior (Sinclair 1991).

The further development of corpus linguistics provided an even stronger methodological foundation for the study of collocations. Particularly noteworthy in this context is the creation of the Sketch Engine platform, one of the major achievements of corpus linguistics. The program was developed in 2004 by Adam Kilgarriff (Kilgarriff et al. 2004) and his team and currently includes around 800 corpora covering more than 100 languages. It is equipped with advanced search tools, and it was this platform that introduced collocational search functionality into corpus analysis (Kilgarriff & Rundell 2002). The Macmillan Dictionary was the first to make use of this functionality, adding the most frequently used collocations to learner dictionary entries (Rundell & Kilgarriff 2011).

The issue of collocations is also central to Igor Mel'čuk's work, both in his theory of the Explanatory Combinatorial Dictionary (Mel'čuk 2012) and in his theory of lexical functions (Mel'čuk 1996). According to the theory of lexical functions, words are interconnected within different semantic contexts, and these relationships can be described as lexical functions. The theory identifies around 70 lexical functions, which Mel'čuk classifies into different categories, including syntagmatic and paradigmatic functions. Syntagmatic lexical functions primarily describe relationships between different parts of speech and specify the nouns, adjectives, adverbs, or verbs that typically co-occur with a given lexeme. Thus, these functions provide semantically expected and appropriate word combinations. Paradigmatic lexical functions (such as synonyms, antonyms, conversives, and others) describe words that belong to the same semantic category and can potentially replace a given lexeme. Syntagmatic lexical functions, by contrast, are used together with the given word in actual usage.

Within the framework of the present research, a new methodology was developed following a review of the theoretical literature on collocations, the principles of compiling collocations dictionaries, and Igor Mel'čuk's theory of lexical functions. According to this methodology, collocations in the entries of the Georgian-English learner's dictionary of collocations are classified according to parts of speech and are supplemented with the corresponding lexical functions. This approach enables users to gain a detailed understanding of the combinatorial potential of lexical units and, when necessary, to construct well-formed syntactic structures independently. From Mel'čuk's system of lexical functions, only those syntagmatic and paradigmatic functions were selected that are relevant to Georgian-English lexicographic goals within the context of a learner's dictionary of collocations. The article provides a detailed discussion of a noun-entry model compiled according to this methodology.

The chosen approach is considered to facilitate the creation of an active, text-synthesis-oriented dictionary, which significantly enhances the effective use of lexicographic resources in both teaching and research contexts.

Each dictionary entry includes not only collocations but also authentic usage examples accompanied by English translations, thereby providing language learners with a realistic understanding of how particular words function in everyday communication. The structure of the dictionary entry is based on the principles of collocations dictionaries employed in the *Oxford Collocations Dictionary* and the *Macmillan Collocations Dictionary for Learners of English*. In these dictionaries, collocations within each entry are grouped according to parts of speech, and in some cases extended phrases are also provided, which makes this type of dictionary particularly practical for language-learning purposes. A collocation-based dictionary model enriched with lexical functions represents a novel approach in the field of Georgian-English learner lexicography.

2 Collocation Dictionaries as a Lexicographic Genre

Corpus linguistics gave new impetus to collocation dictionaries as an independent lexicographic genre. Such dictionaries describe not only individual words and their definitions, but also collocational behavior – the ways in which words combine with one another in different contexts (Moon 1998). Collocation dictionaries are largely based on large-scale language corpora, which makes such dictionaries more reliable and empirically grounded (Sinclair 1991; Stubbs 2001).

The first collocations dictionary is generally considered to be *The BBI Combinatory Dictionary of English*, compiled by Morton Benson, Evelyn Benson, and Robert Ilson, first published in 1986 (Benson, Benson, & Ilson 1986). The BBI dictionary distinguishes between grammatical and lexical collocations. It does not include classical idioms, but rather focuses on word combinations in which the meanings of the components remain at least partially transparent. Grammatical collocations consist of a headword combined with prepositions or other grammatical elements, whereas lexical collocations mainly involve combinations of nouns, verbs, adjectives, and adverbs. It is noteworthy that the dictionary conforms to Firth's theoretical principles (Poulsen 1991).

In fact, the BBI dictionary was not the first of its kind. Its predecessors include Albrecht Reum's *Dictionary of English Style* (1961), which was intended to assist German learners of English in their language learning. Another work that can be classified within this genre is Joseph Isaac Rodale's *The Word Finder* (1947), which provides information useful for young writers.

The publication of the BBI dictionary marked an important milestone in collocations lexicography, as it established a standard for the presentation of collocations in dictionaries. From the 1980s onwards, other collocations dictionaries began to appear. Among the relatively newer generation of such dictionaries is the *Oxford Collocations Dictionary for Students of English*, first published in 2002 and based on the British National Corpus (BNC). Another example is the *Longman Collocations Dictionary and Thesaurus* (2013), which combines the principles of a collocations dictionary and a thesaurus. It provides users not only with infor-

mation on collocations, but also on synonym sets and semantic relations. The *Macmillan Collocations Dictionary* is another representative of this genre. These modern dictionaries have become important resources not only for language learners, but also for translators, lexicographers, and linguists in general.

2.1 The Oxford and Macmillan Collocations Dictionaries

In order to develop the research methodology, initially the principles underlying the compilation of existing collocation dictionaries and the methods used to present collocations in them were examined. For this purpose, two dictionaries were selected. The first is the *Oxford Collocations Dictionary* (online edition) (see Figure 1). In this dictionary, collocations are organized according to parts of speech: in the case of nouns, the dictionary lists the adjectives that typically modify them, as well as nouns that are used attributively with the given noun, and verbs that either precede or follow the noun. In some cases, collocations are accompanied by examples in the form of full sentences or extended phrases.

Below is an example from the aforementioned dictionary illustrating the noun entry for the word *book*. It's collocates are grouped according to parts of speech. For example, the entry presents adjectives that frequently occur in collocation with the noun *book*, such as *latest*, *new*, *recent*, *forthcoming*, *rare*, etc. It also provides an authentic example sentence, for example: *There's nothing like curling up with a mug of tea and a good book.* In some cases, the position of the headword within the collocation is also indicated. For example, the entry presents verbs that occur before the noun *book*, such as *look at* and *read*, as well as verbs that occur after it, including *appear* and *come out* (see Figure 1).

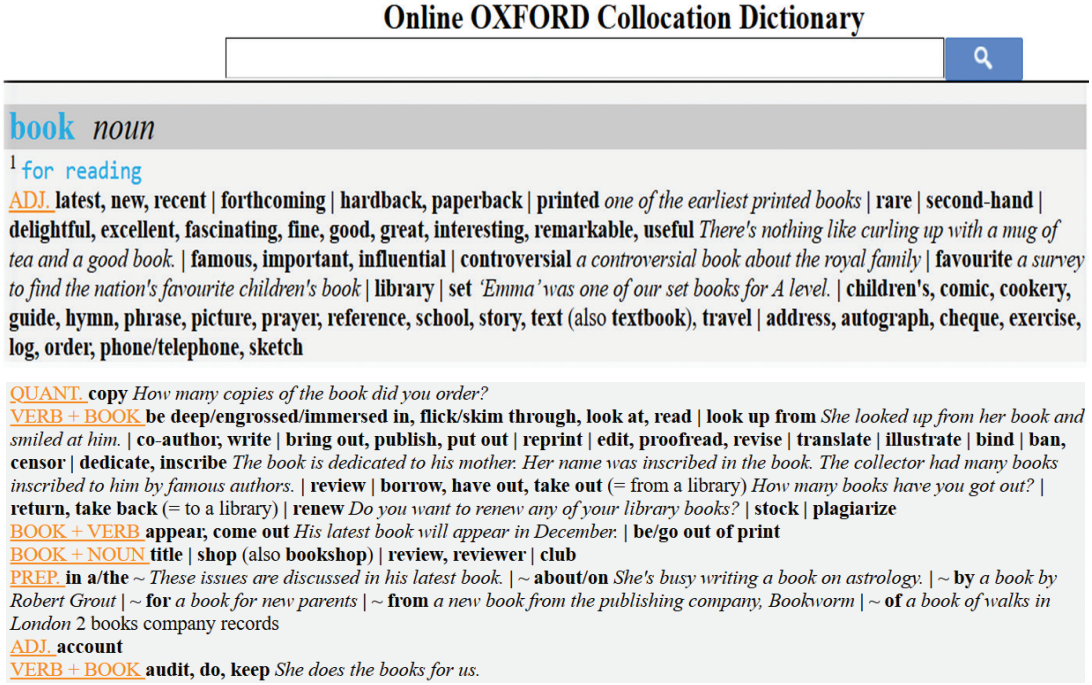
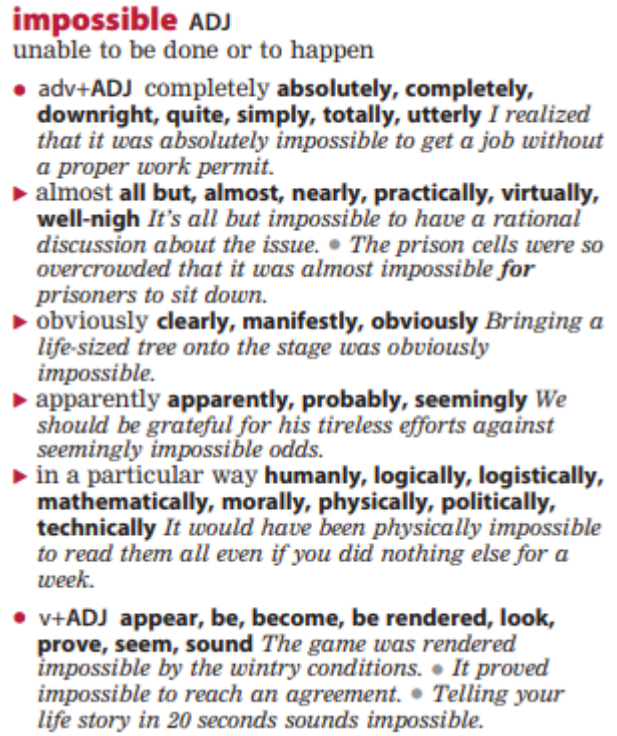


Figure 1: *Oxford Collocations Dictionary* (online), entry "book".

The second dictionary selected for the same purpose is the *Macmillan Collocations Dictionary for Learners of English* (see Figure 2). Like the *Oxford Collocations Dictionary*, the Macmillan dictionary organizes collocations according to parts of speech, which makes it easier for users to understand and remember related word combinations. However, the Macmillan dictionary not only groups collocations by parts of speech, but also separates collocates with similar meanings into distinct groups (see Figure 2).

The example from the *Macmillan Collocations Dictionary* illustrates the adjective entry for the word *impossible*. The collocations are organized according to the part of speech with which the headword combines.

In the section labeled *adv + ADJ*, the dictionary presents adverbs that frequently occur with the adjective *impossible*. For example, intensifying adverbs such as *absolutely*, *completely*, and *totally* are used in combinations like *absolutely impossible* and *totally impossible*. Some collocations are also illustrated through example sentences, such as: *I realized that it was absolutely impossible to get a job without a proper work permit.* The dictionary further groups collocations according to semantic similarity. For instance, adverbs such as *almost*, *nearly*, and *virtually* are grouped together because they express partial possibility, as in *almost impossible*. Similarly, adverbs such as *obviously* and *clearly* indicate certainty, producing collocations such as *obviously impossible* (see Figure 2).



impossible ADJ
unable to be done or to happen

- adv+ADJ completely **absolutely, completely, downright, quite, simply, totally, utterly** *I realized that it was absolutely impossible to get a job without a proper work permit.*
- ▸ almost **all but, almost, nearly, practically, virtually, well-nigh** *It's all but impossible to have a rational discussion about the issue.* • *The prison cells were so overcrowded that it was almost impossible* ***for*** *prisoners to sit down.*
- ▸ obviously **clearly, manifestly, obviously** *Bringing a life-sized tree onto the stage was obviously impossible.*
- ▸ apparently **apparently, probably, seemingly** *We should be grateful for his tireless efforts against seemingly impossible odds.*
- ▸ in a particular way **humanly, logically, logistically, mathematically, morally, physically, politically, technically** *It would have been physically impossible to read them all even if you did nothing else for a week.*

- v+ADJ **appear, be, become, be rendered, look, prove, seem, sound** *The game was rendered impossible by the wintry conditions.* • *It proved impossible to reach an agreement.* • *Telling your life story in 20 seconds sounds impossible.*

Figure 2: *Macmillan Collocations Dictionary for Learners of English*, entry "impossible".

Another feature that distinguishes the *Macmillan Collocations Dictionary* from the *Oxford Collocations Dictionary* concerns the way semantic groups are labeled at the beginning of dictionary entries. In the Oxford Learner's Dictionary, the position of the headword within a collocation is indicated by patterns such as verb + headword, whereas in the Macmillan Dictionary, both elements of the collocation are explicitly identified by their parts of speech (e.g., adj + N).

2.2 Lexical Functions of Igor Mel'čuk

The theory of lexical functions is part of Igor Mel'čuk's Meaning-Text Theory (MTT) and describes the semantic and combinatorial relationships between lexical units. A lexical function (LF) is a key word associated with a specific lexeme that expresses its typical collocational behavior. It has no independent meaning and acquires its functional value only in relation to a given lexical item (Mel'čuk 1996).

The theory of LFs is presented as the principal method for compiling the Explanatory Combinatorial Dictionary (ECD). Unlike traditional dictionaries, the ECD is focused not only on explaining the meaning of a word, but also on describing its use in different contexts. This approach is particularly important for the creation of an active, text-synthesis-oriented dictionary, as it provides users with information about a word's typical collocational patterns (Mel'čuk 2012).

Mel'čuk's theory describes approximately 70 LFs, which are divided into syntagmatic and paradigmatic functions. Syntagmatic LFs describe typical word combinations and indicate the adjectives, verbs, or adverbs that commonly occur with a given lexeme. For example, the LF *Magn* denotes an intensifying function (*to condemn* - ***strongly*** *condemn*, patience - **infinite** patience), the LF *Son* refers to the characteristic sound associated with a lexeme (a dog **barks**, wind **howls**), and the LF *Figur* expresses figurative or metaphorical usage (**wall** of fog, **flames** of passion). Paradigmatic lexical functions, by contrast, describe words that belong to the same semantic category, such as synonyms, antonyms, collective nouns and some others. For example, *Syn* expresses a synonymic relationship (to telephone - to phone), *Anti* denotes a semantic opposite (victory - defeat), and *Mult* refers to a collective noun (ship - fleet).

3 The Structure of a Dictionary Entry

After reviewing the models of collocations dictionaries and the principles underlying their compilation, a new approach to the presentation of collocations in the dictionary was developed. Specifically, collocations in the entries of the Georgian-English learner's dictionary are presented according to a mixed principle: each entry follows the structure of collocations dictionaries by grouping the collocations of a headword according to parts of speech, while also being enriched with the syntagmatic and paradigmatic LFs proposed by Igor Mel'čuk.

Using this method, entry models were developed for nouns, adjectives, and verbs. However, the present article focuses exclusively on the noun-entry model. Below is the dictionary entry for the noun *წვიმა/ts'vima* 'rain' (see Figure 3).

3.1. Fields of a dictionary entry

The dictionary entry for *rain* consists of three fields. The first field includes collocations grouped according to parts of speech: adjective + rain, rain + noun, and rain + verb (e.g. constant rain, beating of rain, rain started, etc.). Syntagmatic lexical functions in the entry are marked in red (see Figure 3). *Magn* is one of the syntagmatic lexical functions identified by Igor Mel'čuk (as mentioned above), meaning an "intensifier" (e.g. pelting rain). When the cursor is placed over the term, the meaning of the lexical function appears on the screen (see Figure 4).

Son denotes the syntagmatic lexical function referring to the "sound produced by rain" (e.g. beating of rain, drumming of rain), while *Figur* refers to the lexical function "figurative" (e.g. rain curtain).

Figure 3: The entry for *rain* in the Georgian-English learner's dictionary of collocations.

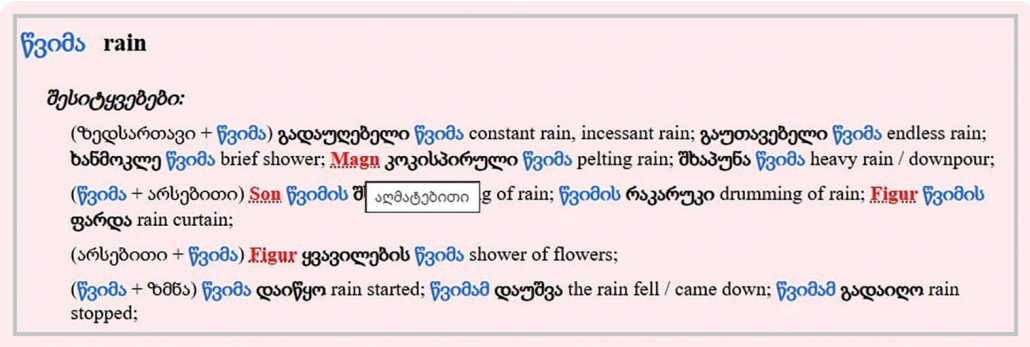


Figure 4: A fragment of the entry for *rain* in the Georgian-English dictionary of collocations.

The second field of the entry is devoted to paradigmatic lexical functions. This section can be toggled on or off (see Figure 5). It includes synonyms of *rain* (e.g. fine drizzle, heavy downpour), a hypernym (meteorological phenomenon), co-hyponyms (snow, hail, wind, storm), and derivatives (rain-washed, rainy, etc.) (see Figure 5). Although paradigmatic functions are not collocations, their presentation in this format was considered important in a learner's dictionary.

The third field is the field of idioms.

Figure 5: The full entry for *rain* with toggled on field of paradigmatic lexical functions.

3.2 Related words

In addition to the fields discussed above, some entries also include a field of related words associated with particular syntagmatic lexical functions. These fields can be toggled on or off. For example, consider the entry for მგელი / *mgeli* 'wolf', which contains two lexical functions: *Mult*, meaning "collective noun," and *Son*, denoting "the sound produced by a wolf" (see Figure 6). This entry therefore includes two corresponding fields of related words linked to the lexical functions *Mult* and *Son*. When activated, the first field displays words associated with the lexical function *Son*, i.e. sounds produced by a wolf. These include expressions referring to sounds produced by other animals and birds, such as the grunting of a pig, cackling of a hen, whinnying of a horse, and twittering of birds, etc (see Figure 7).

The second field of related words is associated with the lexical function *Mult*, meaning "a collective noun," or "a pack of wolves" in this case. When toggled on, it displays collective nouns for different animals and birds, such as a flock of sheep, a herd/drove of horses, a herd of swine/pigs, a flock of deer, and a flock of birds (see Figure 7).

Figure 6: The entry for *wolf* in the Georgian–English dictionary of collocations.

Figure 7: The entry for *wolf* with toggled on fields of words related to the lexical functions *Son* and *Mult*.

The entry for *wolf* also includes a field of paradigmatic lexical functions, comprising a generic term, co-hyponyms, derivatives and so on (see Figure 8).

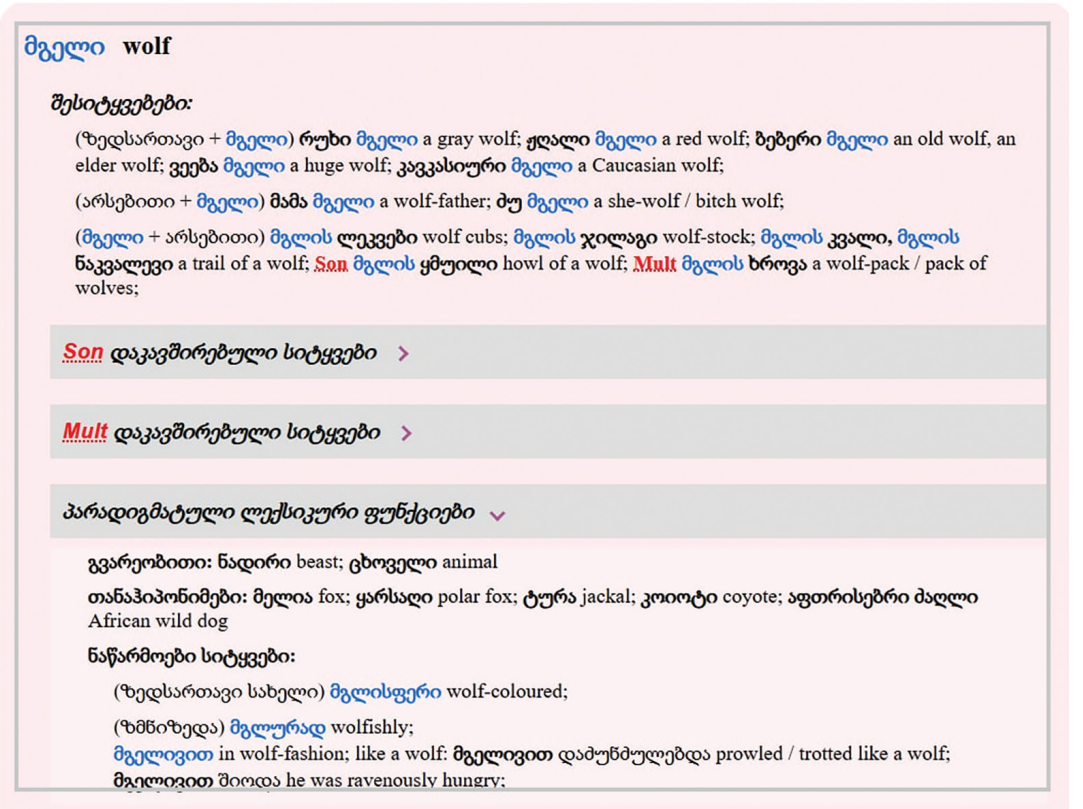


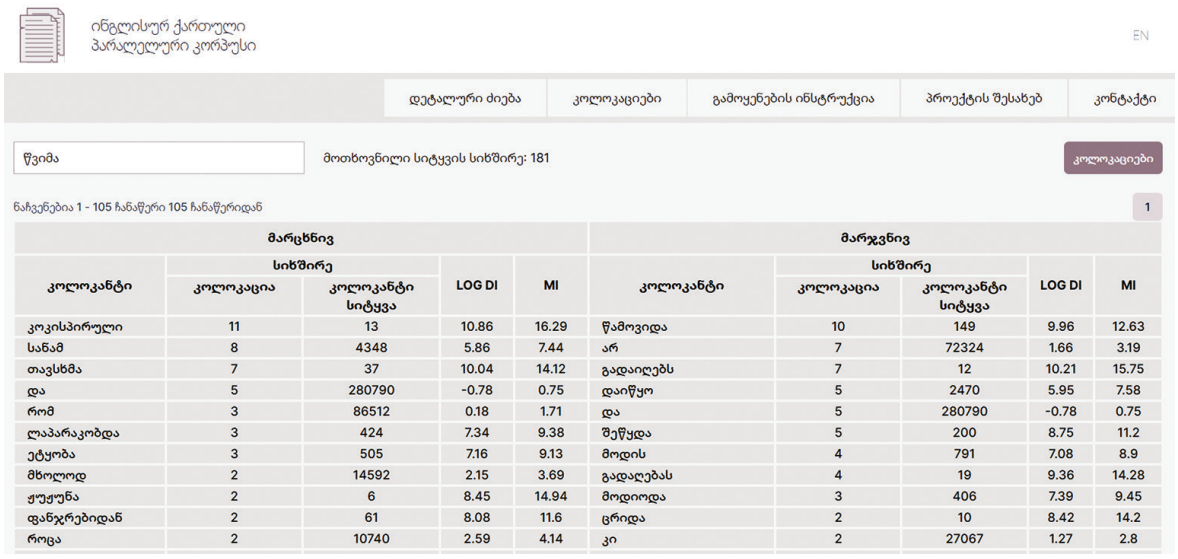
Figure 8: The entry for *wolf* with the paradigmatic lexical functions toggled on.

4 Sources of the *Georgian-English Online Learner's Dictionary of Collocations*

The main sources of the *Georgian-English Online Learner's Dictionary of Collocations* are the following:

The English-Georgian Parallel Corpus.[1] The corpus consists of scientific texts and works of fiction, comprising texts available in both Georgian and English (Margalitadze, Meladze, & Furtskhvanidze 2022). At present, the corpus contains up to 158 text groups, 6,062 text sets, 668,048 manually aligned sentence pairs and 17 million tokens. It also incorporates material from *The Comprehensive English–Georgian Dictionary*. Since 2024, the corpus has been available on a new platform that includes a collocational search function. This platform proved particularly useful, as it allows collocations to be retrieved together with their English translations (see Figure 9), thereby serving as the primary source for the present study.

1 https://enkacorpus.iliauni.edu.ge/en/



მარცხნივ					მარჯვნივ				
კოლოკანტი	სიხშირე		LOG DI	MI	კოლოკანტი	სიხშირე		LOG DI	MI
	კოლოკაცია	კოლოკანტი სიტყვა				კოლოკაცია	კოლოკანტი სიტყვა		
კოკისპირული	11	13	10.86	16.29	წამოვიდა	10	149	9.96	12.63
სანამ	8	4348	5.86	7.44	არ	7	72324	1.66	3.19
თავსხმა	7	37	10.04	14.12	გადაიღებს	7	12	10.21	15.75
და	5	280790	-0.78	0.75	დაიწყო	5	2470	5.95	7.58
რომ	3	86512	0.18	1.71	და	5	280790	-0.78	0.75
ლაპარაკობდა	3	424	7.34	9.38	შეწყდა	5	200	8.75	11.2
ეტყობა	3	505	7.16	9.13	მოდის	4	791	7.08	8.9
მხოლოდ	2	14592	2.15	3.69	გადაღებას	4	19	9.36	14.28
ჟუჟუნა	2	6	8.45	14.94	მოდიოდა	3	406	7.39	9.45
ფანჯრებიდან	2	61	8.08	11.6	ცრიდა	2	10	8.42	14.2
როცა	2	10740	2.59	4.14	კი	2	27067	1.27	2.8

Figure 9: English-Georgian Parallel Corpus - Collocates of the Word *rain*.

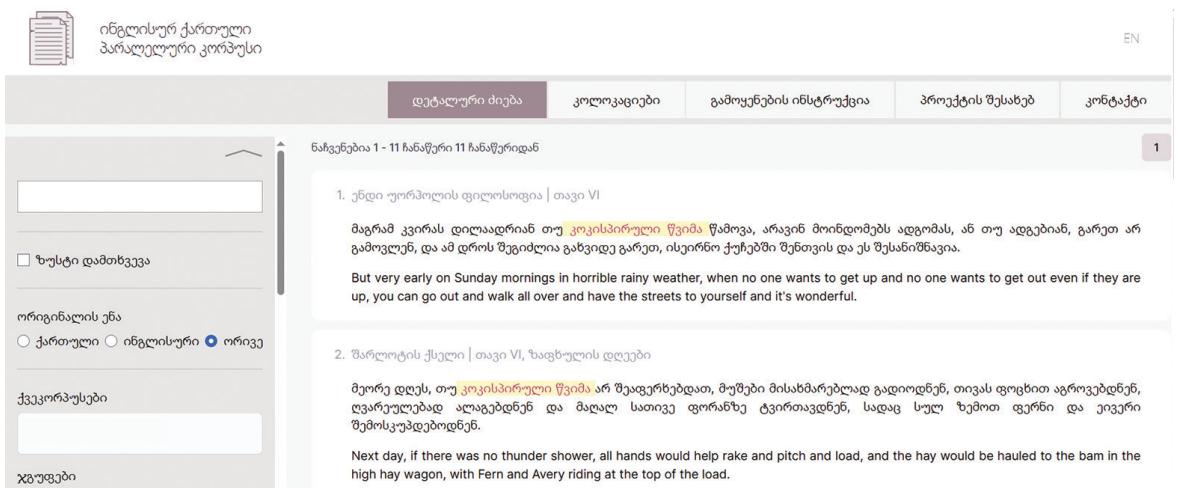


Figure 10: English-Georgian Parallel Corpus - Collocates of the Word *rain* - detailed search function.

The Corpus of Georgian Neologisms[2] (Laluashvili & Margalitadze 2025) - a resource developed in 2024 at Ilia State University. It contains textual material from Georgian online newspapers and magazines, television and other media websites, as well as the websites of governmental agencies and non-governmental organizations. Like the *English-Georgian Parallel Corpus*, this corpus also includes a collocational search function.

2 https://neologism.iliauni.edu.ge/

The Georgian National Corpus[3] (Gippert & Tandashvili 2015)**.** The National Corpus of the Georgian Language includes several different corpora, among them the corpora of Old and Middle Georgian, the Dialect Corpus of the Georgian Language, the Corpus of Political and legal texts, also the Megrelian and Svan corpora. Within the framework of the present study, the Reference Corpus of the Georgian Language (217,130,870 tokens) was used to search for collocations. The aforementioned corpus and the Georgian Neologism Corpus were used as supplementary sources, while the primary source for identifying collocations was the *English-Georgian Parallel Corpus*.

The research showed that, these sources are sufficient for obtaining the relevant data necessary for the compilation of a *Georgian-English Online Learner's Dictionary of Collocations*.

5 Conclusion

The article examines the microstructure of a *Georgian-English Online Learner's Dictionary of Collocations*, which is based on the compilation principles of the Oxford and Macmillan collocations dictionaries, as well as on Igor Mel'čuk's theory of lexical functions.

A collocation is a fixed or semi-fixed combination of words that occurs naturally and recurrently in a language. The study of collocations is important because they reflect the lexical and semantic characteristics of a language, as well as the natural relationships between words. In modern lexicography, dictionaries of collocations have developed into an independent genre, as they play a significant role in language teaching, translation, and linguistic research. From this perspective, the Georgian language remains insufficiently described, therefore, the systematic study of collocations and the compilation of a corresponding dictionary are of particular importance for revealing the lexical potential and usage patterns of the Georgian language for the study process.

The proposed approach involves grouping collocations within dictionary entries according to the principles used in the Oxford and Macmillan collocations dictionaries and describing them through the application of lexical functions.

The research results demonstrate both the effectiveness of the proposed mixed approach in the context of learner lexicography, and the role and reliability of existing lexicographic resources used in the study, the *English-Georgian Parallel Corpus*, the *Corpus of Georgian Neologisms*, and the *Georgian National Corpus*. Moreover, collocations were retrieved semi-automatically, which indicates the potential for further automation of the process; work in this direction would significantly accelerate the compilation of the dictionary.

The study also addresses an important theoretical issue, namely the extent to which Igor Mel'čuk's theory of lexical functions can be adapted and applied in the context of the Georgian language. The results confirm that this theoretical framework can be successfully integrated into Georgian learner lexicography.

3 http://gnc.gov.ge/gnc/page

References

Benson, M., E. Benson, & R. Ilson. (1997). *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. John Benjamins Publishing Company. https://doi.org/10.1075/z.bbi [7/01/2025].

Corpus of Neologisms. Accessed at: https://neologism.iliauni.edu.ge/ [30/11/2025].

Cowie, A. P. (1999). *English Dictionaries for foreign learners: A History.* Oxford University Press.

English–Georgian Parallel Corpus (new platform). Accessed at: https://enkacorpus.iliauni.edu.ge/ [30/11/2025].

Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930-55. In *Studies in Linguistic Analysis*, by John R. Firth, pp. 1-31.

Firth, J. R. (1957). *Papers in linguistics, 1934 – 1951.* London: Oxford University Press.

FreeCollocation.com. (n.d.). *Free English collocations dictionary*. Accessed at: https://www.freecollocation.com/ [7/01/2025].

Georgian National Corpus. Accessed at: http://gnc.gov.ge/gnc/page?page-id=gnc-main-page [9/01/2023].

Gippert, J., & M. Tandashvili. 2015. Structuring a diachronic corpus: The Georgian National Corpus project. In *Historical Corpora: Challenges and Perspectives. Corpus Linguistics and Interdisciplinary Perspectives on Language*, Vol. 5, edited by Jost Gippert and Ralf Gehrke, 305–22. Tübingen: Narr.

Kilgarriff, A., & M. Rundell. (2002). Lexical Profiling Software and its Lexicographic Applications: A Case Study. In *Proceedings of the Euralex 2002 Conference, ELX02-090, 13-17 August 2002.* Denmark: Center for Sprogteknologi, pp. 807-818. Available at: https://euralex.org/publications/lexical-profiling-software-and-its-lexicographic-applications-a-case-study/ [7/01/2025].

Kilgarriff, A., P. Rychlý, P. Smrž, & D. Tugwell. (2004). The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress, 6-10 July 2004.* France: Université de Bretagne-Sud, Faculté des lettres et des sciences humaines, pp. 105-115. Available at: https://euralex.org/publications/the-sketch-engine/ [7/01/2025].

Laluashvili, T. & T. Margalitadze. (2025). Semi-Automatic Detection of New Words in Modern Georgian. *Lexikos* 35 (1), pp. 399-413. https://doi.org/10.5788/35-1-2044.

Longman Collocations Dictionary and Thesaurus. (2013). Pearson Education Ltd., ISBN 9781408252253.

Macmillan Collocations Dictionary. (2010). *Macmillan Collocations Dictionary for Learners of English*. Macmillan Education.

Margalitadze, T., G. Meladze, & Z. Purtskhvanidze. (2022). English–Georgian Parallel Corpus and Its Use in Georgian Lexicography. In *Lexikos*, 32(2), pp. 158-176. https://doi.org/10.5788/32-2-1701 [7/01/2025].

Mel'cuk, I. (1996). Lexical functions in Lexicography and Natural Language Processing. In *Lexical Functions: A Tool for the description Lexical Relations in the Lexicon*, pp. 37-102.

Mel'čuk, I. (2012). Phraseology in the language, in the dictionary, and in the computer. Yearbook of Phraseology 3(1): 31–56. https://doi.org/10.1515/phras-2012-0003.

Poulsen, S. (1991). The BBI Combinatory Dictionary of English: A Guide to Word Combinations. Compiled by Morton Benson, Evelyn Benson and Robert Ilson. John Benjamins Publishing Company/Munksgaards Ordbøger, Amsterdam/Philadelphia 1986. In *HERMES - Journal of Language and Communication in Business*, 4(7), pp. 129–137. https://doi.org/10.7146/hjlcb.v4i7.21483 [7/01/2025].

Reum, A. (1961). *Dictionary of English Style*. Philosophical Library.

Rodale, J. I. (1947). *The Word Finder*. Funk & Wagnalls.

Rundel, Michael, & Adam Kilgariff. 2011. Automating the creation of dictionaries: where will it all end? In *A taste for Corpora: In honor of Sylviane Granger*, edited by Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin, and Magali Paquot, pp. 257–82. https://doi.org/10.1075/scl.45.15run [7/01/2025].

Rundell M., & G. Fox. (2002). *Macmillan English Dictionary : For Advanced Learners.* London: Macmillan.

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

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Blackwell Publishers.