International bilingual journal in
Humanities & social Sciences

Millennium
მილენიუმი

Academy for Digital
Humanities - Georgia

Vol. 4
2026

The Bilingual Scientific Online Journal of the Academy for Digital Humanities

in Humanities and Social Science

დიგიტალური ჰუმანიტარიის აკადემია - საქართველოს

ორენოვანი საერთაშორისო ახალგაზრდული სამეცნიერო ონლაინჟურნალი

მილენიუმი

Millennium

Volume 4

Tbilisi/თბილისი

2026

The peer-reviewed bilingual scientific online journal **Millennium** was initiated by the founders of the Academy for Digital Humanities and is designed for young researchers working in Humanities and Social Sciences – master students, doctoral students and postdocs. Millennium aims to promote the academic development of a new generation of scientists by creating an academic platform for young researchers working in the Humanities and Social Sciences to publish scientific papers.

დიგიტალური ჰუმანიტარიის აკადემიის რეცენზირებადი ელექტრონული ბილინგვური სამეცნიერო ჟურნალი **„მილენიუმი“** დაარსდა დიგიტალური ჰუმანიტარიის აკადემიის დამფუძნებელთა მიერ და განკუთვნილია ჰუმანიტარიაში და სოციალურ მეცნიერებებში მოღვაწე ახალგაზრდა მკვლევართათვის - მაგისტრანტების, დოქტორანტებისა და პოსტდოქტორანტებისათვის. ჟურნალი „მილენიუმი“ მიზნად ისახავს მეცნიერთა ახალი თაობის აკადემიური განვითარების ხელშეწყობას - ჰუმანიტარულ დარგებსა და სოციალურ მეცნიერებებში მომუშავე ახალგაზრდა სპეციალისტებისათვის აკადემიური პლატფორმის შექმნას სამეცნიერო ნაშრომების გამოსაქვეყნებლად.

ISSN 2960-9887 (Online)

Millennium, Volum 4

მილენიუმი, ტომი 4

Content / სარჩევი

V. Communication Studies / მედია და კომუნიკაცია

VI. Political Linguistics / პოლიტიკური ლინგვისტიკა

Analysis of Function Words in Georgian Political Discourse: A New Perspective

Anastasia Kamarauli
(Batumi Shota Rustaveli State University)

DOI: 10.62235/mln.4.2026.12163
a.kamarauli@web.ge || ORCID: 0009-0008-2258-8307

Abstract: This study presents a quantitative linguistic and register-analytic investigation into the political rhetoric of Georgia's first president, Zviad Gamsakhurdia. Analysing a corpus of 25 selected texts divided into three functional registers (official announcements, direct addresses, and interviews), the paper explores the systematic variation of the Type-Token Ratio (TTR), Guiraud's index, and lexical density across different modes of discourse. Drawing on Halliday's register theory and Gregory's diatypic variation model, the research examines how the cognitive constraints of spoken and written channels shape the micro-linguistic patterns of political language, demonstrating that political discourse occupies a hybrid position between spoken and written modalities. The mathematical dependency of lexical diversity measures on text length is methodologically addressed through corrective transformations like the Guiraud index. Furthermore, the study contextualises the empirical results within the morphosyntactic typology of the Georgian language. The findings reveal that while Gamsakhurdia's spontaneous discourse in interviews shifts toward a function-word-rich, relational clausal grammar, his texts overall maintain an exceptionally high lexical density, demonstrating a deeply compressed, nominal phrasal grammar traditionally attributed to highly planned written language. Methodologically, the analysis indicates that standard quantitative metrics exhibit significant limitations when applied to Georgian as a synthetic language. Consequently, the study argues for the development of language-specific adaptation models to ensure the valid application of metrics such as TTR and lexical density to morphosyntactically complex languages.

Keywords: Corpus Linguistics; Political Linguistics; Georgian Political Discourse; Function Words.

Introduction

The study of political rhetoric has traditionally relied primarily on qualitative, hermeneutic methods. However, with the establishment of quantitative linguistics and corpus linguistics, a paradigm shift has occurred: linguistic structures, stylistic patterns, and rhetorical strategies can now be analysed mathematically and statistically. Political speeches, especially those from periods of existential national crises, offer an ideal field of study for these quantitative approaches.

A central starting point for any quantitative text analysis is the fundamental distinction between types and tokens (the so-called type-token distinction). While tokens denote the total number of all word forms occurring in the text, types represent the number of unique, non-redundant word forms. Based on this, the lexical structure can be further refined by dividing the texts into content words (lexical units) such as nouns, main verbs, and adjectives, and function words (structural units) such as conjunctions, pronouns, and prepositions. This functional division forms the basis for calculating lexical density and grammatical weight, which serve as primary indicators of a text's information content, administrative character, and syntactic complexity. Early studies concluded that high lexical density is primarily found in

highly condensed written texts, while spoken language requires more function words and thus has a comparatively lower lexical density.

However, applying these standard quantitative metrics to real-world corpora presents significant mathematical and typological challenges. First, classic metrics like the TTR suffer from a severe text-length dependency: as a text grows longer, words inevitably repeat, causing the vocabulary growth curve to flatten degressively. Consequently, raw TTR values across texts of unequal length are not directly comparable. Second, and more profoundly, standard quantitative metrics exhibit significant limitations when applied to synthetic and agglutinative languages like Georgian. While analytic languages like English encode grammatical relationships through separate, independent function words, synthetic languages morphologically integrate case, postpositions, tense, and even copula verbs directly into a single nominal or verbal stem. Standard whitespace-based tokenizers treat these highly synthesized clusters as single graphemic tokens, thereby systematically underrepresenting functional elements, artificially inflating lexical density, and distorting lexical richness indices.

This study addresses these theoretical and methodological challenges by conducting a register-analytic and quantitative investigation into the political rhetoric of Zviad Gamsakhurdia, Georgia's first president, during the turbulent period of 1991–1993. Gamsakhurdia's speeches occupy a unique, hybrid register-theoretic position between written and spoken modalities. By evaluating a curated corpus of 25 texts, this paper explores the systematic variation of lexical diversity and density across different modes of political communication.

The article is structured as follows: Chapter 1 (Theoretical Framework) establishes the register-linguistic foundation of the study. Chapter 2 (Methodology) describes the corpus of 25 selected texts and details the computerlinguistic workflow, explaining the comparative use of the GNCAnalyzer and the FunctionalWordAnalyzer. Chapter 3 (Empirical Findings and Register-Analytic Results) presents the quantitative results. It provides an overview before dissecting the statistical variations within the three specific registers: Official Announcements, Direct Addresses, and Interviews. Chapter 4 (Discussion and Theoretical Implications) critically evaluates the typological limitations of the data. Using a comparative case study of Text #10 and Text #24, it demonstrates the precise mathematical distortion caused by postpositions and copulas, contextualizing the findings within modern Georgian NLP schemas. Chapter 5 (Conclusion) summarises the key insights, reflects on the methodological integration of corpus statistics and discourse analysis, and outlines future directions for digital Kartvelology.

1. Theoretical Framework

1.1 Characteristics of spoken and written language

In the early stages of research on written versus spoken language, written language was often considered the true, logical, and normative standard, while spoken language was seen as incomplete, flawed, or structurally "weak". Halliday (1989) demonstrated in his book that this view is incorrect. Spoken and written language are structurally and functionally equivalent, but they operate on entirely different principles because they fulfil different evolutionary and social roles.

Halliday examines the process of writing, in which spoken language is transformed into written language (Halliday 1989: 92ff.). He demonstrates how one and the same content can be translated step by step from a typically spoken (congruent) to a typically written (metaphorical) structure. On the way from spoken to written language, the syntactic architecture changes systematically: verbs become nouns (nominalisation), conjunctions become prepositions, and adverbs become adjectives.

Halliday defines this change as grammatical metaphor. A dynamic process is not represented grammatically as a verb, but metaphorically as a noun (a thing). From this grammatical difference, Halliday derives two fundamentally different ways in which people experience and linguistically structure reality.

Spoken language (*World of Happening*; Halliday 1989: 93) is dynamic and process-oriented. It presents the world as a constant flow of events, actions, and interactions. Its complexity lies in its grammatical intricacy—that is, in how many clauses are flexibly, fluently, and dynamically linked through logical connections (Halliday 1989: 76ff.).

Written language (*World of Things*; Halliday 1989: 93) is static and product-oriented. It freezes the flow of events by nominalizing processes and transforming them into fixed, measurable "objects." This allows for a synoptic view: The world is represented as a stable system of things and categories that are logically related to one another (Halliday (1989: 61ff.).

This functional and situational understanding of written and spoken language inevitably has profound implications for the syntactic structure and grammatical architecture of a text, as the different communicative conditions are directly reflected in lexico-grammatical patterns. This structural realization of register variation manifests itself at the micro-linguistic level in a changed distribution of vocabulary as well as in the density of specific word classes – dimensions that can be empirically measured using quantitative measures. As Stubbs notes *"*Speech and writing are polarized terms referring to clusters of characteristics which typically occur together, but which are logically independent. Thus much written language is standard, formal, planned, edited, public and non-interactive, whereas spoken language is typically casual, spontaneous, private and face-to-face*"* (Stubbs 1996: 64).

The quantitative research of political speeches requires a theoretical model that understands linguistic variation as a functional correlate of the respective communication situation. A central foundation for this is provided by register theory, which assumes that lexico-grammatical patterns are significantly shaped by the situational context and the communicative purpose. According to Biber (2012), fundamental differences in syntactic and lexical organisation can be demonstrated along the continuum between conversational orality and informational written prose.

While spoken registers rely heavily on a clause-based grammar, informative, written registers are characterized by a phrasal grammar. The clausal grammar of spoken language manifests itself in a significantly higher frequency of finite clauses (such as adverbial and complement clauses) as well as pronouns and adverbials. In interactive exchange, these grammatical structures serve to dynamically establish syntactic relationships and secure argumentative connections. In contrast, the phrasal grammar of written, informative texts

condenses information into complexly modified noun phrases. Typical features of this style are a high density of nouns, attributive adjectives, and prepositional modifiers (Biber 2012: 24f).

These dimensions of register variation are further explored at the text level through the concept of text cohesion. Max Louwerse et al. (2004) demonstrate that spoken and written language differ particularly in their global and local cohesion mechanisms. While purely word-based analyses often blur the distinction between orality and literacy, the multidimensional analysis of cohesive features (Dimension 1: Speech versus writing) reveals a clear demarcation. Spoken registers are characterized by a high pronoun density, concrete, figurative language, and a high degree of ambiguous quantification. Written registers, on the other hand, rely on a dense network of local and global cohesion, realized through semantic sentence and paragraph similarities and explicit argument overlaps (Louwerse et al. 2004: 845).

1.2 Political language between spoken and written language

In Michael Gregory's (1967) system for differentiating varieties, linguistic variations based on the situational circumstances of actual language use are classified as diatypic varieties (Gregory 1967: 194). These reflect recurring functional features of how language is used in specific situations.

To systematically describe these socio-linguistical conditions, Gregory uses a model consisting of three situational categories and their corresponding contextual equivalents at the linguistic level: the purposive role of the user determines the field of discourse, the addressee relationship determines the tenor of discourse, and the medium relationship determines the mode of discourse (Gregory 1967: 184f).

The differentiation of political language into the "language of politicians" on the one hand and the "language of politics" on the other can be precisely categorized within Gregory's diatypical variation model from a register-theoretical perspective. While the person-centred, situational "language of politicians" is primarily linked to a phonic medium relationship, the highly institutionalized "language of politics" manifests itself predominantly in the graphic medium of writing. This dual orientation results in a hybrid positioning of political language, which, within the mode of discourse, spans a broad continuum of transitional forms.

Gregory's systematic sub-classification of the user's medium relationship allows for a precise representation of this functional range: it extends from forms of spontaneous speaking, such as dialogic conversation or argumentative monologuing in debates and interactive interviews, to the planned speaking of what is written, which in prepared political speeches deliberately imitates a rhetorical immediacy in the mode *"written to be spoken as if not written"*, to purely administrative texts, contracts, or legal texts, which are primarily conceived as "*written not necessarily to be spoken*" for silent reading (Gregory 1967: 189). The concrete lexico-grammatical form of these diverse varieties is determined in Gregory's paradigm by the continuous interaction of the discourse mode with the discourse field (shaped by the pragmatic intention or purposive role) and the discourse tenor (determined by the social role and the addressee relationship between actors and the public).

Political speeches are generally prepared and thus pre-written, even if the speaker may stage an artificial spontaneity. Depending on how closely the written draft is aligned with the actual speaking situation, political speeches can be classified into two subcategories:

A. Written to be spoken as if not written

The speech largely creates its own situation and its own contextual relationship patterns, has a clearly defined beginning and end, and is significantly more compact and self-contained than spontaneous conversations.

These texts exhibit a statistically higher occurrence of the grammatical patterns favoured by spontaneous speech than other written texts. At the same time, they differ from genuine spontaneous speech in that they use pronouns and deictics without intra-textual references far less frequently, since the text largely autonomously constitutes its situation and must be self-sufficient.

B. Written to be spoken[1]

Other speeches make no attempt to conceal their written origin. Since listeners cannot "turn back a page" during an oral presentation to review what has been said, these texts must take into account the irreversible, sequential nature of auditory information processing. Linguistically, this manifests itself in a high frequency of planned, deliberate, and varied repetitions (premeditated varied repetition) at the grammatical and, in particular, lexical levels, distributed throughout the text to ensure comprehensibility and rhetorical effectiveness.

Beyond mere mode, political speeches are structured both in terms of content and social context by the dimensions of field and tenor. Correlating with the speaker's purposive role, the Field of Discourse determines whether the speech assumes a specialized (technical) or non-specialized role. Specialized roles require a restricted collocation range and the frequent occurrence of specific grammatical patterns (such as passive constructions or complexly modified noun phrases for precise and economical definition; Gregory 1967: 186).

1.3 Token-type-relation and lexical density

The central starting point for any quantitative text analysis is the fundamental distinction between types (unique word forms) and tokens (the total number of all running word forms). The standard type-token-ratio (TTR) expresses the percentage ratio of unique word forms to the total number of words in a text.

Research distinguishes between two interpretations of the TTR, as a characteristic of vocabulary richness of a text or as a model for information flow in a text. However, in modern quantitative and applied linguistics, the TTR is considered highly problematic and of only very limited value for measuring actual lexical richness, which is why the TTR is used more as a metric for the information flow of a text (Wimmer 2005: 361f; Popescu et al. 2009: 232f).

[1] Gregory (1967, 192f).

In its standard form, however, the TTR has a significant deficiency: it does not distinguish between semantically meaningful content words and purely grammatical function words, which diminishes its significance for the actual lexical richness.

To capture the lexical structure of a text more precisely, vocabulary is functionally divided into two main categories (Johansson 2008: 66).

- **Content words (autosemantic lexemes)**: nouns, main verbs, and adjectives that convey the actual conceptual core and semantic content.
- **Function words (synsemantic lexemes)**: conjunctions, pronouns, prepositions, auxiliary verbs and articles that establish syntactic relationships and ensure grammatical cohesion.

In stylometrics it is assumed that high-frequency function words function as unconscious "minor encoding habits" of the author, remaining stable completely independent of the text's content. Therefore, statistical analyses enable corresponding author identification (Tuldava 2005). However, Fred J. Damerau (1975) proposes a significant limitation to this assumption. Using Poisson distribution tests[2], he demonstrates that the frequency of function words in real texts is by no means purely random or mechanically distributed. Instead, many function words (especially pronouns like "he" or "his") exhibit considerable context and topic dependence (Damerau 1975: 272).

Furthermore, open and closed word classes behave fundamentally differently with regard to their lexical variability: While open word classes fluctuate strongly thematically and thus exhibit a high and volatile TTR, closed word classes show an extremely low, almost constant TTR and a flat distribution structure across all registers (Van Gijsel et al. 2006: 959ff.).

This leads to a syntactic paradox: Complex sentence structures (such as causal connectives) necessitate a high frequency of function words.

These massively inflate the number of tokens in the denominator but provide relatively few new types for the numerator, which statistically pushes the overall TTR downwards.

The systematic alternation between a nominally dense, function-word-poor phrasal grammar (high lexical density) and a syntactically linked, function-word-rich clausal grammar (low lexical density) thus directly reflects the specific rhetorical and pragmatic purpose of a speech situation.

From this functional division of vocabulary, the measure of lexical density is directly derived, which expresses the ratio of content words to function words or to the total text length and serves as a primary indicator of the information content, the administrative character and the syntactic complexity of a text. Early linguistic theoretical studies showed that a high lexical density primarily occurs in highly condensed written non-fiction texts, while spoken language requires a significantly higher proportion of structural and function words and therefore has a

[2] The Poisson distribution models the probability of a given number of discrete events occurring within a fixed interval or sample space under the assumption of complete independence and a constant average rate.

comparatively lower lexical density (Johansson 2008: 65). More generally, the following can be said: written texts have a much higher lexical density, while spoken language usually has a low density (Stubbs 1996: 72ff.).

Stubbs uses the following formula for lexical density: Lexical density = 100 × L/N[3] (Stubbs 1996: 72). This normalises the value and makes it directly comparable across texts of different lengths. However, Halliday takes a different approach. According to him, the clause is the primary grammatical unit in which human experience and statements (processes, actions or events) are linguistically structured. Therefore, he relies on the clause when calculating the lexical density. Lexical density as the number of content words per clause (instead of the total number of words) directly measures "information packaging" – i.e. the question of how tightly a speaker or writer packs information into a single cognitive processing unit. At the same time, he also mentions that the definition of a partial sentence can also be problematic (Halliday 1989: 61ff.).

Despite its widespread use, the classic type-token relation as well as the lexical density suffers from a fundamental mathematical weakness: it is extremely dependent on the total length of the text. Since words inevitably repeat themselves over the course of a longer text, the vocabulary growth curve flattens out degressively. As a result, longer texts always have a lower TTR than shorter texts, purely mathematically speaking, because the growth rate of new words eventually converges to zero in sufficiently long texts (Popescu 2009: 36). Raw TTR values from corpora or text segments with unequal text lengths are therefore not directly comparable statistically. Moreover, this distribution is not a static phenomenon but is subject to mathematical regularities, which are described, among other things, by Zipf's law. While extreme statistical distortions can occur in very short texts, and these relationships only stabilize in longer texts, increasing text length simultaneously necessitates more complex syntactic constructions. To ensure argumentative coherence across longer stretches of text, an increasing number of functional "bracket" words (structural function words) are required. Because these closed word classes recur frequently in a continuous text, they massively inflate the number of tokens in the denominator without providing any significant new types for the numerator. This grammatical redundancy artificially lowers the TTR with increasing text volume and thus systematically distorts the measurement of actual lexical richness.

To conclude, both metrics suffer from a fundamental mathematical weakness: they are extremely text length dependent. Since words inevitably repeat themselves as the length of the text increases, the curve of vocabulary growth flattens out, which means that, purely mathematically, longer texts always have a lower TTR than shorter texts. Since the growth rate of new words converges to zero for sufficiently long texts, raw TTR values across texts of unequal length are not directly statistically comparable.

[3] L = Number of lexical items, N = Number of all items (Token).

1.4 Methodological limitations and challenges of classic quantitative approaches

The quantitative modelling of linguistic structures encounters a fundamental methodological hurdle in empirical practice: almost all statistical measures based on word frequency distributions vary systematically and non-linearly with text length. Koplenig (2024) demonstrates that supposedly universal linguistic constants such as "information density" or "semantic density" are, in reality, almost entirely mathematically determined by text volume. Since vocabulary grows less than proportionally but steadily with text length according to Herdan's law, an increasing number of tokens leads, purely statistically, to a changed frequency structure. If this length effect is not methodically controlled (for example, through mathematical normalisation, residual analysis, or dividing the corpus into equally sized segments/chunks) massive spurious correlations threaten to artificially generate linguistic effects.

In order to compensate for this length dependence of the TTR, various formulas and mathematical transformations for linearisation have been proposed in quantitative linguistics.

Van Hout and Vermeer (2007) compare various TTR alternatives designed to reduce the influence of token size. These include:

- **Index of Guiraud (V / √N)**[4]: This formula uses the square root of the token count in the denominator.
- **Index of Herdan (log V / log N)**: A logarithmic transformation.
- **D-Measure / VOCD**: A computer-based resampling method that compares the empirical TTR curve of a text with a mathematical model curve as token size increases, in order to estimate the stable parameter D.
- **Advanced Guiraud (AG) and Measure of Lexical Richness (MLR)**: Frequency band-based correction methods that weight words based on their rarity in large reference corpora, since rare words have been mathematically proven to indicate greater richness.

In the empirical comparisons of Van Hout and Vermeer (2007), Guiraud often performs best, since taking the square root represents a suitable middle ground, as it does not overemphasize the denominator in contrast to the uncorrected TTR and at the same time avoids the excessive logarithmic damping that occurs in Herdan's method, which artificially flattens stylistic differences (Van Hout and Vermeer 2007: 136).

Besides mentioning the Index of Guiraud and VOCD, Johansson additionally names Theoretical Vocabulary as an alternative. This is a random sampling method in which samples of a fixed size (e.g. 100 words) are repeatedly drawn from the text to compare the number of types, independent of length (Johannson 2008: 63f).

These TTR corrections primarily aim to linearize the purely mathematical dependence of vocabulary diversity on text length, but they all share a fundamental conceptual weakness: they treat vocabulary as a structurally homogeneous mass, although the TTR behaves

[4] N is the number of words (Token) and V is the number of different words (Type).

differently when broken down to POS (part of speech) (Van Gijsel et al. 2006: 959ff.). These purely quantitative indices do not differentiate between the semantic function of individual word forms and neglect the fact that closed word classes (function words) and open word classes (content words) follow completely different mathematical and syntactic rules in discourse.

Although lexical density provides a more robust picture of the informational character, its empirical implementation suffers from the need for prior, often error-prone syntactic segmentation or automated word-class classification. However, the problem can begin with what is even defined as a word and therefore as a POS (Popescu 2009: 5ff.). Since the TTR seems to be influenced by POS, considerations regarding the POS should not be neglected when dealing with TTR and lexical density (Van Gijsel et al. 2006: 962). Furthermore, the boundary between grammatical and lexical units remains linguistically fluid, which makes an objective, distribution-independent separation difficult.

A mathematically elegant way out of this dilemma is offered by Popescu with the concept of the h-point. The h-point is based on the fixed-point theory of continuous functions and is defined for discrete frequency distributions as the point on a rank-frequency curve at which the rank *r* of a word exactly corresponds to its absolute frequency *f(r)* (r = f(r)) (Popescu 2009: 17ff.).

From a linguistic-theoretical perspective, the h-point functions as an objective, intrinsic dividing line of the vocabulary, which does not require external dictionaries or subjective POS categorisations.

The pre-h area (r ≤ h): By definition, the high-frequency grammatical function words (synsemantics such as articles, pronouns, and prepositions) gather in the foremost positions before the point h, ensuring the syntactic cohesion of the text.

The post-h area (r > h): Behind this lies the much larger class of rarer content words (autosemantics), which carry the actual semantic load and theme of the text (Popescu 2009: 17).

Popescu demonstrates that the h-point not only objectively divides the functional vocabulary but can also be used for the precise mathematical modelling of text length dependency. Since the h-point is related to the total text length (N), the disruptive length effect can be completely neutralized via the invariant parameter $a = N/h^2$. On this basis, Popescu defines a length-independent measure of actual vocabulary richness with the index R_1 ($R_1 = 1 (F(h)^5 – (h^2 / 2N))$) (Popescu 2009: 33).

By subtracting the cumulative relative frequency of function words up to the h-point the index R_1 measures the pure diversity of the thematic vocabulary beyond the grammatical auxiliary constructions.

Another fundamental methodological challenge for quantitative measures such as the TTR and lexical density arises from the significant typological differences between synthetic

[5]F(h) is the cumulative relative frequency up to the h-point.

and analytic languages. While analytic languages like English primarily encode grammatical relationships through word order and separate, independent function words, synthetic and agglutinative languages (such as Georgian) rely on a complex internal word structure. Here, tense, aspect, subject, or case information is morphologically integrated directly into a single word form. As Koplenig demonstrates, this means that synthetic languages require significantly fewer words (tokens, N) than analytic languages to convey the same propositional message (Koplenig 2024: 18). Since classical unigram-based approaches and information-theoretic measures completely ignore internal morphological complexity and word structure, this leads to systematic distortions in cross-linguistic comparisons. For TTR, this means that synthetic languages, due to their smaller number of tokens, can systematically exhibit different (often higher) TTR values for the same content, as shorter texts mathematically achieve higher ratios than longer ones. A similar situation exists with lexical density as it is highly language-dependent (Johannson 2008: 65f). Languages with a strongly bound morphology exhibit a systematically higher proportion of content words, since grammatical relations are realized as affixes directly attached to the word and do not appear as separate function words in the text, artificially inflating the density. Popescu arrives at similar results when comparing various synthetic and analytical languages, the differences are mathematically visible (Popescu 2009: 24, 43, 47f).

Additionally, the linguistic phenomenon of homonymy represents a fundamental source of error in the automated calculation of quantitative metrics like TTR and lexical density. Since standard, unigram-based word counts segment texts purely graphemically at the token level, identical orthographic forms are counted as one and the same type, even if they belong to completely different classes syntactically and semantically. If measures such as TTR are to serve as a representation of semantic information flow or lexical diversity, such homonymous forms must necessarily be counted as distinct types. In languages where homonyms occur more frequently, the systematic oversight of these subtle semantic distinctions leads to an artificial levelling of the type count in raw word form counts without prior morphosyntactic tagging, thus distorting the measurement results (Popescu 2009: 234).

In corpus linguistic practice and in the standard calculation of the type-token ratio (TTR), punctuation is conventionally ignored or completely removed during text cleaning, since a word form is usually strictly defined graphemically as a sequence of letters between two spaces. However, for the analysis of political language, it can be argued that this methodological exclusion distorts the rhetorical and structural reality of the text. Political speeches occupy a special hybrid position: They belong to the register of “written to be spoken”. To recap what has been said, the communication of information must be structured in a highly sequential and rhythmic manner.

In this rhetorical framework, punctuation assumes a fundamental functional role in the written concept of speech: it is the graphic correlate of acoustic-prosodic design. Punctuation marks encode pauses, speech rhythms, changes in pace, and rhetorical caesuras, which are essential for creating an artificial sense of spontaneity.

Although punctuation marks are not autosemantic lexemes with independent conceptual content, they thus have a direct impact on text semantics and cognitive information processing. They mark the boundaries of clauses, which, according to Halliday, represent the primary units of meaning for structuring human experience. The placement of periods, commas, or dashes controls logical relationships (such as causality or contrast), disambiguates meanings, and shifts semantic weight within the sentence. They function as visual and rhetorical brackets that structure the flow of argumentation across larger sections of text and support text cohesion.

The explicit consideration of punctuation in quantitative analyses of political discourses (not as lexical units, but as structural and semantically distinct functional elements) therefore offers a methodological advantage. It makes it possible to operationalize the complex interplay of rhythmic structure, syntactic complexity, and rhetorical pragmatism that fundamentally characterises the strategic architecture of political language.

Nevertheless, the investigation of punctuations is extremely difficult to systematize in empirical practice. Punctuation marks cannot be counted as homogeneous units analogous to simple orthographic word forms, since they fulfil completely different pragmatic functions depending on the syntactic context. Therefore, in order to validly include these structural elements in quantitative analyses, a theoretically and methodologically sound model must be developed first.

2. Methodology

The empirical basis of this study is a text selection from 25 speeches by Zviad Gamsakhurdia.[6] Methodologically, this work follows a "Type B" study design (Biber 2012: 12f). This means that the register-analytic comparison does not take place at an abstract system level, but rather focuses on the quantitative density and distribution of lexico-grammatical features in real, individual texts as primary units of observation.

In order to systematically capture the diatypical variation in the corpus, the 25 texts are assigned to three functional registers. These can be classified according to the taxonomy of the medium relationship and the mode of discourse established by Gregory (1967).

1. Official announcements (declarations): These are highly formal, written administrative texts with a highly informative and impersonal character. In Gregory's scheme, these texts fall under speaking non-spontaneously, or more precisely under the category of "the speaking of what is written".

2. Direct addresses (appeals): These speeches, aimed at the general public or the Georgian people, have a strong persuasive, mobilizing and appealing character. They are also essentially part of speaking what is written. In contrast to purely administrative announcements, however, they attempt to fake conventional immediacy and rhetorical spontaneity. They are therefore classified as "written to be spoken as if not written". According to Gregory, political-rhetorical speeches and sermons explicitly fall under this type because, although they represent planned,

[6] These and other speeches can be found in the political subcorpus of the Georgian National Corpus (http://gnc.gov.ge/).

prepared behaviour, they specifically imitate the typical favourite grammatical patterns of spontaneous speech.

3. Interviews: These interactive, more spontaneous or semi-structured question-and-answer scenarios with journalists have a distinct argumentative justification structure. In contrast to prepared speeches, they are based on the situational primary category of speaking spontaneously.

For the quantitative analysis of the texts, an analysis program implemented in Python (Kamarauli & Tsetskhladze 2026) and a tool for analysing function words written in Java (Kamarauli & Kamarauli 2025). were used. In the GNCAnalyzer, which is based on the GNC parsing API, the punctuation marks had to be retained in the first step because the downstream sentiment model required the continuous text to be segmented into syntactically complete sentences. However, this inclusion of punctuation marks was no longer necessary to determine the function words and the lexical variety, which is why the punctuation was subsequently removed for the final calculation of the statistical key figures here.

In addition, minimal methodological differences in tokenisation and classification led to minor deviations in the raw data generated: In the GNCAnalyzer, the recognition of the function words was based on an automated stop word list of the statistically most frequently occurring word forms of the GNC, whereas the FunctionalWordAnalyzer operates on the basis of a manually compiled, dedicated list of Georgian function words. The absolute total token count also differs slightly because the GNCAnalyzer uses the tokenisation of the GNC parser, while the FunctionalWordAnalyzer used a simple graphemic filter that extracted all pure word forms without taking complex morphosyntactic structures into account. For the final evaluation, the punctuation-corrected token and type values from the GNCAnalyzer were used and merged with the more precise function word frequencies from the FunctionalWordAnalyzer. These discrepancies can be used to clearly illustrate how sensitive quantitative style measures and information-theoretic indices react to even minimal, seemingly trivial deviations in data processing.

#	Text	Text Type
1	მიმართვა ქართველი ერისადმი 24.06.1992	Direct address
2	მიმართვა 21.06.1992	Direct address
3	მიმართვა 12.07.1992	Direct address
4	განცხადება 24.06.1992	Official announcements
5	განცხადება 26.06.1992	Official announcements
6	მიმართვა რუსეთის პრეზიდენტს 07.07.1992	Direct address
7	მიმართვა	Direct address

8	პასუხი შეკითხვებზე 31.07.1992	Interview
9	განცხადება 07.08.1992	Official announcements
10	მიმართვა 05.08.1992	Direct address
11	მიმართვა 09.08.1992	Direct address
12	განცხადება 12.07.1992	Official announcements
13	განცხადება 14.08.1992	Official announcements
14	მიმართვა 21.08.1992	Direct address
15	მიმართვა 23.08.1992	Direct address
16	მიმართვა 05.08.1992	Direct address
17	მიმართვა 19.08.1992	Direct address
18	მიმართვა 19.08.1992	Direct address
19	განცხადება 05.09.1992	Official announcements
20	მიმართვა 07.09.1992	Direct address
21	ინტერვიუ „ნაროდნაია პრავდა", №41	Interview
22	პასუხი შეკითხვებზე 19.06.1993	Interview
23	ღია წერილი 05.02.1992	Direct address
24	ინტერვიუ 12.06.1992	Interview
25	პრესკონფერენცია 03.07.1991	Interview

Table 1: Analysed Texts

Based on these values, the simple TTR was calculated for the respective text, then the Guiraud index and the lexical density were calculated. Only Gamsakhurida's text passages were taken into account in the interviews, so the values do not refer to the entire interviews. In the following chapter the results will be presented and discussed.

3. Empirical Findings and Register-Analytic Results

To provide a systematic overview of the quantitative and structural composition of the corpus under investigation, the following table presents the basic statistical parameters for all 25 analysed texts by Zviad Gamsakhurdia.

In addition to the absolute frequencies of word occurrences (tokens) and unique word forms (types), the table also records the differentiated proportions of function and content words, as well as the readability score, the classic TTR, the length-corrected Guiraud index, and the percentage lexical density for each individual text contribution.

#	Token	Type	Funct. Words	Lex. Words	Readability	TTR	Guiraud	Lexical Density
1	147	117	15	133	65.81	0.80	9.65	90.48
2	363	198	42	308	78.19	0.55	10.39	84.85
3	176	110	28	145	76.45	0.63	8.29	82.39
4	171	117	26	141	81.32	0.68	8.95	82.46
5	218	152	24	190	69.41	0.70	10.29	87.16
6	128	86	15	113	80.95	0.67	7.60	88.28
7	219	141	23	196	87.45	0.64	9.53	89.50
8	2132	824	397	1729	57.29	0.39	17.85	81.10
9	555	297	91	466	74.36	0.54	12.61	83.96
10	592	352	98	497	68.11	0.59	14.47	83.95
11	539	307	75	455	66.58	0.57	13.22	84.42
12	95	65	12	83	62.44	0.68	6.67	87.37
13	131	101	13	118	81.84	0.77	8.82	90.08
14	36	30	3	33	87.42	0.83	5.00	91.67
15	108	73	15	93	77.00	0.68	7.02	86.11
16	609	345	96	516	70.24	0.57	13.98	84.73
17	118	99	13	107	73.34	0.84	9.11	90.68
18	597	319	87	506	75.78	0.53	13.06	84.76
19	251	178	42	209	73.93	0.71	11.24	83.27
20	534	291	65	467	85.62	0.54	12.59	87.45
21	3232	1164	546	2663	61.32	0.36	20.47	82.39
22	913	471	173	750	67.94	0.52	15.59	82.15
23	1452	658	229	1206	73.85	0.45	17.27	83.06
24	648	369	102	538	58.35	0.57	14.50	83.02
25	3224	996	619	2594	54.71	0.31	17.54	80.46

Table 2: Quantitative Analysis

The data clearly illustrates the declining behaviour of TTR: The highest TTR values are found in the shortest texts, such as #14 (36 tokens, TTR = 0.83) and #17 (118 tokens, TTR = 0.84).

Conversely, the longest texts exhibit the lowest TTR values, particularly #25 (3,224 tokens, TTR = 0.31) and #21 (3,232 tokens, TTR = 0.36). This systematic decrease in TTR is purely mathematical, as grammatical function words inevitably repeat themselves with increasing text length.

The Guiraud index, which attempts to mitigate this length effect by taking the square root of the denominator, significantly alters the picture:

While text #14, despite its extremely high TTR of 0.83, achieves only a very low Guiraud value of 5.00, the longest text (#21), despite a low TTR of 0.36, exhibits the highest vocabulary richness according to Guiraud (20.47).

This empirically demonstrates that while longer texts show a lower relative growth rate of new words due to statistical saturation, the absolute diversity of the developed vocabulary is, as expected, far greater.

The lexical density is high in all texts and almost consistently exceeds 80%. This suggests that Gamsakhurdia's political language, even in spoken context, possesses a highly information-rich, nominally structured form typical of carefully prepared written texts.

To systematically capture the diatypical variation, the 25 texts were divided into three functional registers. This classification reflects different characteristics of the medium and the rhetorical purpose.

3.1 Direct Addresses

Text	Token	Type	Funct. Words	Lex. Words	Readability	TTR	Guiraud	Lexical Density
1	147	117	15	133	65.81	0,80	9,65	90,48
2	363	198	42	308	78.19	0,55	10,39	84,85
3	176	110	28	145	76.45	0,63	8,29	82,39
6	128	86	15	113	80.95	0,67	7,60	88,28
7	219	141	23	196	87.45	0,64	9,53	89,50
10	592	352	98	497	68.11	0,59	14,47	83,95
11	539	307	75	455	66.58	0,57	13,22	84,42
14	36	30	3	33	87.42	0,83	5,00	91,67
15	108	73	15	93	77.00	0,68	7,02	86,11
16	609	345	96	516	70.24	0,57	13,98	84,73
17	118	99	13	107	73.34	0,84	9,11	90,68
18	597	319	87	506	75.78	0,53	13,06	84,76
20	534	291	65	467	85.62	0,54	12,59	87,45
23	1452	658	229	1206	73.85	0,45	17,27	83,06

Table 3: Quantitative Analysis - Direct addresses

The direct addresses and appeals (a total of 14 texts) are characterized by a highly persuasive and mobilizing nature. They are formally prepared, but deliberately imitate rhetorical immediacy in the "*written to be spoken as if not written*" mode.

This register exhibits the greatest internal variance. It ranges from extremely short like #14 (36 tokens) to comprehensive, programmatic speeches like #23 (1,452 tokens).

The degressive vocabulary growth can be seen in this register. The ultra-short address (#14) achieves a TTR of 0.83, while the longest appeal (#23) drops to a TTR of 0.45, but increases to a significantly higher Guiraud score of 17.27.

Despite its persuasive, orally oriented character, the lexical density remains extremely high, ranging from 82.39% (#3) to 91.67% (#14). Gamsakhurdia achieves a rhetorical balancing act here: he conveys highly condensed, metaphorical, and patriotic core concepts (high lexical density), yet clothes them in the rhythmic and syntactic garb of emotionalizing oral speech.

3.2 Official Announcements

Text	Token	Type	Funct. Words	Lex. Words	Readability	TTR	Guiraud	Lexical Density
4	171	117	26	141	81.32	0,68	8,95	82,46
5	218	152	24	190	69.41	0,70	10,29	87,16
9	555	297	91	466	74.36	0,54	12,61	83,96
12	95	65	12	83	62.44	0,68	6,67	87,37
13	131	101	13	118	81.84	0,77	8,82	90,08
19	251	178	42	209	73.93	0,71	11,24	83,27

Table 4: Quantitative Analysis - Official Announcements

The official announcements (#4, #5, #9, #12, #13, #19) represent administrative, highly informative written texts. The texts in this register are consistently short to medium in length (95 to 555 tokens). The lexical density is exceptionally high and stable on average, ranging from 82.46% (#4) to 90.08% in #13.

This finding supports Biber's model of a strongly nominal phrasal grammar. To formulate administrative matters precisely and impersonally, these texts rely on complex noun phrases and high-frequency content words, while the proportion of structural function words (such as pronouns or conjunctions) is deliberately minimized.

3.3 Interviews

Text	Token	Type	Funct. Words	Lex. Words	Readability	TTR	Guiraud	Lexical Density
8	2132	824	397	1729	57.29	0,39	17,85	81,10
21	3232	1164	546	2663	61,32	0,36	20,47	82,39
22	913	471	173	750	67.94	0,52	15,59	82,15
24	648	369	102	538	58.35	0,57	14,50	83,02
25	3224	996	619	2594	54,71	0,31	17,54	80,46

Table 5: Quantitative Analysis - Interviews

The interviews (#8, #21, #22, #24, #25) fall under the category of spontaneous or semi-structured speech in dialogic or argumentative contexts. Since only Gamsakhurdia's own contributions were analysed, the data reflect his genuine oral argumentation practice.

These texts represent the most extensive documents (three of the five texts significantly exceed the 2,000-token threshold).

Due to their considerable length, the interviews exhibit the statistically lowest TTR values (0.31 in #25; 0.36 in #21; 0.39 in #8). However, when this length effect is adjusted using the Guiraud Index, it becomes clear that Gamsakhurdia's active vocabulary was exceptionally rich in these free-speaking situations, as evidenced by the peak values of 20.47 (#21) and 17.85 (#8).

Significantly, the interviews exhibit the lowest lexical density, consistently ranging from 80.46% (#25) to 83.02% (#24). This empirically confirms Halliday's theory of spoken language: in spontaneous dialogue, interactive relationship building and dynamic argumentative linking require a clausal grammar. Gamsakhurdia must increasingly utilize functional linking elements such as pronouns, conjunctions, and relative pronouns, which inflates the number of tokens in the denominator and systematically reduces the lexical density compared to the prepared declarations.

4. Discussion and Theoretical Implications

The empirical application of quantitative metrics to Georgian is further complicated by language-specific morphosyntactic features that systematically distort basic statistical metrics, such as the TTR and Lexical Density, which rely on precise proportions of lexical (content) and functional (grammatical) words. A primary source of statistical bias is the behaviour of Georgian postpositions, which function as the structural equivalents of English prepositions. In Georgian, these elements are split into two categories: cliticised (bound) postpositions, which are orthographically merged with the preceding noun, and free-standing (unbound) postpositions. Word-level tokenizers which segment texts purely based on whitespace and punctuation, fail to capture cliticised functional elements as distinct grammatical units.

This phenomenon can be illustrated in Zviad Gamsakhurdia's speeches:

(1) Text #24

sakartveloši arasodes šeċq̇veṭila brʒola damouḳideblobisatvis.

sakartvelo-ši	*arasodes*	*šeċq̇veṭila*	*brʒola*
Georgia.DAT.SG-in	never	stop.S3SG.PERF	fight.NOM.SG

damouḳideblob-is-a-tvis

independence-GEN.SG-EMPH.V-for

'In Georgia, the fight for independence has never ceased.'

(2)

(2) Text #24

mosḳovs ʒalian sc̣irdeba isini damonebul respu̇bliḳebši baṭonobisatvis.

mosḳov-s	*ʒalian*	*sc̣irdeba*	*isini*	*damonebul*
Moscow-DAT.SG	very	need.S3SG.PRES	they.NOM.SG	enslaved.DAT.SG

respu̇bliḳ-eb-ši	*baṭonobisatvis*
republic-PL-in	domination-GEN.SG-EMPH.V-for

‘Moscow needs them very much to rule in the enslaved republics.’

From a graphemic perspective, sentence (1) consists of exactly five words. However, a morphosyntactic analysis reveals that two essential functional elements are statistically “lost” because they are bound as suffixes: the locative postposition *-ši* in *sakartveloši* (საქართველოში) and the causal-final postposition *-tvis* in *damouḳideblobisatvis* (დამოუკიდებლობისათვის). Similarly, sentence (2) contains seven orthographic words, yet it conceals two functional clitics: the locative *-ši* in *respu̇bliḳebši* (რესპუბლიკებში) and the final *-tvis* in *baṭonobisatvis* (ბატონობისათვის). Consequently, simple quantitative evaluations artificially deflate the functional token count in the denominator, which in turn distorts the TTR and inflates the Lexical Density index.

Furthermore, Georgian morphosyntax permits the coordinate ellipsis of postpositions. When two coordinate nouns are linked by the conjunction *da* (და/“and”), the postposition is dropped from the first conjunct and attached exclusively to the second noun. Gamsakhurdia utilises this rhetorical structure to appeal to dual audiences:

(3) Text #10

mimartva kartveli xalxisa da msoplio sazogadoebriobisadmi.

mimartva	*kartvel-i*	*xalx-is-a*	*da*
appeal.NOM.SG	Georgian-GEN.SG	people-GEN.SG-EMPH.V	and

msoplio	*sazogadoebriob-is-a-dmi*
world.GEN.SG	community-GEN.SG-EMPH.V-to

‘Appeal to the Georgian people and the world community.’

An underlying structural decomposition of this phrase reveals a deep-level syntactic symmetry: მიმართვა ქართველი ხალხის-ა-**დმი** და მსოფლიო საზოგადოებრიობისადმი.

Functionally and semantically, the postposition *-dmi* governs both noun phrases. However, on the surface morphosyntactic level, it is realized only once. The first noun *xalxisa* (ხალხისა) is grammatically marked with the genitive case ending followed by the adpositional agreement suffix *-a*, which acts as a structural placeholder for the elided postposition. While a human reader semantically reconstructs the first phrase as *xalxisadmi* (ხალხისადმი), standard automated tokenizers and POS-taggers register only one postposition (*-dmi* as an adposition) and treat *xalxisa* as a bare inflected noun. This structural asymmetry demonstrates that raw graphemic text processing is fundamentally inadequate for synthetic-agglutinative languages like Georgian. Without a deep-level parser that is capable of segmenting multi-word tokens into their constituent lexical and functional morphemes, statistical measures of lexical richness and syntactic weight remain heavily biased.

4.1 Manual Comparison

To empirically test the theoretical necessity of clitic separation and to measure the precise mathematical distortion caused by standard graphemic tokenisation, a manual comparative analysis was conducted on two representative texts – #10 (direct address, "written to be spoken") and #24 (interview).

The contrastive dataset compares the raw graphemic metrics against a clitic-separated model, where bound postpositions are treated as independent functional tokens. This morphological expansion adds exactly 8 unique postpositional types to both texts.

Metric	#10	#24
Token	592	648
Type	325	369
TTR	0.59	0,57
Functional Words	98	102
Guiraud	14,47	17,27
Bound Postpositional Forms	27	18
Postpositions / Token	≈ 4.56 %	≈ 2.77 %
Postpositions / Type	≈ 8.31%	≈ 4, 88 %
Token with Postpositions	619	666
Type with Postposition	333	377
TTR with Postpositions	0,47 (- 0,12)	0,55 (- 0,02)
Guiraud with Postpositions	13,38 (- 1,09)	14,61 (- 2,66)

Table 6: Statistic with postpositions

Although both texts are of highly comparable graphemic length (592 vs. 648 tokens), they display a striking divergence in their reliance on bound postpositions. In #10 cliticised postpositions constitute 4.56 % of all tokens and 8.31 % of types. In #24 this density drops sharply to 2.77 % of tokens and 4.88 % of types.

This difference strongly correlates with Douglas Biber's register model. #10 utilises an elaborated phrasal grammar to package complex socio-political concepts and patriotic appeals into dense nominal phrases, which rely on postpositional case-marking and spatial-causal modifiers. #24 shifts toward a verbal clausal grammar characteristic of oral production, where logical relations are expressed through coordinate clauses and independent finite verbs rather than highly synthesized nominal suffixes.

When clitics are separated from their nominal hosts, we witness a counter-intuitive decline in both TTR and Guiraud values. While clitic separation increases the unique vocabulary size slightly (adding the 8 postpositional types to the numerator), it inflates the running token count (the denominator) by a far greater margin (+27 tokens for #10 and +18 tokens for #24). This disproportionate growth in the denominator systematically flattens the vocabulary growth curve, causing both indices to decrease.

However, the magnitude of this drop seems to be register-sensitive. In #10 TTR drops by -0.12 (from 0.59 to 0.47), and Guiraud drops by -1.09. In the verb-heavy #24 TTR only drops by a marginal -0.02 (from 0.57 to 0.55), though Guiraud decreases by -2.66 due to the larger overall text length.

4.2 Functional Phrases

Beyond the morphosyntactic level of individual word forms, a further challenge for computational linguistics arises from the existence of functional phrases and multi-word functional expressions in Georgian.

(4) Text #8

ra tkma unda, aseti saprtxe arsebobs.

ra tkma unda	*aset-i*	*saprtxe*	*arsebobs*
of course	such-NOM.SG	threat.NOM.SG	exist.S3SG.PRES

'Of course, there is such a threat.'

A good example of this is the focus phrase *ra tkma unda* (რა თქმა უნდა "of course"), which is frequently used in political discourse and functions semantically and pragmatically as a fixed unit in the sense of "natural" or "self-evident."

Graphemically, however, this phrase consists of three syntactically independent words: the interrogative pronoun *ra* (რა / "what"), the verbal noun *tkma* (თქმა / "say"/ "speak"), and the modal verb *unda* (უნდა / "must"/"want").

In a standard, word-based computational linguistic count, this leads to a double distortion:

- The verbal noun *tkma* is incorrectly recorded as an independent content word, even though it is in the grammaticalisation process to a function word unit;
- The phrase artificially inflates the number of tokens in the denominator to three units, even though it should pragmatically be considered a single structural focus element.

Since the current empirical study analysed only isolated function words at the unigram-based level, this phenomenon remains statistically unaccounted for. Therefore, to obtain a methodologically complete and unbiased picture of the actual functional density and vocabulary richness in Georgian, it is essential for future research to systematically identify and disambiguate such formulaic function phrases using dedicated multi-word expression filters and to incorporate them as individual functional units into the statistical models.

4.3 Copula in Georgian

Besides the merging of postpositions, the morphosyntactic realisation of the copula in Georgian represents another systematic source of error for graphemic word counts and POS analyses. The high-frequency auxiliary verb *aris* (არის / "is"; 3rd person singular) can be systematically shortened in Georgian to the enclitic *-a* (-ა) and directly suffixed to the preceding nominal predicate (a noun, adjective, or pronoun) (Hardie & Daraselia 2026, 126f).

(5) Text #8

vpikrob, adamianis bedi missave xelšia.

vpikrob	*adamian-is*	*bed-i*	*mis-s-a-ve*
think.S1SG.PRES	human-GEN.SG	fate-NOM.SG	his/her-DAT.SG-EMPH.V-FOC

xel-ši-a

hand.DAT.SG-in-COP

'I think a person's fate is in their own hands.'

In a sentence like the one above, the verb thus graphemically merges into the single orthographic word form *xelšia* (ხელშია / is in the hand).

This cliticisation has immediate, distorting effects on quantitative text measures. Standard word class taggers or simple graphemic tokenizers incorrectly record the merged form *xelšia* as a single noun. As a result, the copula verb "disappears" completely from the statistical analysis.

It is particularly important to emphasize that this phenomenon is by no means a feature solely of informal, spoken language. The reduction of the copula to the suffix *-a* is a fixed, grammatically fully accepted, and standard feature of both spoken and formal written Georgian. It therefore occurs systematically in administrative declarations as well as in spontaneous speech. Without in-depth morphological segmentation that specifically isolates these enclitics,

any unigram-based determination of word class distribution and lexical density remains systematically flawed.[7]

5. Conclusion

This study presents a quantitative and register-analytical investigation of the political rhetoric of Georgia's first president, Zviad Gamsakhurdia, based on 25 selected texts.

The empirical analysis reveals a systematic variation in the TTR, the Guiraud index, and lexical density across different functional registers. While Gamsakhurdia's extemporaneous contributions in interviews exhibit a shift toward a function-word-rich, relationally linked clausal grammar, his texts as a whole are characterized by a high lexical density. This demonstrates the presence of a dense, nominalized phrasal grammar, a characteristic traditionally attributed in register linguistics to highly planned, written registers. The quantitative shift between a nominally dense, function-word-poor structure in official proclamations and a relationally linked structure in interviews thus reflects the speaker's deliberate functional shifts within political discourse. Gamsakhurdia's rhetoric is the result of dynamic self-regulation processes in which he clothes highly condensed, metaphorical and patriotic core concepts in the rhythmic garb of oral speech.

A methodological contribution of this work lies in the identification and systematisation of the typological barriers that arise when classical quantitative metrics are directly applied to Georgian as a highly synthetic language. While the mathematical dependence of the uncorrected TTR on text length has been successfully mitigated by the Guiraud index, empirical findings and theoretical analyses nevertheless demonstrate that unigram-based graphemic counts systematically distort the actual lexical structure of Georgian. The morphological complexity of Georgian leads to significant measurement errors on four levels:

a) **Cliticisation of Postpositions**: Postpositions (such as *-shi*, *-tvis*) graphemically merge with their nominal stem and are captured by standard, whitespace-based tokenizers as a single lexical token (noun). This "disappearance" of functional units artificially inflates the lexical density and depresses the TTR, as functional tokens are mathematically lost in the denominator. Therefore, treating enclitic elements separately as distinct tokens is essential for adequate quantitative modelling of the Georgian vocabulary.

b) **Coordinated Ellipsis of Postpositions**: Georgian syntax permits the coordinated ellipsis of postpositions, in which a postposition (such as *-dmi*) is realized only at the last element, even though it semantically governs both noun phrases (e.g., *xalxisa* da *sazogadoebriobisadmi*). Because automated taggers only capture the first noun as a genitive form without a postposition, a structural asymmetry arises that distorts quantitative frequency data.

c) **Copula cliticisation**: The high-frequency copula verb *aris* ("is") merges directly with nominal predicates (e.g., *xelšia,* strictly speaking, this is a location.) in its enclitic form

[7] See also Tandaschwili (2024) on the limitations of standard tools for Caucasian Languages.

-a. This renders the auxiliary verb statistically invisible in unigram-based POS counting, artificially overemphasizing the nominal nature of a text.

d) Functional phrases: Functional phrases like *ra tkma unda* ("naturally") are incorrectly counted as three independent words by word-based counters, and a nominal element (*tkma*) is incorrectly recorded as a content word, even though the phrase pragmatically functions as a single modal particle.

These findings support the methodological warnings of Tandaschwili (2024), who, using Caucasian languages as an example, demonstrates that conventional tools produce highly inaccurate statistics without adaptation to the grammatical properties of the respective target language. They also underscore the relevance of modern corpus-linguistic annotation models, which prescribe consistent multi-word tokenisation for the precise capture of clitics and postpositions, thus aligning theoretical modelling more closely with linguistic reality.

This leads to clear research objectives for future computational and register-linguistic investigations into Georgian. To enable valid comparisons of lexical richness and density, a two-stage statistical data processing procedure must be established as part of the pre-processing. At the lower level, the linguistic peculiarities of each language must be systematically controlled through morphosyntactic annotations and token splitting before cross-linguistic comparisons are made at the upper level. The development of automated filters for identifying and disambiguating formulaic multi-word expressions is necessary in order to feed these into the computational models as individual functional units.

In summary, this study demonstrates that purely statistical modelling of text data without a linguistically sensitive grammatical theory remains incomplete and leads to misinterpretations. Only the close interplay of quantitative metrics, precise computational linguistic tokenisation, and functional register theory allows for a valid reconstruction of the subtle strategic architecture of political language within the tension between nominal condensation and argumentative linking.

Abbreviations

3	3rd person
COP	copula
DAT	dative
EMPH.V	emphatic vowel
FOC	focus
GEN	genitive

NOM	nominative
S	subject
SG	singular
PERF	perfect
PRES	present
PL	plural

References

Biber 2012: Biber, Douglas, 'Register as a predictor of linguistic variation', in *Corpus linguistics and linguistic theory*, 8(1). Berlin/New York: De Gruyter, 9–37. https://doi.org/10.1515/cllt-2012-0002.

Damerau 1975: Damerau, Fred J., 'The use of function word frequencies as indicators of style', in *Computers and the Humanities*, 9(6). New York: Springer, 271–280. https://doi.org/10.1007/BF02396290.

Gregory 1967: Gregory, Michael, 'Aspects of varieties differentiation', in *Journal of Linguistics*, 3(2). Cambridge: Cambridge University Press, 177–198. doi:10.1017/S0022226700016601

Halliday 1989: Halliday, Michael A. K., *Spoken and written language*. Oxford: Oxford University Press.

Hardie & Daraselia 2026: Hardie, Andrew, & Daraselia, Sophiko, 'A theory for words in Georgian: traditional constructs versus corpus annotation', in Stefanie Wulff & Stefan Th. Gries (eds), *Corpus Linguistics and Linguistic Theory*, 22(1). Berlin/New York: De Gruyter, 115–138. https://doi.org/10.1515/cllt-2023-0107.

Johansson 2008: Johansson, Victoria, 'Lexical diversity and lexical density in speech and writing: A developmental perspective', in *Working papers/Lund University, Department of Linguistics and Phonetics*, 53, 61–79.

Kamarauli & Kamarauli 2025: Kamarauli, Mariam & Kamarauli, Anastasia, 'Unbound and Untamed? A Corpus-Based Exploration of Georgian Function Words', in *Digital Kartvelology*, 4, 160–190. https://doi.org/10.62235/dk.4.2025.10521.

Kamarauli & Tsetskhladze 2026: Kamarauli, Anastasia & Tatia Tsetskhladze, 'Towards Automatic Analysis of Political Speeches: A Prototype Using the GNC Parsing API', in Kamarauli, Mariam (ed.), *Caucasiology in the Digital Era. A Festschrift for Manana Tandaschwili*. Wiesbaden: Reichert Verlag, 111–144. https://doi.org/10.29091/9783752003437.

Koplenig 2024: Koplenig, Alexander, Text length strongly matters when analysing word frequency distributions. https://doi.org/10.31219/osf.io/p5nhd.

Louwerse et al. 2004: Louwerse, Max M., McCarthy, Philip M., McNamara, Danielle S. & Graesser, Arthur C., 'Variation in language and cohesion across written and spoken registers', in *Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 26, No. 26)*, 843–848. https://escholarship.org/uc/item/7d8631cr.

Popescu 2009: Popescu, Ioan I., *Word frequency studies (Vol. 64)*. Berlin/New York: De Gruyter.

Stubbs 1996: Stubbs, Michael, *Text and corpus analysis: Computer-assisted studies of language and culture*. Oxford: Blackwell.

Tandaschwili 2024: Tandaschwili, Manana, 'Statistics in Empirical Translation Studies: Challenges of Rustvelology in the Digital Age', in *Caucasus Journal of Social Sciences*, 17(1), 166–191. https://doi.org/10.62343/cjss.2024.246.

Tuldava 2005: Tuldava, Juhan, 'Stylistics, author identification (Stilistik und Autorenbestimmung)'. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds), *Quantitative Linguistik/Quantitative Linguistics: Ein internationales Handbuch/ An International Handbook*. Berlin/New York: De Gruyter, pp. 368–387. https://doi.org/10.1515/9783110155785

Van Gijsel et al. 2006: Van Gijsel, Sophie, Speelman, Dirk, & Geeraerts, Dirk, 'Locating lexical richness: a corpus linguistic, sociovariational analysis', in *8es Journées internationales d'Analyse statistique des Données Textuelles*. Cambridge: Cambridge University Press, pp. 961–972.

Van Hout et al. 2007: Van Hout, Roeland W. N. M., & Vermeer, Anne R., 'Comparing measures of lexical richness', in Helmut Daller, James Milton & Jeanine Treffers-Daller (eds), *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press, pp. 93–115.

Wimmer 2005: Wimmer, Georg, 'The type–token relation (Das Type–Token-Verhältnis)', in Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds*), Quantitative Linguistik/Quantitative Linguistics: Ein internationales Handbuch/An International Handbook* (pp. 361–368). Berlin/New York: De Gruyter. https://doi.org/10.1515/9783110155785.

ფუნქციური სიტყვების ანალიზი ქართულ პოლიტიკურ დისკურსში: ახალი პერსპექტივა

ანასტასია ყამარაული
(ბათუმის შოთა რუსთაველის სახელმწიფო უნივერსიტეტი)

DOI: 10.62235/mln.4.2026.12163
a.kamarauli@web.ge || ORCID: 0009-0008-2258-8307

შესავალი

პოლიტიკური რიტორიკის შესწავლა ტრადიციულად, ძირითადად თვისებრივ, ჰერმენევტიკულ მეთოდებს ეყრდნობოდა. თუმცა, რაოდენობრივი ლინგვისტიკისა და კორპუსლინგვისტიკის დამკვიდრებასთან ერთად მოხდა პარადიგმის ცვლილება: ენობრივი სტრუქტურების, სტილისტური ნიმუშებისა და რიტორიკული სტრატეგიების ანალიზი ახლა მათემატიკურად და სტატისტიკურად არის შესაძლებელი. პოლიტიკური გამოსვლები, რომლებიც ეროვნული კრიზისის პერიოდში იქმნება, განსაკუთრებით მნიშვნელოვანია პოლიტიკური რიტორიკის რაოდენობრივი ანალიზისათვის, ამიტომ 1991-1992 წლებში ზვიად გამსახურდიას პოლიტიკური ზეპირმეტყველების ნიმუშების კვლევა ტექსტის რაოდენობრივი მიდგომებისთვის კვლევის იდეალურ სფეროს წარმოადგენს.

წინამდებარე ნაშრომში მოცემულია საქართველოს პირველი პრეზიდენტის, ზვიად გამსახურდიას პოლიტიკური რიტორიკის კვლევის შედეგები. კვლევა თანამედროვე, სტატისტიკური მეთოდებით არის განხორციელებული.

1. მეთოდოლოგია

სტატიაში წარმოდგენილია პოლიტიკოსის ზეპირმეტყველების რაოდენობრივ-ლინგვისტური და რეგისტრულ-ანალიტიკური გამოკვლევა. ანალიზისთვის ემპირიულ ბაზად აღებულია 25 ტექსტი ქართული ენის ეროვნული კორპუსის პოლიტიკური ქვეკორპუსიდან (http://gnc.gov.ge/gnc/text-list?session-id=262003632351823), რომელიც დაყოფილია სამ ფუნქციურ რეგისტრად: ოფიციალური განცხადებები, მიმართვები და ინტერვიუები. კვლევის მიზანია დისკურსის სხვადასხვა მოდალობაში ტიპ-ტოკენის თანაფარდობის (TTR), გიროს ინდექსისა (Guiraud index) და ლექსიკური სიმკვრივის სისტემური ცვალებადობის შესწავლა.

კვლევაში გამოყენებულია დუგლას ბაიბერის (2012) რეგისტრული მოდელი, რომელიც ენის რეალიზაციის ორ ძირითად სტრუქტურას აღწერს:

პირველი სტრუქტურა განიხილება, როგორც **კლაუზური გრამატიკა (Clausal Grammar)**, რომელიც დამახასიათებელია ინტერაქციული და სპონტანური ზეპირმეტყველებისთვის. მასში ფიქსირდება ფინიტური ზმნები/ზმნური ფრაზები, ახასიათებს ნაცვალსახელების, ზმნიზედებისა და დაქვემდებარებული კავშირების მაღალი სიხშირე, რაც ემსახურება არგუმენტაციის დინამიკურ განვითარებას.

მეორე სტრუქტურა წარმოადგენს ე.წ. **ფრაზული გრამატიკის (Phrasal Grammar)** შემთხვევას**,** რომელიც ძირითადად საინფორმაციო წერილობით ენაში რეალიზდება. აქ ინფორმაცია „შეკუმშულია“ რთულად მოდიფიცირებულ სახელურ ფრაზებში, სადაც მაღალია არსებითი სახელების, ატრიბუტული ზედსართავებისა და თანდებულიანი მოდიფიკატორების სიმკვრივე.

პოლიტიკური ზეპირმეტყველება ხშირად თავსდება წერილობითი ენისა და ზეპირმეტყველების მიჯნაზე, რაც ვლინდება მიკროლინგვისტურ და სტატისტიკურ კანონზომიერებებში. სტატია ეყრდნობა კორპუსლინგვისტიკისა და რეგისტრის თეორიის კლასიკურ მოდელებს, რათა ავხსნათ პოლიტიკური რიტორიკის ჰიბრიდული ბუნება.

2. თეორიული ჩარჩო

ჩვენი კვლევის თეორიულ ჩარჩოდ აღებულია მაიკლ ჰალიდეის რეგისტრების თეორია და მაიკლ გრეგორის დიატიპური ვარიაციის მოდელი.

რეგისტრების თეორიაში ჰალიდეი (1989) უარყოფს ტრადიციულ შეხედულებას იმის შესახებ, რომ წერილობითი ენა ნორმატიული სტანდარტია, ხოლო ზეპირმეტყველება — მისი ნაკლოვანი ან სტრუქტურულად სუსტი ვერსია. ჰალიდეი ამტკიცებს, რომ ზეპირი და წერილობითი ენა ფუნქციურად თანაბარმნიშვნელოვანია, თუმცა ისინი რეალობის კოდირებას განსხვავებული პრინციპებით ახდენენ. მათ შორის განსხვავების აღსაწერად იგი ორ ფუნდამენტურ კონცეფციას გვთავაზობს:

ზეპირმეტყველებას ჰალიდეი ახასიათებს, როგორც დინამიკურ და პროცესზე ორიენტირებულ **ქმედების სამყაროს** (World of Happening), რომელიც რეალობას წარმოაჩენს, როგორც მოვლენების, მოქმედებებისა და ინტერაქციების უწყვეტ ნაკადს. მისი სირთულე ვლინდება „გრამატიკულ ხლართებში“, ანუ

იმაში, თუ რამდენად მოქნილად უკავშირდება ერთმანეთს წინადადებები (clausal grammar). შესაბამისად, ფუნქციური სიტყვების წილი მასში მაღალია.

წერილობით ენას (წერითმეტყველებას) ჰალიდეი უწოდებს **საგნების სამყაროს** (World of Things). იგი მას ახასიათებს, როგორც სტატიკურ მოცემულობას, რომელიც „ყინავს“ მოვლენათა ნაკადს ნომინალიზაციის მეშვეობით. აქ ინფორმაცია შეკუმშულია რთულ სახელურ ფრაზებში, რაც იწვევს მაღალ ლექსიკურ სიმკვრივეს. დინამიკური, პროცესზე ორიენტირებული ქმედება ნომინალიზაციის შედეგად გარდაიქმნება ფიქსირებულ ობიექტებად, რაც ქმნის მკვრივ ფრაზულ სტრუქტურას (phrasal grammar).

მაიკლ გრეგორის (1967) ვარიანტების დიფერენცირების სისტემაში, ენის რეალური გამოყენების სიტუაციურ გარემოებებზე დაფუძნებული ენობრივი ვარიაციები კლასიფიცირდება, როგორც დიატიპური ვარიანტები. ეს ასახავს კონკრეტულ სიტუაციებში ენის გამოყენების განმეორებად ფუნქციურ მახასიათებლებს.

ამ სოციოლინგვისტური პირობების სისტემურად აღწერისათვის გრეგორი იყენებს მოდელს, რომელიც შედგება სამი სიტუაციური კატეგორიისა და მათი შესაბამისი კონტექსტუალური ეკვივალენტებისგან ენობრივ დონეზე: ა) **მომხმარებლის მიზანმიმართული როლი** განსაზღვრავს დისკურსის სფეროს, ბ) **ადრესატის** ურთიერთობა განსაზღვრავს დისკურსის ტონს და გ) **მედიუმთან ურთიერთობა** განსაზღვრავს დისკურსის რეჟიმს.

პოლიტიკური ენის დიფერენციაცია, ერთი მხრივ, „პოლიტიკოსების ენაზე“ და მეორე მხრივ, „პოლიტიკის ენაზე“, გრეგორის დიატიპური ვარიაციის მოდელში შეიძლება ზუსტად დაიყოს კატეგორიებად რეგისტრულ-თეორიული პერსპექტივიდან. თუ პიროვნებაზე ორიენტირებული, სიტუაციური „პოლიტიკოსების ენა“ ძირითადად ზეპირმეტყველების მეშვეობით არის დაკავშირებული მედიუმთან, უაღრესად ინსტიტუციონალიზებული „პოლიტიკის ენა“ უპირატესად ვლინდება წერის გრაფიკულ საშუალებაში. ეს ორმაგი ორიენტაცია იწვევს პოლიტიკური ენის ჰიბრიდულ პოზიციონირებას, რომელიც დისკურსის რეჟიმში გარდამავალი ფორმების ფართო კონტინუუმს მოიცავს.

ეს კანონზომიერება დაწვრილებით არის აღწერილი და დასაბუთებული დუგლას ბაიბერის მულტიგანზომილებიან მოდელში. ბაიბერის (2012) მიხედვით, რეგისტრული ვარიაციები პირდაპირ აისახება ტექსტის გრამატიკულ არქიტექტურაზე:

- წერილობითი ენა ხასიათდება არსებითი სახელების, ატრიბუტული ზედსართავებისა და თანდებულიანი მოდიფიკატორების მაღალი სიხშირით. ეს სტილი ტიპურია ინფორმაციული, წერილობითი რეგისტრებისთვის.
- ზეპირმეტყველება ეყრდნობა ზმნურ სტრუქტურებს, ნაცვალსახელებსა და ზმნიზედებს. იგი უზრუნველყოფს დინამიკურ კავშირებს და დამახასიათებელია ინტერაქციული, ზეპირი კომუნიკაციისთვის.

3. დისკუსია

3.1. კორპუსის სტატისტიკური მონაცემები და რეგისტრული ანალიზი

კვლევაში გაანალიზებულია 25 ტექსტი, რომლებიც განაწილებულია 3 ძირითად რეგისტრში. ქვემოთ წარმოდგენილია სტატიის ძირითადი შემაჯამებელი ცხრილი. როგორც სტატისტიკური მონაცემების შედარება გვიჩვენებს, ტრადიციული რაოდენობრივი მეტრიკები ზვიად გამსახურდიას ზეპირმეტყველებაში მკვეთრად რეაგირებენ ტექსტის ჟანრსა და სიგრძეზე:

ტექსტის ტიპი / №	ტოკენი (Token)	ტიპი (Type)	ფუნქც. სიტყვა	ლექს. სიტყვა	TTR	გირო (Guiraud)	ლექს. სიმკვრივე
მიმართვა №14 (მოკლე)	36	30	3	33	0.83	5.00	91.67%
მიმართვა №1	147	117	15	133	0.80	9.65	90.48%
მიმართვა №23 (გრძელი)	1452	658	229	1206	0.45	17.27	83.06%
ოფიც. განცხადება №4	171	117	26	141	0.68	8.95	82.46%
ოფიც. განცხადება №13	131	101	13	118	0.77	8.82	90.08%
ინტერვიუ №21 (გრძელი)	3232	1164	546	2663	0.36	20.47	82.39%
ინტერვიუ №25 (გრძელი)	3224	996	619	2594	0.31	17.54	80.46%

ცხრილი 1: შერჩეული რეზონანსული ტექსტების შედარებითი მონაცემები რეგისტრების მიხედვით

ცხრილში მოცემულ სტატისტიკურ მონაცემთა შედარებითი ანალიზი შემდეგი დასკვნის გამოტანის საშუალებას იძლევა:

- **ოფიციალური განცხადებები** ხასიათდება სტაბილური, მაღალი ლექსიკური სიმკვრივით (82%-დან 90%-მდე). აქ სტრუქტურული ფუნქციური სიტყვების რაოდენობა მინიმუმამდეა დაყვანილი, რათა ინფორმაცია გადაიცეს ზუსტად, ოფიციალურად და არაპერსონალურად.
- **ინტერვიუები** წარმოადგენს სპონტანურ ან ნახევრად სტრუქტურირებულ მეტყველებას. მათში ფიქსირდება ყველაზე დაბალი ლექსიკური სიმკვრივე (დაახლოებით 80-83%), რადგან დიალოგი მოითხოვს კლაუზურ გრამატიკას და ფუნქციური სიტყვების (ნაცვალსახელები, კავშირები, მოდალური სიტყვები) ხშირ გამოყენებას.
- თვალსაჩინოა **TTR**-ისა და **გიროს ინდექსის** კორექციული როლი - ტრადიციული TTR (Type-Token Ratio) უკიდურესად მგრძნობიარეა ტექსტის სიგრძის მიმართ. ზვიად გამსახურდიას პოლიტიკურ ზეპირმეტყველებაში ყველაზე მაღალი TTR ფიქსირდება ულტრამოკლე მიმართვებში (მაგ.: ტექსტი №14, 36 ტოკენი, TTR = 0.83) და ყველაზე დაბალი — გრძელ ინტერვიუებში (მაგ.: ტექსტი №25, 3224 ტოკენი, TTR = 0.31). მიუხედავად დაბალი TTR-ისა, ინტერვიუებში გიროს ინდექსი აღწევს პიკურ მაჩვენებლებს (20.47 და 17.54), რაც ამტკიცებს, რომ თავისუფალი მეტყველებისას გამსახურდიას აქტიური ლექსიკა ბევრად უფრო მდიდარია, ვიდრე მოკლე ტექსტებში.

კვლევამ გვიჩვენა, რომ ზვიად გამსახურდიას პოლიტიკური რიტორიკა ამ კონცეფციების ჭრილში უნიკალურ ჰიბრიდულ პოზიციას იკავებს. მიუხედავად იმისა, რომ მისი გამოსვლები ზეპირმეტყველების ნიმუშია (განცხადებები, მიმართვები, ინტერვიუები), მათში შენარჩუნებულია უაღრესად მაღალი ლექსიკური სიმკვრივე და შეკუმშული სახელური ფრაზული სტრუქტურა, რაც ტრადიციულად მხოლოდ წინასწარ დაგეგმილ წერილობით ენას (ანუ დაწერილ ტექსტებს) ახასიათებს.

3.2. მეთოდოლოგიური გამოწვევები ქართული ენისთვის

თეორიული თვალსაზრისით სტატიის ერთ-ერთი უმნიშვნელოვანესი მიღწევაა იმის ჩვენება, რომ დასავლური კომპიუტერული ლინგვისტიკის სტანდარტული ხელსაწყოებით ვერ ხერხდება ქართული ენის სტრუქტურის ზუსტი რაოდენობრივი ანალიზი. ეს კი იწვევს მონაცემთა სისტემურ დამახინჯებას ოთხ ძირითად დონეზე:

ა) **შერწყმული თანდებულები:** სახელზე დართული თანდებულების (Bound Postpositions) სტატისტიკა (ქართული ენის რიგი თანდებულები - მაგ., *-ში, -თვის, -ადმი* მორფოლოგიურად ერწყმიან სახელის ბრუნვიან ფორმას: საქართველო-[ს]-**ში** , დამოუკიდებლობ-ის-ა-**თვის**). სტანდარტული ტოკენიზატორი მათ ერთ სიტყვად (ტოკენად) აღიქვამს, რის გამოც ფუნქციური ერთეულები სტატისტიკიდან იკარგება. ეს ხელოვნურად ზრდის ლექსიკურ სიმკვრივეს და ამცირებს TTR-ს.

ბ) **თანდებულთა კოორდინირებული ელიფსი:** ორი სახელის **და** კავშირით შეერთებისას თანდებული პირველ სახელთან ელიმინირდება და მხოლოდ მეორე სახელს დაერთვის (მაგ.: *ხალხისა და საზოგადოებრიობისადმი*). კორპუსული დამუშავებისას პროგრამები პირველ სიტყვას ჩვეულებრივ ავტოსემანტიკურ ელემენტად აღიქვამენ, რითაც "„იგნორირებულია" ქართული ენის სინტაქსური რეალობა.

გ) **ყოფნა ზმნის (კოპულას) კლიტიზაცია:** დამხმარე ზმნა **არის** შედგენილ შემასმენელში ხშირად **-ა** სუფიქსის სახით არის წარმოდგენილი, რომელიც ერთვის სახელურ ნაწილს (მაგ.: ხელში არის > ხელშია, ჭეშმარიტება არის > ჭეშმარიტებაა). შედეგად, დამხმარე ზმნა „უჩინარი" ხდება ავტომატური პროგრამებისთვის.

დ) **ფუნქციური ფრაზები:** მრავალსიტყვიანი ფუნქციური კონსტრუქციები, როგორიცაა, მაგალითად, ***რა თქმა უნდა***, პრაგმატულად ერთ ფუნქციურ ფრაზას წარმოადგენს, თუმცა კომპიუტერი მას სამ ცალკე სიტყვად აღიქვამს, რაც ორმაგ მათემატიკურ აცდენას იწვევს.

მათემატიკური უზუსტობის შემთხვევები თვალსაჩინოდ არის წარმოდგენილი ქვემოთ მოცემულ ცხრილში 2:

მეტრიკა / მაჩვენებელი	ტექსტი №10 (მიმართვა)	ტექსტი №10 (თანდებულებით)	ტექსტი №24 (ინტერვიუ)	ტექსტი №24 (თანდებულებით)

ტოკენების საერთო რაოდენობა (N)	592	619 (+27)	648	666 (+18)
უნიკალური ტიპები (V)	325	333 (+8)	369	377 (+8)
მარტივი TTR	0.59	0.47 (-0.12)	0.57	0.55 (-0.02)
გიროს ინდექსი (Guiraud)	14.47	13.38 (-1.09)	17.27	14.61 (-2.66)
შერწყმული თანდებულების წილი ტოკენებში	4.56%	—	2.77%	—

ცხრილი 2: შერწყმული თანდებულების გამოცალკევების გავლენა მეტრიკებზე

როგორც ვხედავთ, თანდებულების გამოყოფის შედეგად, როგორც TTR, ისე გიროს ინდექსი პარადოქსულად მცირდება. ეს ფაქტი აიხსნება იმით, რომ თანდებულების ცალკე ტოკენებად აღიარება მნიშვნელოვნად ზრდის მნიშვნელის (Denominator / სიტყვების საერთო რაოდენობა) მოცულობას, რაც მათემატიკურად უფრო მეტად ასწორებს და დაბლა წევს ლექსიკური მრავალფეროვნების მრუდს. ეს კი ცხადად გვიჩვენებს, რომ სტანდარტული დასავლური ტოკენიზატორები ხელოვნურად "აღარიბებენ" ქართულ ტექსტს ფუნქციური სიტყვების რაოდენობის თვალსაზრისით და ტექსტის რაოდენობრივი დამუშავებისას გვაძლევენ დამახინჯებულ სურათს.

4. დასკვნა

როგორც სტატიაში წარმოდგენილმა ზვიად გამსახურდიას პოლიტიკური ზეპირმეტყველების კვლევამ გვიჩვენა, ქართული ენის გამოთვლითი და რეგისტრულ-ლინგვისტური კვლევებისთვის ლექსიკური სიმდიდრისა და სიმკვრივის ვალიდური შედარებების უზრუნველსაყოფად, მომავალში წინასწარ უნდა შეიქმნას ემპირიული მასალის სტატისტიკური დამუშავების ორეტაპიანი პროცედურა. ქვედა დონეზე, თითოეული ენის ენობრივი თავისებურებები სისტემატურად უნდა გაკონტროლდეს მორფოსინტაქსური ანოტაციებისა და ტოკენების

დაყოფის გზით, სანამ ზედა დონეზე ჯვარედინი ენობრივი შედარებები განხორციელდება. ფუნქციური ფრაზების იდენტიფიცირებისათვის აუცილებელია ავტომატური ფილტრების შემუშავება, რათა ისინი გამოთვლით მოდელებში არა ცალკეულ ტოკენებად, არამედ ფუნქციურ ერთეულებად იქნეს მიჩნეული.

წარმოდგენილი კვლევა გვიჩვენებს, რომ ტექსტური მონაცემების წმინდა სტატისტიკური მოდელირება ენობრივად ადაპტირებული გრამატიკული მოდელების გარეშე არაზუსტია და არასწორ თეორიულ ინტერპრეტაციებს იწვევს. მხოლოდ რაოდენობრივი მეტრიკის, ზუსტი გამოთვლითი ენობრივი ტოკენიზაციისა და ფუნქციური რეგისტრის თეორიის მჭიდრო ურთიერთქმედება იძლევა პოლიტიკური ენის დახვეწილი სტრატეგიული არქიტექტურის ვალიდური რეკონსტრუქციის საშუალებას.