Enhancement Possibilities for the Georgian National Corpus

Kamarauli Mariam, PostDoc
University of Hamburg
Hamburg, Germany

ORCID: 0009-0006-0404-4424

DOI: https://doi.org/10.62343/cjss.2024.245

ABSTRACT

The aim of this paper is to make suggestions for improving the Georgian National Corpus based on selected linguistic processes. The Georgian National Corpus is currently the most developed and detailed corpus of the Georgian language. One of the reasons for this is the included annotation of the texts, the variety of text genres, and the size of the corpus. While the morphosyntactic analysis of the texts is great, there is room for improvement in the semantic-pragmatic analysis, especially as far as the semantic-pragmatic analysis of functional elements is concerned. Many factors make this issue very interesting, such as grammaticalisation processes or the fundamental development of language. Implementing this type of analysis is essential, especially when it comes to adequate translations by machine translations. The paper contains an approach for analysing functional elements using the example of the particle xom.

Keywords: *Corpus linguistics, Annotation, Modern Georgian, GNC*

INTRODUCTION

The 21st century, along with the rapid development of information technologies, brought significant changes to any scientific field and, of course, also to linguistics. The classical grouping of languages established in linguistics has been replaced by a new paradigm of classification. If the traditional classification paradigm included genetic (classification of languages into families according to their genetic relationship), typological (classification of languages according to their morphological structure) and relational classification (classification of languages according to their relational type into, e.g. nominative-accusative, ergative-absolutive and active-stative alignment), today the paradigm of language classification has changed and the focus of language classification added to the quality of the languages' digital representation. What is meant here is the existence of big data both from a quantitative point of view (textbases and speech data of hundreds of millions of tokens) and from a qualitative point of view (high level of annotation quality, electronic dictionaries, grammar resources such as bases of grammatical morphemes and rules, sentiment analysis, treebank, etc.). Thus, according to the approach of language classification, languages are grouped into High Resource Languages (HRL) and Low Resource Languages (LRL). Of the alleged 7,000 languages in the world, only 20 languages have sufficient resources to perform the tasks of Natural Language Processing (NLP). Despite the fact that a large number of monolingual and bilingual digital resources have been created for the Georgian language (GNC, 2024; Georgian Dialect Corpus, 2024; Rustaveli Goes Digital - Parallelkorpus, 2024), it is still classified as a low-resource language (see RichardLitt, 2024). To change this status of the Georgian language, a number of tasks need to be solved, such as the enhancement and further development of the Georgian National Corpus (GNC) – some of the proposals will be presented below.

In general, during the construction of a corpus, the general principles of corpus construction (corpus structure) should be considered, on the one hand, and on the other hand, the structural and grammatical features of the language of the resource embedded in the corpus, which will be considered when creating the corpus search system - the corpus manager. For the efficient use of the corpus, the methodological aspect is also important, in particular, the relationship between data and theory (theoretical qualification of data), the so-called 3A perspective (Wallis & Nelson, 2001: 311ff), namely annotation, abstraction and analysis:

- "Annotation consists of the application of a scheme to texts. Annotations may include structural markup, part-of-speech tagging, parsing, and nu-

merous other representations.

- Abstraction consists of the translation (mapping) of terms in the scheme in a theoretically motivated model or dataset. Abstraction typically includes linguist-directed search but may include rule-learning for parsers, for example.
- Analysis consists of statistically probing, manipulating and generalising from the dataset. Analysis might include statistical evaluations, optimisation of rule-bases or knowledge discovery methods" (Agapova, 2014, p. 282).[1]

The advantage of an annotated corpus is that users can use it for a wider range of research issues and conduct experiments using the corpus manager.

The higher the degree of annotation in the corpus, that is, the more annotation levels are provided in the corpus, the more useful the given corpus is for interdisciplinary research, on the one hand. On the other hand, annotated corpora are needed to implement natural language processing (NLP) and to train artificial intelligence (AI) for a given language.

METHOD

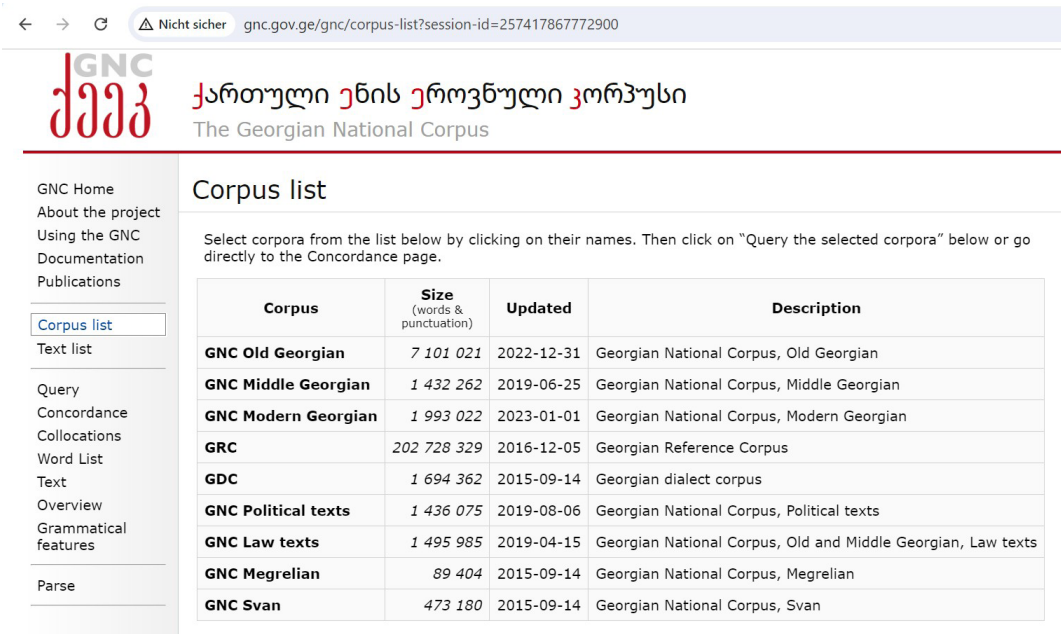
Two extensive databases have to be mentioned when discussing the Georgian language, namely Thesaurus Indogermanischer Text- und Sprachmaterialien (TITUS) (University of Frankfurt, n.d.) and Georgian National Corpus (GNC) (Georgian National Communications Commission, n.d.). The former comprises corpora of ancient Indo-European languages (such as Avestan, Vedic Sanskrit, Phrygian, or Umbrian) and also materials in more recent Indo-European as well as neighbouring languages, among them the South Caucasian languages (such as Georgian, Megrelian, Svan and Laz) but TITUS does not contain as many textual resources for Modern Georgian as GNC. The National Corpus of the Georgian Language (GNC) is the largest corpus created for the Georgian language (more than 202 million tokens),

1 Wallis, S. (n.d.). Annotation takes a set of texts and adds linguistic information to it, enriching it and identifying instances

of linguistically meaningful entities and relations. At this point, the resulting enriched dataset ('corpus') is usually distributed to the research community. Abstraction is the researcher's exploratory process of establishing a mapping between concepts they wish to research, and representations found in the corpus (text + annotation). It also maps the structured corpus to a regular dataset that can be analysed by conventional statistical methods. The key linking element in abstraction is a corpus query. Analysis is the process of applying statistical and other methods to data that has been abstracted in this way. Retrieved from https://www.ucl.ac.uk/english-usage/staff/sean/

which is the reason. GNC belongs to the type of diachronic corpora, which combines Old, Middle, and Modern Georgian language resources. The corpus includes both resources of the written Georgian language from ancient monuments (inscriptions, handwritten sources) to the present day, and samples of oral speech - the Georgian dialect corpus is integrated into the corpus. When it comes to text genres, GNC is a balanced corpus containing religious, historical, juridical and political texts. The latter two genres are also represented as separate sub-corpora. Nevertheless, the corpus requires further development both in terms of genre and quantity.

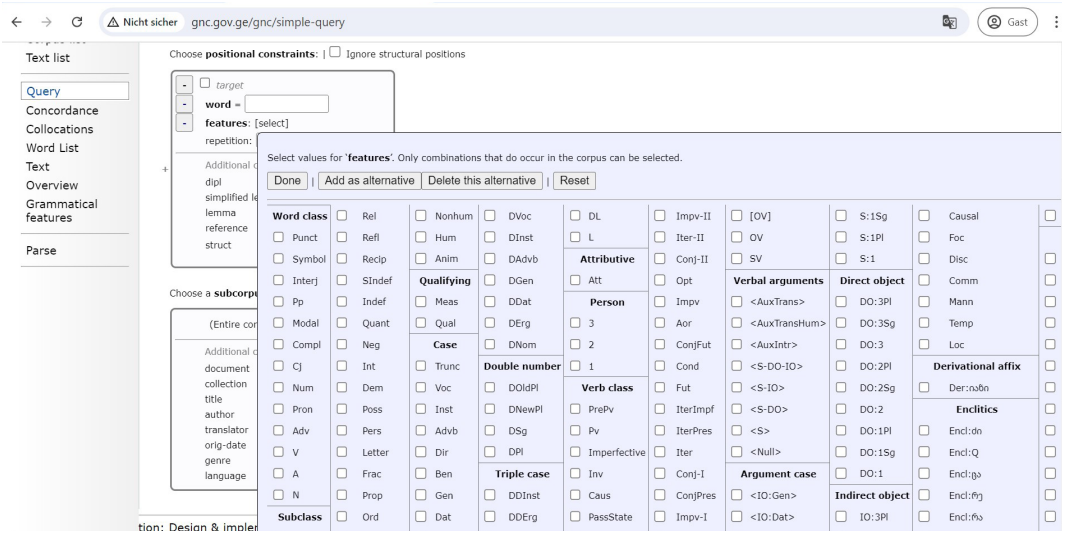
Figure 1: The sub-corpora of the GNC



In addition to the Georgian language, the GNC includes resources for other South-Caucasian languages - Megrelian and Svan. Both the textual material published in these languages and the modern oral resources (which only represent a fraction of what TITUS has to offer) were obtained and processed within the framework of the international scientific projects implemented at the University of Frankfurt (TITUS, ECLinG, SSGG), are presented here. A large Georgian reference corpus (GRC) is included, which contains less thoroughly processed texts from various fictional and non-fictional domains.

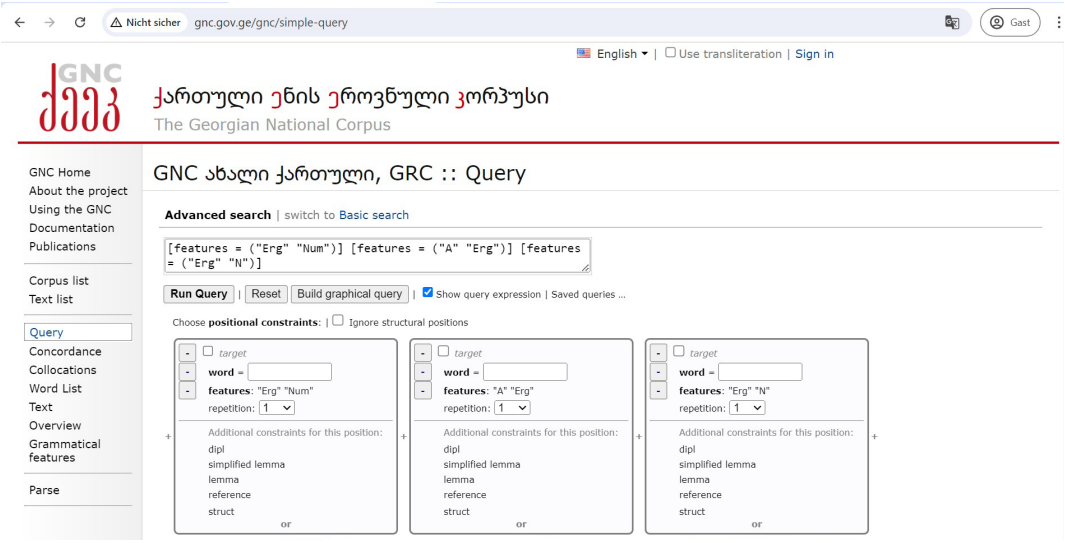
GNC is an annotated corpus - the corpus manager allows for both simple and complex searches in the corpus. In the case of a complex search, it is possible to find a word form according to one or several grammatical features combined.

Figure 2: Example of a complex search in the GNC



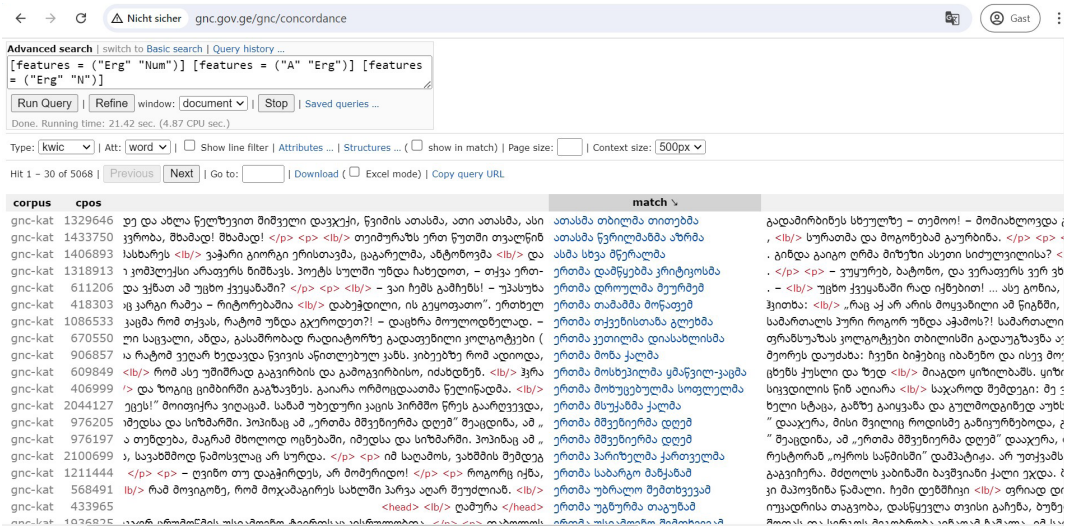
The corpus search engine also allows you to search the corpus for phrasal constructions:

Figure 3: Searching interface for phrasal constructions in the GNC (a phrase containing a numeral, an adjective and a noun in the ergative)



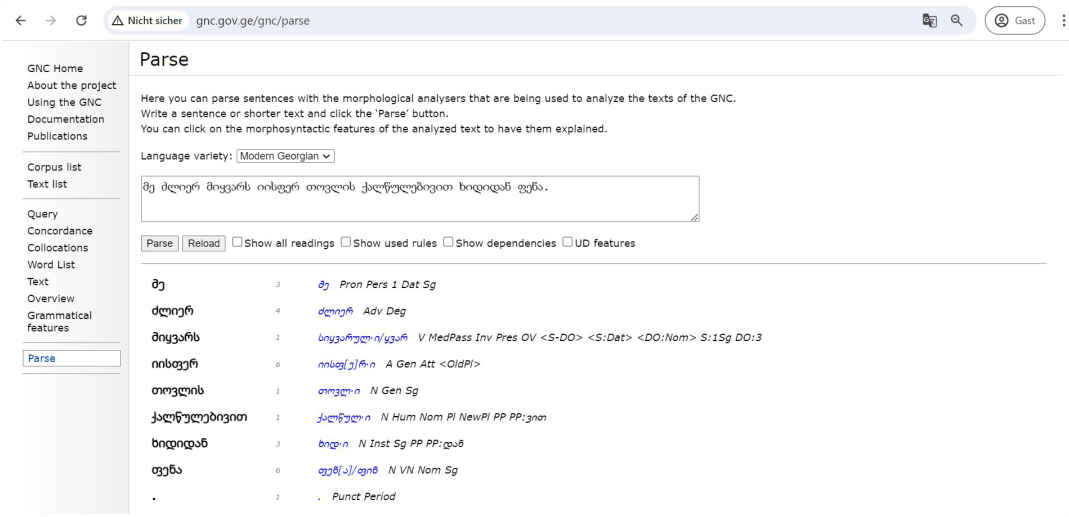
The results of the search are then displayed in the corresponding concordance:

Figure 4: Results of the search



The high degree of annotation in the corpus allows for morphosyntactic and syntactic analysis:

Figure 5: Parsing of a sentence

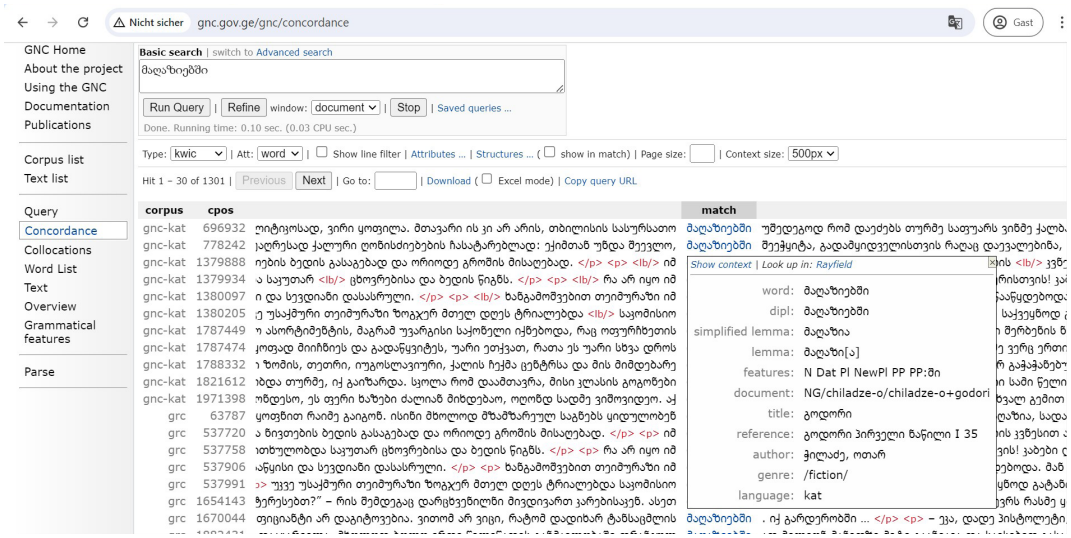


GNC was created within the framework of international scientific cooperation in the years 2012-2019. Both European (Frankfurt University, University of Bergen) and Georgian scientific and educational institutions (Georgian National Communications Commission, n.d.) participated in its creation.

The quality of big data annotation is crucial for AI tasks. The quality of data annotation refers to the accuracy and consistency of data labelling for machine learning models. It is crucial to ensure that the algorithms learn effectively from the annotated data provided. High-quality data annotation leads to more accurate predictions and better model performance. It also implies a multi-level system of analysis, which includes morphological, morphosyntactic, syntactic, pragmatic, and semantic levels. In the case of speech data, in addition to text, audio and video resources, suprasegmental analysis is also provided. Suprasegmental features help to convey meaning, structure and emotional undertones in oral communication. They affect the way syllables, words and sentences are pronounced and influence the meaning and perception of spoken language at a higher level.

The GNC is characterised by a relatively high level of token annotation, which includes both the lemma and grammatical features of the token, as well as other relevant information (source, author, title, date of the text, suprasegmental annotations, etc.). Below, an example from the nominal morphology is provided:

Figure 6: Search result of the noun მაღაზიებში "in the stores"



The same applies to search results from the verbal morphology and uninflectable words:

Figure 7: Search result of the verb ტრიალებდა *"[(s)he] was spinning"*

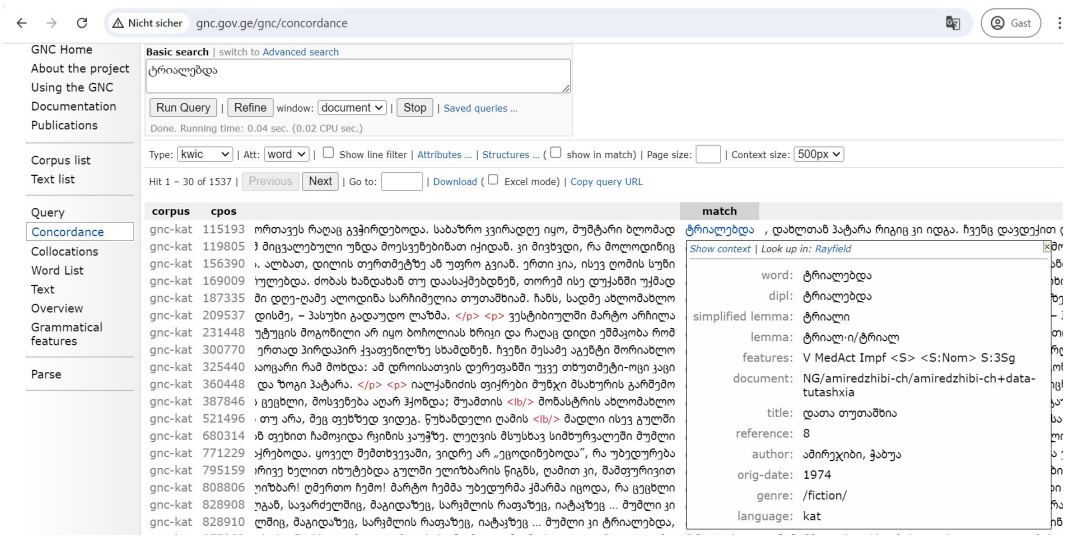
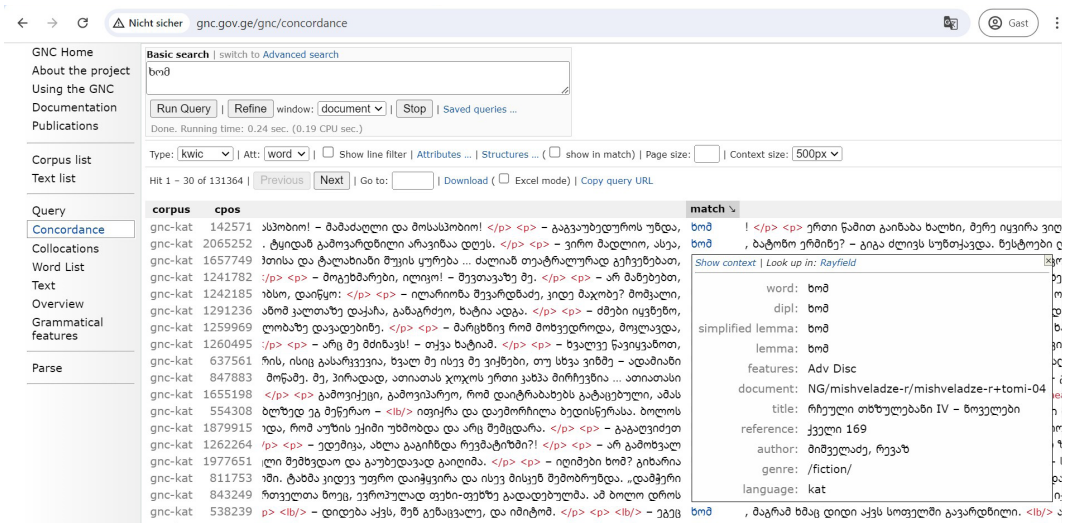


Figure 8: Search result of the affirmative particle xom *(ხომ)*



In the case of uninflectable words, as shown in Fig. 8, the syntactic-pragmatic function is indicated: xom - Adv Disc (discourse adverb). However, a certain part of the tokens in GNC is not annotated, which is due to the fact that the issues of the functional grammar of the Georgian language are still theoretically unresearched and have only been studied in fragments. Accordingly, the grammatical characterisation of some words in the corpus is either inaccurate or the grammatical features are not defined at all - in such a case, only "unknown" is indicated. Below, we

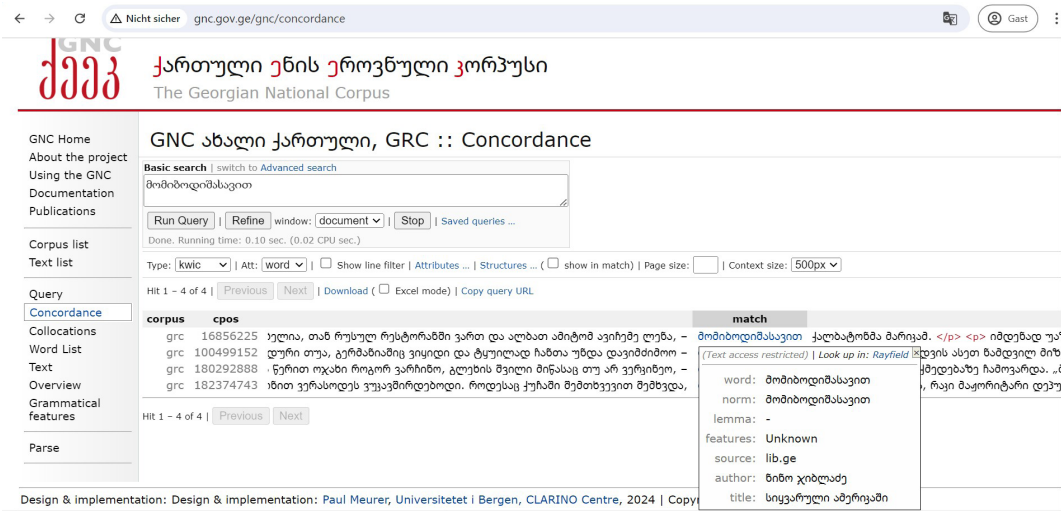
discuss several works related to GNC annotation system improvement and corpus development proposals and present my proposal regarding the annotation of invariant words in the corpus.

DISCUSSION

In order to achieve a high-quality annotation, specific phenomena of any given language must be considered - structural features, grammatical processes in the language, functional-semantic and pragmatic meaning of linguistic elements, and other specific features. This applies not only to simple elements of the corpus, such as word forms, but also to complex structural units, such as phrasal structures. In my opinion, further development of GNC requires the refinement of the specific phenomena of the Georgian language. Below, I will present suggestions for improving the annotation of the Georgian National Corpus, using examples of simple elements and complex constructions.

One of the linguistic phenomena of the Georgian language is approximative verbs; these elements represent a symbiosis of the nominal and verbal domain, as the marker used for approximativeness originates from the nominal domain and is suffixed to a fully inflected verb. The suffix *-vit* ('like, as'), which is typically suffixed to a noun in the nominative or the dative case (the former applies to nouns with consonantal stems, the latter to nouns with vocalic stems), can also be found suffixed to nouns in the genitive case, which is the rarest among the cases in combination with the suffix (Kamarauli, 2023, p.52).

Figure 9: Example of an approximative verb, classified as "unknown"



The GNC has not yet provided a classification for such constructions, so these are labelled as "unknown". What I propose is the following: when verbs are analysed as usual according to the grammatical markers such as person, number, tense, etc., another feature must be added, namely verbal approximativeness (AppV). The morpheme expressing approximativeness (APP: სავით) should be added at the end of the grammatical features:

Cf.:

EXAMPLE	GRAMMATICAL FEATURES
მეცადინეობდა	V MedAct Impf <S> <S:Nom> S:3Sg
vs.	
მეცადინეობდასავით	AppV MedAct Impf <S> <S:Nom> S:3Sg APP: სავით

One of the important challenges in the analysis of the Georgian language is the issue of annotation of uninflectable elements - particles, conjunctions, adverbs, conjunctions. The correct annotation of functional elements is indispensable for solving both semantic analysis and treebank tasks.

Below, I present my annotation approach of functional elements on the example of the functional-semantic analysis of the particle *xom*.

The particle *xom* is analysed as an interrogative particle in scientific literature, in particular as:

- An interrogative particle, which 1. is used in interrogative clauses and denotes confirmation, and 2. is used together with a negative word (არ, ვერ, არავინ...) and indicates doubt (Explanatory Dictionary n.d);
- An interrogative particle-morphemoid, which a) expresses confirmation in interrogative clauses, b) expresses doubt with negative morphemoids (no, can, nobody), c) is used in negative constructions to express the function of a request (Jorbenadze, K'obakhidze & Beridze, 1988: 474-475);
- It is used when asking a question and wanting to have the answer confirmed (Georgian Dictionary n.d);
- It is annotated as a discourse adverb in the National Corpus of the Georgian language (Georgian National Corpus n.d.).

In the reference sub-corpus of GRC, *xom* is statistically one of the most frequently used particles.

Cf.:

Table 1: Frequency of particles in the GNC

PARTICLE	FUNCTION	HITS
ar	negation (neutral)	1821586
ki	affirmation	784365
tu	condition	738435
ver	negation (potential)	350503
xom	affirmation	128056
nu	negation (prohibitive)	37945
gana	elicitation	14520
ho	affirmation	10984
nutu	elicitation	8778
aki	evidentiality	4025

The functional-semantic analysis of the particle *xom*, which is presented below, relies on the resources provided by the GNC. Both classic research methods and corpus linguistic research methods are used to analyse the examples. Additionally, substitution, elimination, permutation and paraphrasing tests were also used in the research. The corpus linguistic analysis showed that the particle can convey more functional semantics than in the definitions presented above. In addition, the conducted analysis showed that the following parameters are crucial for determining the functional semantics of the particle *xom*, which will be introduced below:

- Clause type (declarative, interrogative, imperative, etc.),
- Its position in the sentence (initial, midfield, final position),
- Ability to transpose and the resulting scope effects,
- Ability to combine with other uninflectable words in a sentence.

The particle *xom* usually appears in interrogative clauses and is used with an interrogative-affirmative function. It can be placed as sentence-initial, mid-sentence, or sentence-final. Below, every mentioned instance is shown.

- Initial position:

(1a)	***xom***	*ḳarg-i*	*azr-i-a?*
	AFF	good-NOM.SG	idea-NOM.SG-COP

‘It is a good idea, right?’

(1b)	*ḳarg-i*	*azr-i-a*	***xom****?*
	good-NOM.SG	idea-NOM.SG-COP	AFF

‘It is a good idea, right?’

(1c)	*ḳarg-i*	*azr-i-a?*
	good-NOM.SG	idea-NOM.SG-COP

‘Is it a good idea?’

As the examples above show, it is possible to transpose the particle *xom* in (1a-b) and even omit (1c) from the sentence. In the case of transposition, the sentence maintains the semantics of confirmation (affirmativeness). Therefore, the probable answer is ‘yes’. In the case of omission, affirmativeness is lost, and the sentence becomes a ‘yes/no’ question - the answer can be either positive or negative.

Both sentences (1a) and (1b) require a positive answer. The difference between them is the speaker’s attitude: in (1a), the speaker offers his opinion to the listener, which is affirmative and conveys the speaker’s position; as a result of the transposition of the particle in (1b), the speaker expects the listener to confirm the opinion expressed by him.

The following example confirms that the particle *xom* placed in the final position expresses the expectation of confirmation from the listener:

(2a)	*ramden-ze*	*gagvarige*	*me*	*da*	*besarion-i?*
	how much.DAT.SG-on	settle.S2SG.O1PL. AOR	I.NOM.SG	and	Besarion-NOM.SG

	otxas-i	*manet-i*	*unda*	*moeca*	***xom****?*
	fourhundred-NOM.SG	Manet-i-NOM.SG	MPTCL	give.S3SG.PLU-PERF	AFF

‘How much money did me and Besarion agree on thanks your help? He should have given me 400 Manetis, right?’ (Davit kldiašvili, *Soloman Morbelaʒe*)

When the particle *xom* is placed in the initial position, the speaker expects the listener to confirm the amount of money:

(2b)	*ramden-ze*	*gagvarige*	*me*	*da*	*besarion-i?*
	how much. DAT.SG-on	settle.S2SG.O1PL.AOR	I.NOM.SG	and	Besarion-NOM. SG

xom	*otxas-i*	*manet-i*	*unda*	*moeca?*
AFF	fourhundred-NOM. SG	Maneti-NOM. SG	MPTCL	give.S3SG.PLUPERF

'How much money did me and Besarion agree on thanks your help? He should have given me 400 Manetis, right? '

Example (2a) is an interrogative clause, and the answer requires specifying the amount. In the following example, (2b), the speaker states the amount himself and waits for the addressee to confirm it. Both sentences are affirmative sentences, but in the second example, the affirmation is given from the perspective of the speaker, and in the first case, the affirmation requires confirmation from the perspective of the listener.

- Mid-sentence position:

A similar functional semantics can be observed when the particle is in the second position:

(3a)	*čven*	***xom***	*adre-c*	*ševxvedrivart*	*ertmanet-s?*
	we.NOM.SG	AFF	early-FOC	meet.S1PL.PERF	each other-DAT.SG

'We have met each other before, haven't we?'

(3b)	*čven*	*adre-c*	*ševxvedrivart*	*ertmanet-s*	***xom****?*
	we.NOM.SG	early-FOC	meet.S1PL.PERF	each other-DAT. SG	AFF

'We have met each other before, haven't we?' (confirmation from the listener's perspective)

(3b)	*čven*	*adre-c*	*ševxvedrivart*	*ertmanet-s?*
	we.NOM.SG	early-FOC	meet.S1PL.PERF	each other-DAT. SG

'Have we met each other before?' (neutral semantics - 'yes/no' question)

The particle *xom* can also be used as a discourse element; A relatively extensive context is provided below, where the particle conveys a presupposition:

Table 2: Excerpt from the novella 'The Little Prince', chapter 15

Oḳeaneebi tu aris tkvens ṗlaneṭaze?	"Has your planet any oceans?"
Ver geṭq̇vi, - tkva geograpma.	"I couldn't tell you," said the geographer.
A! - ṗaṭara uplisc̣uli ar moeloda aset ṗasuxs.	"Ah!" The little prince didn't expect such an answer.
Arc mtebi?	"Not even mountains?"
Verc magaze giṗasuxeb.	"I couldn't answer that either."
Kalakebi, mdinareebi an udabnoebi?	"Towns, rivers or deserts?"
Verc magaze geṭq̇vi rames. Rac ar vici, ar vici, - miugo geograpma.	"I couldn't tell you that either. What I don't know, I just don't know" – answered the geographer
*Magram tkven **xom** geograpi xart?*	"But you are a geographer, **right**?"

In this context, the particle *xom* is a pragmatic element, namely a presupposition marker. If we omit the adversative conjunction *magram* 'but' in the last sentence, we get the following expression: *tkven **xom** geograpi xart?* 'You are a geographer, **right**?'. Here, the presupposition is clearly readable, and it is marked in the sentence with the particle *xom*. By eliminating it, the presupposition in the sentence is lost - the sentence turns into a simple 'yes/no' question: *tkven geograpi xart?* 'Are you a geographer?'. The adversative conjunction *magram* 'but' makes the speaker's position even stronger: the geographer's answers in the discourse (lack of geographical knowledge) surprise the speaker since he expects the geographer to have this knowledge. The opinion of the speaker in the last sentence is critical, which is marked by the adversative conjunction *magram* in the initial position, and to convey his position, the speaker uses an affirmative sentence with the particle *xom*.

- Final position and scope effects

The possibility to transpose elements also brings some changes in scope and, therefore, semantics. The following examples have been constructed to demonstrate the functionality and the resulting scope effects of the particle *xom* when transposed:

(4a)	***xom***	*luḳa-m*	*dalia*	*sam-i*	*lud-i?*
	AFF	Luka-ERG.SG	drink.S3SG.AOR	three-NOM.SG	beer-NOM.SG

'Luka drank three beers, right?'

(4b)	*luḳa-m*	***xom***	*dalia*	*sam-i*	*lud-i?*
	Luka-ERG.SG	AFF	drink.S3SG.AOR	three-NOM.SG	beer-NOM.SG

'Luka drank three beers, right?'

(4c)	*luḳa-m*	*dalia*	***xom***	*sam-i*	*lud-i?*
	Luka-ERG.SG	drink.S3SG.AOR	AFF	three-NOM.SG	beer-NOM.SG

'Luka drank three beers, right?'

*(4d)	*luḳa-m*	*dalia*	*sam-i*	***xom***	*lud-i?*
	Luka-ERG.SG	drink.S3SG.AOR	three-NOM.SG	AFF	beer-NOM.SG

'Luka drank three beers, right?'

(4e)	*luḳa-m*	*dalia*	*sam-i*	*lud-i*	***xom****?*
	Luka-ERG.SG	drink.S3SG.AOR	three-NOM.SG	beer-NOM.SG	AFF

'Luka drank three beers, right?'

In (4a), the proper name 'Luka' is inside the scope of the particle xom; the speaker wants to ensure that the mentioned person drinking three beers is Luka and not another person. In (4b), the process of drinking is inside of the scope of the particle *xom*; the speaker wants to make sure that the three beers were drunk and not poured away. In (4c), the numeral *sami* 'three' and the modified head element *ludi* 'beer' are within the scope of the particle *xom*; the speaker wants to make sure that it was three beers that were drunk by the protagonist and not, e.g. four cocktails. At this point, the following conclusion can be made: the particle refers to phrases and not individual elements of the phrase, which is the reason why (4d) is incorrect as *xom* cannot split the phrase, transform it into a discontinuous one and still be grammatically correct. As for the last example (4e), where the particle is placed sentence-final: the protagonist, the act of drinking and also the beverages are all within the

scope of *xom*. Additionally, with the sentence-final positioning of *xom*, the speaker asks for confirmation from the hearer.

The particle *xom* can also be used in declarative clauses, but in such cases, it does not function as an interrogative particle anymore but only expresses the semantics of confirmation (affirmativeness):

(5)	*q̇vela*	*did-i*	***xom***	*bavšv-i*	*iq̇o*	*odesġac*
	every.NOM.SG	big-NOM. SG	AFF	child-NOM. SG	be.S3SG. AOR	at some time

'After all, all adults were children once.' (Antoine de Saint-Exupéry, *The Little Prince*)

(6)	*ṗaṭivmoq̇vare*	*ḳac-is*	*tval-ši*	***xom***	*q̇vela*
	vainglorious.GEN. SG	man-GEN.SG	eye.DAT.SG-in	AFF	every.NOM. SG

	adamian-i	*mis-i*	*taq̇vanismcemel-i-a*
	human-NOM.SG	his-NOM.SG	worshipper-NOM.SG-COP

'In the eyes of a respectful man, every human is his worshipper.' (Antoine de Saint-Exupéry, *The Little Prince*)

Declarative clauses with the particle *xom* are often used as an argument that reinforces/justifies the statement expressed in the discourse. These sentences show an unmarked argumentative structure since they do not contain argumentation markers. These types of sentences mainly use the verb *q̇opna* 'to be' - they are copula sentences and convey conventional or conversational implications.

The opinion that such sentences serve as argumentations is methodologically difficult to justify in the case of simple sentences, but in case of more complex syntactic constructions, we can the method of paraphrasing:

(7a)	*ċarmodgena-c*	*ara*	*akvs*	*mosalodnel*	*saprtxe-ze,*
	idea.NOM.SG-FOC	NEG	have.S3SG.PRES	expecting.DAT.SG	danger.DAT.SG-on

	gavipikre	*me.*	*mas*	***xom***	*arasodes*	*gamoucdia*
	think.S3SG. AOR	I.NOM.SG	he.NOM. SG	AFF	never	experience.S3SG.PERF

šimšil-i	*da*	*ċq̇urvil-i*
hun-ger-NOM.SG	and	thirst-NOM.SG

'He has no idea about the impending danger, I thought. - He has never experienced hunger and thirst.' (Antoine de Saint-Exupéry, *The Little Prince*)

→ Paraphrasing the second sentence

(7b)	***vinaidan***	*mas*	*araso-des*	*gamoucdia*	šimšil-i
	as	he.NOM.SG	never	experience.S3SG.PERF	hunger-NOM.SG

da	*ċq̇urvil-i*
and	thirst-NOM.SG

'As he has never experienced hunger and thirst.'

(8a)	*me*	*unda*	*vizruno*	*mas-ze.*	*igi*	***xom***
	I.NOM.SG	MPT-CL	care.S1SG.OPT	(s)he.DAT.SG	(s)he.NOM.SG	AFF

iset-i	*susṭ-i*	*da*	*iset-i*	*gulubrq̇vilo-a*
such-NOM.SG	weak-NOM.SG	and	such-NOM.SG	naïve.NOM.SG-COP

'I have to care about her/him. He is so weak and so naïve.' (Antoine de Saint-Exupéry, *The Little Prince*)

→ Paraphrasing the second sentence

(8b)	***vinaidan***	*igi*	*iset-i*	*susṭ-i*	*da*
	as	(s)he.NOM.SG	such-NOM.SG	weak-NOM.SG	and

iset-i	*gulubrq̇vi-lo-a*
such-NOM.SG	naïve.NOM.SG-COP

'As he is so weak and so naïve.'

As shown in the examples (7a-b) and (8a-b), we can consider that the particle xom is used as an argumentation marker when it is realised in the midfield of declarative sentences.

In interrogative sentences, the particle can be realised in combination with the modal word *šeiʒleba* ‘can’ (1635 such cases are confirmed in the GNC) and conveys possibility, permission or assumption in all three positions:

(9)	*[…]*	***xom***	*šeiʒleba*	*tan*	*raġac*	*gḳitxot?*
	[…]	AFF	can	at the same time	something.NOM.SG	ask.S1SG.O2PL.OPT

‘[…] I can ask you something at the same time, right?’ (Journal *Liṭeraṭuruli ṗaliṭra,* 2008)

(10)	*magram*	*kac-i-c*	***xom***	*šeiʒleba*	*iġos*	*meṣġvile!*
	but	man-NOM.SG-FOC	AFF	can	be.S3SG.OPT	partner.NOM.SG

‘But a man can also be a partner, can’t he!’ (Tariel Ča̱nṭuria, *Orni ḳuṗeši*)

(11)	šen-tan	*ertad*	*rom*	*ṭrailer-it*	*vimgzavro,*	***xom***	*šeiʒleba?*
	you.DAT.SG-WITH	together	that	trailer-INST.SG	travel.S1SG.OPT	AFF	can

‘Is it possible for me to travel with you in a trailer?’ (Aḳaḳi Gegenava, *Mogzauris dġiurebi*)

The combination *xom šeiʒleba* can also be used in declarative clauses:

(12a)	*magram*	*zogžer*	*vpikrob:*	***xom***	*šeiʒleba*	*rom*
	but	sometimes	think.S1SG.PRES	AFF	can	that

	adamian-s	*sakme*	*daavicṣdes.*
	human-DAT.SG	business.NOM.SG	forget.S3SG.O3SG.OPT

‘But sometimes I think: a human can forget about the business, can’t he.’ (Antoine de Saint-Exupéry, *The Little Prince*)

(12b)	*magram*	*zogžer*	*vpikrob:*	*šeiʒleba*	***xom***	*rom*
	but	sometimes	think.S1SG.PRES	can	AFF	that

	adamian-s	*sakme*	*daavicṣdes?*
	human-DAT.SG	business.NOM.SG	forget.S3SG.O3SG.OPT

‘But sometimes I think: a human can forget about the business, right?’

(12c)	*magram*	*zogǯer*	*vpikrob:*	*šeiʒleba*	*rom*	*adamian-s*
	but	sometimes	think.S1SG.PRES	can	that	human-DAT.SG

sakme	*daavicq̇des*
business. NOM.SG	forget.S3SG.O3SG.OPT

‘But sometimes I think: can a human forget about the business?’

In the case of elimination of the particle *xom* as shown in (12c), the dependent clause requires a transformation into an interrogative clause, which can function as a rhetorical question. The paraphrase of this sentence would be: “A man cannot forget his work.” In the case of the transposition of the particle *xom* in the second position in (12b), the affirmative sentence with the semantics of possibility is preserved, but the perspective changes: the speaker expects to receive confirmation from the listener.

The combination *xom šeiʒleba* can also be in the second position as in the next example, and here too the particle *xom* conveys the expectation of the speaker to receive confirmation:

(13)	*ese-c*	***xom***	*šeiʒleba*	*iq̇os*	*liṭeraṭura?*
	this.NOM.SG-FOC	AFF	can	be.S3SG.OPT	literature.NOM.SG

aman-a-c	***xom***	*šeiʒleba*	*besṭseler-is*	*saxel-i*
this.ERG.SG-EMPH.V-FOC	AFF	can	bestseller-GEN.SG	name-NOM.SG

moixveč̣os?

gain.S3SG.O3SG.OPT

‘This can also be literature, right? This can also gain the title of a bestseller?’ (Nene Ḳviniḳaʒe, *Iaguarebis tekno*)

Declarative clauses with the particle *xom* are characterised by more intensity, the persuasive power of the opinion expressed by the speaker is greater, which is strengthened by the repetition method used in this case. Accordingly, these types of sentences are often found in the speeches of politicians.

In interrogative sentences, the particle *xom* is often found in combination with the

negation particle *ar*, although the negation particle itself is not desemanticised (also called semantic bleaching), but the meaning of the sentence does not convey negation on a pragmatic level. In such sentences, both particles *xom* and *ar* should be considered as one functional element '*xom+ar'*. In case of transposition and elimination, they are moved or eliminated together. The combination of *xom* and *ar* is used in the initial position during a polite question:

(14a) *xom* *ar* *gciva?* vs. (14b) *gciva?*
AFF NEG being cold.S2SG.PRES being cold.S2SG.PRES
'You are not feeling cold, are you?' 'Are you cold?'

(15a) *xom* *ar* *dagaviçqdeba?* vs. (15b) *dagaviçqdeba?*
AFF NEG forget.S2SG.O3SG.FUT forget.S2SG.O3SG.FUT
'You won't forget, will you?' 'Will you forget?'

(16a) *xom* *ar* *geçqineba?* vs. (16b) *geçqineba?*
AFF NEG being offended.S2SG.O3SG.FUT being offended.S2SG.O3SG.FUT
'You won't feel offended, will you?' 'Will you feel offended?'

The combination *xom+ar* is mostly found in the second position, and depending on which verb it is combined with, it conveys different semantics:

- Questions with propositional semantics:

(17a) *rame* ***xom*** ***ar*** *ginda?*
something.NOM.SG AFF NEG want.S2SG.O3SG.PRES
'Do you want anything?' (Aḳaḳi Gegenava, *Mogzauris dġiurebi*) → offering to bring/buy

vs.

(17b) *rame* *ginda?*
something.NOM.SG want.S2SG.O3SG.PRES
'Do you want something?' → yes/no question

- Question with the semantics of doubt:

(18a) *brma* ***xom*** ***ar*** *aris?*
blind.NOM.SG AFF NEG be.S3SG.PRES
'(S)he isn't blind, is (s)he?' (Niḳo Lomouri, *Ṗaçia mego-brebi*) → expressing doubt

vs.

(18b) *brma* *aris?*
blind.NOM.SG be.S3SG.PRES
'Is (s)he blind?' → yes/no question

- Question with clarification/inquiring semantics:

(18a) *Pikria* ***xom*** ***ar*** *ginaxavs?*
Pikria.NOM.SG AFF NEG see.S3SG.O2SG.PRES
'You haven't seen Pikria, have you?' (Mixeil Ǯavaxišvili, *Arsena marabdeli*) → inquiring

vs.

(18b) *Pikria* *ginaxavs?*
Pikria.NOM.SG see.S3SG.O2SG.PRES
'Have you seen Pikria?' → yes/no question

- Rhetorical question:

(19a) *umizezo-d* ***xom*** ***ar*** *gaġares?*
groundless-ADV.SG AFF NEG expell.S3PL.O3PL.AOR
'They weren't expelled without reason, were they?' (Radio *Tavisupleba,* 18.02.2004) → rhetorical

vs.

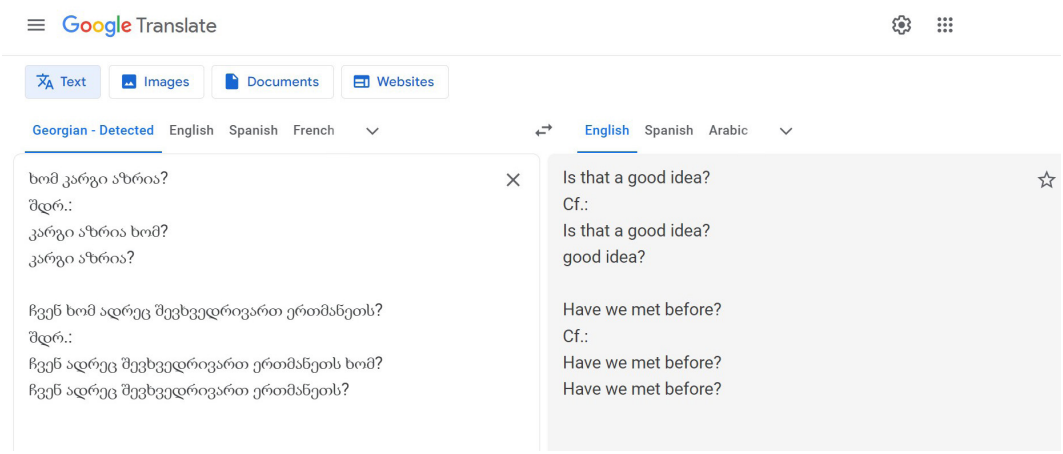
(19b) *umizezo-d* *gaġares?*
groundless-ADV.SG expell.S3PL.O3PL.AOR
'Were they expelled without reason?' → yes/no question

Undoubtedly, there are more combination possibilities and more semantic classifications, which will be dealt with in an upcoming work, as it would go beyond the scope of this paper.

RESULTS

The functional-semantic analysis of uninflectable elements is utterly significant not only for refining the theoretical model and for creating a functional grammar of the Georgian language, but it also has practical significance for the development of language technologies, especially for the improvement of automatic translation. None of the currently available translation programs can adequately convey the semantic difference in Georgian sentences from a functional-semantic point of view:

Figure 10: How Google translates sentences with and without the particle xom



As shown in Fig. 10, Google Translate does not differentiate between the meaning of sentences with or without the particle *xom*, which makes the significance of such an analysis all the more necessary.

CONCLUSIONS

The analysis and the variety of examples in the present paper have shown that the particle *xom*, even though invariant, can trigger different readings depending on the position and combination of other elements. Several relevant factors such as clause type (declarative, interrogative), the position of the particle in the sentence (initial, midfield, final position), the ability to transpose and the resulting scope effects or the combination ability with other uninflectable words in a sentence, determine the functionality and the semantics of the particle in relation to the sentence.

From the presented analysis in this paper leaves, I can conclude as follows:

- In initial or final position, the particle *xom* refers to the whole sentence but triggers different readings:

a. In initial position, the affirmation requires confirmation from the perspective of the listener;
b. In final position, the affirmation is given from the perspective of the speaker;

- The particle *xom* refers to entire phrases and not to single elements of phrases;
- When combined with the negation particle *ar*, the combination *xom+ar* has to be considered one functional element;
- Depending on the position *xom+ar*, the sentence can have different semantics:
  a. In initial position, the sentence can convey politeness;
  b. In midfield position, the following semantics can be conveyed:
     i. Propositional semantics,
     ii. Semantics of doubt,
     iii. Clarification/inquiring semantics,
     iv. Rhetorical question.

The analysis of the particle *xom* showed that in order to accurately understand and translate Georgian, not only a morphosyntactic but additionally a semantic-pragmatic analysis should be implemented. Of course, there are still many relevant aspects left to research; this paper served to present a first approach and to open the topic for future research.

ABBREVIATIONS

ADV	adverbial case	MPTCL	modal particle
AFF	affirmative	NEG	negation
AOR	aorist tense/aspect	NOM	nominative case
COP	copula	O	object
DAT	dative case	OPT	optative
EMPH.V	emphatic vowel	PERF	perfect tense/aspect
ERG	ergative case	PLUPERF	plusquamperfect
EXT.V	extensional vowel	PRES	present tense
FOC	focus	PL	plural
FUT	future tense	S	subject
GEN	genitive case	SG	singular
INST	instrumental case	1/2/3	1st/2nd/3rd person

REFERENCES

Agapova, S. (2014, June). On text linguistics. In *Collected Articles of the 3rd International Linguistics Conference* (Taganrog, Russia) (p. 265). Cambridge Scholars Publishing.

Association of language modelling. (n.d.). Enis modelirebis asociacia. Retrieved on June 15, 2024, from https://www.ena.ge/explanatory-online

Big Georgian dialect base and the electronic atlas of Georgian dialects GEDA. (n.d.). *Didi kartuli dialect'uri baza da kartuli dialekt'ebis elekt'ronuli at'lasi GEDA*. Retrieved on June 15, 2024, from http://www.corpora.co

Georgian Lexicon. (n.d.). *Kartuli leksik'oni*. Retrieved on June 16, 2024, from https://www.ganmarteba.ge/word/ბოძ

GNC. (2024, June 15). *Georgian Dialect Corpus*. K'art'uli enis erovnuli k'orp'usi. Retrieved on June 15, 2024 from http://gnc.gov.ge/

Jorbenadze, B., K'obakhidze, M., & Beridze, M. (1988). *Kartuli enis morpemebisa da modaluri element'ebis leksik'oni (Masalebi kartuli enis sist'emat'uri k'ursistvis)* [The lexicon of morphemes and modality elements of the Georgian language (Collection for the systematic course of the Georgian language)]. Tbilisi: Publishing House "Metsniereba".

Kamarauli, M. (2023). Approximative verbs: A symbiosis of the nominal and the verbal domain (on the example of the Georgian language). *Millennium, 1*, 50-70.

Low resource languages. (n.d.). Retrieved on June 15, 2024, from https://github.com/RichardLitt/low-resource-languages

TITUS - Thesaurus Indogermanischer Text- und Sprachmaterialien. (n.d.). Retrieved on June 15, 2024, from https://titus.fkidg1.uni-frankfurt.de/

Wallis, S., & Nelson, G. (2001). Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery, 5*, 305-335.

Rustaveli Goes Digital - Parallelkorpus. (n.d.). *Parallel Corpus of the Translations of the epic "The Knight in the Panther's Skin"*. Retrieved on June 15, 2024, from https://rustaveli-goes-digital.de