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.2023.245

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 The classical grouping of languages established in linguistics has been replaced by a new paradigm of classification.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.

METHODS

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), which is the reason. GNC belongs to the type of diachronic corpora, which com-

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/

bines 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

\rightarrow C \land Nich	ht sicher gnc.gov.ge/gnc/corpu	us-list?session-id	=25741786777	2900				
JJJJJ	<mark>ქ</mark> ართული ენის ეროვნული კორპუსი The Georgian National Corpus							
GNC Home	Corpus list							
Using the GNC Documentation	Select corpora from the li directly to the Concordan	st below by clic ce page.	king on their	names. Then click on "Query the selected corpora" below or go				
Publications Corpus list	Corpus	Size (words & punctuation)	Updated	Description				
Text list	GNC Old Georgian	7 101 021	2022-12-31	Georgian National Corpus, Old Georgian				
Query	GNC Middle Georgian	1 432 262	2019-06-25	Georgian National Corpus, Middle Georgian				
Concordance	GNC Modern Georgian	1 993 022	2023-01-01	Georgian National Corpus, Modern Georgian				
Collocations Word List	GRC	202 728 329	2016-12-05	Georgian Reference Corpus				
Text	GDC	1 694 362	2015-09-14	Georgian dialect corpus				
Overview	GNC Political texts	1 436 075	2019-08-06	Georgian National Corpus, Political texts				
Grammatical features	GNC Law texts	1 495 985	2019-04-15	Georgian National Corpus, Old and Middle Georgian, Law texts				
Parce	GNC Megrelian	89 404	2015-09-14	Georgian National Corpus, Megrelian				
	GNC Svan	473 180	2015-09-14	Georgian National Corpus, Svan				

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*

→ C ▲ Ni	icht sicher	gnc.gov.ge/g	inc/sin	nple-qu	ery														Gast @ Gast)
Text list	С	hoose positional	constr	aints:		gnore structi	ural	positions												
Query Concordance Collocations Word List		target word = features: [repetition:	select]	ct values	for 'f	eatures'. OI	nly c	ombination	s tha	t do occur in t	ne c	orpus can be s	elect	ed.						
Overview	+	dipl	Do	ne /	Add a	is alternativ	e	Delete this	s alte	ernative F	Res	et								
Grammatical features		simplified le lemma	Wo	rd class	0	Rel		Nonhum		DVoc		DL		Impv-II	[0V]		S:1Sg		Causal	
		reference		Punct		Refl		Hum		DInst		L		Iter-II	🗆 ov		S:1Pl		Foc	
Parse	l	struct		Symbol		Recip		Anim		DAdvb		Attributive		Conj-II	🗆 sv		S:1		Disc	
	с	hoose a subcorp i		Interj Pp		SIndef Indef	Q	ualifying Meas		DGen DDat		Att Person		Opt Impv	Verbal arguments <a href="https://www.sciencescommutation-commutatio-commutation-commutation-commutation-commutation-commutation-commutation-commutation-commutation-commutation-commutation-commutation-commutation-commutation-commutation-commutation-commutation-commut-commutation-commutatio-commutatio-c</td> <td>Di</td> <td>DO:3Pl</td> <td></td> <td>Comm Mann</td> <td></td>	Di	DO:3Pl		Comm Mann	
	ſ	(Entire cor		Modal		Quant		Qual		DErg		3		Aor	AuxTransHum>		DO:3Sg		Temp	0
		Additional c		Compl		Neg		Case		DNom		2		ConjFut	AuxIntr>		DO:3		Loc	
		document		Cj		Int		Trunc	Do	ble number		1		Cond	S-D0-I0>		DO:2PI	D	erivational affix	
		collection		Num		Dem		Voc		DOIdPI		Verb class		Fut	S-IO>		DO:2Sg		Der:na6n	
		author		Pron		Poss		Inst		DNewPl		PrePv		IterImpf	S-D0>		DO:2		Enclitics	
		translator		Adv		Pers		Advb		DSg		Pv		IterPres	□ <\$>		DO:1PI		Encl:do	
		genre		V		Letter		Dir		DPI		Imperfective		Iter	<null></null>		DO:1Sg		Encl:Q	
		language		А		Frac		Ben	Т	riple case		Inv		Conj-I	Argument case		DO:1		Encl:80	
				N		Prop		Gen		DDInst		Caus		ConjPres	<io:gen></io:gen>	Inc	lirect object		Encl:რე	
	tion: D	esian & impler	Su	ıbclass		Ord		Dat		DDErg		PassState		Impv-I	<io:dat></io:dat>		IO:3PI		Encl:რა	

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)

	ht sicher gnc.gov.ge/gnc/simple-query			🕼 🙆 Gast 🗄
JONC JJJJJ	<mark>ქართული ენის ეროვნ</mark> უ The Georgian National Corpus	🏾 _{Englis} ული <mark>კ</mark> ორპუსი	h ▼ □ Use transliteration Sign in	
GNC Home About the project Using the GNC Documentation Publications Corpus list Text list Query Concordance Collocations Word List Text	GNC ວຽວლი ქართული, G Advanced search switch to Basic search [features = ("Erg" "Num")] [feature = ("Erg" "N")] Run Query Reset Build graphical query choose positional constraints:] Ignore struc . vord =	h h ss = ("A" "Erg")] [features ltral positions ltral positions ltratures: "A" "Erg" repetition: [v	 rarget word = features: "Erg" "N" repetition: 1 • 	
Overview Grammatical features Parse	+ Additional constraints for this position: dipl simplified lemma lemma reference struct or	Additional constraints for this position: dipl simplified lemma lemma reference struct or	+ Additional constraints for this position: + dipl simplified lemma lemma reference struct or	

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

Figure 4 *Results of the search*

$\leftarrow \rightarrow$	C Z	Nicht sicher gnc.gov.ge/gnc/concordance		Gast :
Advanced	search swi	tch to Basic search Query history		
[feature = ("Erg	es = ("Er " "N")]	g" "Num")] [features = ("A" "Erg")] [features		
Run Que	ery Refi	ne window: document V Stop Saved queries		
Done, Run	ning time: 21	.42 sec. (4.87 CPU sec.)		
- 0.0			500	
Type: KWI	C ♥ Att	: Word V U Show line filter Attributes Structures (U show in match) Page size	e: Context size: 500px V	
Hit 1 - 30	of 5068 P	revious Next Go to: Download (Excel mode) Copy query URL		
corpus	cpos		match 5	
gnc-kat	1329646	დე და ახლა წელზევით შიშველი დავჯექი, წვიმის ათასმა, ათი ათასმა, ასი	ათასმა თბილმა თითებმა	გადამირბიწეს სხეულზე – თემოო! – მომიახლოვდა ;
gnc-kat	1433750	კვრობა, მხამად! მხამად! <lb></lb> თეიმურაზს ერთ წუთმი თვალწინ	ათასმა წვრილმანმა აზრმა	, <lb></lb> სურათმა და მოგონებამ გაურბინა. <
gnc-kat	1406893	პასხარეს <ib></ib> კაჭარი გიორგი ერისთავმა, ცაგარელმა, ანტონოვმა <ib></ib> და	ასმა სხვა მწერალმა	. გინდა გაიგო ღრმა მიზეზი ასეთი სიძულვილისა? <
gnc-kat	1318913	აკომპლექსი არაფერს წიმწავს. პოეტს სულში უნდა ჩახედოთ, – თქვა ერთ-	ერთმა დამწყებმა კრიტიკოსმა	. – ვუყურებ, ბატოზო, და ვერაფერს ვერ ვხ
gnc-kat	611206	და ვქნათ ამ უცხო ქვეყანაში? <ib></ib> – ვაი ჩემს გამჩენს! – უპასუხა	ერთმა დროულმა მეურმემ	. – <lb></lb> უცხო ქვეყაწაში რად იქწებით! ასე გოწია,
gnc-kat	418303	»ც კარგი რამეა – რიტორებაშია <lb></lb> დაბეჭდილი, ის გეყოფათო". ერთხელ	ერთმა თამამმა მოწაფემ	ჰეითხა: <lb></lb> "რაც აქ არ არის მოყვაწილი ამ წიგწში,
gnc-kat	1086533	კაცმა რომ თქვას, რატომ უნდა გჯეროდეთ?! – დაცხრა მოულოდწელად. –	ერთმა თქვენისთანა გლეხმა	სამართალს პური როგორ უწდა აჭამოს?! სამართალი
gnc-kat	670550	ლი საცვალი, ანდა, გასაშრობად რადიატორზე გადაფენილი კოლგოტკები (ერთმა კეთილმა დიასახლისმა	ფრანსუაზას კოლგოტკები თბილისში გადაუგზავნა აქ
gnc-kat	906857	აა რატომ ვეღარ ხედავდა წვივის აწითლებულ კანს. კიბეებზე რომ ადიოდა,	ერთმა მოწა ქალმა	მეორეს დაუძახა: ჩვენი ბიჭებიც იბანენო და ისევ მო
gnc-kat	609849	რომ ასე უშიშრად გაგვირბის და გამოგვირბისო, იძახდნენ. <lb></lb> ჰკრა	ერთმა მოსხეპილმა ყმაწვილ-კაცმა	ცხენს ქუსლი და ზედ <lb></lb> მიაგდო ყიზილბაშს. ყიზი
gnc-kat	406999	/> და ზოგიც ციმბირში გაგზავწეს. გაიარა ორმოცდაათმა წელიწადმა. <lb></lb>	ერთმა მოხუცებულმა სოფლელმა	სიკვდილის წიზ აღიარა <lb></lb> საჯაროდ შემდეგი: მე ვ
gnc-kat	2044127	ეცეს!" მოიფიქრა ვიღაცამ. საზამ უბედური კაცის პირმშო წრეს გაარღვევდა,	ერთმა მსუქაწმა ქალმა	ხელი სტაცა, განზე გაიყვანა და გულმოდგიწედ აუხს
gnc-kat	976205	იმედსა და სიზმარში. პოპინაც ამ "ერთმა მშვენიერმა დღემ" შეაცდინა, ამ "	ერთმა მშვეწიერმა დღემ	" დააჯერა, მისი შვილიც როდისმე გაწიკურწებოდა, გ
gnc-kat	976197	ა თენდება, მაგრამ მხოლოდ ოცნებაში, იმედსა და სიზმარში. პოპინაც ამ "	ერთმა მშვეწიერმა დღემ	″ შეაცდინა, ამ "ერთმა მშვენიერმა დღემ″ დააჯერა, -
gnc-kat	2100699	ა, სავახშმოდ წამოსვლაც არ სურდა. იმ საღამოს, ვახშმის შემდეგ	ერთმა პარიზელმა ქართველმა	რესტორან "ოქროს საწმისში" დამპატიჟა. არ უთქვამს
gnc-kat	1211444	– ღვინო თუ დაგჭირდეს, არ მომერიდო! როგორც იქნა,	ერთმა საბარგო მაწქაწამ	გაგვიჩერა. მძღოლს კაბინაში ბავშვიაწი ქალი ეჯდა. ≀
gnc-kat	568491	lb/> რამ მოვიგონე, რომ მოჯამაგირეს სახლში პარვა აღარ შეუძლიან. <lb></lb>	ერთმა უბრალო შემთხვევამ	კი მაპოვწიწა წამალი. ჩემი დეწშჩიკი <lb></lb> ფრიად დი
gnc-kat	433965	<head> <lb></lb> ლამურა </head>	ერთმა უგნურმა თაგუნამ	იუკადრისა თაგვობა, დასწყევლა თვისი გაჩენა, ბუნე
ane-kat	1036975	มากลังพิศภิณิสิกใส่กุ่ม ๆ เกมส์คาลิก สิกกลักประมาณได้การการการ 200 200 การการการการการการการการการการการการการ	Sanda min and man and man a	amoul on light mil an manna van ma a son a lig

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

Figure 5

Parsing of a sentence

← → C ▲	Nicht sicher gnc.gov.ge/	gnc/pars	2	© Gast :					
GNC Home	Parse								
About the project Using the GNC Documentation Publications	Here you can parse sentences with the morphological analysers that are being used to analyze the texts of the GNC. Write a sentence or shorter text and click the 'Parse' button. You can click on the morphosyntactic features of the analyzed text to have them explained.								
Corpus list Text list	Language variety: Mod	Language variety: [Modern Georgian v] მე მლიერ მიყვარს იისფერ თოვლის ქალწულებივით ხიდიდან ფენა.							
Query Concordance Collocations Word List	Parse Reload Show all readings Show used rules Show dependencies UD features								
Text	მე	3	ag Pron Pers 1 Dat Sg						
Overview Grammatical	ძლიერ	4	dლიერ Adv Deg						
features	მიყვარს	1	სიყვარულ ი/ყვარ V MedPass Inv Pres OV <s-do> <s:dat> <do:nom> S:1Sg DO:3</do:nom></s:dat></s-do>						
Parse	იისფერ	6	nnlug[j]@•n A Gen Att <oldpl></oldpl>						
	თოვლის	1	თოვლ-ი N Gen Sg						
	ქალწულებივით	1	ျမင္လာေတြက္လာက္ N Hum Nom PI NewPI PP PP:3001						
	ხიდიდან	3	ხიდ-ი N Inst Sg PP PP:დამ						
	ფენა	6	ങ്ങള്(ച)/തൃന്ദ് N VN Nom Sg						
		1	. Punct Period						

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

→ C ▲	cht sicher gnc.gov.ge/gnc/concordance		B <u></u> ((Gast
iNC Home	Basic search switch to Advanced search			
bout the project	მაღაზიებში			
sing the GNC		li li		
ocumentation	Run Query Refine window: document v Stop Saved gueries			
ublications	Done. Running time: 0.10 sec. (0.03 CPU sec.)			
	Type: kwic Y Att: word Y Show line filter Attributes Structure	es (show in match) Page size:	Context size: 500px Y	
orpus list		onit (= onon in match) (rage one)	1 context biter compare	
ext list	Hit 1 - 30 of 1301 Previous Next Go to: Download (Exc	al mode) Copy query URL		
luery	corpus cpos	,	match	
oncordance	gnc-kat 696932 ლიტიკოსად, ვირი ყოფილა. მთავარი ის კი არ	არის, თბილისის სასურსათო მაღ	ღაზიებში უშედეგოდ რომ დაეძებს თურმე საფუაჩ	რს ვინმე ქაღ
ollocations	gnc-kat 778242 კაღრესად ქალური ღონისძიებების ჩასატარებღ	აად: ექიმთან უნდა შეევლო, მაღ	ღაზიებში - შეეჭყიტა, გადამყიდველისთვის რაღაც (დაევალებინ
/ord List	gnc-kat 1379888 იების ბედის გასაგებად და ორიოდე გროშის მ	ისაღებად. <lb></lb> nd <u>Shr</u>	how context Look up in: Rayfield	🗵 🕹 🕹 🕹 🕹 🕹
evt	gnc-kat 1379934 ა საკუთარ <ib></ib> ცხოვრებისა და ბედის წიგნს.	<lb></lb> ഗാ ഗ് വുന വി		რისთვის!
exe	gnc-kat 1380097 ი და სევდიანი დასასრული.	ნგამოშვებით თეიმურაზი იმ	word: asasanjoan	ააწყდებო
weiview	gnc-kat 1380205 ე უსაქმური თეიმურაზი ზოგჯერ მთელ დღეს	ჭრიალებდა <ib∕> სა₃ოⴋისიო</ib∕>	dipl: მაღაზიებში	საქვეყნო
eatures	gnc-kat 1787449 უ ასორტიმენტის, მაგრამ უვარგისი საქონელი	იქწებოდა, რაც ოფურჩხეთის si	implified lemma: მაღაზია	ი შერბენი
	gnc-kat 1787474 კოფად მიიჩზიეს და გადაწყვიტეს, უარი ეთქვა	თ, რათა ეს უარი სხვა დროს	lemma: მაღაზი[ა]	ე ვერც ერ
arse	gnc-kat 1788332) ზომის, თეთრი, იუგოსლავიური, ქალის ჩექმა	ცენტრსა და მის მიმდებარე	features: N Dat PI NewPI PP PP:00	რ გაჭაჭან
	gnc-kat 1821612 აბდა თურმე, იქ გაიზარდა. სკოლა რომ დაამთ	ავრა, მისი კლასის გოგოწები	documente NG/chiladae e/chiladae e ceder	in სამი წეე
	gnc-kat 1971398 ონდესო, ეს ფერი ხაზები ძალიან მიხდებაო, ი	ლოწდ სადმე ვიშოვიდეო. აქ	uocument. NG/chiladze-o/chiladze-o+godol	' ხვალ გემი
	grc 63787 ყოფნით რაიმე გაიგონ. ისინი მხოლოდ მზამზ	არეულ საგწებს ყიდულობეწ	title: გოდორი	ღაზია, სა
	grc 537720 ა წივთების ბედის გასაგებად და ორიოდე გრო	.შის მისაღებად. იმ	reference: გოდორი პირველი ნაწილი I 35	ის კვწესიძ
	grc 537758 ათხულობდა საკუთარ ცხოვრებისა და ბედის წ	იგნს. რა არ იყო იმ	author: ჭილაძე, ოთარ	ვის! კაბებ
	grc 537906 აწყისი და სევდიანი დასასრული. ხა	ნგამოშვებით თეიმურაზი იმ	genre: /fiction/	ხებოდა. მ
	grc 537991 ა> უკვე უსაქმური თეიმურაზი ზოგჯერ მთელ (<u></u> ღეს ტრიალებდა საკომისიო	language: kat	ყნოდ გატ
	grc 1654143 ჭერესებთ?" – რის შემდეგაც დარცხვენილნი მ	ივდივართ კარებისაკენ. ასეთ	laliyuaye. Kat	ქვრს რასმ
	grc 1670044 ფიციანტი არ დაგიტოვებია. ვითომ არ ვიცი, რ	ატომ დადიხარ ტანსაცმლის მაღ	ღაზიებში 🛛 . იქ გარდერობში – ეკა, დად	ე პისტოლე

Search result of the noun dsgsvogodo "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"

CNC Hama	un sicher gric.go	.ge/git/concordance) :
About the project	Basic search sw ტრიალებდა	ch to Advanced search			
Documentation	Run Query I	Refine window: document V Stop Saved queries			
Publications	Done. Running tim	e: 0.04 sec. (0.02 CPU sec.)			
Corpus list	Type: kwic 🗸	Att: word 🗸 🗋 Show line filter Attributes Structures (🗋 show in match) Page siz	e: 📃 Context size	e: 500px 🗸	
Text list	Hit 1 - 30 of 1537	Previous Next Go to: Download (Excel mode) Copy query URL			
Query	corpus cpo		match		
Concordance	gnc-kat 1151	3 ორთავეს რაღაც გვჭირდებოდა. საბაზრო კვირადღე იყო, მუშტარი ბლომად	ტრიალებდა , დ	ახლთან პატარა რიგიც კი იდგა. ჩვენც დავდ	ქით (
Collocations	gnc-kat 1198	5 ჰ მიცვალებული უნდა მოესვეწებინათ იქიდან. კი მივხვდი, რა მოლოდინიც	Show context Look	up in: Rayfield	×ðr
Word List	gnc-kat 1563	0 ა. ალბათ, დილის თერთმეტზე აწ უფრო გვიაწ. ერთი კია, ისევ ღომის სუწი	wor	d: @@a.vmabca.v	þð
Text	gnc-kat 1690	9 ულებდა. ძობას ხანდახან თუ დაასაქმებდნენ, თორემ ისე დუქანში უქმად	WOI		b
Overview	gnc-kat 1873	5 მი დღე-ღამე ალოდინა სარჩიმელია თუთამხიამ. ჩანს, სადმე ახლომახლო	aip	ი: ტოიალეიდა	δJ
Grammatical	gnc-kat 2095	/ დისძე, – პასუზი გადაუდო ლაზძა. ვესტიბიულში ძარტო არჩილა	simplified lemma	a: ტრიალი	- 1
features	gnc-kat 2314	8 უტუცის მოგონილი აო იყო ბოიოლიას ხოივი და ოაღაც დიდი ემმავობა ოომ	lemma	a: ტრიალ•ი/ტრიალ	n
	gnc-kat 3007	0 ერთად პირდაპირ ქვაფენილზე სხაძდნენ. ჩვენი ძესაძე აგენტი ძორიანლო 0 აკლანი ნაშ მოხლა ამ ლნოვს ლიის დინილაწში ლაი თხლიშიზი ლი აკლი	feature	s: V MedAct Impf <s> <s:nom> S:3Sg</s:nom></s>	me
Parse	gnc-kat 3254	 Songson as another set and the set as a set and the set as a set and the set as a set as	documen	t: NG/amiredzhibi-ch/amiredzhibi-ch+data	-
	gnc-kat 3878	6 (do bright begans, c/p> cp2 hat good of bright of both for both gond bright of both gond bright and brigh		tutashxia	I'Gr
	gnc-kat 5214	$6 \rightarrow 0$ (jgg/gm, chog) (jgg/gm) and (co, co) (control < 0) $>$ (h)	title	e: დათა თუთაშხია	0
	gnc-kat 6803	$4 \rightarrow 5$ anthrow hydrayona finitial symbols with the momenta during the second state of the second state o	reference	e: 8	m
	gnc-kat 7712	9 አრებოდა, სოველ შემთხვევაში, ვიდრე არ "ეცოდინებოდა", რა უბედურება	autho	ი: ამირიჯიბი, მაბუა	5.
	gnc-kat 7951	9 არივე ხელით იხუტებდა გულში ელიზბარის წიგნს, ღამით კი, მამთურივით	orig_date	a: 1974	bn
	gnc-kat 8088	6 უიზბარ! ღმერთო ჩემო! მარტო ჩემმა უბედურმა ქმარმა იცოდა, რა ცეცხლი	ong-uad	c. 19/4	þn
	gnc-kat 8289	8 უგან, სავარძელშიც, მაგიდაზეც, სარკმლის რაფაზეც, იატაკზეც მუმლი კი	genro	e: /fiction/	რა
	gnc-kat 8289	0 ლშიც, მაგიდაზეც, სარკმლის რაფაზეც, იატაკზეც მუმლი კი ტრიალებდა,	language	e: kat	nõ

Figure 8

Search result of the affirmative particle xom (bmd)

GNC Home	Basic sear	ch switch t	Advanced search			
About the project	ხომ					
Using the GNC			1			
Documentation	Run Que	ery Refi	e window: document v Stop Saved gueries			
Publications	Done. Run	ning time: 0.	24 sec. (0.19 CPU sec.)			
Corpus list	Type: kwi	c 🗸 Atl	: word 🗸 🗆 Show line filter Attributes Structures (🗆 show in	match) Page siz	e: Context size: [5	500px 🗸
Text list	Hit 1 - 30	of 131364	Previous Next Go to: Download (Excel mode) Copy	y query URL		
Query	corpus	cpos			match 5	
Concordance	gnc-kat	142571	ასპობიო! – მამაძაღლი და მოსასპობიო! – გაგვაუბე	დუროს უნდა,	bmð !	> ერთი წამით გაინაბა ხალხი, მერე იცვირ
Collocations	gnc-kat	2065252	. ტყიდან გამოვარდნილი არავინაა დღეს. – ვირო მ	მადლიო, ასეა,	ხომ , ბატოწო	ერმიწე? – გიგა ძლივს სუწთქავდა. წესტო
Word List	gnc-kat	1657749	პთისა და ტალახიანი მუკის ყურება ძალიან თეატრალურაც	დ გეჩვენებათ,	Show context Look up	o in: Rayfield
	gnc-kat	1241782	:/p> - მოგეხმარები, ილიკო! - შევთავაზე მე. -	არ მანებებთ,		
lext	gnc-kat	1242185	იბსო, დაიწყო: – ილარიონა შევარდნაძე, კიდე მაჯო	ობე? მომკალი,	word:	ნომ
Overview	gnc-kat	1291236	აწომ კალთაზე დაქაჩა, განაგრძეო, ხატია ადგა. – ძ	ქმები იყვწეწო,	dipl:	ხომ
Grammatical	gnc-kat	1259969	ლობაზე დავადებიწე. – მარცხწივ რომ მოხვედროდ	<u></u> და, მო <u>კ</u> ლავდა,	simplified lemma:	ხომ
leatures	gnc-kat	1260495	:/p> – არც მე მძინავს! – თქვა ხატიამ. – ხვალვ	ე წავიყვანოთ,	lemma:	bmð
Parce	gnc-kat	637561	რის, ისიც გასარკვევია, ხვალ მე ისევ მე ვიქნები, თუ სხვა ვინ	მე – ადამიაწი	featuree	Adv Disc
	gnc-kat	847883	მოწამე. მე, პირადად, ათიათას ჯოჯოს ერთი კახპა მირჩევწი	ია ათიათასი	le cucures.	NG/mishusis day n/mishusis day notami
	gnc-kat	1655198	გამოვიქეცი, გამოვიპარეო, რომ დაიტრაბახებს გატა	აცებული, ამას	document:	NG/mishveladze-r/mishveladze-r+tomi-
	gnc-kat	554308	ბლზედ ეგ მეწერაო - <ib></ib> ითექრა და დაემორჩილა ბედისწ	ერასა. ბოლოს	title:	რჩეული თხზულებანი IV – ნოველები
	gnc-kat	1879915	ⴈდა, რომ აუზის ექიმი უხმობდა და არც შემცდარა.	– გაგაღვიძეთ	reference:	ქველი 169
	gnc-kat	1262264	/p> – ედემიკა, ახლა გაგიჩნდა რევმატიზმი?! –	- არ გამოხვალ	author:	მიშველაძე, რევაზ
	gnc-kat	1977651	ლი შემხვდაო და გაუბედავად გაიღიმა. – იღიმები	ხომ? გიხარია	genre:	/fiction/
	gnc-kat	811753	იმი. ტახმა კიდევ უფრო დაიჭყვირა და ისევ მისკენ შემობრუნ	ნდა. "დამჭერი	languago	kat
	gnc-kat	843249	რთველთა წოეც, ევროპულად ფეხი-ფეხზე გადადებულმა. ა	მ ბოლო დროს	language:	Kat
	anc-kat	538239	n> <lb></lb> - momoha stals, doñ anñagasimo, ma odoðind, <n< td=""><td>> <1/></td><td>ხომ მაგრამ ხ</td><td>പ്രത്തന്നെ പ്രപ്രവത്തന്നു പ്രവത്തന്നെ പ്രവത്തന്നെ പ്രവത്തം പ്രവത്തം പ്രവത്തം പ്രവത്തം പ്രവത്തം പ്രവത്തം പ്രവത്ത</td></n<>	> <1/>	ხომ მაგრამ ხ	പ്രത്തന്നെ പ്രപ്രവത്തന്നു പ്രവത്തന്നെ പ്രവത്തന്നെ പ്രവത്തം പ്രവത്തം പ്രവത്തം പ്രവത്തം പ്രവത്തം പ്രവത്തം പ്രവത്ത

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.

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.

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: bs3000) should be added at the end of the grammatical features:

Cf.:

EXAMPLE	GRAMMATICAL FEATURES
მეცადინეობდა	V MedAct Impf <s> <s:nom> S:3Sg</s:nom></s>
VS.	
მეცადინეობდასავით	AppV MedAct Impf <s> <s:nom> S:3Sg APP: სავით</s:nom></s>

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. Table 1 (see next page)

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:

Cf.:

Table 1

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

Frequency of particles in the GNC

- 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	ķarg-i	azr-i-a?
	AFF	good-NOM.SG	idea-NOM.SG-COP
'It is	a good idea, right?'		
(1b)	ķarg-i	azr-i-a	xom?
	good-NOM.SG	idea-NOM.SG-COP	AFF
'It is	a good idea, right?'		
(1c)	ķarg-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.o	1pl. aor	I.NOM.	and	Besarion-NOM.
				SG		SG
	otxas-i	manet-i	unda	тоеса		xom?
	fourhundred-NOM.SG	Mane- ti-NOM.SG	MPTCL	g i v e . s PLUPERF	3 s g .	AFF

'How much money did me and Besarion agree on thanks your help? He should have given me 400 Manetis, right?' (Davit kldiašvili, *Soloman Morbela*3e)

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

(2b)	ramde	n-ze gagvarige	me		da	besarion-i?
	how	much. settle.s2sg.o1	pl.aor I.nom.	SG	and	Besarion-NOM.
	DAT.SG	-on				SG
	xom	otxas-i	manet-i	unda	moe	ca?
	AFF	fourhundred-NOM.	Maneti-NOM.	MPTCL	give	.s3sg.pluperf
		SG	SG			

'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)čvenxomadre-cševxvedrivartertmanet-s?we.NOM.SGAFFearly-FOCmeet.S1PL.PERFeach other-DAT.SG'We have met each other before, haven't we?'

(3b)	čven	adre-c	ševxvedrivart	ertma	inet-s	xom?
	we.NOM.SG	early-FOC	meet.s1pl.perf	each	other-DAT.	AFF
				SG		

'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

Oķeaneebi tu aris tkvens planețaze?	"Has your planet any oceans?"
Ver getąvi, - tkva geograpma.	"I couldn't tell you," said the geographer.
A! - pațara uplisçuli ar moeloda aset pasuxs.	"Ah!" The little prince didn't expect such an answer.
Arc mtebi?	"Not even mountains?"
Verc magaze gipasuxeb.	"I couldn't answer that either."
Kalakebi, mdinareebi an udabnoebi?	"Towns, rivers or deserts?"
Verc magaze getąvi 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	luķa-m	dalia	sam-i	lud-i?
	AFF	Luka-erg.sg	drink.s3sg.aor	three-NOM.SG	beer-NOM.SG
' <u>Luka</u>	drank three be	ers, right?'			
(4b)	luķa-m	xom	dalia	sam-i	lud-i?
	Luka-erg.sg	AFF	drink.s3sg.aor	three-NOM.SG	beer-NOM.SG
'Luka	drank three be	ers, right?'			
(4c)	luķa-m	dalia	xom	sam-i	lud-i?
	Luka-erg.sg	drink.s3sg.aor	AFF	three-NOM.SG	beer-NOM.SG
'Luka	drank three be	ers, right?'			
*(4d)	luķa-m	dalia	sam-i	xom	lud-i?
	Luka-erg.sg	drink.s3sg.aor	three-NOM.SG	AFF	beer-NOM.SG
'Luka	drank three be	ers, right?'			
(4e)	luka-m	dalia	sam-i	lud-i	xom?
()	Luka-ERG.SG	drink.s3sg.aor	three-NOM.SG	beer-NOM SG	AFF
' <u>Luka</u>	drank three be	ers, 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 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

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)	<i>ġvela</i>	did-i	xom	bavšv-i	iġo	odesġac
	every.NOM.SG	big-noм.	AFF	child-NOM.	be.s3sg.	at some time
		SG		SG	AOR	

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

(6)	<i>p</i> ațivmo <i>q</i> vare	ķac-is	tval-ši	xom	<i>q</i> vela
	vainglorious. GEN.SG	man-GEN.SG	eye.DAT.SG- in	AFF	every.NOM. SG
	adamian-i	mis-i	taġvanismce	mel-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-Ex-

upéry, *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 *qopna* '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)	çarmodgena [.]	-c a	ra	akvs		mosalodnel		saprtxe-ze,
	idea.NOM.SG-F	OC N	EG	have.s3sc	.PRES	expecting.DA	AT.SG	danger.DAT.SG-on
	<i>gavipikre</i> think.s3sg. AOR	<i>те</i> . І.пом	.SG	<i>mas</i> he.nom. sg	xom AFF	<i>arasodes</i> never	gama expei	oucdia rience.s3sg.perf
	šimšil-i hun- ger-NOM.SG	<i>da</i> and	<i>çqur</i> thirst	<i>vil-i</i> t-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*)

 \rightarrow Paraphrasing the second sentence

(7b)	vinaidan	mas	araso- des	gamoucdia	šimšil-i
	as	he.nom.sg	never	experience.s3sg.perf	hunger-NOM. sG
	da	çqurvil-i			
	and	thirst-nom.sg			
'As h	e has never e	xperienced hung	ger and thi	irst.'	
(8a)	те	unda vizruno)	mas-ze. igi	xom
	I.NOM.SG	MPT- care.sl	SG.OPT	(s) he. (s)he.nom.sg DAT.SG	AFF
	iset-i	sus ț- i	da	iset-i	gulubr <i>qvilo-a</i>
	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*)

 \rightarrow Paraphrasing the second sentence

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

iset-i gulubrqvilo-a such-NOM. naïve.NOM. SG 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	šei z leba	tan	raġac	gķitxot?
	[]	AFF	can	at the	something.NOM.SG	ask.s1sg.o2pl.opt
				same time		

'[...] I can ask you something at the same time, right?' (Journal *Liţeraţuruli paliţra*, 2008)

(10)	magram	kac-i-c	xom	šei z leba	iġos	mecġvile!
	but	man-NOM.SG-FOC	AFF	can	be.s3sg.	partner.NOM.SG
					OPT	

'But a man can also be a partner, can't he!' (Tariel Čanțuria, Orni kupeši)

(11)	šen-tan	ertad	rom	țrailer-it	vimgzavro,	xom	šei z leba?
	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?' (Akaki Gegenava, Mogzauris dgiurebi)

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

(12a)	magram	zogžer	vpikrob:	xom	šei 3 leba	rom
	but	sometimes	think.s1sg.pres	AFF	can	that
	adamian-s	s sakme	daavicģdes.			
	human-DA	T. business.	forget.s3sG.	03sg.opt		
	SG	NOM.SG				

'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:	šei z leba	xom	rom
	but	sometimes	think.s1sg.pres	can	AFF	that
	adamian-s	sakme	daavicqdes	?		
	human-DA	T. business.	forget.s3sg.	03sg.opt		
	SG	NOM.SG				

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

(12c)	nagram zogžer	vpikrob:	šei z leba	rom	adamian-s
	out sometimes	think.s1sg.pres	can	that	human-DAT.SG
	<i>akme daavicqu</i> pusiness. forget.s3	daavicą́des forget.s3sG.03sG.0PT			
	vusiness. forget.s3	sg.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	šei 3 leba	iqos	lițerațura?
	this.nom.sg-foc	AFF	can	be.s3sg.opt	literature.NOM.SG
	aman-a-c	xom	šei z leba	besțseler-is	saxel-i
	this.erg.sg-emph.v	-FOC AFF	can	bestseller-GEN	NSG name-NOM.SG
	moixvečos?				
	gain.s3sG.03sG.0P	Г			
'This	can also be literatur	e right? Th	is can also ga	in the title of a be	estseller?' (Nene

'This can also be literature, right? This can also gain the title of a bestseller?' (Nene Kvinika3e, *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 elim-

ination, 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 a	re not f	eeling	cold, are you?'		'Are y	ou cold?'
(15a)	xom	ar	dagaviçqdeba?	VS.	(15b)	dagavicqdeba?
	AFF	NEG	forget.s2sG.o3sG.FUT			forget.s2sg.o3sg.fut
'You w	von't fo	rget, v	vill you?'		'Will y	vou forget?'
(16a)	xom	ar	geçqineba?	VS.	(16b)	geçqineba?
	AFF	NEG	being offended.s2sG. 03sG.FUT			being offended.s2sG.o3sG.FUT
'You w	von't fe	el offe	nded, will you?'		'Will y	vou 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.NO	OM.SG	AFF	NEG	want.s2sg.o3sg.pres
'Do you <i>urebi</i>)	want anything?' (A	Akaki Geger	nava, <i>Mogz</i>	auris dġi-	\rightarrow offering to bring/buy
VS.					
(17b)	rame		ginda?		
	something.NC	OM.SG	want.s2so	G.O3SG.PRES	5
'Do you	want something?'				\rightarrow yes/no question
• (Question with the	semantics	of doubt:		
(18a)	brma	xom	ar	aris?	
	blind.nom.sg	AFF	NEG	be.s3sg.	PRES
'(S)he is	sn't blind, is (s)he?	' (Niķo Lom	ouri, <i>Þaçid</i>	a megobreb	<i>i</i>) \rightarrow expressing doubt
VS.					
(18b)	brma	aris?			
	blind.nom.sg	be.s3sg.pr	ES		
'Is (s)he	blind?'				\rightarrow yes/no question

• Question with clarification/inquiring semantics:

(18a)	Pikria	xom		ar	ginaxav	s?
	Pikria.nom.sg	AFF		NEG	see.s3sG	.02sg.pres
'You ha <i>Arsena</i>	aven't seen Pikria <i>marabdeli</i>)	, have	you?' (]	Mixeil Š ava	xišvili,	\rightarrow inquiring
VS.						
(18b)	Pikria	ginax	avs?			
	Pikria.NOM.SG	see.s3	sg.o2s	G.PRES		
'Have y	you seen Pikria?'					\rightarrow yes/no question
•	Rhetorical questi	on:				
(19a)	umizezo-d		xom	ar	gaġares	?
	groundless-ADV	.SG	AFF	NEG	expell.s	3pl.o3pl.aor
'They weren't expelled without reason, were they?' (Radio <i>Tavisupleba</i> , 18.02.2004)					?'	\rightarrow rhetorical
VS.						
(19b)	umizezo-d		gaġare	es?		
	groundless-ADV	.SG	expell	.s3pl.o3pl.4	AOR	
'Were t	hey expelled with	nout rea	ason?'			\rightarrow 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.

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
СОР	copula	0	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	1 st /2 nd /3 rd 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/bmd

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