Key Challenges in AI-Powered Translation in the Context of the Georgian Language

Maia Nachkepia

PhD, Assistant Professor Department of English Philology, Akaki Tsereteli State University, 599 30 32 14/593 205 433; maia.natchkepia@atsu.edu.ge; maianatchkepia@gmail.com; Kutaisi, Georgia

Abstract

The given article examines key challenges facing AI-powered translation systems when processing Georgian, a language with unique linguistic and cultural characteristics but referred to as a low-resource language in terms of the lack of computerized data. The study identifies two primary categories of translation barriers: cultural and linguistic challenges.

Cultural challenges include mistranslation of culturally specific concepts, inadequate handling of Georgian idioms and proverbs, and failure to capture figurative language. Linguistic challenges stem from Georgian's agglutinative nature, ergative case system, flexible word order, and polypersonal agreement patterns. Georgian verbs encode subject, object, tense, aspect, and directionality in single morphological units, creating difficulties for AI systems trained primarily on structurally different languages.

Current neural machine translation systems, including Google Translate and DeepL, consistently struggle with Georgian due to limited training data and inadequate handling of morphological complexity, resulting in awkward or culturally inappropriate translations. The findings highlight the need for specialized NLP approaches for morphologically complex, low-resource languages and emphasize the importance of incorporating cultural context into AI translation systems.

Keywords: machine translation, low-resource languages, neural machine translation, AI translation challenges, cross-cultural communication

Introduction

It is already no wonder that AI-powered translation is taking reigns, but there are still some challenges that need to be addressed - and probably will remain so as long as human thinking continues to exist and develop. It all started in the 1950s when **machine translation by computers** came into operation. The technology used initially was rule-based, but **statistical machine translation** emerged in the late 1980s, followed by **neural machine translation** in the 2010s (Science Co., Ltd. 2022). However, cultural and linguistic context still remains a hurdle concerning many languages with one of them being Georgian.

History of AI translation

In rule-based machine translation, translation rules are manually created based on dictionaries and grammar. It was necessary to create a vast number of rules, which made development and updates for new terms labor-intensive. The accuracy was not high, and it could mainly translate only formulaic sentences.

In statistical machine translation, computers learn the rules instead of humans. They read a large number of pairs of original texts and their translations (such as one million sentences) and learn to statistically associate words and phrases from the original text and the translation based on that data (called a corpus). However, translation between languages with significantly different grammar, such as English and Georgian is difficult, and the translation accuracy was not yet practical.

In addition, there are technologies such as hybrid machine translation that combines statistical machine translation and rule-based methods, as well as example-based machine translation that extracts similar parts from existing pairs of source texts and translations for use in translation. These machine translation methods have improved translation accuracy compared to rule-based and statistical machine translation.

Neural machine translation, like statistical machine translation, learns by feeding a large number of pairs of original texts and their translations to a computer. However, by using neural networks and deep learning, which are types of machine learning, it extracts and utilizes far more information for translation than statistical machine learning. Compared to traditional machine translation, the accuracy of translations has significantly improved. The translations are characterized by their fluency, producing natural translations that resemble human translations. With the advent of neural machine translation, machine translation has gained attention and is now widely used in everyday life and business.

Representative machine translation services include DeepL, Google, and Microsoft, all of which utilize neural machine translation technology.

The evolution of AI translation technology has accelerated with the advent of large language models (LLMs) such as OpenAI's ChatGPT. LLMs learn from vast amounts of text data, enabling natural language understanding and generation similar to that of humans. This allows for translations that go beyond simple language conversion, understanding context and nuances, and reflecting cultural backgrounds for natural translations. LLMs can flexibly adapt to the translation style and tone requested by users, providing appropriate translations across a wide range of contexts—from casual conversations to formal business documents, and even emotionally expressive content (Science Co., Ltd. 2022).

Research

As mentioned in the beginning of the article, although machine translation has become a prevalent tool for translation, it still struggles with cultural and linguistic contexts regarding such low-resource languages as Georgian. What is a low-resource language in the context of a machine translation? In machine translation (MT) and natural language processing (NLP), a low-resource language refers to a language that lacks sufficient data (text, speech, or parallel corpora) to train high-quality AI models effectively. These languages pose challenges for statistical and neural machine translation because modern AI systems rely heavily on large datasets.

In the presented article we will break down the challenges regarding the Georgian-into - English language (or vice-versa) translations using AI tools into **two main sections: cultural and linguistic categories.**

I. Cultural challenges

Main cultural hurdles that we come across while translating Georgian texts into English are as follows:

1. Untranslated concepts and culturally specific terms

Like in other languages, there are many words and phrases in Georgian that are connected with its culture or religion. For example: Words like *"სუფრა" [supra]* (a Georgian feast with rituals) or *"მამული"* [mamuli] (homeland with emotional weight) lack direct equivalents. By machine translation they are usually translated as "feast" or "table" or even as "table-cloth" with its direct meaning but none of them communicate its cultural reference as a feast loaded with traditional rituals; the same can be said of *"მამული"* [mamuli]– usually translated as "estate domain", "manor" or "patrimony". This word has several meanings in Georgian: 1. Ancestral land or property passed down through generations; 2. The place where your father was born (cf. *"დედული"* [deduleti] - the place where your mother was born) (Digital Laboratory of the Georgian Language, n.d.) 3. Homeland (Ganmarteba.ge, n.d.). The last one being the most common usage, it is usually

translated into English as 'homeland' (or 'fatherland'), but it carries more profound, historical connotation.

Other examples include: *"თამადა"* [tamada] - literal: "toastmaster"; cultural meaning: The leader of a Georgian feast who orchestrates toasts with poetic or philosophical depth.

"ალილო" [alilo] - literal meaning: a Christmas carol; cultural: a ritual when people, usually children, go door to door on Christmas Eve singing Christmas carols and asking for treats (Ganmarteba.ge, n.d.).

2. Idioms, proverbs and figurative speech

Another challenge that AIs 'struggle' with is connected with idioms that are often used in literary language and carry various connotations. Literary translation does not usually convey the meaning it implies in a given context. Let's take an example: a Georgian idiom: *"გულ–ხელის დაკრეფა"* [gul-khelis dakrefa] is literary translated as - (with one's arms) folded across one's chest (idioms.tsu.ge/?p=20934), through different AI tools it is translated differently: via Claude.ai - **"Gathering courage"** or **"Summoning up courage", through Google translate – 'Heart-hand picking', Deepseek - "typing by heart and hand"** - none of them being correct, even more, they sound absurd or awkward: the correct English equivalents would be: 'put one's feet up' or 'to sit idly' – meaning being inactive or passive, especially when action or intervention is needed or expected (idioms.thefreedictionary.com/sit+(idly)+by). Other examples include: ველოსიპედის *გამოგონება"* {velosipedis gamogoneba] – lit: to recreate the wheel. Re-discover America; fig. to peddle old stuff; to say or claim a general truth or statement; to represent as new what is known to everyone (idioms.tsu.ge/?p=20794). *"პირი იქვალა"* [piri itsvala] – Lit.: to turn round. Fig.: to negate something already done or uttered; to deny one's words.

Provers also fall in this category. Although in comparison with idioms they are easier to understand and we can search for their equivalents in the second language, still it creates barriers in translation due to cultural differences. A Georgian proverb *"ვარდი უეკლოდ და კაცი უნაკლოდ არ არსებობსო"* [vardi ueklod da katsi unaklod ar arsebobso] can literary be translated as "*There is no rose without a thorn and no man without flaws"* (Google Translate) but it has some equivalents in both American and British English, such as *"Every light has its shadow", "Every man has its faults"* or *"Every sky has its cloud"* (Am.), *"No gold (silver) without its dross"* (Br.), "*No garden without its weeds"*, and many more (idioms.tsu.ge/?p=19998). *"ბუ ბუსა შობს და ყვავი ყვავსაო"* [bu busa shobs da kvavi kvavsao] – is literally translated as: '*The owl gives birth to owls and the crow to crows'* through online translation tools, but in the English language they also have their equivalents, like *"Every cask smells of the wine it contains",* or '*Like begets like"* (Am., Br.), "*Like breeds like"* (Br.) (idioms.thefreedictionary.com/like+breeds+like).

Georgian language is rich is figurative use of words and phrases, especially those connected with religion, nature or body parts; cases of metaphors and similes are very frequent that often cause misunderstanding in machine translation. For example, the word "თავი" [tavi], meaning the 'head' in English can be used both with its direct and indirect (or figurative) meaning: 'თავი დაიბანა' [tavi daibana] – in this phrase the word 'tavi' (the head) is used with its direct meaning and is translated as 'he/she washed his/her head', but in the phrases: 'თავი ისახელა' [tavi isakhela], 'თავი შეირცხვინა' [tavi isakhela]– the same word is used figuratively. In the first phrase it means: 'He gained a fame or reputation' but translation tools may make a mistake and just translate it as: "He named himself" (Google Translate), since the phrase includes the word 'ისახელა', deriving from the root word 'სახელი' [sakheli] meaning 'name'. 'თავი შეირცხვინა' – in this idiomatic use of the word 'თავი' 'tavi' – 'the head', it implies the whole personality not just a part of the body and means: 'He/she disgraced oneself', 'acted in a way that brings dishonor'.

II. Linguistic barriers

Machine translation often struggles with translating Georgian texts into English or communicating the message of an English text into the Georgian language in the proper way. This is due to the fact that Georgian is considered a ''low-resource language'' in the AI/NLP world, meaning there are relatively few high-quality parallel corpora (e.g., Georgian-English texts) available for training machine learning models. According to the definition of a Georgian scholar, M. Tandaschvili, "While the traditional classification paradigm included genetic, typological, and relational classifications, a new paradigm has emerged. This new paradigm categorizes languages according to different parameters: 1) the legal status of the language, 2) the viability of the language and its areas of use, and 3) the degree of operation of languages in the digital age. According to this new classification, languages are now categorized into high-resource languages (HRL) and low-resource languages (LRL). Consequently, the traditional Language Atlas, which reflects the genealogical classification or distribution area of languages, has been replaced by a new Language Atlas that shows the percentage of digital resources available for natural language processing (NLP) in relation to the number of languages spoken worldwide.

According to theoretical scientific literature, LRLs can be understood as few studied, resource-scarce, less computerized, less privileged, and less commonly taught (Cieri et al., 2016:4543-44; Magueresse et al., 2020:1). Artificial intelligence experts define LRLs as languages with a limited amount of linguistic data and resources for natural language processing (NLP) tasks" (Tandaschvili, 2025).

Another obstacle in machine translation is **the complex grammar and syntax** of the Georgian language. It will take us too far to analyse them in detail here, so we will just single out some of them.

Agglutinative nature

Verb morphology. Georgian adds multiple prefixes and suffixes to root words, creating long, complex words that can be difficult for AI to segment correctly. B. G. Hewitt explores Georgian grammar in his book Georgian: A Structural Reference Grammar (1995) and provides a detailed analysis of Georgian agglutinative morphology and its linguistic complexities, which directly contribute to challenges in NLP and machine translation.

For example, a Georgian verb

ა-წერ-ინ-ებ-და [a-tser-in-eb-da] combines several elements:

- ○ ა-[a] (preverb: indicates perfective aspect or directional meaning)
- ○ - წერ- [tser] (root, meaning "write")
- ○ - ინ-[in] (causative suffix (makes someone else do the action))
- ○ - ებ- [eb] (present/imperfective stem formant)
- ○ - და [da] (imperfect tense marker/3 rd person)

So, the structure in this single verb looks quite complex: Preverb-Root-Causative-Stem-Formant-Past/Imperfect that could be translated as: "He/she/it was making him/her/them write" or "He/she/it had (someone) write".

This is an example of a rich verbal morphology of the Georgian language, where multiple grammatical meanings are expressed through affixes attached to the verb root. It is no doubt that when lacking digital data, machines would find it very difficult to translate the verb that encodes subject, object, tense/aspect/mood, directionality, causation, polite forms in just one form.

Georgian syntax creates barriers that AI systems hurdle with due to its flexible **word order** where grammatical roles are marked morphologically (via case endings or verb agreement) rather than by their position in a sentence in contrast to rigid word-order languages like English.

For example,

1. ბავშვი წიგნს კითხულობს [bavshvi tsigns kitkhulobs] (SOV)

    "The child is reading the book."

2. წიგნს ბავშვი კითხულობს [tsigns bavshvi kitkhulobs] (OSV)
   [**The book the child is reading**] (Same meaning, emphasis changes)

    "The child is reading the book."

3. კითხულობს ბავშვი წიგნს [kitkhulobs bavshvi tsigns](VSO)
   [**is reading the child the book**]
   "The child is reading the book."

Most MT systems (e.g., Google Translate) assume fixed word order (English: SVO). When Georgian uses OSV or VSO, AI often misassigns subject/object roles.

Ergative case system. Unlike most Indo-European languages, Georgian uses an ''ergative-absolutive'' case system in certain tenses, which can confuse AI models trained on nominative-accusative languages (like English). Georgian is a rare case of split ergativity that means that it exhibits both ergative-absolutive and nominative-accusative alignment patterns within its grammar. For example, in Georgian transitive verbs take the subject with the –მ/–მა (-m/-ma) ending (Ergative) in the aorist and perfective series, while intransitive verb subjects are unmarked (Nominative).

- Transitive (Aorist):
  გიორგიმ (Giorgi-**m**, ERG) წერილი (ts'erili, NOM) დაწერა (dats'era, wrote).
  →*"Giorgi wrote the letter."*
  (Lit. "By Giorgi the letter was written.")

- Intransitive (Aorist):
  ○ გიორგი (Giorgi, NOM) წავიდა (ts'avida, left).
  →*"Giorgi left."*

- Present (Nominative-Accusative):
  ○ გიორგი (Giorgi, NOM) წერს (ts'ers, writes) წერილს (ts'eril-**s**, DAT).
  →*"Giorgi writes the letter.*

In English all subjects are unmarked, as in the translation of this Georgian example - "Giorgi" (the subject) but Georgian marks the agent in transitive with -მ (-m) prefix. It switches from NOM (present) to ERG (aorist), while the object switches from DAT (present) to NOM (aorist). English does not recognize this rule.

Regarding translation issues, AI models often rely on word order and prepositions to identify subjects/object in nominative-accusative languages (like English). However, Georgian's ergative-nominative alignment breaks theses expectations. Besides, AI lacks deep tense/aspect awareness in Georgian and defaults to literal case mappings. Polypersonal agreement in Georgian, may further confuse AI models – Georgian verbs encode subject, object and indirect object in a single word, while English requires separate pronouns, so AI often drops or misplaces arguments (Iliauni eprints, n.d.).

For example, the Georgian verb - *გადმოგცემთ [gadmogtsemt]* is translated into English as - *I (or we) will give it (or them) to you* encoding the doer, the object and the action all together in one verb:

Georgian	Morpheme	Function	Meaning/Role
გად-	*gad-*	Preverb	Implies **transfer across a boundary** ("over," "across")
-მო-	-mo-	Thematic suffix	Often marks **directionality** (no direct English equivalent)
-გ-	-g-	Indirect object marker	"**to you**" (2nd person singular)
-ცემ-	*-tsem-*	Root	"**give**"
-თ	-t	Subject marker (or Indirect object marker in the polite form)	"**we**" ["I"} (1-st person plural or singular) (May refer to indirect object as well if used in a polite context)

In translating this verb and many others like this, standard translating machines like Google Translate usually gives translation based on its encoded system, or on what patterns it is trained on, but it lacks the awareness of the context which often plays a vital role in translation.

Conclusion

The analysis of AI-powered translation challenges in the context of the Georgian language reveals a complex landscape of both cultural and linguistic barriers that significantly impact translation quality and accuracy. Despite remarkable advances in neural machine translation and the emergence of large language models, Georgian remains a low-resource language in terms of computerized language bases that poses unique challenges for automated translation systems.

The cultural challenges identified in this study demonstrate that AI translation tools consistently fail to capture the deep cultural significance embedded in Georgian concepts, idioms, and proverbs. Terms like "სუფრა" (supra), "მამული" (mamuli), and "თამადა" (tamada) lose their rich cultural connotations when reduced to literal translations such as

"feast," "estate," and "toastmaster." Similarly, Georgian idioms and figurative expressions often result in awkward or completely incorrect translations that fail to convey their intended meanings, highlighting the limitations of current AI systems in understanding contextual and metaphorical language use.

The linguistic barriers present even greater challenges due to Georgian's unique grammatical structure. As an agglutinative language with complex verb morphology, Georgian encodes multiple grammatical functions within single words, creating difficulties for AI systems trained primarily on languages with different structural patterns. The ergative case system, flexible word order, and polypersonal agreement further complicate translation processes, as demonstrated by the analysis of verbs like "გადმოგცემთ" (gadmogtsemt), which encodes subject, object, and action in a single morphological unit.

These findings underscore the critical need for developing specialized approaches to machine translation for morphologically complex and culturally rich languages like Georgian. Future improvements will require not only expanded parallel corpora and better training data but also fundamental advances in AI systems' ability to understand cultural context, handle complex morphological structures, and recognize the nuanced relationships between language and culture.

 As AI translation technology continues to evolve, addressing these linguistic and cultural barriers will be essential for creating truly effective cross-cultural communication tools that preserve the richness and nuance of human language diversity.

References

Cieri, C., Maxwell, M., Strassel, S., & Tracey, J. (2016). *Challenges and opportunities in language resource development for low-resource languages*. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 4543-4550. https://aclanthology.org/L16-1720/

Digital Laboratory of the Georgian Language. (n.d.). University of Georgia. Retrieved August 3, 2025 from https://dlab.ug.edu.ge

Ganmarteba.ge. (n.d.). *Georgian Dictionary*. Retrieved July 20, 2025, from https://www.ganmarteba.ge

Hewitt, B. G. (1995). *Georgian: A structural reference grammar*. John Benjamins Publishing. https://doi.org/10.1075/loall.2

Makharoblidze T. (n.d.). *Polypersonal verbal categories in Georgian sign language*. Ilia University Repository. Retrieved July 30, 2025 from https://eprints.iliauni.edu.ge/3051/1/polypersonalism%205.pdf

Magueresse, A., Carles, V., & Heinecke, J. (2020). *Low-resource languages: A review of past work and future challenges.* arXiv preprint arXiv:2006.07264. https://arxiv.org/abs/2006.07264

Thuleen N. (n.d.). *The Verbal Syntax and Ergativity of Georgian.* Retrieved from https://www.nthuleen.com/papers/L12paper.html

Science Co., Ltd. (2022). *The evolution of machine translation: From rule-based to neural machine translation.* Science.co.jp. Retrieved July 28, 2025 from https://www.science.co.jp/en/nmt/blog/32553/

Shota Rustaveli National Science Foundation. (n.d.). Fund of Georgian idioms and proverbs. Tbilisi State University. Retrieved July 26, 2025 from https://idioms.tsu.ge/

Tandaschvili, M. (2025). *Low-resource languages in the digital age: A new paradigm for linguistic classification* [Unpublished manuscript]. Ilia State University. https://eprints.iliauni.edu.ge/

The Free Dictionary. (n.d.). Sit (idly) by. In Idioms. Retrieved July 25, 2025, from https://idioms.thefreedictionary.com

ხელოვნური ინტელექტით შესრულებულ თარგმანთან დაკავშირებული გამოწვევები ქართული ენის კონტექსტში

მაია ნაჭყეპია

აკაკი წერეთლის სახელმწიფო უნივერსიტეტის ინგლისური ფილოლოგიის დეპარტამენტის ასისტენტ–პროფესორი; 599 30 32 14/593 205 433; maia.natchkepia@atsu.edu.ge; maianatchkepia@gmail.com ქუთაისი, საქართველო

აბსტრაქტი

წარმოდგენილი სტატია განიხილავს იმ ძირითად გამოწვევებს, რომლებსაც ხელოვნური ინტელექტით თარგმნის სისტემები აწყდება ქართული ენის დამუშავებისას - ენისა, რომელსაც გააჩნია უნიკალური ლინგვისტური და კულტურული მახასიათებლები, თუმცა მას დაბალრესურსიან ენად მოიხსენიებენ კომპიუტერიზებული მონაცემების ნაკლებობის თვალსაზრისით. კვლევა განსაზღვრავს თარგმნის ბარიერების ორ ძირითად კატეგორიას: კულტურულ და ლინგვისტურ გამოწვევებს.

კულტურული გამოწვევები მოიცავს კულტურულად სპეციფიკური ცნებების არასწორ თარგმნას, ქართული იდიომებისა და ანდაზების არაადეკვატურ დამუშავებას და გადატანითი ენის ვერგადმოცემას. ლინგვისტური ხასიათის სირთულეებს განაპირობებს ქართული ენის აგლუტინაციური ბუნება, ერგატიული ბრუნების სისტემა, სიტყვათა მოქნილი რიგი და პოლიპერსონალური შეთანხმების მოდელები. ქართული ზმნის ერთ მორფოლოგიურ ერთეულში კოდირებულია სუბიექტი, ობიექტი, დრო, ასპექტი და მიმართულება, რაც სირთულეებს ქმნის ხელოვნური ინტელექტის ისეთი სისტემებისთვის, რომლებიც ძირითადად სტრუქტურულად განსხვავებულ ენებზეა აგებული.

ნეირონული მანქანური თარგმნის არსებული სისტემები, მათ შორის Google Translate და DeepL, მუდმივად სირთულეებს აწყდება ქართულთან მიმართებაში, რაც გამოწვეულია მონაცემთა შეზღუდული ბაზებითა და ქართული ენის მორფოლოგიური სირთულის არაადეკვატურად დამუშავებით, რის შედეგადაც მიიღება უხერხული ან კულტურულად შეუფერებელი თარგმანები. კვლევის შედეგები ხაზს უსვამს ბუნებრივი ენის დამუშავების (NLP) სპეციალიზებული მიდგომების აუცილებლობას ისეთი მორფოლოგიურად რთული, დაბალრესურსიანი (ციფრული მონაცემების თვალსაზრისით) ენებისთვის, როგორიცაა ქართული და ხაზგასმით აღნიშნავს ხელოვნური ინტელექტით თარგმნის სისტემებში კულტურული კონტექსტის ინტეგრირების მნიშვნელობას.

საკვანძო სიტყვები: მანქანური თარგმანი, დაბალრესურსიანი ენები, ნეირონული მანქანური თარგმანი, ხელოვნური ინტელექტით თარგმნის გამოწვევები, კულტურათაშორისი კომუნიკაცია