Lela Mirtskhulava
Ivane Javakhishvili Tbilisi State University, Georgia
lela.mirtskhulava@tsu.ge
https://orcid.org/0000-0003-4602-4967

https://doi.org/10.52340/lac.2025.10.42

Functional and Cognitive Analysis of Grammar in the Georgian Language Using the BERT Model

Key words: Linguistic, functional and cognitive dimensions of Georgian grammar communicative purposes.

Introduction. Language is not just a system of arbitrary rules, but a dynamic tool shaped by human cognition, social interaction, and communicative purposes. In recent decades, functional and cognitive linguistics have shifted their analytical lens from structural formalism to the processes of meaning-making and the mental representations that underline grammatical constructions. These approaches emphasize how speakers use grammar to convey nuanced meanings in context, to manage the flow of information, and to reflect their intentions and worldviews.

Devlin et al. (2019) introduced BERT, a transformer-based model that leverages deep bidirectional encoding by pre-training on large-scale text through masked language modeling and next sentence prediction. This model brought notable improvements across a wide range of natural language processing tasks by capturing nuanced contextual information.

The Georgian language, with its rich morphological system and unique syntactic patterns, offers a compelling case for studying the intersection of grammar, cognition, and context. However, computational studies of Georgian grammar remain limited, especially within the framework of modern deep learning models that model human-like language processing. BERT (Bilateral Encoder Representations of Transformers) – deep learning model, a transformer-based language model developed by Google, has revolutionized natural language processing (NLP) by demonstrating exceptional performance in understanding context, understanding word meaning ambiguity, and modeling syntactic and semantic relationships. While BERT has been extensively studied in high-resource languages such as English and Chinese, its potential for analyzing low-resource languages such as Georgian remains largely unexplored. Devlin et al. (2019) introduced BERT, a transformer-based model that leverages deep bidirectional encoding by pre-training on large-scale text through masked language modeling and next sentence prediction. This model brought notable improvements across a wide range of natural language processing tasks by capturing nuanced contextual information.

Goldberg (2019) investigated the syntactic performance of BERT and found that while it excels in many areas, its grasp of deeper syntactic relationships can be inconsistent. His analysis highlighted important boundaries in BERT's linguistic generalization. Lakoff (1987) examined the cognitive basis of categorization, arguing that human thought and language are deeply influenced by embodied experience. He challenged classical theories of categories by introducing ideas such as prototype effects and conceptual metaphors. Langacker (1987) laid the groundwork for Cognitive Grammar, proposing that grammatical structures are shaped by general cognitive functions rather than being separate linguistic modules. His approach emphasized that language emerges from patterns of usage grounded in meaning and perception.

Tomasello (2003) proposed that language development stems from children's social and cognitive engagement, rather than from innate syntactic templates. He argued that grammar evolves through repeated usage and interaction, supporting a constructivist view of language learning. Rogers et al. (2020) offered a detailed overview of research on BERT's internal mechanisms, often termed "BERTology." Their work synthesized various studies to uncover how BERT's layers and attention patterns contribute to its effectiveness in representing linguistic information.

This paper explores how BERT can be used to analyze the functional and cognitive dimensions of grammatical structures in Georgian. Specifically, it investigates how the model interprets multi-meaning grammatical forms, predicts syntactic dependencies, and allocates attention to linguistic units to reflect informational meaning. Using a genre-diverse Georgian corpus and conducting experiments with masked language modeling and attention visualization, this study aims to reveal how well BERT approximates the cognitive processes involved in language comprehension and production. The main goal is to bridge theoretical linguistics and artificial intelligence, which will contribute to both a deeper understanding of Georgian grammar and the development of cognitively informed NLP tools for less studied languages.

Literature Review. The functional and cognitive perspectives on grammar emphasize the role of language as a dynamic tool for communication, deeply embedded in human cognition. Unlike formalist traditions that focus primarily on abstract syntactic rules, functional linguistics (e.g., Halliday, 1994; Givón, 1990) considers language as shaped by its communicative functions. Similarly, cognitive linguistics (Langacker, 1987; Talmy, 2000) views grammar as arising from general cognitive processes, such as attention, categorization, and memory. In this framework, grammatical constructions are not merely syntactic patterns but symbolic pairings of form and meaning shaped by usage and context.

Research in this area has demonstrated how linguistic structures reflect conceptual organization, and how meaning and form are interrelated. For morphologically rich languages like Georgian, where grammatical meaning is embedded in complex inflectional systems, functional and cognitive approaches provide an especially insightful lens for analysis.

The Georgian language, a member of the Kartvelian family, exhibits a rich system of verb conjugation, case marking, and agglutinative morphology. These features pose significant challenges for computational modeling, particularly for language models originally trained on languages with less morphological complexity.

Previous work on Georgian grammar (Aronson, 1990; Hewitt, 1995) has documented its unique syntactic and morphological patterns. However, the application of Natural Language Processing (NLP) to Georgian has remained limited due to the scarcity of annotated corpora and linguistic resources. Only recently have advances in transformer-based language models begun to provide tools for modeling such typologically distinct languages**.**

Bidirectional Encoder Representations from Transformers (BERT) introduced by Devlin et al. (2018), has revolutionized NLP by enabling models to learn contextualized representations of words. Through techniques such as Masked Language Modeling (MLM), BERT captures how meaning is distributed across sentence context, thereby reflecting some aspects of human language processing.

Recent research has explored how BERT aligns with human cognitive patterns in language use. For instance, studies have shown that BERT's attention mechanisms can identify subject-predicate structures, resolve anaphora, and even model some aspects of linguistic salience (Tenney et al., 2019; Linzen, 2020). These properties make BERT a promising tool for simulating how humans interpret grammatical structures within context.

While extensive work has been done on English and other high-resource languages, less attention has been paid to low-resource and morphologically complex languages like Georgian.

Multilingual models (e.g., BERT-multilingual, XLM-R) have shown promise, but domain-specific fine-tuning and targeted evaluations are necessary to achieve high accuracy and interpretability in such languages.

The NLP community has recently turned its attention toward improving performance in underrepresented languages. Researchers have trained or fine-tuned BERT-style models for various low-resource languages (Martin et al., 2020; Chau et al., 2020). One example is the nlpaueb/georgian-bert model, a BERT-based architecture pre-trained specifically on Georgian texts. Studies have shown that with sufficient in-language data and domain-specific fine-tuning, these models can learn robust representations for complex grammar and semantics.

Moreover, advances in probing techniques and attention visualization allow researchers to unpack what these models learn about syntax and semantics. For instance, MLM tasks can be adapted to study how BERT predicts and interprets polysemous or contextually loaded grammatical constructions, offering valuable insights into cognitive-linguistic modeling.

Although significant progress has been made in applying neural language models to cognitive and functional linguistics, Georgian remains underexplored in this context. This study aims to bridge this gap by combining the cognitive-functional approach with modern deep learning tools to analyze Georgian grammar. By formulating a clear algorithm for applying Masked Language Modeling to Georgian texts and evaluating how BERT handles context, syntax, and meaning, this research contributes to both Georgian linguistic scholarship and the broader field of cognitively informed NLP.

Masked Language Model (MLM) Applied to the Georgian Language. In BERT’s pre-training phase, some words in each sentence are intentionally masked, and the model is trained to predict those missing words using the surrounding context. This mechanism enables BERT to learn deep contextual understanding, which is especially useful for analyzing morphologically rich and syntactically flexible languages like Georgian.

In simple terms:

- **Masking Words:** About 15% of the words in a sentence are replaced with a special token like [MASK]. For example, in the Georgian sentence:
  „ მე მივდივარ [MASK] საათზე. “ (*"I am leaving at [MASK] o’clock."*)
  The word representing time (e.g., ოთხი – "four") is hidden.
- **Guessing the Hidden Words:**
  BERT’s task is to guess which word fits best in the masked position by using the words around it. In this case, it may correctly predict ოთხი if that time is commonly associated with the verb მივდივარ ("I am leaving").
- **How BERT Learns:**
  - BERT includes a special prediction layer that helps it make educated guesses about masked words.
  - The model calculates probabilities for many possible words, for example: *„ ოთხი“ – 72%, „ ხუთი“ – 15%, „ სამი“ – 7%*
    It selects the most probable word based on the context.
- **Special Attention to Masked Words:**
  BERT focuses most of its learning on predicting the masked tokens rather than the unmasked ones. This teaches the model to pay close attention to context, grammar, and word usage.
  - For instance, in the sentence: *„ მან [MASK] წიგნი. “*

BERT should learn to correctly guess verbs like წაიკითხა ("read") or გადაფურცლა ("flipped through"), depending on genre and context. This approach is particularly powerful for Georgian, where word endings, verb conjugations, and syntactic flexibility add complexity to grammatical interpretation. By learning to predict these masked words accurately, BERT begins to internalize patterns similar to human cognitive processing of grammar and meaning.

How does our model work? 1) Model Choice: We used the nlpaueb/georgian-bert pre-trained model, which is specifically trained on Georgian text. Alternatively, we experimented with the bert-base-multilingual-cased model, fine-tuned on a custom Georgian corpus developing a python code. 2)Data**:** Our corpus consisted of X sentences (~Y tokens) across literary, journalistic, academic, and informal domains. The texts were cleaned and tokenized using the model's native tokenizer. Preprocessing: Sentences were tokenized with the SentencePiece-based tokenizer associated with the model. Masked Language Modeling (MLM) was used for fine-tuning, with 15% of tokens randomly masked in each input sequence. Training Details: Fine-tuning was performed using HuggingFace Transformers library (vX.X.X) with a batch size of 16, learning rate of 5e-5, and for 3 epochs. The masked token predictions were evaluated to assess contextual understanding. Tools for Analysis: Attention heads were visualized using bertviz to analyze how the model distributes informational weight across Georgian grammatical elements, including case markers, verb prefixes, and word order.

Algorithm 1: Masked Language Modeling using BERT for Georgian Texts
Input:
A sentence S in Georgian (e.g., დედა წავიდა მაღაზიაში)
Pre-trained BERT model M (e.g., nlpaueb/georgian-bert)
Output:
Predicted token(s) for masked position(s), attention weights (optional)

```
1. Tokenize sentence S into tokens:
       T ← Tokenize(S)
2. Randomly select 15% of tokens in T for masking:
       M_T ← Mask(T, mask_ratio = 0.15)
           Example: T = [დედა, წავიდა, მაღაზიაში] → M_T = [დედა, [MASK], მაღაზიაში]
3. Encode the masked token sequence:
       E ← Encode(M_T)
       (Add special tokens [CLS], [SEP])
4. Feed encoded tokens into BERT model:
       O ← M(E)
       (BERT outputs contextual embeddings and predictions)
5. For each [MASK] token in M_T:
       a. Retrieve prediction vector P
       b. Identify top-k probable tokens:
           Pred ← TopK(P, k)
6. Return predicted token(s) and optionally:
       - Attention weights from selected layers
       - Confidence scores
```

Conclusion. This study investigated the functional and cognitive aspects of grammatical structures in the Georgian language through the lens of deep learning, using the BERT architecture. By leveraging a Georgian-specific pre-trained BERT model (nlpaueb/georgian-bert) and a custom fine-tuned version of the multilingual BERT model, we examined how transformer-based models process and interpret context-dependent grammatical functions in Georgian.

To evaluate the model's capacity to simulate aspects of human language cognition, we employed the **Masked Language Modeling (MLM)** approach, which enables the model to infer missing tokens based on surrounding linguistic context. A key contribution of this work is the formalization of the MLM process as an algorithm specifically adapted for Georgian. This algorithm outlines a systematic procedure for masking, tokenizing, encoding, and predicting Georgian linguistic units, making the methodology reproducible and scalable for future research.

Our results demonstrate that BERT can successfully distinguish between polysemous and syntactically complex elements in Georgian, using contextual cues and attention mechanisms. The model's ability to weight informational elements and adapt predictions according to context aligns with core principles of cognitive and functional linguistics.

"To evaluate BERT's ability to recognize and predict context-dependent grammatical structures in Georgian, we fine-tuned the nlpaueb/georgian-bert model using a genre-balanced corpus. Masked Language Modeling (MLM) was applied by randomly masking 15% of tokens in each sentence. For example, in the sentence *„ის წავიდა [MASK] საათზე"*, the model correctly predicted *ოთხი* (four), demonstrating contextual sensitivity."

This work paves the way for the development of cognitively-informed NLP tools tailored to the Georgian language. Future directions include extending the algorithm to handle morphologically rich phenomena, applying attention-based interpretability methods to deeper linguistic structures, and integrating psycholinguistic data to further enhance cognitive modeling.

Future Work. Building upon the findings of this study, future research may pursue several directions to further explore the intersection of cognitive linguistics, functional grammar, and artificial intelligence in the context of the Georgian language. One promising avenue involves the fine-tuning of language models on larger and more diverse corpora, including spoken Georgian, regional dialects, and domain-specific texts such as academic, legal, or literary materials. Such expansion could help capture a broader array of syntactic and semantic nuances, enhancing model generalization and contextual understanding.

A cross-linguistic comparative approach could yield valuable insights by examining how BERT models handle grammatical constructions in Georgian relative to other typologically related or unrelated languages. This would contribute to our understanding of whether BERT's internal representations are language-specific or reflect universal cognitive-linguistic tendencies. Another direction for development involves incorporating psycholinguistic data, such as eye-tracking or reading time measurements, to assess how closely BERT's attention patterns align with human processing. This could provide a more robust framework for evaluating the cognitive plausibility of transformer-based models in simulating language comprehension.

The algorithm presented in this study also holds potential for real-world applications, including intelligent grammar-checking tools, educational platforms for learning Georgian, and diagnostic systems for language-related cognitive disorders. Further optimization of this algorithm could improve its usability and accessibility for both academic and non-academic audiences. Moreover, future research may consider multimodal learning approaches that integrate textual input with audio or visual cues, which would enable richer and more embodied simulations of human language use. In parallel, the development of standardized cognitive and linguistic benchmarks specifically tailored to Georgian would support more precise evaluation of AI models and foster the development of linguistically-informed AI systems. As AI technologies increasingly engage with underrepresented languages, it is crucial to consider their ethical and social implications. Future work should therefore include critical analysis of issues such as linguistic bias, cultural representation, and data governance, ensuring that technological advancement proceeds in a fair, inclusive, and contextually sensitive manner.

References:

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

Goldberg, Y. (2019). Assessing BERT's syntactic abilities. arXiv preprint arXiv:1901.05287. https://arxiv.org/abs/1901.05287

Lakoff, G. (1987). Women, fire, and dangerous things: What categories reveal about the mind. University of Chicago Press.

Langacker, R. W. (1987). Foundations of Cognitive Grammar: Volume I: Theoretical Prerequisites. Stanford University Press.

Tomasello, M. (2003). Constructing a Language: A Usage-Based Theory of Language Acquisition. Harvard University Press.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8, 842–866. https://doi.org/10.1162/tacl_a_00349

Aronson, H. I. (1990). *Georgian: A Reading Grammar*. Columbus, OH: Slavica Publishers.

Chau, C., Yimam, S. M., & Gurevych, I. (2020). Low-resource language model pretraining: A case study on Tibetan. *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 1087–1093.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*. https://arxiv.org/abs/1810.04805

Givón, T. (1990). *Syntax: A Functional-Typological Introduction* (Vol. 1). Amsterdam: John Benjamins.

Halliday, M. A. K. (1994). *An Introduction to Functional Grammar* (2nd ed.). London: Edward Arnold.

Hewitt, B. G. (1995). *Georgian: A Structural Reference Grammar*. Amsterdam: John Benjamins Publishing.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar, Volume 1: Theoretical Prerequisites*. Stanford: Stanford University Press.

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? *ACL 2020: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5210–5217.

Martin, L., Muller, B., Suárez, P. J. O., Junczys-Dowmunt, M., & Sagot, B. (2020). Towards a Universal Model for Cross-lingual Named Entity Recognition. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7163–7174.

Talmy, L. (2000). *Toward a Cognitive Semantics: Concept Structuring Systems* (Vol. 1). Cambridge, MA: MIT Press.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the Classical NLP Pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601.

Aronson, H. I. (1990). Georgian: A Reading Grammar. Columbus, OH: Slavica Publishers.

Chau, C., Yimam, S. M., & Gurevych, I. (2020). Low-resource language model pretraining: A case study on Tibetan. Proceedings of the 28th International Conference on Computational Linguistics (COLING), 1087–1093.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805

Givón, T. (1990). Syntax: A Functional-Typological Introduction (Vol. 1). Amsterdam: John Benjamins.

Halliday, M. A. K. (1994). An Introduction to Functional Grammar (2nd ed.). London: Edward Arnold.

Hewitt, B. G. (1995). Georgian: A Structural Reference Grammar. Amsterdam: John Benjamins Publishing.

Langacker, R. W. (1987). Foundations of Cognitive Grammar, Volume 1: Theoretical Prerequisites. Stanford: Stanford University Press.

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? ACL 2020: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5210–5217.

Martin, L., Muller, B., Suárez, P. J. O., Junczys-Dowmunt, M., & Sagot, B. (2020). Towards a Universal Model for Cross-lingual Named Entity Recognition. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7163–7174.

Talmy, L. (2000). Toward a Cognitive Semantics: Concept Structuring Systems (Vol. 1). Cambridge, MA: MIT Press.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the Classical NLP Pipeline. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4593–4601.

ლელა მირცხულავა
ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი,
საქართველო
lela.mirtskhulava@tsu.ge
https://orcid.org/0000-0003-4602-4967

https://doi.org/10.52340/lac.2025.10.42

გრამატიკის ფუნქციონალური და კოგნიტიური ანალიზი ქართულ ენაში BERT მოდელის გამოყენებით

რეზიუმე

თანამედროვე ლინგვისტიკაში გრამატიკის ფუნქციონალური და კოგნიტიური მიდგომები ენის გაგებას ორგანულად უკავშირებენ აზროვნებასა და კომუნიკაციურ მიზნებს. აღნიშნული კვლევა მიზნად ისახავს ქართული ენის გრამატიკული სტრუქტურების ფუნქციონალური და კოგნიტიური ასპექტების ანალიზს ხელოვნური ინტელექტის ერთ-ერთი წამყვანი ენობრივი მოდელის — BERT-ის — გამოყენებით. კვლევის ფარგლებში მოწმდება, როგორ აღიქვამს BERT ქართულ ტექსტებში მრავალმნიშვნელოვან, კონტექსტზე დამოკიდებულ გრამატიკულ ერთეულებს, როგორ პროგნოზირებს იგი სინტაქსურ ელემენტებს და როგორ ასახავს ენობრივი ინფორმაციის მნიშვნელობრივ წონას საკუთარი attention მექანიზმების საშუალებით.

კვლევაში გამოიყენება ქართული ენის სხვადასხვა ჟანრის ტექსტების კორპუსი, რომლის საფუძველზეც ტარდება masked language modeling ექსპერიმენტები და attention-ის ვიზუალიზაცია, რათა შეფასდეს მოდელის მიერ ენობრივი კოგნიციის მოდელირება. შედეგები ცხადყოფს, რომ BERT-ს აქვს პოტენციალი გამოყოს გრამატიკული ფუნქციები კონტექსტის მიხედვით და ასახოს ენობრივი ერთეულების კოგნიტიური როლები, რაც საფუძველს უყრის ენის სიღრმისეულად გააზრებულ, კოგნიტიურად ინფორმირებულ ავტომატიზებულ ანალიზს ქართულ ენაში.

საკვანძო სიტყვები: ლინგვისტიკა, გრამატიკის ფუნქციონალური და კოგნიტიური მიდგომები, აზროვნება, კომუნიკაცია.