Unbound and Untamed? A Corpus-Based Exploration of Georgian Function Words

Mariam Kamarauli (Hamburg / Frankfurt)
Anastasia Kamarauli (Hagenberg / Batumi)

DOI: https://doi.org/10.62235/dk.4.2025.10521
mariam.kamarauli@uni-hamburg.de/
ma.kamarauli@em.uni-frankfurt.de || ORCID: 0009-0006-0404-4424
a.kamarauli@web.de || ORCID: 0009-0008-2258-8307

Abstract: This paper explores the role and function of Georgian function words from a corpus-linguistic perspective, focusing on their morphosyntactic as well as syntactic and typological qualities. Particular attention is paid to the formal and functional properties of prepositions, postpositions, particles, conjunctions, and other elements and their usage in both the nominal and verbal domain. The study is grounded on evidence from a representative Georgian corpus, the GNC, which facilitates the empirical investigation of syntactic environments and distributional patterns. In addressing the challenge of function word classification in a morphologically complex language such as Georgian, the study confronts several theoretical models of classification of synsemantics and autosemantics. To support this analysis and enable further empirical exploration, a lightweight Java tool has been developed as part of the study. The tool allows users to supply a predefined list of Georgian function words and analyse their occurrence within any given input text. It automatically identifies which function words are present, counts their frequency, and optionally visualises the results. This practical component demonstrates how computational methods can complement theoretical linguistic investigation and highlights the importance of corpus-based, tool-supported methodologies in advancing our understanding of function word systems in typologically complex languages such as Georgian.

Keywords: Functional Grammar; corpus linguistics; computational tools; Georgian language; Java

1. Introduction

Function words (also known as *synsemantic elements*) are words that are largely devoid of independent semantic content, serving only to express grammatical relationships between words in a sentence. They include articles (*the, a*), adpositions (*in, on*), conjunctions (*and, but*), auxiliary verbs (*to be, to have*), modal verbs (*can, must*), and others. Unlike content (or *autosemantic*) words, which carry clear, independent meaning (e.g., nouns and full verbs), function words primarily fulfil grammatical, structural, and interactional roles. They are characterised by a high frequency in discourse and play a strategic role in communication. Coherence and cohesion in discourse are largely supported by function words, which help signal logical relationships between ideas, establish connections between sentences, organise arguments, moderate interaction, and enable speakers and listeners to express and interpret complex ideas effectively. Function words are essential for indicating cause and effect, contrast, conditions, and overall coherence. Beyond structuring grammar, function words are also crucial in conveying tone and first speaker intention. For example, a function word such as a negation particle can completely alter the meaning of a sentence. These words can also soften statements, add emphasis, or qualify meaning. Shifts in meaning are particularly significant in spoken discourse, where intonation and delivery often carry as much interpretive weight as the lexical content. In addition to individual function words, function phrases are frequently used to convey a speaker's stance or attitude (e.g., *it seems, of course, by the way, I*

think). These expressions help speakers e.g. assert their position, express uncertainty or confidence, or transition between topics. In conclusion, function words are far more than mere "fillers" in texts. They are powerful tools for rhetorical strategy and interpersonal communication. Mastery of their use is essential for expressing ideas clearly, accurately, and persuasively.

In Georgian linguistic literature, function words have typically been studied from a purely formal perspective, often neglecting their functional and semantic roles. Compared to autosemantic words, function elements have received limited attention in lexicography; they are underrepresented as independent units in most dictionaries. The only dictionary that includes functional elements more comprehensively is the *Dictionary of Morphemes and Modal Elements of the Georgian Language* by Jorbenadze *et al*., published in 1988. However, this resource no longer meets contemporary needs due to the following limitations:

a) its content was processed manually and lacks systematic organisation
b) it exists solely in printed form and is incompatible with digital research tools
c) it reflects theoretical frameworks that were current only until the early 1990s
d) the functional analysis it offers requires revision and clarification based on more recent theoretical advancements.

2. Theoretical Framework

Linguistic theory has traditionally placed a strong emphasis on the difference between autosemantic (content) and synsemantic (function) words. This contrast, which has historically been presented as a binary opposition or dichotomy, has influenced how we understand morphosyntactic organisation, lexical structure, and grammar in various languages. But as time has gone on, researchers have come to see the relationship as a continuum that is impacted by both the structural characteristics of individual languages and diachronic processes like grammaticalisation.

In the upcoming part, several theories are introduced which concern the distinction and/or entanglement of autosemantics and synsemantics. It offers a comparative overview and introduces analytical tools for modelling the continuum between lexical and grammatical elements, drawing on ideas from structuralism, generative grammar, functional-typological models, cognitive-constructional and pragmatic approaches.

2.1 Foundations in Early Linguistic Thought

The roots of this distinction can be traced back to the early 20th century, when linguistic and philosophical grammar (as introduced by Wittgenstein) began to take shape. Otto Jespersen was among the pioneers to define this concept, making a clear distinction between "notional words", which carry their own meaning, and "form words", which mainly serve to express grammatical relationships. Jespersen pointed out that content words stand on their own in terms of meaning, while function words depend heavily on their syntactic context for understanding.[1] In 1934, Karl Bühler, in his significant work *Sprachtheorie*, proposed the so-called "organon model" that connected autosemantic words to a "representation function" and synsemantic

[1] Jespersen 1924: 73–75.

words to an "expressive function" and a "conative function". According to Bühler, function words play a crucial role in organising discourse, rather than just acting as syntactic fillers.[2]

2.2 Structuralist and Distributional Approaches

Within American structuralism, Leonard Bloomfield (1933) placed significant emphasis on the distributional behaviour of words. He categorised "full words" – typically nouns, verbs, and adjectives – as autosemantic due to their ability to stand alone and contribute referential meaning. In contrast, "function words" were seen as dependent items that appeared in limited syntactic slots.[3] Zellig Harris extended this distributional approach by proposing formal methods to categorise words based on their positional behaviour and frequency within corpora. For Harris, function words are characterised by high frequency, syntactic dependency, and constrained positional freedom.[4] In the European dependency tradition, Lucien Tesnière's *Éléments de syntaxe structurale* likewise opposed "mots pleins" and "mots vides":[5] in his stemma diagrams, content words form the nuclei of constructions, while function words serve as relational connectors, anticipating later structural and functional treatments of the autosemantic–synsemantic divide.

The core distinctions between autosemantic and synsemantic words can be summarised as illustrated in Table I.

Table I: Core distinctions between autosemantic and synsemantic words

Property	Autosemantic Words	Synsemantic Words
Semantic Autonomy	High	Low
Grammatical Function	Minimal	Central
Distributional Flexibility	Broad	Restricted
Phonological Independence	Often independent	Often clitic or bound
Frequency	Typically lower	Generally higher

2.3 Generative Grammar and Formal Syntactic Categories

The generative grammar framework brought a fresh, more abstract way of looking at how we categorise words. In his books, *Syntactic Structures* (1957) and *Aspects of the Theory of Syntax* (1965), Noam Chomsky made a clear distinction between lexical and functional categories. Lexical items, which are usually autosemantic, are kept in the lexicon and carry semantic meaning (e.g. *run*, *house*, *child*). On the other hand, functional elements (e.g. *the*, *will*, *of*) act as the structural heads of phrases, such as Determiner Phrases (DPs), Tense Phrases (TP), and Complementiser Phrases (CPs).[6] This distinction became even more significant in *The Minimalist Program* (1995), where the syntactic spine is often made up entirely of functional projections. Components like T (Tense), C (Complementiser), and D (Determiner) illustrate synsemantic elements that, whereas they may not contribute much to meaning, play a vital role in the process of syntactic derivation.[7]

[2] Bühler 1934: 28–34.
[3] Bloomfield 1933: 178–180.
[4] Harris 1951: 122–126.
[5] Tesnière 1959: 53–55.
[6] Chomsky 1965: 68–77.
[7] Chomsky 1995: 177–184.

2.4 Functional Grammar and Diachronic Change

Functionalist approaches offer a new perspective, highlighting the communicative and historical aspects of word categories. Simon Dik (1978, 1997) introduced the autosemantic-synsemantic division within a larger functional grammar framework. For Dik, content words are all about encoding new, referential information, while function words help organise that information within discourse.[8] Talmy Givón (1979, 1984) took this concept even further by emphasising grammaticalisation as the main process that drives the transformation of autosemantic items into synsemantic ones. His famous saying, "Today's morphology is yesterday's syntax" captures the essence of how grammatical markers often evolve from complete lexical items over time.[9]

Syntactic reanalysis,[10] semantic bleaching,[11] and phonological reduction[12] have been pinpointed as crucial mechanisms that turn autosemantic roots into grammatical elements. Building on these earlier insights, Bernd Heine and his colleagues explored this historical development in their research on grammaticalisation pathways.[13]

Example:

Latin *habere* ("to have") → French future tense auxiliary *-ai* in *chanterai* "(I) will sing"

2.5 Typological and Cognitive-Constructional Models

From a typological perspective, Martin Haspelmath (2000, 2011) suggests that we should view the autosemantic-synsemantic contrast as a scale rather than a strictly binary opposition. He presents a variety of diagnostic criteria like obligatoriness, semantic generality, and phonological integration to determine where a particular item fits in the lexical-grammatical spectrum.[14] In cognitive linguistics, Ronald Langacker (1987, 2008) also breaks away from rigid categorisations. He analyses all linguistic expressions as meaningful, even the most grammaticalised elements, though these are regarded in more schematic and abstract ways.[15]

Adele Goldberg's *Construction Grammar* (1995, 2006) provides yet another viewpoint. In this framework, meaning doesn't just lie in individual words but in constructions, i.e. combinations of form and function. Function words are essential within these constructions, as they help shape argument patterns and discourse routines.[16] Fig. 1 illustrates the lexical-grammatical continuum in English, showing how items range from fully lexical words like *run* to highly grammatical elements such as the plural suffix *-s*.

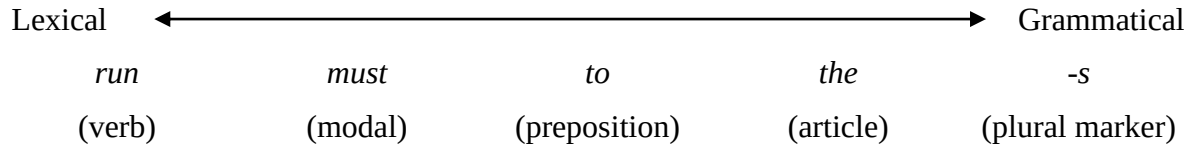


Fig. 1: Continuum Representation

[8] Dik 1997: 120–124.
[9] Givón 1971: 413.
[10] E.g. Langacker 1977: 59.
[11] E.g. Givón 1981: 51.
[12] E.g. Bybee & Pagliuca 1985: 76.
[13] Heine *et al.* 1991: 17–36.
[14] Haspelmath 2011: 41–44.
[15] Langacker 1987: 58–64.
[16] Goldberg 1995: 9–12.

2.6 Gradient Classification Models

The idea of a lexical-grammatical continuum has inspired some researchers to suggest classifications that go beyond a simple two-way split, proposing instead three-way or even multi-dimensional frameworks. A notable model comes from Haspelmath (2011), who advocates for a graded typology, while Lehmann (1982) introduces scalar parameters in the process of grammaticalisation. These frameworks often identify an "intermediate" category that encompasses auxiliaries, modals, or aspectual markers – elements that sit somewhere between being fully lexical and fully grammatical. A summary is provided in Table II.

Table II: Expanded Diagnostic Table (based on Lehmann 1982; Haspelmath 2011)

Property	Autosemantic	Intermediate	Synsemantic
Semantic specificity	high	moderate	low
Syntactic obligatoriness	optional	variable	required
Phonological independence	full	partial	reduced/clitic
Frequency in discourse	low–moderate	moderate–high	high
Position in clause	flexible	mid–clause	fixed/pre/postposed
Diachronic stability	high	moderate	low

2.7 Pragmatic Approaches to Function Words

Beyond their grammatical behaviour, function words are also essential for structuring discourse and directing interpretation in context. Pragmatically, they can be used as discourse markers, modal particles, and focus or topic indicators. Deborah Schiffrin (1987) suggested the concept of discourse markers as items that structure spoken language, marking coherence and speaker intention. Words such as *well, so, you know*, and *but* serve not grammatical but interpersonal and organisational functions in conversation.[17] Similarly, Fraser (1999) categorises discourse markers as lexical items used to signal a relationship between the discourse segment they precede and the prior discourse. These words do not contribute propositional meaning but are crucial to the pragmatic interpretation of speech.[18]

In Germanic languages, modal particles like *doch, ja,* and *mal* convey speaker attitude or epistemic certainty. Diewald (2006) views these as grammatical elements with pragmatic functions, specifically in dialogic contexts.[19] Topic and focus markers are likewise pragmatic function words. Lambrecht (1994) describes how topic-comment structures in languages like Hungarian or Japanese are encoded by way of particles like *wa* or *ga*.[20]

Searle in his theory of *Speech Acts* (1969) highlights the function of words like *please, let's,* and modal auxiliaries to create performative functions. These words are very important in projecting illocutionary force, enabling utterances to have commanding, requesting, or declarative power.[21] Levinson (1983) extends this within pragmatics, noting that function words often communicate the speaker's implicatures of mutual knowledge and discourse relevance. For example, the use of focus-sensitive particles such as *even* or *only* demonstrates

[17] Schiffrin 1987: 31–40.
[18] Fraser 1999: 931.
[19] Diewald 2006: 407–410.
[20] Lambrecht 1994: 117–124.
[21] Searle 1969: 63–70.

pragmatic scope and presuppositional structure.[22] Table III presents cross-linguistic examples of major types of pragmatic function markers – discourse markers, modal particles, focus/topic markers, and speech act markers – along with illustrative items, together with the languages in which they occur and key references.

Table III: Pragmatic Functions of Function Words

Function Type	Examples	Languages	Key References
Discourse Markers	*well, so, anyway*	English, Spanish	Schiffrin (1987), Fraser (1999)
Modal Particles	*doch, ja, mal*	German	Diewald (2006), Abraham (1991)
Focus/Topic Markers	*wa, ga, mo*	Japanese, Hungarian	Lambrecht (1994), Givón (1983)
Speech Act Markers	*please, let's, sorry*	English, Korean	Searle (1969), Levinson (1983)

According to these pragmatic approaches, functional elements ought to be categorised according to their function in discourse and communicative intent as well as their involvement in syntax or morphology. By including a dimension that represents speaker-hearer interaction, they enhance the conventional autosemantic-synsemantic split.

3. Function words in Georgian

In Georgian, syntactic and pragmatic functions can manifest themselves in function words or function constructions. The latter consist of a function word combined either with other function words or with different elements. Function words are quite versatile: depending on the element they determine, their function changes. A good example is the grading adjective *upro* 'more', which can be used to construct the 1st and 3rd stage comparatives in the analytic gradation of descriptive adjectives.[23] Examples (1–3) show the different use, function and possible combination of *upro* with other elements.

(1) *Leo-s* *ak* ***upro*** ***ʒvel-i*** *ʒmaḳac-eb-i* *hqavs*
Leo-DAT.SG here **more** **old-NOM.SG** male friend-PL-NOM have.S3SG.PRES
'Leo has **older** (male) friends here […]' (Revaz Mišveladze, *Rčeuli txzulebani IV - novelebi*)

(2) *brʒol-is* *survil-i* *aġaravis* *aġmoačnda*
fight-GEN.SG wish-NOM.SG no one.FOC.GEN.SG discover.S3SG.AOR

upro ***imiṭom*** *rom* *brʒola* *uazroba* *iqo*
more **because** that fight.NOM.SG meaninglessness.NOM.SG be.S3SG.AOR
'No one wanted to fight anymore, **mostly because** fighting was pointless.' (Journal *Axali taoba*, 2000)

(3) *ar* *šeiʒleba* *gačereba* ***mit*** ***upro*** *axla*
NEG be possible.S3SG.PRES stop.INF **that.INST.SG** **more** now
'It must not stop, **especially** now.' (Journal *Axali epoka*, 2003)

[22] Levinson 1983: 204–211.

[23] Georgian can differentiate between three levels of comparatives: 1st level comparatives are built with the adverb *upro* 'more' (e.g. *upro lamazi* 'more beautiful'), 2nd level comparatives are constructed with *bevrad* 'much (more)' (e.g. *bevrad lamazi* 'much more beautiful'), and 3rd level comparatives with both *bevrad* 'much (more)' and *upro* 'more' (e.g. *bevrad upro lamazi* 'by far more beautiful'; Kamarauli 2022: 113.

In example (1), the adverb *upro* ‘more’ is paired with a lexical adjective *ʒveli* ‘old’ and functions as a comparative grading adverb (‘friends older than others’); in (2), *upro* is combined with another function word, namely the causal subordinator *imiṭom* ‘because’, which is intensified through this combination and triggers the following argumentative structure; and lastly, in (3), *upro* is part of a lexicalised phrase, which has focusing function (*mit upro* ‘especially’).

Another good example is *ra* ‘what’, which is quite diverse in its meaning and function; examples (4–6) showcase *ra* ‘what’ in combination with different auto- and synsemantics.

(4)	***es***	***ra***	*gaaḳete*	*rom*	*icode*
	this.NOM.SG	**what.NOM.SG**	do.S2SG.AOR	that	know.S2SG.CONJ

‘[If] you [only] knew what you did with this.’ (Revaz Mišveladze, *Rčeuli txzulebani IV - novelebi*)

(5)	***ḳarg-i***	***ra***	*gexveċebi*	*sxva*	*rame-ze*
	good-NOM.SG	**what.NOM.SG**	beg.S1SG.PRES	other	something.DAT.SG-ON

	vilaṗaraḳot
	speak.S2PL.OPT

‘**Okay**, I beg you… let’s talk about something else.’ (Revaz Mišveladze, *Rčeuli txzulebani I - novelebi*)

(6)	***ra***	***tkma***	***unda***	*cud-ad*	*iq̇o*
	what.NOM.SG	**say.INF**	**MOD**	bad-ADV.SG	be.S3SG.AOR

‘Of course, he was feeling unwell.’ (Revaz Mišveladze, *Rčeuli txzulebani I - novelebi*)

In (4), *ra* has a referencing function: it refers to an action prior to the utterance and the speaker evaluates the action of the hearer. In (5), *ra* is paired with the adjective *ḳargi* ‘good’ and has a convincing, admitting function; lastly, in (6), *ra* is part of the grammaticalised function phrase *ra tkma unda* ‘of course’ (lit. ‘what talk does it need’), today written as one word; syntactically, it functions as a clausal adverb (modifying the whole clause) and adopts the meaning of an obvious conclusion (presupposing the previous expectation/knowledge of the speaker about the state of the referred person).

Another illustration of the multifunctionality of function words – and thus the need for a multi-layered approach – is provided by *erti* ‘one’; cf. examples (7–10).

(7)	*[…]*	*ševedi*	*da*	***ert-i***	***cal-i***	*viq̇ide*
	[…]	go in.S1SG.AOR	and	**one-NOM.SG**	**piece-NOM.SG**	buy.S1SG.AOR

	arada	*saxl-ši*	*uḳve*	*oc-amde*	*mkonda*
	even though	house.DAT.SG-in	already	twenty-until	have.S1SG.IMPF

‘[…] I went in and bought **one piece**, even though I already had about twenty at home.’ (Journal *11x11*, 2010)

(8)	***ert***	***dġe-s***	*movedi*	*saxl-ši*	*da*
	one.DAT.SG	**day-DAT.SG**	come.S1SG.AOR	house.DAT.SG-in	and

	iaṭaḳ-ze	*goraobda*	*tiḳa-s-tan*	*ertad*
	floor.DAT.SG-on	roll around.S1SG.IMPF	Tika-DAT.SG-with	together

‘**One day** I came home and he was rolling around on the floor with Tika.’ (Journal *Axali ṭaoba*, 2006)

(9)	***ǯer erti***	*uxerxul-i*	*ikneboda*	*morcxv-ad*
	first of all	awkward-NOM.SG	be.S3SG.COND	shy-ADV.SG
	stkva	*Andriḳo-m*	*[…]*	
	say.S3SG.AOR	Andriko-ERG.SG	[…]	

‘“**First of all**, it would be awkward” Andriko said shily […]’ (Guram Dočanašvili, *Čvens ezoši ċvima modis*)

(10)	***ert-i***	***es-e-c***	*mitxari*	*bazar-ši*	*rogor*
	one-NOM.SG	**this-EMPH.V-FOC**	say.S2SG.IMP	bazaar.DAT.SG-in	how
	moxvdi				
	turn up.S2SG.AOR				

‘**Now** tell me this, how did you turn up at the bazaar?’ (Journal *Sakartvelos respubliḳa*, 2005)

In (7), *erti* functions as a numeral, and together with the numeral classifier *cali* ‘piece’, the phrase denotes a definite quantity. The opposite happens in example (8), where *ert* is paired with *dġes* ‘day’, which entails the meaning of ‘one day’ and functions as an unspecific and indefinite temporal phrase. In contrast, examples (9) and (10) demonstrate more multi-layered functions of *erti*: in (9), together with *ǯer*, *erti* triggers an argumentative structure and introduces a listing (*first of all X and secondly, Y*), which carries a focusing function, whereas in (10), *erti* can be considered to have an adhortative function, changing the topic and intensifying the focus given by *esec* ‘this’.

All these examples demonstrate the urgency of introducing a multi-layered approach that includes not only syntax but also semantics and pragmatics. This will be discussed in the following Chapter.

4. Analysis

4.1 Linguistic approach

For the present paper, the linguistic approach includes several subfields of linguistics:

- syntax, in particular syntactic roles, e.g. what grammatical function does the word fulfil (connector, modifier, etc.)?
- semantics, in particular semantic autonomy, e.g. does the word carry standalone meaning, or is it dependent?
- pragmatics, in particular pragmatic function, e.g. does it manage discourse, express stance, or organise information?

For this analysis, we have chosen 100 of the most frequent function words found in the Georgian National Corpus (hereafter: GNC),[24] more precisely in the subcorpus of Modern Georgian (GNC-NG). The following Tables are a first attempt at classifying and explaining function words according to their syntactic (Table IV), pragmatic (Table V), and semantic (Table VI) functions. In Table IV, the roles and grammatical functions of these 100 most frequent function words are given.

[24] http://gnc.gov.ge/. This and all other URLs quoted in this article were last accessed on 30 December 2025.

Table IV: Classification of the 100 most frequent function words in GNC-NG			
Word	**Translation**	**Role**	**Grammatical function**
da	and	connector	coordinating conjunction
ar	not	negation	negative adverb
rom	that, if	subordinator	complementiser/subordinating conjunction
magram	but	connector	coordinating conjunction
tu	if	conditional/subordinator	conjunction
ḳi	yes, well	contrastive/affirmative	affirmative word
ar	not	negation	negative adverb
rom	that, if	subordinator	complementiser/subordinating conjunction
magram	but	connector	coordinating conjunction
tu	if	conditional/subordinator	conjunction
ḳi	yes, well	contrastive/affirmative	affirmative word
ra	what	interrogative	pronoun/ WH-word
unda	must	modal verb	auxiliary
ara	no	negation	negative word
erti	one	quantifier/numeral	indefinite numeral
ver	not (potential)	inability marker	negative auxiliary
arc	not even	negative coordination	negative conjunction
axla	now	time adverb	temporal adverb
mere	then, after	time adverb	temporal adverb
mainc	however	concessive marker	particle
ase	this way	manner adverb	modal adverb
ise	that way	manner adverb	modal adverb
rogorc	as	comparison/subordination	comparative conjunction
xom	after all, well	question tag/emphasis	particle
upro	more	comparative degree	adverb/ degree modifier
ḳidev	again	additive/focus	focus particle/adverb
rac	what	relative pronoun	WH-word/relativiser
tavi	head	reflexive noun	grammaticalised noun
isev	as before	repetition/focus	adverb
ḳaci	man (general subject, expletive)	generic subject	grammaticalised noun
aġar	not anymore	temporal/negative	negative particle
žer	first	temporal adverb	temporal adverb
rogor	how	interrogative adverb	WH-word (manner)
roca	when	temporal subordinator	subordinating conjunction
titkos	as if	hypothetical/evidential/ modal	modal particle
ras	what	interrogative object	WH-pronoun
mxolod	only	focus marker	focus particle
an	or	alternative connector	coordinating conjunction
šemdeg	after	time adverb/postposition	temporal adverb
uḳve	already	perfectivity marker	aspectual adverb
tavs	head	reflexive form	grammaticalised noun
ak	here	locative adverb	spatial adverb
nu	not (prohibitive)	prohibitive particle	negation/imperative particle
mašin	then	temporal adverb	temporal adverb
mašinve	instantly	temporal adverb	temporal adverb
sul	always	emphasis/frequency	intensifier/adverb
radgan	because	causal subordinator	conjunction
marṭo	alone	focus/quantification	adverb/focus marker
ċin	before	direction/postposition	adverb/postposition
ert	one	indefinite numeral	quantifier
xolme	sometimes	habitual marker	aspectual particle

Table IV: Classification of the 100 most frequent function words in GNC-NG			
Word	**Translation**	**Role**	**Grammatical function**
aba	well then	turn-taking/irony marker	discourse particle
albat	probably	epistemic modality	modal particle
coṭa	few	quantity	quantifier
raṭom	why	interrogative	reason WH-word
meṭi	more	comparative degree	quantifier
ʒalian	very	adverb	degree modifier
martla	truly	particle/adverb	emphatic
ertxel	once	adverb	temporal adverb
sad	where	interrogative	WH-locative
ik	there	adverb	locative adverb
vidre	than	conjunction	comparative subordinator
tviton	self	pronoun/focus	reflexive/emphatic
sċored	truly	focus marker	emphatic particle
ertad	together	adverb	manner/coordination
tumca	but	conjunction	adversative subordinator
ḳargad	well	adverb	manner adverb
ai	after all, well	particle	demonstrative/emphatic
uḳan	back	adverb	locative/directional
šina	in	postposition/locative	locative adverb
romelic	which	relative pronoun	WH-word
saertod	generally	adverb	scope/generalisation
imiṭom	because	subordinator (causal)	subordinating conjunction
xolo	but	conjunction	contrastive
iseti	that kind of	pronoun/adjective	descriptive/demonstrative degree
aseti	this kind of	pronoun/adjective	descriptive/demonstrative degree
raġac	anything	pronoun/indefinite	thing/something
ḳidec	yet again	particle	additive particle
sanam	until	conjunction	temporal subordinator
uceb	suddenly	adverb	temporal/manner
xan	sometimes	particle/temporal	iteration
rame	something	indefinite pronoun	something
ram	something	indefinite pronoun	variant of above
ġvelas	all	pronoun/quantifier	universal
verc	not even (potential)	negative auxiliary	verb-related negation
veġar	not anymore	negative auxiliary	inability marker
ġovel	every	quantifier	universal
torem	or else	conjunction	conditional/contrastive
šoris	between	postposition	locative (between)
ertmanets	each other	pronoun	reciprocal
vitom	as if	particle	hypothetical
bevri	much	quantifier	lexical
ikneb	maybe	modal particle	possibility
aravin	no one	pronoun	indefinite negative
xans	time	noun (temporal use)	temporal
tan	at the same time	particle/adverb	accompaniment
saḳutari	own	adjective/pronoun	reflexive possessive
ḳai	good, okay	adjective (colloquial)	description of quality
ġvelaze	most	quantifier (superlative)	degree
ġoveli	every	quantifier	universal
gamo	because of	postposition/causal marker	causal adverb
turme	apparently	modal particle	evidential
martalia	it is true	concessive marker	modal/contrastive conjunction
romelsac	which	relative pronoun	WH-relative
sadac	where	relative pronoun	WH-locative

The analysis of syntactic functions and roles maps structural dependency. Function words cluster around clausal structure, e.g.:

- conjunctions (*da* 'and', *rom* 'that', *radgan* 'because', *tumca* 'but') mark syntactic linking
- sentence particles (*ḳi* 'yes', *aba* 'well') operate at sentence or discourse level, often outside argument structure
- adverbs and pronouns (*ak* 'here', *ik* 'there', *ase* 'in this way', *romelic* 'which') serve as intermediate links – they connect content to structure.

Syntactic functions constitute the formal backbone of Georgian syntax. Function words are crucial scaffolding elements that carry syntactic but no propositional meaning – they structure grammar rather than content.

As for the pragmatic aspect of function words, several aspects need to be considered:

- discourse markers: do these words organise a turn or indicate how the discourse is structured (e.g., *aba* 'well', *xom* 'after all')?
- modal particles; do these words express speaker stance or attitude (e.g., *albat* 'maybe', *titkos* 'as if')?
- focus/emphasis markers: do these words highlight or limit scope (e.g., *mxolod* 'only', *ḳidev* 'again')?
- topic/frame markers: do these words set up contrasts or frame shifts (e.g., *ise* 'like that', *aba* 'well then')?
- illocutionary markers: do these words indicate a speech act type (e.g., *nu* 'well', *ḳi* 'yes')?
- rhetorical markers: are these words used in argumentation, irony, questioning (e.g., *raṭom* 'why', *ḳaci* 'one' (generic subject))?

Not all 100 words have pragmatic functions. The 23 that do are explained in Table V according to their pragmatic function.

For the analysis of the pragmatic function of some function words, the implementation of interpersonal and discourse layers is needed. Words like *xom, ḳi, aba, tumca, albat* show that many function words serve pragmatic rather than purely grammatical purposes, such as guiding the listener's interpretation:

- modal particles: *albat* 'probably', *turme* 'apparently' (speaker stance)
- discourse markers: *xom, aba, ai* (interactional control)
- focus particles: *mxolod* 'only', *scored* 'truly', *ḳidec* 'yet again' (information structure).

This confirms that *functionality* in language is not purely syntactic – it can extend into discourse management and intersubjective meaning, implying that function words may be procedural rather than conceptual.[25]

Lastly, semantics needs to be included to the analysis of function words, to be more precise, the level of their semantic autonomy:

- high (autosemantic): the word has a referential or lexical meaning; it is interpretable in isolation

[25] Cf. Blakemore 1987: 75.

Table V: Classification of the 23 function words with pragmatic function			
Word	**Translation**	**Pragmatic function**	**Notes**
ḳi	yes, well	illocutionary/emphatic/ contrastive	used for contrast, affirmation
xolme	sometimes	aspectual/framing	indicates habitual action
aba	well then	discourse marker/framing	turn-taking, irony, emphasis
mainc	however	concessive marker	implies contrast or unexpectedness
mxolod	only	focus marker	restricts the scope of assertion
albat	maybe	modal particle	expresses epistemic uncertainty/probability
coṭa	few	quantitative emphasis	often mitigates or softens assertions
raṭom	why	rhetorical/interrogative	signals justification or challenges
rogor	how	interrogative (pragmatic)	also used rhetorically, not just for inquiry
ar	not	illocutionary marker	negates propositions, can mark prohibitive tone
titkos	as if	modal particle	used in hedging, hypothetical framing
nu	not (prohibitive)	illocutionary/directive	used in prohibitions, soft commands
xom	after all, well	discourse/tag particle	used to confirm shared knowledge or expectation
ai	after all, well	discourse marker	introduces examples or emphasis
martla	truly	emphatic marker	speaker stance
ikneb	maybe	modal particle	possibility
tumca	but	concessive marker/ discourse-level adversative	often rhetorical, same as ‘but’ in argumentation
imiṭom	because	rhetorical/causal	explains cause
vitom	as if	hypothetical/ironic	hedging function
torem	or else	rhetorical connector	expresses warning or contrast
sc̈ored	truly	focus particle	highlights specific constituent
ḳidec	yet again	additive particle	reinforces previous constituent
sanam	until	temporal discourse marker	frames time of main action

- medium (intermediate): the word has a limited standalone meaning; it is sometimes interpretable without context
- low (synsemantic): the word lacks a standalone meaning; it functions only in relation to other elements.

Table VI shows the 100 most frequent function words, categorised according to their semantic autonomy (from low to high) and relevant notes.

The semantic autonomy criterion thus reveals a continuum, not a binary opposition. Many words in Georgian do not fit cleanly into the categories *autosemantic* (content) or *synsemantic* (function), instead, they form a gradient:

- high autonomy: lexical or quasi-lexical items (*erti* ‘one’, *bevri* ‘many’, *saḳutari* ‘own’, *ḳargad* ‘well’)
- medium autonomy: adverbs and pronouns (*ase* ‘in this way’, *ise* ‘in that way’, *ik* ‘there’, *isev* ‘again’)
- low autonomy: particles, conjunctions, and negators (*da* ‘and’, *ar* ‘not’, *ḳi* ‘yes’, *tu* ‘if’, *radgan* ‘because’, *tumca* ‘but’).

The semantic continuum which Georgian function words show supports the theories by Haspelmath (2011) and Lehmann (1982): grammatical and lexical elements form a scalar hierarchy, not a dichotomy.

Table VI: Classification of the function words according to their semantic autonomy			
Word	**Translation**	**Autonomy**	**Notes**
da	and	low	has no meaning without linking two elements
ar	not	low	pure negation, context-bound
rom	that, if	low	grammatical subordinator
magram	but	low	logical connector, not lexical
tu	if	low	conditional/focus, highly context-dependent
unda	must	low	modal auxiliary without independent meaning
ara	no	low	pure negation
ver	not (potential)	low	grammaticalised inability marker
arc	not even	low	coordinated negation
rogorc	as	low	comparative marker, dependent on clause
xom	after all, well	low	discourse particle, context-driven
aġar	not anymore	low	composite negation and aspect marker
roca	when	low	subordinator
an	or	low	logical disjunction, purely structural
nu	not (prohibitive)	low	directive/prohibitive, lacks standalone meaning
radgan	because	low	subordinator, non-lexical
aba	well then	low	discourse-only use
vidre	than	low	subordinator; purely comparative in function
tumca	but	low	adversative conjunction; has little lexical content
ai	after all, well	low	emphatic/discourse function; no lexical reference
šina	in	low	postpositional; cannot appear in isolation
xolo	but	low	logical connector; no semantic autonomy
ḳidec	yet again	low	additive/focus particle; not interpretable alone
sanam	until	low	subordinator; only meaningful with full clause
verc	not even (potential)	low	negative clitic; syntactically and semantically dependent
veġar	not anymore	low	composite negation + aspect; non-autonomous
torem	or else	low	discourse connective; only meaningful in clause structure
šoris	between	low	postposition; semantically empty without complement
vitom	as if	low	hypothetical/discourse use; no stable referent
ikneb	maybe	low	modal particle; epistemic, speaker-oriented
gamo	because of	low	postpositional causal; semantically empty alone
turme	apparently	low	evidential particle; relies entirely on speaker stance
ḳi	yes, well	low–medium	affirmative or contrastive; may carry stance
mainc	however	low–medium	pragmatic concession, vague semantics
upro	more	low–medium	comparative degree, no standalone referent
titkos	as if	low–medium	modal/hypothetical frame, no concrete referent
mxolod	only	low–medium	focus marker, vague semantics
uḳve	already	low–medium	perfectivity marker, aspectual nuance
xolme	sometimes	low–medium	aspectual particle, pragmatically loaded
albat	probably	low–medium	modal, speaker-oriented; interpretable in vague sense
martalia	it is true	low–medium	fixed concessive form; modal-discourse with partial meaning
ra	what	medium	WH-word with referential potential
axla	now	medium	temporal adverb, somewhat interpretable alone
mere	then, after	medium	time-related, needs discourse anchor
ase	this way	medium	modal adverb, deictic, partially interpretable
ise	that way	medium	similar to *ase*
ḳidev	again	medium	additive, context-enhanced meaning
rac	what	medium	relative pronoun, semantically active
isev	as before	medium	temporal iteration, moderately autonomous
žer	first	medium	temporal nuance, vague alone
rogor	how	medium	WH-adverb, interpretable in questions
ras	what	medium	interrogative pronoun, referential
šemdeg	after	medium	adverbial/postpositional, partially lexical
ak	here	medium	spatial deictic, interpretable alone

Table VI: Classification of the function words according to their semantic autonomy			
Word	**Translation**	**Autonomy**	**Notes**
mašin	then	medium	temporal reference, discourse-anchored
mašinve	instantly	medium	temporal adverb; semantically specific in discourse
sul	always	medium	adverbial, quantifying, vague stand-alone
ċin	before	medium	adverb/postposition, spatial reference
martla	truly	medium	emphatic stance marker; vague without context
coṭa	few	medium	quantifier, meaning is scalar
raṭom	why	medium	interrogative, semantically oriented
meṭi	more	medium	quantifier, relational but partly referential
sad	where	medium	WH-word with referential potential
ik	there	medium	deictic; interpretable but needs discourse anchor
tviton	self	medium	reflexive pronoun; requires antecedent
sċored	truly	medium	focus marker; semantically weak but locatable
ertad	together	medium	manner adverb; dependent but partly interpretable
uḳan	back	medium	spatial adverb; interpretable with spatial context
romelic	which	medium	WH-relative; needs antecedent for full interpretation
sadac	where	medium	WH-locative, needs antecedent for full interpretation
saertod	generally	medium	generalising adverb; vague alone, clear in context
imiṭom	because	medium	causal phrase; compositional meaning with *imis gamo, rom* ‘because of this’
iseti	that kind of	medium	degree expression; requires comparative reference
aseti	this kind of	medium	demonstrative; needs a referent to specify
raġac	anything	medium	indefinite pronoun; referential but vague
xan	sometimes	medium	temporal/discourse use; vague and context-sensitive
rame	something	medium	indefinite pronoun; weak referential value
ram	something	medium	variant of rame; also vague but referential
ġvelas	all	medium	quantifier/pronoun; requires context for scope
ġovel	every	medium	quantifier; needs noun to specify scope
ertmanets	each other	medium	reciprocal pronoun; contextually anchored
aravin	no one	medium	negative pronoun; referential but polarity-bound
xans	time	medium	noun of time; vague without construction
tan	at the same time	medium	focus/discourse marker; context-dependent
ġvelaze	most	medium	superlative adverb; dependent on comparative frame
ġoveli	every	medium	quantifier; general scope without specific referent
romelsac	which	medium	relative pronoun; dependent on antecedent
marṭo	alone	medium–high	adverb/quantifier, semantically rich
erti	one	high	lexical numeral
tavi	head	high	lexical noun, even when grammaticalised
ḳaci	man (general subject, expletive)	high	lexical noun, semantically full
tavs	head	high	lexical noun inflected
ert	one	high	numeral, lexical
ʒalian	very	high	lexical adverb; expresses intensity independently
ertxel	once	high	temporal adverb; specific lexical meaning (“once”)
ḳargad	well	high	lexical adverb (manner); semantically rich
uceb	suddenly	high	temporal/manner adverb; interpretable in isolation
bevri	much	high	quantifier/lexical; has referential content
saḳutari	own	high	possessive adjective; strong lexical meaning
ḳai	good, okay	high	adjective (colloquial); referential

How the three introduced dimensions (semantic autonomy, syntactic role, and pragmatic function) interact, can be summarised as shown in Table VII and visualised as in Fig. 2.

Table VII: Interaction of the three dimensions

Dimension	Level of Description	Role in Functional System
Semantic autonomy	lexical–grammatical	degree of meaning dependency
Syntactic role	structural	position and combinatorial function
Pragmatic function	communicative	interpretation and discourse management

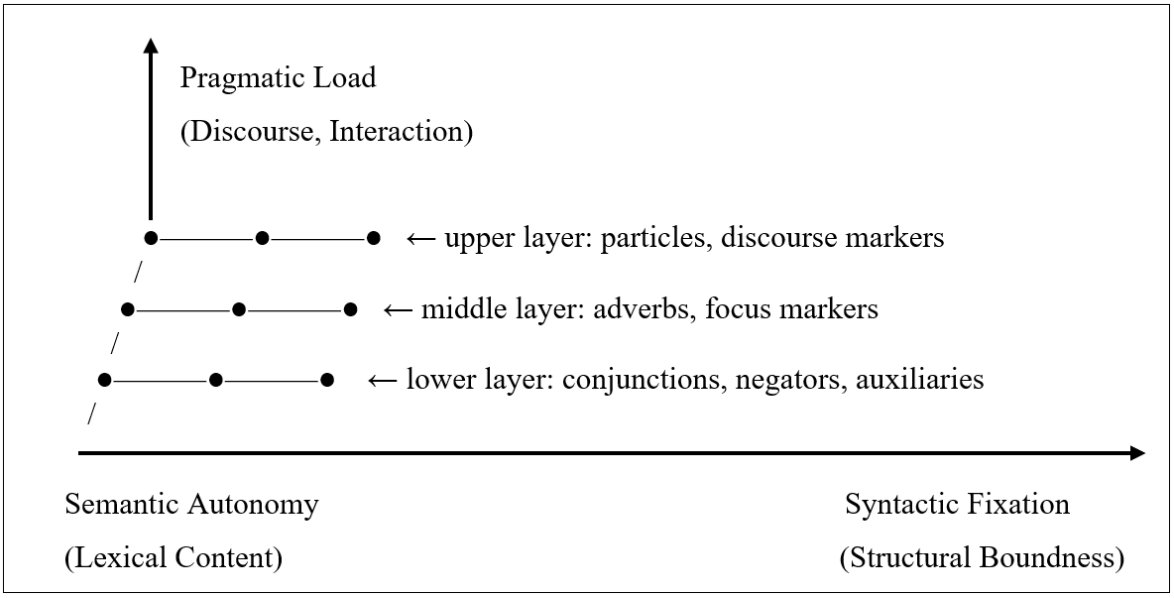


Fig. 2: The three-dimensional model of Georgian function words

The model shown in Fig. 2 visualises the interplay between semantic autonomy, syntactic fixation, and pragmatic load as a dynamic continuum rather than a categorical split. Function words are distributed within a conceptual space in which semantic autonomy decreases as syntactic fixation increases, while pragmatic load rises orthogonally, reflecting discourse-level functions. Elements such as *da* ('and') and *ar* ('not') cluster in the grammatical core, characterised by low semantic autonomy and high syntactic dependency. By contrast, discourse particles like *ḳi* 'well, yes', *xom* 'after all', and *albat* 'probably' occupy the upper pragmatic layer, where speaker stance and interactional meaning dominate. Adverbs and focus markers, including *ḳargad, upro*, and *marṭo*, lie between these poles, mediating between lexical content and structural function. The model thus captures the continuum nature of Georgian function words as multi-dimensional operators balancing meaning, structure, and discourse.

4.2 Computational Approach: Functional Elements Analyser

4.2.1 Development and application of an analysis tool

To automate the identification of functional elements in Georgian and to visualise the results in an intuitive format, an analysis tool was developed. The software is implemented as a lean, standalone Java program with a clear separation of data management (I/O), logical processing, and presentation.

4.2.2 System architecture and implementation

The system follows a classic three-tier application pattern adapted to a desktop tool (Fig. 3):

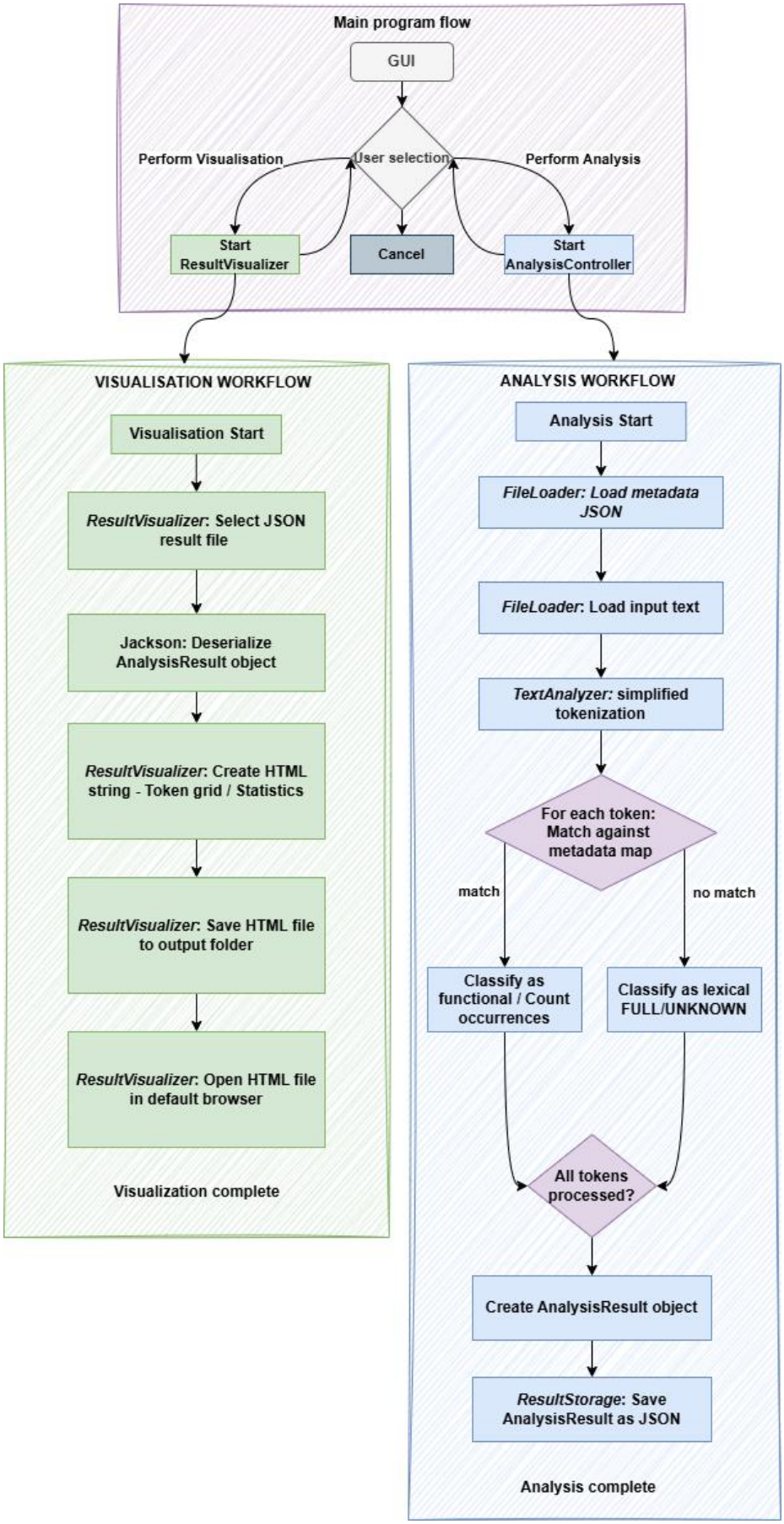


Fig. 3: Program flow

1) **data access (I/O)**

FileLoader reads input texts and function-word metadata. *ResultStorage* serialises analysis results in machine-readable JSON to ensure reusability and long-term archiving.

2) **analysis logic**

Implemented in *TextAnalyzer* and *AnalysisController*. This layer performs tokenisation, comparison against the metadata base, and statistical evaluation.

3) **presentation/visualisation**

User interaction is handled via simple Swing[26] dialogs. *ResultVisualizer* generates an interactive, platform-independent HTML file that opens automatically in the default browser.

Java provides high portability and robust error handling. The Jackson library[27] is used for efficient JSON serialisation and deserialisation.

The structure of the Functional Elements Analyser and the relationship between the individual classes can be seen in the UML diagram[28] (Fig. 4).

4.2.3 The metadata base (function-word corpus)

The analysis relies on a function-word metadata base that translates the theoretical classification into a processable data model. Each function word is represented by the *FunctionWord* data model with the following attributes:

- **word:** the function word itself (primary key for matching)
- **role** and **pragmaticFunction:** linguistic categorisation for contextualisation
- **semanticAutonomy:** central classification level (e.g., LOW, MEDIUM_HIGH, HIGH)
- **pragmaticFunctionNote/semanticAutonomyNote:** optional explanatory notes

Semantic autonomy is the central feature. The model differentiates words along a spectrum rather than a binary functional/lexical split, from **LOW** (purely grammatical or discourse-structuring, no independent meaning) to **HIGH** (a function word with notable lexical colouring).

The metadata are stored in JSON and loaded at start-up into a Map<String, FunctionWord>.[29] This enables O(1) retrieval[30] of metadata for tokens found in the text.

---

[26] Swing is a Java-based GUI toolkit that provides lightweight, platform-independent components for building desktop applications. For the documentation see https://docs.oracle.com/en/java/javase/21/docs/api/java.desktop/javax/swing/package-summary.html.

[27] The Jackson library (*FasterXML/Jackson*) is the de facto standard in Java for processing JSON data. For the documentation see https://github.com/FasterXML/jackson.

[28] *Unified Modeling Language* (UML) is a standardised graphical modeling language for the specification, construction, and documentation of software systems.

[29] A *Map* is a collection type that stores key–value pairs and allows efficient lookup of values based on their associated keys.

[30] *O(1) retrieval* refers to constant-time access in algorithmic complexity, meaning that the lookup time remains the same regardless of the size of the dataset.

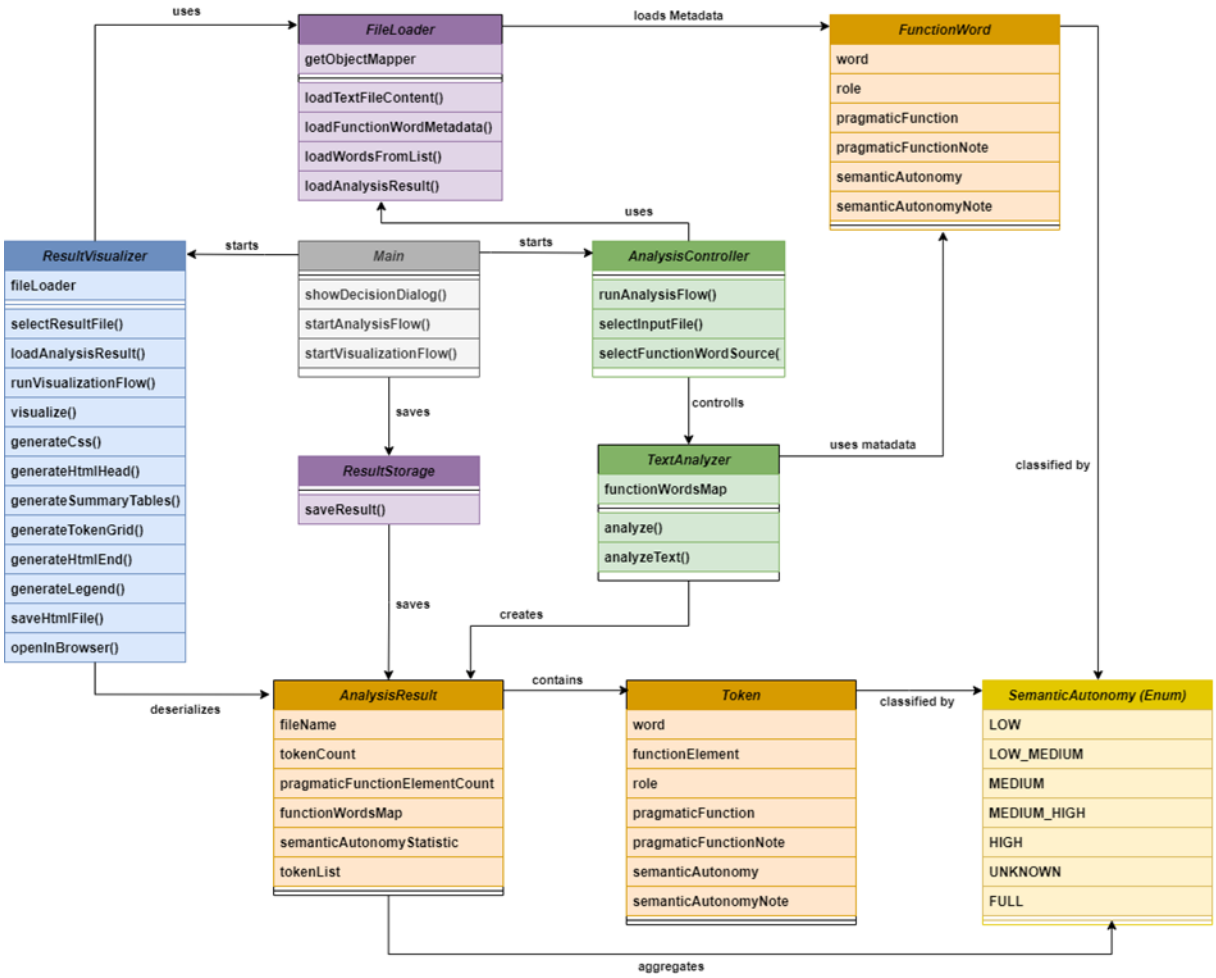


Fig. 4: UML Model

4.2.4 Analysis workflow

The workflow is organised into three steps coordinated by *TextAnalyzer*. The goal is to produce a token list enriched with function-word metadata and to compute summary statistics.

The autonomy levels (**LOW, LOW_MEDIUM, MEDIUM, MEDIUM_HIGH, HIGH, UNKNOWN, FULL**) are defined as an Enum[31] (*SemanticAutonomy*) to ensure consistency and prevent assignment errors.

4.2.5 Text processing and simplified tokenisation

Processing begins with simplified tokenisation. Regular expressions replace all characters that are neither letters nor spaces (e.g., commas, periods, parentheses) with spaces. The cleaned string is then split on one or more spaces.

This simplified tokenisation deliberately ignores more complex phenomena such as compounds, clitics/apostrophes, or word-internal punctuation. It was chosen for the prototype stage.

[31] An *Enumeration (Enum)* defines a fixed set of named constant values. It is used to create a type-safe collection of constants that prevents the use of invalid input values.

4.2.6 Metadata assignment and counting

After tokenisation, each token is matched against the *FunctionWord* map:

1) **matching:** for each token, the tool checks whether it appears as a key in the metadata map
2) **classification:**
   - **functional:** if found, the token is marked *isFunctionalElement* = true. The corresponding metadata (*role*, *pragmaticFunction*, *semanticAutonomy*) are copied into the token object, and the counter for that function word is incremented in *functionWordCounts*.
   - **lexical:** if not found, the token is marked *isFunctionalElement* = false and assigned **FULL** semantic autonomy as a default (potentially lexical with independent meaning).
3) **statistical analysis:** in parallel, occurrences per autonomy level are aggregated in *semanticAutonomyStatistic*. Regardless of type frequency, the total number of tokens assigned to each autonomy level (e.g., LOW or MEDIUM_HIGH) is recorded, enabling percentage distributions across the text.

4.2.7 Data model and result storage

After processing, a comprehensive *AnalysisResult* object is created and stored:

1) ***AnalysisResult* encapsulates:**
   - *tokenList* – the complete sequential list of tokens with metadata
   - *tokenCount* – the total number of tokens
   - *functionWordCounts* – frequency distribution for each function word in the metadata base
   - *semanticAutonomyStatistic* – aggregated counts per autonomy level
2) **persistent storage:** *ResultStorage* writes the *AnalysisResult* to JSON. Filenames are generated from the original input name plus a timestamp. JSON preserves the structure and classifications for later reuse (e.g., additional visualisations, cross-text comparisons, or external analyses).

4.2.8 Interactive data visualisation

Visualisation is essential for making the classification interpretable. *ResultVisualizer* deserialises the JSON output and generates a self-contained, interactive HTML page.

4.2.9 Technical concept (HTML generation)

Instead of a native Java GUI, the visualisation is produced as a complete HTML file, which offers:

1) **platform independence:** viewable in any modern browser without additional dependencies
2) **interactivity:** HTML/CSS allow tooltips and flexible layouts that are cumbersome in basic Java UI components
3) **archiving:** a static document that preserves results independently of the analysis tool.

The method *saveHtmlFile* reads the JSON results and builds a full HTML string with inline CSS and the token grid.

4.2.10 The colour-coded token grid

The central element is a colour-coded token grid (Fig. 5) that supports rapid, holistic assessment:

- each word is rendered as a .[32]
- only tokens classified as functional elements receive a coloured background; lexical tokens (autonomy **FULL**) remain uncoloured
- background colours map to autonomy levels, increasing in intensity from grammatical to lexically stronger functions:
    - **blue (LOW):** purely structural/grammatical
    - **green (LOW_MEDIUM):** weak lexical/modal function
    - **yellow (MEDIUM):** low autonomy, context-dependent interpretation
    - **orange (MEDIUM_HIGH):** pronounced lexical or pragmatic role
    - **red (HIGH):** significant independent lexical colouring
    - **grey (UNKNOWN):** in metadata but autonomy unresolved

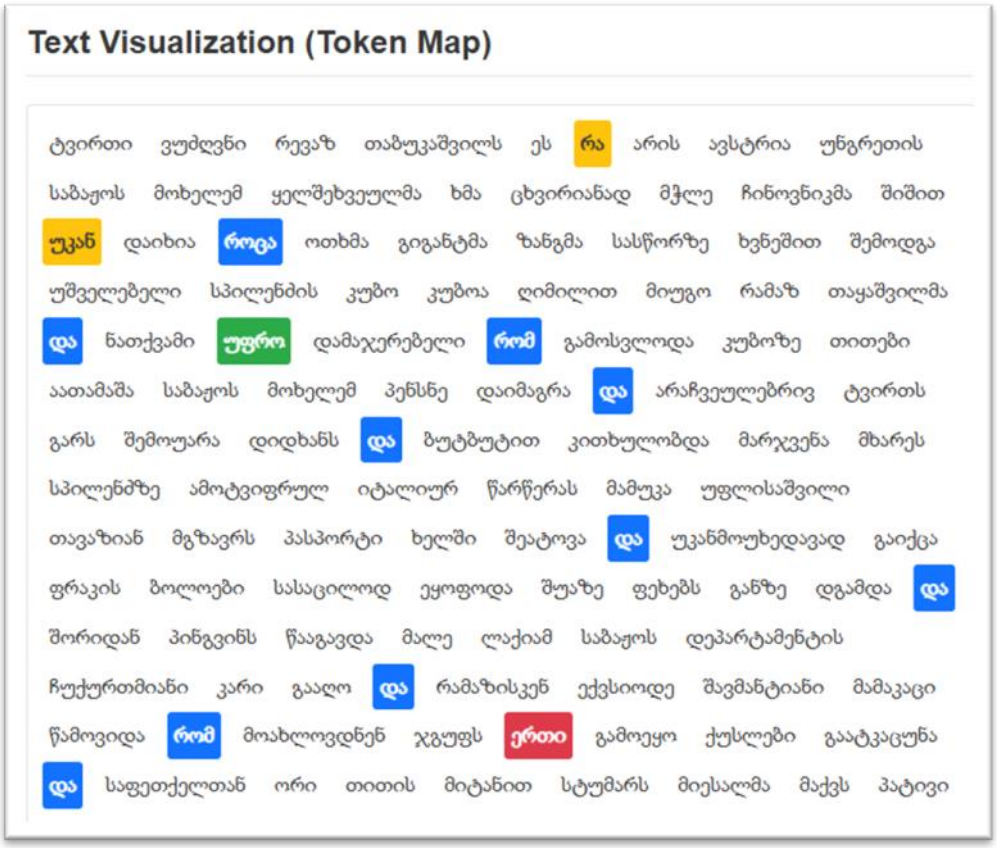


Fig. 5: Color-Coded Token Grid

[32] The element in HTML is an inline container used to group text or other inline elements for styling or scripting purposes without affecting the document's layout and does not convey any functional meaning.

Each coloured token includes a tooltip[33] with role, *pragmaticFunction*, *semanticAutonomy*, and additional notes, making the classification transparent (Fig. 6).

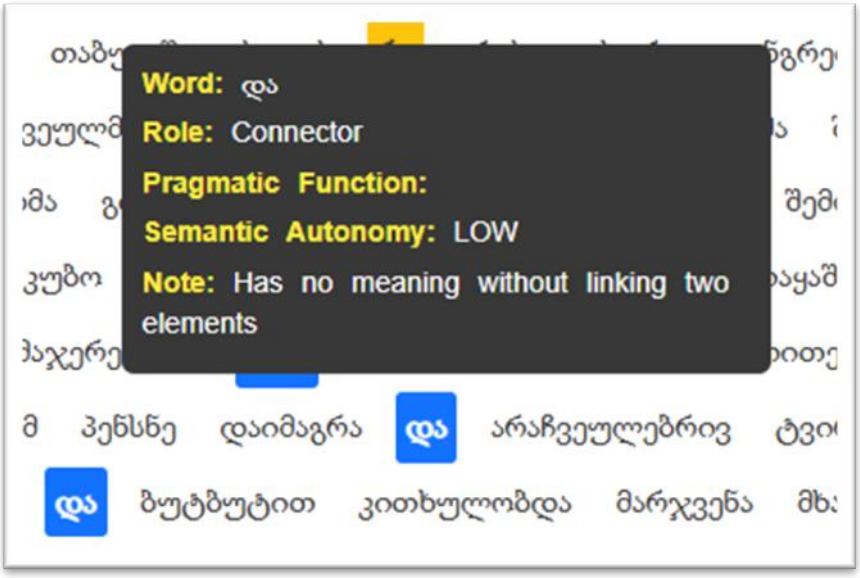


Fig. 6: Tooltip

4.2.11 Statistical overview

The HTML page also provides a statistical summary:

1) **general text statistics** (Fig. 7): total number of tokens and number of items classified as functional/pragmatic
2) **autonomy distribution** (Fig. 7): a Table with absolute and percentage shares across autonomy levels (LOW to HIGH, FULL, UNKNOWN)
3) **frequency list** (Fig. 7): all function words found in the text (present in the metadata base) with absolute frequencies, supporting frequency-based analysis.

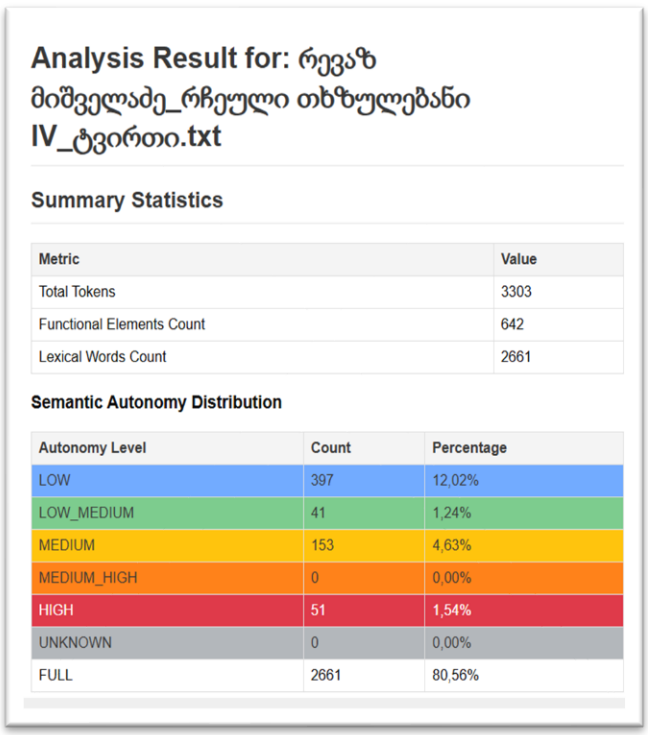
4.3 Discussion of Limitations and Outlook

The tool efficiently identifies, classifies, and visualises function words using a semantic-autonomy model. Clear separation of data, logic, and presentation yields a robust, portable Java application that converts linguistic classification into statistics and readable visual patterns.

4.3.1 Benefits for future research

1) **efficiency:** the colour grid enables an immediate qualitative assessment of functional density and autonomy levels. Researchers can quickly locate passages or documents with specific profiles (e.g., high shares of higher-autonomy function words)
2) **transparency:** tooltips expose token-level metadata, ensuring the traceability of decisions and links to the theoretical model

[33] A tooltip is a small, contextual pop-up text box that appears when a user hovers over an element, providing additional information without cluttering the main interface.



Analysis Result for: რევაზ მიშველაძე_რჩეული თხზულებანი IV_ტვირთი.txt

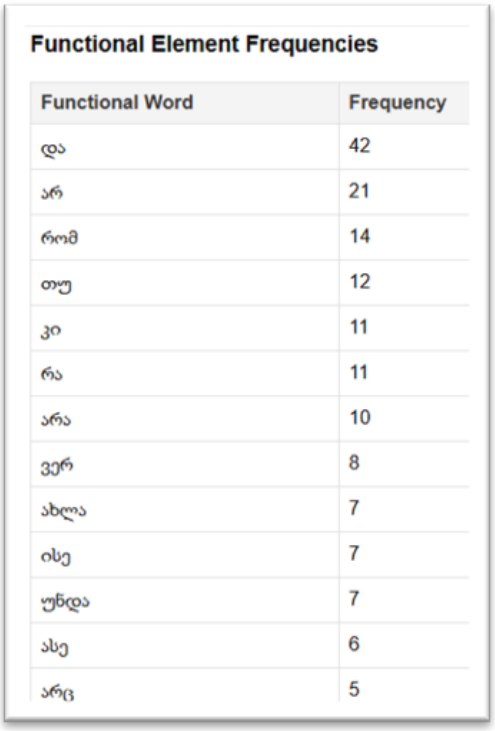
Summary Statistics

Metric	Value
Total Tokens	3303
Functional Elements Count	642
Lexical Words Count	2661

Semantic Autonomy Distribution

Autonomy Level	Count	Percentage
LOW	397	12,02%
LOW_MEDIUM	41	1,24%
MEDIUM	153	4,63%
MEDIUM_HIGH	0	0,00%
HIGH	51	1,54%
UNKNOWN	0	0,00%
FULL	2661	80,56%

Fig. 7: General text statistics



Functional Element Frequencies

Functional Word	Frequency
და	42
არ	21
რომ	14
თუ	12
კი	11
რა	11
არა	10
ვერ	8
ახლა	7
ისე	7
უნდა	7
ასე	6
არც	5

Fig. 8: Frequency list

3) **comparability:** absolute and percentage metrics support a quantitative comparison across text types, authors, or periods. The JSON output integrates easily with external statistical tools.

4.3.2 Limitations of the tool

Despite its strengths, the limitations of the analysis tool primarily lie in its deliberately simplified design choices and its reliance on external data:

1) **rudimentary tokenisation:** RegEx[34]-based splitting may mishandle clitics, compounds, or internal punctuation, leading to misclassification as UNKNOWN or FULL
2) **no POS or syntactic disambiguation:** the identification is string-based; homonyms that can be functional or lexical (e.g., *tavi*) are not distinguished via the context
3) **dependence on the metadata base:** the quality of the result depends on the coverage and accuracy of the classification. Revisions of the theoretical model require manual updates to the metadata.

[34] *Regular Expressions* are a sequence of characters that define a search pattern. In this context it is utilised for rudimentary tokenisation by systematically removing punctuation and splitting the text content based on whitespace. This method provides a lightweight, language-agnostic approach to segmentation but does not account for complex linguistic phenomena such as clitics or compound words.

4.4 Outlook

In summary, the prototype analysis tool shows that a scalar model of semantic autonomy can be operationalised for Georgian, turning abstract classifications of function words into transparent visualisations and comparable statistics, which allow for both the internal frequency comparison of functional elements within a single text and the external comparison of functional profiles across different texts. Its modular Java architecture and JSON-based metadata make it portable and extensible, but current limitations in tokenisation, lack of POS- and syntax-based disambiguation, and dependence on a hand-crafted metadata base still constrain coverage and precision. Future work will focus on integrating a more fine-grained tokenisation adapted to Georgian orthography, lightweight syntactic and POS cues for resolving homonymy and scope, and enhanced visual and statistical modules that enable systematic comparison across larger corpora, text types, and time periods. In this way, the tool can evolve from a proof of concept into a broader platform for quantitative and qualitative research on function words in Georgian and beyond.

5. Conclusion

Over the past century, the autosemantic-synsemantic contrast was mostly treated as a dichotomous, hierarchical opposition. More recently, this has shifted towards a scalar and dynamic conception. Older linguistic models as presented in Chapter 2, ranging from Jespersen to Bloomfield and Diewald, emphasised the opposition of formal and semantic functions. Other approaches, such as generative grammar, supplemented abstract syntactic functions, while functionalist and typological approaches introduced communicative and diachronic considerations. Nowadays, the content-function distinction is increasingly regarded as gradient and dynamic, being shaped not only by usage but also by structure, diachronic evolution, and pragmatic function.

The analysis conducted in this study has led to the following conclusions:

1) there are functional overlaps across domains: words like *ḳidev* ‘again’, *marṭo* ‘only’, *sul* ‘always’ blur syntax, semantics, and pragmatics, acting simultaneously as focus markers, intensifiers, and adverbs
2) cross-linguistic parallels are created: the scalar relationship matches patterns in German, Japanese, and English, suggesting universality in how languages encode pragmatic force through semantically “light” items
3) the necessity for reconceptualising arises for the GNC concerning classification, which must be multi-dimensional: instead of a single tag (“function word”), we need layered tagging (semantic, syntactic, pragmatic).

The three classification levels (syntactic, pragmatic and semantic) together show that function words in Georgian form a dynamic continuum linking meaning, structure, and use: semantically, they range from lexical to fully grammatical; syntactically, they anchor clause architecture; and pragmatically, they orchestrate interaction, focus, and stance.

This confirms that function words are not a homogeneous category – they represent multifunctional, context-sensitive operators that integrate semantics, syntax, and pragmatics into a cohesive linguistic system.

The analysis tool developed by Anastasia Kamarauli is a first computational approach and will certainly need enhancements. These specifically include improved tokenisation tailored to Georgian orthography, POS tagging with light syntactic cues to resolve homonymy/syncretic

forms and identify scope-sensitive categories (e.g., negation, complementisers), and a module for direct comparison of multiple documents to support quantitative studies.

Abbreviations

ADV	adverbial case	INST	instrumental case
AOR	aorist tense	MOD	modal
COND	conditional	NEG	negation
DAT	dative case	NOM	nominative case
EMPH.V	emphatic vowel	OPT	optative
ERG	ergative case	PL	plural
FOC	focus	PRES	present tense
GEN	genitive case	S	subject
IMP	imperative	SG	singular number
IMPF	imperfect tense	1/2/3	1st/2nd/3rd person

References

Abraham (1991): Werner A. “Discourse particles in German: How does their illocutive force come about?”, in Werner A. (ed.), *Discourse Particles: Descriptive and theoretical investigations on the logical, syntactic and pragmatic properties of discourse particles in German*, Amsterdam: John Benjamins, 203–252.

Blakemore (1987): Diane B., *Semantic Constraints on Relevance*. Oxford: Blackwell.

Bloomfield (1933): Leonard B., *Language*. New York: Henry Holt. https://archive.org/details/in.ernet.dli.2015.147712

Bühler (1934): Karl B., *Sprachtheorie: Die Darstellungsfunktion der Sprache*. Jena: Fischer. https://hdl.handle.net/11858/00-001M-0000-002A-F88A-A

Bybee & Pagliuca (1985): Joan L. B. and William P., “Cross-linguistic comparison and the development of grammatical meaning”, in Jacek Fisiak (ed.), *Historical Semantics – Historical Word Formation* (Trends in Linguistics, Studies and Monographs, 29), 59–83.

Chomsky (1957): Noam C., *Syntactic Structures*. The Hague: Mouton.

— (1965): id., *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

— (1995): id., *The Minimalist Program*. Cambridge, MA: MIT Press.

Diewald (2006): Gabriele D., “Discourse particles and modal particles as grammatical elements”, in Kersti Fischer (Ed.), *Approaches to Discourse Particles*. Amsterdam: Elsevier, 403–425.

Dik (1978): Simon C. D., *Functional Grammar*. Amsterdam: North-Holland.

— (1997): id., *The Theory of Functional Grammar.* Part I: *The Structure of the Clause*. 2nd ed. Berlin: Mouton de Gruyter.

Fraser (1999): Bruce, F., “What are discourse markers?”, *Journal of Pragmatics* 31/7, 931–952. https://www.researchgate.net/profile/Bruce-Fraser-2/publication/222481598.

Givón (1971): Talmy G., “Historical syntax and synchronic morphology: An archaeologist’s field trip”, in *Papers from the 7th Meeting of the Chicago Linguistic Society*, 394–415. https://www.researchgate.net/publication/272816328.

— (1979): id., *On Understanding Grammar*. New York: Academic Press.

— (1981): id., "On the development of the numeral 'one' as an indefinite marker", *Folia Linguistica Historica* 2/1, 35–53.

— (1983): *Topic Continuity in Discourse: A quantitative cross-language study*. Amsterdam: John Benjamins.

Goldberg (1995): Adele E. G., *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

— (2006): ead., *Constructions at Work: The Nature of Generalisation in Language*. Oxford: Oxford University Press.

Harris (1951), Zellig H., *Methods in Structural Linguistics*. Chicago: University of Chicago Press.

Haspelmath (2000): Martin H., "Parts of speech", in Geert Booij, Christian Lehmann & Joachim Mugdan (eds), *Morphology: An International Handbook on Inflection and Word Formation*. I. Berlin: De Gruyter, 10–28.

— (2011): id., "The indeterminacy of word segmentation and the nature of morphology and syntax", *Folia Linguistica* 45/1, 31–80. https://www.researchgate.net/publication/267242558.

Heine, Claudi & Hünnemeyer (1991): Bernd H., Ulrike C. and Friederike H., *Grammaticalisation: A conceptual framework*. Chicago: University of Chicago Press.

Jespersen (1924): Otto J., *The Philosophy of Grammar*. London: Allen & Unwin. https://archive.org/details/in.ernet.dli.2015.282299.

Jorbenadze, Kobaidze & Beridze (1988): ბესარიონ ჯორბენაძე, მანანა კობაიძე, მაია ბერიძე, *ქართული ენის მორფემებისა და მოდალური ელემენტების ლექსიკონი*. თბილისი: მეცნიერება.

Kamarauli (2022): Mariam K., *The Nominal Domain in Georgian: A diachronic analysis* (Folia Caucasica, 5). Wiesbaden: Reichert.

Lambrecht (1994): Knud L., *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge: Cambridge University Press.

Langacker (1977): Ronald W. L., "Syntactic reanalysis", in Charles N. Li (ed.), *Mechanisms of Syntactic Change*, Austin: University of Texas Press., 57–139).

— (1987): id., *Foundations of Cognitive Grammar*. I: *Theoretical Prerequisites*. Stanford: Stanford University Press.

— (2008): id., *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press.

Lehmann (1982): Christian L., *Thoughts on Grammaticalisation*. Munich: LINCOM Europa. 3rd edition 2015. https://library.oapen.org/bitstream/20.500.12657/32841/1/603353.pdf.

Schiffrin (1987): Deborah Sch., *Discourse Markers*. Cambridge: Cambridge University Press.

Searle (1969): John R. S., *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Tesnière (1959): Lucien T., *Éléments de syntaxe structurale*. Paris: Klincksieck. https://archive.org/details/LucienTesniereElementsDeSyntaxeStructurale.

ქართული ენის ფუნქციური სიტყვების კორპუსზე დაფუძნებული კვლევა

მარიამ ყამარაული (ფრანკფურტი / ჰამბურგი)

ანასტასია ყამარაული (ჰაგენბერგი / ბათუმი)

DOI: https://doi.org/10.62235/dk.4.2025.10521
mariam.kamarauli@uni-hamburg.de/
ma.kamarauli@em.uni-frankfurt.de || ORCID: 0009-0006-0404-4424
a.kamarauli@web.de || ORCID: 0009-0008-2258-8307

შესავალი

წინამდებარე სტატიის „ქართული ფუნქციური სიტყვების კორპუსზე დაფუძნებული კვლევა“ მიზანია ქართული ენის ფუნქციური სიტყვების სისტემური, კორპუსლინგვისტური კვლევა, კერძოდ, იმის აღწერა, თუ როგორ იქცევიან ქართულ ენაში ფუნქციური სიტყვები (ანუ სინსემანტიკური ელემენტები) ბუნებრივ გარემოში და რატომ არის რთული მათი ცალსახა კლასიფიკაცია მორფოლოგიურად მდიდარ ენაში. ნაშრომი ორი ურთიერთდაკავშირებული ამოცანის გადაჭრას ისახავს მიზნად: (1) შეიმუშაოს ქართული ენის ფუნქციური სიტყვების კორპუსზე დაფუძნებული, ლინგვისტურად დეტერმინირებული კლასიფიკაციის სისტემა, რომელიც სცილდება მარტივ დიქოტომიას „შინაარსი vs. ფუნქცია“ და (2) აჩვენოს, თუ როგორ შეიძლება ამ კლასიფიკაციამ ოპერაციონალიზაცია მოახდინოს კომპიუტერული ლინგვისტიკის ფარგლებში საგანგებოდ შექმნილი მსუბუქი ინსტრუმენტის მეშვეობით, რომელიც ამოიცნობს ფუნქციურ სიტყვებს ტექსტში და უზრუნველყოფს შედეგების ვიზუალიზაციას.

მოტივაცია და წინაპირობა

ავტორთა დაკვირვებით, ფუნქციური სიტყვები, მიუხედავად იმისა, რომ ისინი ხშირად აღწერილია, როგორც სემანტიკურად „მსუბუქი“ ელემენტები, გადამწყვეტ როლს თამაშობენ გრამატიკული მნიშვნელობის, დისკურსის კოჰერენტულობისა და ინტერაქციის თვალსაზრისით. ისინი მიუთითებენ ლოგიკურ ურთიერთობებზე (მიზეზი, კონტრასტი, პირობა), მართავენ თანმიმდევრობას წინადადებებსა და ფრაზებს შორის და შეუძლიათ ილოკუციური ძალის შეცვლა (მაგ., უარყოფის ან მოდალობის გამოხატვა). ქართულ საენათმეცნიერო ლიტერატურაში ფუნქციური სიტყვები ძირითადად ფორმალური პერსპექტივიდან განიხილება და შედარებით ნაკლებად არის წარმოდგენილი ლექსიკოგრაფიაში; სპეციალიზებულ ლექსიკონებშიც კი (მაგ., ჯორბენაძე და სხვ. 1988), რომლებიც მანუალურად დამუშავებულ რესურსებს ეყრდნობიან და

მხოლოდ ბეჭდური ფორმით არიან გამოცემული, თეორიულად ვერ ფარავენ ფუნქციური ელემენტების კვალიფიკაციისათვის აუცილებელ თანამედროვე მოთხოვნებს. ამიტომ, არსებული გარემოება მოითხოვს ფუნქციური სიტყვების განახლებულ, პრაქტიკულ გამოყენებაზე დაფუძნებულ მიდგომას, რომელიც კორპუსული კვლევის შედეგებით იქნება გამყარებული.

თეორიული ჩარჩო: დიქოტომიიდან გრადაციამდე

სტატიის თეორიულ ნაწილში, რომელიც კვლევის ბირთვს წარმოადგენს, დაწვრილებით განვიხილავთ ლინგვისტურ თეორიებში მოცემულ ავტოსემანტიკური და სინსემანტიკური ელემენტების შეპირისპირებით ანალიზს. ადრეულ თეორიულ ჩარჩოებში (მაგ., იესპერსენი, ბიულერი) ხაზგასმულია ის გარემოება, რომ ფუნქციური სიტყვები დამოკიდებულია კონტექსტზე და ემსახურება დისკურსის ორგანიზებას. სტრუქტურალისტური/დისტრიბუციული მიდგომა (მაგ., ბლუმფილდი, ჰარისი) ხაზს უსვამს ფუნქციურ ელემენტთა განაწილების შეზღუდვას და მათი გამოყენების მაღალ სიხშირეს. დამოკიდებულების გრამატიკა (ტენიერი) ფუნქციურ სიტყვებს განიხილავს, როგორც შინაარსობრივი სიტყვის „ბირთვების“ გარშემო არსებულ რელაციურ შემაერთებლებს. გენერატიული გრამატიკა (ჩომსკი) კი მათ ლექსიკურ და ფუნქციურ კატეგორიებად ყოფს შესაბამისი ფუნქციური თავებით (D, T, C), რომლებიც წინადადების სტრუქტურის ჩამოყალიბებას უზრუნველყოფენ. ფუნქციონალისტურ და გრამატიკალიზაციაზე ორიენტირებულ ნაშრომთა ავტორები (მაგ., დიკი, გივონი, ჰაინე და სხვ.) ხაზს უსვამენ ცვლილებას დიაქრონულ ასპექტში, რის შედეგადაც ლექსიკური ერთეულებიდან მიიღება გრამატიკული მარკერები რეანალიზის, სემანტიკური გაუფერულებისა და ფონოლოგიური შემცირების გზით. ტიპოლოგიურ/კოგნიტურ/კონსტრუქციულ მოდელებში (მაგ., ჰასპელმათი, ლანგაკერი, გოლდბერგი) ავტორები გვთავაზობენ გრადუირებული, სხვადასხვა კრიტერიუმებისაგან შემდგარი კონტინუუმის არსებობას ბინარული სისტემის სანაცვლოდ. პრაგმატული მიდგომები (მაგ., შიფრინი, ფრეიზერი, დივალდი, ლამბრეხტი, სირლი) კი გვიჩვენებს, რომ ბევრი „მცირე სიტყვა“ ძირითადად დისკურსს მართავს და სტრუქტურული და გამოყენებითი ცვლილების გზით არის ჩამოყალიბებული, და რომ პრაგმატული დატვირთვა ხშირად ცალკე, დამოუკიდებელ განზომილებად უნდა განვიხილოთ.

ქართული ენის მონაცემები: მრავალფუნქციურობა და კონსტრუქციული ქცევა

ცალკე ქვეთავში განვიხილავთ საკითხს, თუ რატომ მიგვაჩნია განსახილველ ოდენობებთან მიმართებით ერთი ეტიკეტი („ფუნქციური სიტყვა“) არაადეკვატურად. როგორც კვლევამ გვიჩვენა, ერთი და იგივე ელემენტი შეიძლება მონაწილეობდეს სხვადასხვა კონსტრუქციაში, სხვადასხვა პოზიციაში, რის

შედეგადაც იცვლება მისი წვლილი სინტაქსში, სემანტიკასა და პრაგმატიკაში. მაგალითად:

- „უფრო“ შეიძლება იყოს ა) შედარებითი ხარისხის მოდიფიკატორი ზედსართავ სახელებთან, ბ) ინტენსიფიკატოuრი, რომელიც აძლიერებს მიზეზობრივ დაქვემდებარებას ან გ) იყოს ლექსიკალიზებული ფოკუსირებული გამოთქმის ნაწილი (მაგ., „მით უფრო“).
- „რა“ მერყეობს კითხვითი ელემენტიდან დისკურსზე დაფუძნებულ ინტერაქტიულ გამოყენებამდე და ასევე გვხვდება გრამატიკალიზებულ ფუნქციურ ფრაზაში „რა თქმა უნდა“, რომელიც წინადადებაში ფუნქციონირებს როგორც მოდალური ფრაზა.
- „ერთი“ იქცევა ა) როგორც რიცხვითი სახელი კლასიფიკატორებთან, ბ) როგორც განუსაზღვრელობითი ნაცვალსახელად („ერთ დღეს“), გ) დისკურსში ფოკუსირებადი სტრუქტურის ნაწილი, დ) როგორც გამხსნელი ფრაზის ნაწილი ჩამონათვალში („პირველ რიგში...“) ან გამოიყენება სასაუბრო თემის სხვა თემაზე გადამტანი ელემენტის ფუნქციური ოპერატორი.

ეს მაგალითები საშუალებას იძლევა ჩავატაროთ მრავალდონიანი ანალიზი: ფუნქციური სიტყვები ქართულში ხშირად მოქმედებენ როგორც ოპერატორები, რომელთა როლი დამოკიდებულია მათ სინტაქსურ გარემოსა და კონსტრუქციულ შეფუთვაზე.

კორპუსზე დაფუძნებული კლასიფიკაცია სამი განზომილების მიხედვით

სტატიის ემპირიულ ნაწილში, საკვლევად შევარჩიეთ 100 ყველაზე მაღალი სიხშირის ფუნქციური სიტყვა ქართული ენის ეროვნული კორპუსის თანამედროვე ქართული ენის ქვეკორპუსიდან (GNC-NG) და მოვახდინეთ მათი კლასიფიკაცია სამი განზომილების მიხედვით - სინტაქსური როლი, პრაგმატული ფუნქცია და სემანტიკური ავტონომია - რომლებიც წარმოდგენილია ნაშრომის IV–VII ცხრილებში.

განზომილება	რას მოიცავს	ტიპიური ქართული ერთეულები (მაგალითები)	რატომ არის მნიშვნელოვანი
სინტაქსური როლი	სტრუქტურული ფუნქცია წინადადების არქიტექტურაში (შემაერთებელი, მაქვემდებარებელი, მოდიფიკატორი და ა.შ.)	შემაერთებელი სიტყვები/მაქვემდებარებელი სიტყვები (მაგ., „და“, „მაგრამ“, „რომ/თუ“, „რადგან“), ნეგატორები, wh-ფორმები, ზმნიზედები, თანდებულები	ასახავს, თუ როგორ ქმნიან ფუნქციური სიტყვები წინადადების სტრუქტურას.

პრაგმატული ფუნქცია	დისკურსის მართვა, ურთიერთქმედების ჩარჩოები (პოზიცია, ფოკუსი), რიტორიკული ძალა	დისკურსის მარკერები, მოდალური/ევიდენციალური ნაწილაკები („შესაძლოა/როგორც ჩანს“),ფოკუსის შემზღუდველები („მხოლოდ“).	განსაზღვრავს ფუნქციურ სიტყვებს, რომლებიც გრამატიკის მიღმა ინტერპრეტაციას წარმართავენ.
სემანტიკური ავტონომია	„დამოუკიდებელი მნიშვნელობის“ ხარისხი კონტექსტუალურ დამოკიდებულებასთან შედარებით	LOW: კავშირები, ნაწილაკები, ნეგატორები; MEDIUM: ზმნიზედები, ნაცვალსახელები; HIGH: უფრო მეტად ლექსიკური შეფერილობის ერთეულები (მაგ., კვანტიფიკატორები, ზოგიერთი ზმნიზედა, გრამატიკალიზებული არსებითი სახელი, როგორიცაა „თავი/თვით“).	წარმოაჩენს კონტინუუმს არა როგორც ბინარულ, არამედ როგორც გრადაციულ სისტემას.

კვლევის ძირითადი თეორიული შედეგი ის არის, რომ ქართული ფუნქციური სიტყვები ნაწილდება სემანტიკურ-ავტონომიურ გრადაციულ ველში: დაბალი ავტონომიის ერთეულები განთავსდებიან გრამატიკულ ბირთვში (მაღალდამოკიდებული, სტრუქტურული), საშუალო ავტონომიის ერთეულები ხშირად მოიცავს დეიქტურ ზმნიზედებს და ნაცვალსახელის ფორმებს, ხოლო მაღალი ავტონომიის „ფუნქციური“ ერთეულები ინარჩუნებენ უფრო ძლიერ ლექსიკურ შინაარსს. ნაშრომში წარმოდგენილ კლასიფიკაციას განვიხილავთ, როგორც სკალარული მოდელის მტკიცებულებას და არა მკაცრ ავტოსემანტიკურ vs. სინსემანტიკურ დიქოტომიის. წარმოდგენილი კვლევის პროცესში შევქმენით ინსტრუმენტი „ფუნქციური ელემენტების ანალიზატორი“ (Java), რომელიც კვლევის კომპიუტერული ლინგვისტიკის ნაწილს წარმოადგენს და განახორციელებს კლასიფიკაციის ოპერაციულ რეალიზებას. მომხმარებლები წინასწარ ადგენენ ფუნქციური სიტყვების ან ფუნქციურსიტყვიანი კონსტრუქციების სიას (JSON მეტამონაცემების სახით) და შეჰყავთ იგი საანალიზო ტექსტთან ერთად; შემდეგ ინსტრუმენტი განსაზღვრავს, თუ რომელი ფუნქციური სიტყვები გვხვდება მოცემულ ტექსტში, ითვლის ფუნქციური სიტყვების სიხშირეს და დამატებით ახდენს მიღებული შედეგების ვიზუალიზაციას.

არქიტექტურა და სამუშაო პროცესი (რომელიც აღწერილი და დიაგრამირებულია ნაშრომის ვიზუალიზაციაში) ხორციელდება სამ ეტაპად - (1) მონაცემების მიწოდება (ტექსტის + მეტამონაცემების ჩატვირთვა, შედეგების JSON-ად შენახვა), (2) ტექსტის ანალიზი (ტოკენიზაცია, შესაბამისობა, სტატისტიკური დამუშავება) და (3) პრეზენტაცია (Swing დიალოგები და HTML ვიზუალიზაცია, რომელიც იხსნება ბრაუზერში).

შეზღუდვები და პერსპექტივები

ნაშრომში ნათლად არის მითითებული, რომ ჩვენ მიერ შექმნილი ინსტრუმენტი პროტოტიპულია: ტოკენიზაცია განზრახ არის გამარტივებული და შესაძლოა არასწორად გააანალიზოს ნაწილაკები ან რთული სიტყვები; იდენტიფიკაცია სტრიქონებს ეფუძნება და არ გააჩნია POS/სინტაქსური დისამბიგვირება, რის გამოც ომონიმურობა, მაგალითად, „თავი“, როგორც ლექსიკური ერთეული (არსებითი სახელი) და უკუქცევითი ნაცვალსახელი, ავტომატურად ვერ გაანალიზდება; კვლევის შედეგების საერთო ხარისხი დამოკიდებულია ხელით შედგენილი მეტამონაცემების სიის სისრულესა და სისწორეზე. დაგეგმილი სამომავლო სამუშაოები მოიცავს ქართული ენის ბუნების გათვალისწინებით განხორციელებულ ტოკენიზაციას, მსუბუქ POS ანალიზს, დისამბიგვირების განხორციელებას და უკეთეს მხარდაჭერას სისტემური ჯვარედინი ტექსტური შედარებებისთვის.

დასკვნა

ნაშრომის მთავარი შედეგი არის დებულება, რომ ქართული ფუნქციური სიტყვები არ ქმნიან ჰომოგენურ კლასს. ისინი წარმოადგენენ მრავალგანზომილებიან სისტემას, რომელიც აბალანსებს (i) სტრუქტურულ სინტაქსს, (ii) გრადუირებულ სემანტიკურ ავტონომიას და (iii) პრაგმატულ-დისკურსურ ფუნქციას. კორპუსზე დაფუძნებული კლასიფიკაცია (100 მაღალი სიხშირის ერთეული; მათგან 23 მონიშნულია, როგორც პრაგმატულად მნიშვნელოვანი) ოპერაციულ ინსტრუმენტთან ერთად ხელს უწყობს ქართული ენის კორპუსებში ქართული ენის ფუნქციური სიტყვების უფრო მდიდარ ანოტაციასა და ანალიზს, ასევე თეორიული ლინგვისტიკისა და მსუბუქი გამოთვლითი მეთოდების გაერთიანებას.