OPEN ACCESS

European Journal of Engineering Science and Technology
ISSN 2538-9181

Automated Development of a Grammatical Dictionary for Georgian Dialects

Liana L. Lortkipanidze[1*] and Anna R. Chutkerashvili[2]

[1] Ivane Javakhishvili Tbilisi State University, Georgia
[2] Archil Eliashvili Institute of Control Systems of the Georgian Technical University, Georgia

ARTICLE INFO

Keywords:
Acquisition of Lexicon,
Agglutinative Languages,
Language Modelling,
Lemmatization Rules,
Morphological Analysis

ABSTRACT

This paper presents an automated system for compiling grammatical dictionaries of the Georgian language and its dialects. Unlike traditional dictionaries, grammatical dictionaries include not only base word forms but also complete paradigms, offering detailed morphological and syntactic information. This is particularly crucial for agglutinative-inflectional languages such as Georgian, where word forms vary significantly depending on context. The system applies a dictionary-based approach to expand lexical resources by identifying words with shared grammatical markers and integrates an innovative lemmatization algorithm capable of processing unknown words, automatically generating their base forms and paradigms. The methodology builds upon prior research in dialectal lexicography and syntactic annotation within Georgian corpora, while introducing comparative insights from similar linguistic technologies applied to other agglutinative languages. The developed system demonstrated high efficiency in automating the creation of grammatical dictionaries. Testing on Georgian literary corpora revealed that only 2% of non-dictionary word forms required manual correction post-lemmatization. The affix-based algorithm significantly outperformed traditional suffix-only methods, particularly in handling complex morphological structures. These results confirm the system's effectiveness in expanding lexical resources and highlight its adaptability for other Kartvelian languages. The study emphasizes the value of integrating linguistic theory with computational approaches to address challenges in morphological processing and lexicon development, offering both theoretical contributions and practical applications in language technology.

1. Introduction

The distinction between a grammatical dictionary and conventional dictionaries lies in the fact that a grammatical dictionary provides not only headword forms but also the complete paradigms derived from those words. It presents comprehensive morphological and partial

* Corresponding Author's E-Mail Address: liana.lortkipanidze@tsu.ge, https://orcid.org/0000-0003-0974-1668

Cite this article as:
Lortkipanidze, L., & Chutkerashvili, A. (2025). Automated Development of a Grammatical Dictionary for Georgian Dialects. *European Journal of Engineering Science and Technology, 8*(1): *13*-25. https://doi.org/10.33422/ejest.v8i1.1553

syntactic characteristics of language units. Traditional dictionaries often fall short in assisting users to fully understand texts, especially in agglutinative-inflectional languages like Georgian, where words frequently change form in context, making it difficult to identify their base forms.

Grammatical dictionaries have been successfully utilized in various languages for tasks such as text annotation and natural language processing (NLP). An effective system for compiling and expanding grammatical dictionaries must accommodate unknown words appearing in corpora and account for dialectal variations in orthography, morphology, and vocabulary. To address these challenges, dictionary compilers and grammatical analyzers are essential tools.

2. Methodology

The methodology for developing the grammatical dictionary compiler for the Georgian language is based on a hybrid approach that combines automated processes with linguistic supervision to ensure accuracy and adaptability across diverse language forms.

2.1. Paradigm Formation Techniques

The system supports three methods for constructing paradigms:

- Unsupervised Method: The program automatically identifies word forms that belong to the same paradigm by grouping lexemes sharing a lemma and part of speech.
- Supervised Method: Linguists manually define paradigms to ensure proper alignment of lexemes and their morphological forms.
- Hybrid Method: A combination of automated detection and manual correction is employed to achieve optimal results, particularly when addressing irregular forms or dialectal variations.

All entries in the Georgian grammatical dictionary are introduced as headwords—nouns in the nominative case and verbs in the third-person singular present tense. Each lexical entry includes:

- Lemma
- Grammatical Features:
    - Part of speech
    - For nouns: animacy, concreteness
    - For verbs: transitivity, voice, version, aspect, mood, etc.
    - For unchangeable words: classification as postpositions, conjunctions, particles, or interjections
- Root Type: Identification of vowel or consonant endings for nouns, and thematic markers for verbs
- Declension or Conjugation Type
- Automated Lexicon Expansion

The system employs corpus-based acquisition to expand the dictionary using lemmatized lists of unknown word forms. Given the complexity of Georgian morphology, which involves both prefixation and suffixation, a trainable affix-based lemmatizer was developed. This lemmatizer is capable of handling:

- Prefixes (primarily for verbs)
- Suffixes (for most parts of speech)
- Infixes where applicable

Following the methodology outlined by Karttunen (Karttunen et al. 1996) and Forsberg (Forsberg & Ranta 2004), the lemmatizer generates transformation rules based on full form–lemma pairs. The system prioritizes longer suffixes to resolve ambiguities effectively.

2.2. Rule Generation and Ranking

Using the GeoTrans morphological generator, paradigms for over 35,000 verbs and 65,000 other word forms were produced to train the lemmatizer. The rule generation algorithm identifies common substrings and formulates lemmatization rules, such as:

*-ebshi → -*i*

*-shi → -*i*

A ranking algorithm, inspired by approaches in probabilistic morphology (Mikheev, 1997), evaluates potential lemmas based on their frequency and rule reliability within the corpus. Ambiguous cases are flagged for manual validation, ensuring high precision in lexicon expansion.

2.3. Integration of Semantic Resources

To enhance the system's capability for semantic processing and text generation, GeWordNet—an adaptation of WordNet for Georgian—is integrated. This allows for richer contextual understanding and supports dialog systems by providing semantic relations such as synonymy and hyponymy.

3. Algorithm of Ranking

The process of lemmatization, particularly for agglutinative-inflectional languages like Georgian, often results in multiple possible lemmas for a single word form due to morphological ambiguity. To address this, the system employs a probabilistic ranking algorithm designed to select the most likely lemma from a set of candidates generated during morphological analysis.

The core principle of the ranking algorithm is based on frequency analysis within the corpus and rule reliability. Each lemmatization rule, generated during the training phase, is assigned a probability score reflecting its accuracy and frequency of correct application across the dataset (Mikheev, 1997). This probability is calculated using the formula:

$$PR_i=1/NR_i$$

Where PRi is the probability of the rule Ri and NRi represents the number of alternative rules applicable to similar affix patterns. The more unique and precise a rule is, the higher its probability.

Once all possible lemmas for a given word form are generated, the system compiles a matrix where each entry includes:

- The word form W_i
- The candidate lemma L_i
- The applied rule R_i
- The frequency of W_i in the corpus OC_i
- The probability PR_i of the applied rule

The overall probability P_i that a word form corresponds to a particular lemma is calculated as:

$$P_i = \frac{\sum_{i=1}^{n} PR_i * OC_i}{\sum_{i=1}^{m} PR_i * OC_i}$$

Where:

- n is the number of occurrences of identical lemmas,
- m is the total number of possible lemma interpretations for the word form.

This statistical approach allows the system to resolve ambiguities by favoring lemmas that are both more frequent and associated with more reliable rules. For example, in cases where a suffix like *-is* could indicate either a genitive noun form or be part of a different morphological structure, the algorithm prioritizes the lemma that historically appears more frequently in similar contexts.

An additional feature of the ranking algorithm is its capacity to handle **newly derived words** by recognizing productive morphological patterns. Particularly in Georgian, where prefixation is highly active in verb formation, the system assigns higher probabilities to lemmas following productive derivational models, as supported by corpus evidence (Hajic, 2000).

The ranking process concludes with a validation phase. Lemmas with low confidence scores or those derived from rare rules are flagged for manual review. This ensures that the lexicon expansion maintains high accuracy, especially when integrating previously unseen forms or dialectal variations.

By implementing this probabilistic ranking mechanism, the system significantly reduces the rate of incorrect lemma assignments, enhancing both the precision and reliability of the grammatical dictionary compilation process.

3.1. Morphological Disambiguation

Each lemmatization rule generated during training is assigned a probability (P_R) calculated as:

P_R = 1 / N_R

Where N_R represents the number of times a particular rule applies to different lemmas sharing identical affixes. This approach is aligned with statistical methods used in morphological disambiguation (Yarowsky & Wicentowski, 2000).

For each word form (W_i) found in the corpus, the system records:

- The number of occurrences (OC_i)
- The hypothetical lemma (L_i)
- The applied rule and its probability (P_R)

The overall probability (P_i) that a word form corresponds to a specific lemma is determined by:

P_i = Σ(P_R × OC_i) / Σ(P_R × OC_i)_{total}

This calculation ensures that the lemma with the highest likelihood is selected. In cases where the probabilities are too close or insufficient data exists, the system flags the entry for manual review, maintaining a balance between automation and accuracy.

This ranking mechanism is particularly effective for Georgian, where homonymy and morphological variation are prevalent due to its agglutinative-inflectional nature. By leveraging corpus statistics, the system improves lemmatization precision without extensive manual intervention.

Finally, once the ranking is completed, duplicate lemmas are removed, and the validated entries are incorporated into the expanding grammatical dictionary.

4. Orthographic Considerations for Dialectal Variations

One of the significant challenges in developing grammatical dictionaries for the Georgian language and its dialects is addressing orthographic inconsistencies and variations. Georgian dialects often exhibit deviations from the standardized literary language in terms of phonology, morphology, and orthographic representation. These differences can lead to difficulties in accurately processing and lemmatizing dialectal texts using tools designed exclusively for standard Georgian.

To overcome this, the system incorporates a normalization module that aligns dialectal orthographic forms with their standard equivalents before morphological analysis. This pre-processing step is essential to ensure that variant spellings, phonetic shifts, and dialect-specific morphological markers do not hinder the lemmatization and lexicon expansion processes.

For instance, certain dialects may replace standard Georgian vowels or consonants within word stems or suffixes, resulting in forms that are unrecognizable to a system trained solely on literary Georgian. The normalization algorithm applies a set of transformation rules based on documented phonetic and orthographic patterns observed across different Georgian dialects. These rules are derived from linguistic studies on dialectal variation (Robins & Waterson, 1952) and adapted for computational implementation.

Furthermore, the system leverages the flexibility of the GeoTrans morphological generator by extending its capabilities to accommodate dialectal paradigms. This involves integrating additional dialect-specific affixes and root variations into the morphological templates. As a result, the system is capable of recognizing and processing forms that deviate from standard norms while preserving linguistic accuracy.

The orthographic module also plays a crucial role in minimizing false negatives during corpus lemmatization. By standardizing input forms, it ensures that words which would otherwise be treated as unknown (OOV) due to minor dialectal differences are correctly identified and linked to their base forms in the dictionary.

This approach not only enhances the robustness of the grammatical dictionary compiler but also supports linguistic diversity by systematically incorporating dialectal data. It facilitates the creation of comprehensive lexical resources that reflect the full spectrum of Georgian language usage, which is particularly valuable for linguistic research, language preservation efforts, and educational applications.

5. Methodology

The paradigms used to construct the dictionary may be derived as follows:

- Automatically – Unsupervised method. The program can define the word-form in the paradigm that combines the lexemes with one lemma and one part of speech in the boundaries of one paradigm (Archvadze & Pkhovelishvili, 2012).

- Manually – Supervised method. Linguist decides how to form paradigms in order to unite the lexemes and their corresponding lexical forms (Lortkipanidze, 2006).
- Mixed method – Combination of automatic and manual approaches, often necessary to achieve optimal results (Amirezashvili et al., 2017).

All words in the Georgian grammatical dictionary are input as headwords: nouns in the nominative case, verbs in the present tense, third person singular form.

We use automatic enlargement of the dictionary from the Georgian corpus through lemmatized lists of unknown word forms (Lortkipanidze & Gegechkori, 2016). The tag set applied is developed specifically for the Georgian corpus and consists of about 100 morpho-syntactic tags (Lortkipanidze et al., 2015).

Given the complexity of Georgian morphology, advanced affix-based methods were implemented, surpassing simple suffix replacement techniques (Archvadze, Pkhovelishvili, & Shetsiruli, 2017). The lemmatizer is capable of handling prefixes, suffixes, and infixes, tailored for Georgian's agglutinative-inflectional structure (Archvadze et al., 2014).

The **GeoTrans** system was employed to generate paradigms for over 35,000 verbs and 65,000 words from other parts of speech, covering all morphological templates (Lortkipanidze & Gegechkori, 2016). The lemmatization rules were derived by identifying the longest common substrings between full forms and lemmas.

For ranking possible lemmas, we applied statistical methods based on corpus frequency and rule probability, following approaches used in corpus-based lexicon acquisition (Amirezashvili et al., 2017).

5.1. Data Collection and Preparation

The foundation of the system is a curated lexical database derived from existing Georgian corpora, including dialectal texts. Initial lexical resources were enriched using semi-automated extraction techniques, followed by manual validation by linguists to ensure accuracy in dialectal variations.

5.2. Lemmatization Algorithm

A novel affix-based lemmatization algorithm was developed, tailored specifically for Georgian. Unlike conventional suffix-stripping methods, this algorithm processes prefixes, suffixes, and infixes, which are characteristic of Georgian word formation. The lemmatizer was trained using a dataset generated by the GeoTrans morphological generator, covering over 100,000 word forms across various parts of speech.

To address out-of-vocabulary (OOV) words, the system applies a rule-ranking mechanism that selects the most probable lemma based on corpus frequency and morphological patterns. This probabilistic approach reduces ambiguity, especially in cases where identical word forms correspond to multiple lemmas.

5.3. Paradigm Generation

For each lemma, the system automatically generates full paradigms using predefined morphological templates. These templates were designed to reflect both standard Georgian and dialect-specific inflectional patterns. Linguistic experts reviewed and adjusted templates to capture irregular forms and dialectal nuances.

5.4. System Architecture

The compiler is implemented as a modular system, consisting of:

- Morphological Analyzer and Generator: Processes input word forms and produces corresponding lemmas and paradigms.
- Lexical Database Manager: Handles storage, retrieval, and updating of grammatical entries.
- Semantic Integration Module: Incorporates GeWordNet to enhance semantic relations within the dictionary.
- User Interface: Allows linguists to review, edit, and expand dictionary entries interactively.

5.5. Validation Process

The system underwent iterative testing on literary and dialectal corpora. Each cycle involved automatic processing followed by manual verification of ambiguous cases. Performance metrics focused on lemmatization accuracy, paradigm completeness, and adaptability to unseen dialectal forms.

This methodology ensures that the compiler not only automates dictionary creation but also maintains linguistic integrity, making it a scalable solution for Georgian and potentially other Kartvelian languages.

6. Results and Discussion

The developed grammatical dictionary compiler was evaluated using a comprehensive set of Georgian literary and dialectal corpora to assess its effectiveness in morphological processing and lexicon expansion.

The lemmatizer was evaluated on a large Georgian text corpus to assess its accuracy and effectiveness. We used a corpus compiled from the novels of Otar Chiladze, containing **95,224** unique word forms. Of these, **74,900** forms (approximately 79%) were already present in the system's dictionary (i.e., known lemmas), derived from existing lexical resources such as the GeoTrans morphological lexicon. The remaining **20,324** were out-of-vocabulary (OOV) or *non-dictionary* word forms that required lemmatization. The lemmatizer processed all OOV forms, proposing base forms (lemmas) and full paradigms for each. After automatic lemmatization, only about **2%** of the OOV cases (≈406 word forms) needed manual disambiguation or correction. In other words, the system correctly lemmatized roughly **98%** of new word forms without human intervention, a very high success rate for an inflectionally complex language. These results indicate that the approach can rapidly expand the grammatical dictionary with minimal manual edits, a significant efficiency gain for lexicographers and corpus annotators.

The Table 1 summarizes the number of word forms analyzed and the lemmatization accuracy of the system.

Table 1. Lemmatization results on the test corpus

Category	Count	Percentage (relative)
Total unique word forms in corpus	95,224	100% (corpus)
– Known forms (in dictionary)	74,900	78.6% (of corpus)
– Unknown OOV forms (to lemmatize)	20,324	21.4% (of corpus)
Lemmatization of OOV forms:		
– Correctly lemmatized (auto)	~19,918	98% (of OOV forms)
– Requiring manual correction	~406	2% (of OOV forms)

The high accuracy of the affix-based lemmatizer demonstrates its effectiveness in handling Georgian's rich morphology. Notably, the system was able to infer the correct lemmas for the vast majority of previously unseen words, including those from regional dialects or archaic vocabulary, with only a small fraction requiring human review. In many of the latter cases, the need for manual disambiguation arose from homographs or rare morphological patterns where contextual or semantic information was necessary to choose the correct lemma. Overall, a **manual correction rate of just 2%** is an excellent outcome, highlighting the system's strength in generalizing morphological rules. Because the grammatical dictionary stores full paradigms for each lemma, these results translate into the automatic addition of thousands of fully inflected word paradigms to the lexicon. This capability addresses a known resource gap – previously, no automated grammatical dictionary compiler existed for Georgian – by enabling rapid, data-driven enrichment of the lexicon across different Georgian dialects.

Crucially, the **affix-based** lemmatization approach outperformed a traditional **suffix-only** method in our experiments. The affix-based lemmatizer, which learns to handle prefixes and infixes in addition to suffixes, achieved higher precision and recall in identifying correct lemmas for OOV words. In contrast, a suffix-only baseline (one that strips or replaces only word endings) struggled with many Georgian word forms. For example, Georgian verbs often employ preverbal prefixes and vowel alternations that a suffix-only algorithm cannot capture. Consistent with observations by Kanis and Müller (Kanis and Müller, 2015) that simple suffix rules are insufficient for complex OOV morphology, our affix-enabled system correctly lemmatized forms that the suffix method mis-analyzed. Empirically, the affix algorithm produced **fewer ad-hoc rules** based on single instances from the training corpus than the suffix-only algorithm. This means the affix-based lemmatizer derived more general and well-founded transformation rules, effectively handling small groups of words with exceptional morphology (as often seen in Georgian) without overfitting. The result is a more robust lemmatization: the affix approach required fewer manual corrections and resolved ambiguities more reliably than the suffix-only approach. This finding underscores the importance of modeling the full range of affixation (prefixes, infixes, suffixes) in Georgian, in line with linguistic studies of Georgian morphology that note its complex inflectional processes.

In summary, the proposed system shows strong performance in automatically building a grammatical dictionary for Georgian. It successfully expands the lexicon by detecting and lemmatizing new word forms (including dialectal variants) with high accuracy. The affix-based method offers clear advantages over simpler methods, as it captures Georgian's non-linear morphological patterns and minimizes error rates. These results confirm that an automated, corpus-driven approach can significantly aid the development of comprehensive grammatical dictionaries for Georgian and its dialects, with only minimal human intervention needed. The system's ability to learn from corpora and accurately guess unseen word forms

exemplifies a valuable step forward in Georgian language technology, supporting more efficient corpus annotation, lexicon development, and linguistic research.

6.1. Lemmatization Accuracy

The system processed a corpus containing 95,224 unique word forms, of which 74,900 were identified as existing lemmas. For the remaining out-of-vocabulary (OOV) forms, the affix-based lemmatization algorithm achieved a **98% accuracy rate**, with only **2%** of cases requiring manual disambiguation. This significantly outperforms traditional suffix-only lemmatizers, which struggled with Georgian's complex morphological structures.

The success of the affix-based approach demonstrates its ability to generalize across diverse morphological patterns, including:

- Handling of compound affixes.
- Recognition of dialect-specific inflectional variations.
- Effective processing of verbs with complex prefixation and infixation.

6.2. Paradigm Generation Effectiveness

The automatic generation of paradigms covered both standard Georgian and dialectal forms. Linguistic validation confirmed that over **95%** of generated paradigms adhered to correct grammatical rules, reducing the need for manual adjustments. This highlights the system's robustness in managing irregular forms and dialectal diversity.

6.3. Semantic Integration Outcomes

The incorporation of **GeWordNet** enhanced the semantic depth of the compiled dictionaries. The system successfully linked lemmas through synonymy, hyponymy, and other semantic relations, enabling:

- Improved context-aware text analysis.
- Enhanced functionality for future dialogue systems and NLP applications.

6.4. Practical Implications

While the system is rooted in theoretical linguistic models, it offers clear practical applications:

- **Educational Tools**: The compiler can support language learning platforms by providing dynamic grammatical resources.
- **NLP Solutions**: Integration into machine translation, spell-checkers, and AI-driven dialogue systems.
- **Dialect Preservation**: Facilitates the documentation and digitalization of endangered Georgian dialects.

6.5. Limitations and Future Work

Despite its high performance, the system faces challenges in:

- Processing extremely rare dialectal forms not represented in existing corpora.
- Semantic disambiguation in polysemous words, which may require deeper contextual AI models.

Future development will focus on:

- Expanding the training dataset with more dialectal sources.
- Incorporating deep learning techniques for enhanced semantic understanding.

- Adapting the system framework for other Kartvelian languages such as Mingrelian and Laz.

7. Lexicon Expansion and Validation

The effectiveness of any grammatical dictionary compiler depends not only on the accuracy of lemmatization but also on its ability to dynamically expand the lexicon while ensuring the integrity of newly integrated entries. In this project, lexicon expansion is driven by automated processes supported by linguistic validation to maintain high-quality lexical resources.

7.1. Automated Lexicon Expansion

The system continuously scans large Georgian corpora, identifying out-of-vocabulary (OOV) word forms through morphological analysis. Utilizing the affix-based lemmatization algorithm combined with the ranking mechanism described earlier, the system generates candidate lemmas for these unknown forms. Each candidate lemma is stored along with metadata, including:

- Source corpus and frequency data
- Applied morphological rules
- Confidence scores derived from the ranking algorithm

This automated process allows for rapid growth of the lexicon, capturing both standard Georgian forms and dialectal variants, which are often underrepresented in traditional dictionaries (Tiberius & Schoonheim, 2014).

7.2. Semantic Integration with GeWordNet

Once new lemmas are generated, they are cross-referenced with the GeWordNet semantic network to determine potential synonym sets (synsets), hypernyms, and other semantic relations. This step enriches the lexical entries, transforming them from simple lemma lists into interconnected semantic structures that support advanced applications such as semantic search, contextual text generation, and intelligent dialogue systems.

7.3. Validation Process

Despite automation, human-in-the-loop validation remains essential, especially for ambiguous cases and dialectal entries. The system flags:

- Lemmas with low confidence scores
- Forms generated by rarely applied morphological rules
- Dialect-specific forms lacking sufficient corpus frequency

Linguists review these flagged entries through a dedicated interface, where they can approve, modify, or reject suggestions. This hybrid approach balances efficiency with linguistic accuracy, ensuring that the lexicon remains both comprehensive and reliable.

7.4. Iterative Improvement

Each validation cycle feeds back into the system, updating rule probabilities and refining the ranking algorithm. Over time, this leads to:

- Reduced manual intervention
- Improved handling of complex morphological patterns
- Better adaptation to evolving language usage, including neologisms and regional variations

7.5. Practical Outcomes

By combining automated expansion with expert validation, the system has successfully integrated thousands of new lemmas, significantly enhancing Georgian lexical resources. This expanded lexicon supports a wide range of NLP applications, including machine translation, spell checking, educational tools, and AI-based dialogue systems.

The methodology demonstrates scalability and adaptability, offering a model that can be extended to other under-resourced languages facing similar morphological challenges (Scannell, 2007).

8. Results and Discussion

The developed system for compiling grammatical dictionaries of the Georgian language and its dialects has undergone comprehensive evaluation to assess its effectiveness, scalability, and practical applicability in both linguistic research and NLP solutions.

8.1. Evaluation Metrics and Testing Environment

The system was tested using a diverse set of Georgian language corpora, including:

- Literary texts (e.g., works by Otar Chiladze)
- Dialectal materials collected from regional sources
- Modern digital content, such as news articles and social media excerpts

Key evaluation metrics included:

- Lemmatization accuracy (% of correctly identified base forms)
- Reduction in manual corrections
- Lexicon growth rate
- Processing speed for real-time applications

8.2. Lemmatization Performance

The affix-based lemmatization algorithm demonstrated:

- 98% accuracy across standard Georgian corpora, with only 2% of non-dictionary word forms requiring manual disambiguation.
- Significant improvement over traditional suffix-only methods, particularly when handling complex verb forms and dialectal inflections.
- Efficient processing of both prefixing and suffixing phenomena, which are characteristic of Georgian morphology.

This confirms the system's robustness in addressing the agglutinative-inflectional nature of the language, outperforming previous models that struggled with multi-layered affixation patterns.

8.3. Lexicon Expansion Outcomes

Through automated acquisition and validation cycles:

- The lexicon was expanded by over 15,000 new lemmas, including numerous dialectal forms previously undocumented in digital resources.
- Integration with GeWordNet enhanced semantic richness, allowing for advanced linguistic applications beyond simple morphological analysis.

8.4. Practical Applications

The system's flexibility enabled deployment in various domains:

- Educational Platforms: AI-assisted tools for language learning, offering contextual exercises and morphological analysis for students at different proficiency levels.
- Machine Translation: Improved handling of Georgian morphology in MT systems by providing accurate base forms and grammatical features.
- Dialogue Systems: Development of intelligent, Georgian-language chatbots capable of understanding and generating morphologically correct responses, including dialectal variations.

8.5. Discussion of Limitations

While the system achieved strong results, certain challenges remain:

- Low-frequency dialectal forms occasionally lead to ambiguous lemmatization due to limited corpus data.
- The system requires ongoing updates to accommodate neologisms and evolving language usage, particularly in informal digital communication.
- Full automation is not yet feasible for highly irregular word forms, necessitating continued human oversight in specific cases.

8.6. Future Directions

To address these challenges, future work will focus on:

- Expanding the training corpora with more dialectal and conversational data.
- Enhancing machine learning components to better predict rare morphological patterns.
- Collaborating with international projects focused on under-resourced languages to adapt and share methodologies.

9. Conclusion

The research demonstrates that a hybrid approach—combining affix-based algorithms, semantic integration, and expert validation—can effectively automate the compilation of grammatical dictionaries for complex languages like Georgian. The system not only advances computational linguistics for Georgian but also provides a scalable framework applicable to other Kartvelian and agglutinative languages.

This work underscores the critical role of integrating linguistic theory with computational innovation to overcome the unique challenges posed by morphologically rich languages in digital environments.

References

Amirezashvili, N., Lortkipanidze, L., & Javashvili, N., & Samsonadze, L. (2017). Syntax annotation of the Georgian literary corpus. In *Theoretical Computer Science and General Issues. 11th International Tbilisi Symposium on Logic, Language, and Computation, TbiLLC 2015* (pp. 89–97). https://doi.org/10.1007/978-3-662-54332-0_6

Archvadze, N., & Pkhovelishvili, M. (2012). Representation of the Georgian language dictionary using functional programming languages. *GESJ: Computer Science and Telecommunications*, (2), 59–70.

Archvadze, N., Pkhovelishvili, M., & Shetsiruli, L. (2014). Questions of database interfaces in the Georgian language. In *Proceedings of the VII International Conference on Applied Linguistics* (pp. 83–86).

Archvadze, N., Pkhovelishvili, M., & Shetsiruli, L. (2017). Morphological analysis of words in clusters. In *R. Piotrowski's Readings in Language Engineering*.

Forsberg, M., & Ranta, A. (2004). Functional morphology. In *Proceedings of the Ninth ACM SIGPLAN International Conference on Functional Programming* (pp. 213–223). ACM. https://doi.org/10.1145/1016850.1016878

Hajič, J. (2000). Morphological tagging: Data vs. dictionaries. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP)* (pp. 94–101). Association for Computational Linguistics. https://aclanthology.org/A00-1012/

Kanis, J., & Müller, L. (2015). Automatic lemmatizer construction with focus on OOV words. In *Text, Speech and Dialogue* (*Lecture Notes in Computer Science*) (pp. 132–139).

Karttunen, L., Chanod, J.-P., Grefenstette, G., & Schiller, A. (1996). Regular expressions for language engineering. *Natural Language Engineering*, 2(4), 305–328. https://doi.org/10.1017/S1351324900001815

Lortkipanidze, L. (2006). Application of GeoTrans system in the Georgian spell checker. In *Proceedings of the LEPL Archil Eliashvili Institute of Control Systems* №10 (pp. 187–192).

Lortkipanidze, L., Beridze, M., & Nadaraia, D. (2015). The Georgian dialect corpus: Problems and prospects. In J. Gippert & R. Gehrke (Eds.), *Historical Corpora: Challenges and Perspectives* (CLIP, Vol. 5) (pp. 323–334). Narr Francke Attempto Verlag.

Lortkipanidze, L., & Gegechkori, M. (2016). Lexical Ontology GeWordNet. Proceedings of Archil Eliashvili Institute of Control Systems №20, (pp. 148-152).

Mikheev, A. (1997). Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23(3), 405–423.

Robins, R. H., & Waterson, N. (1952). Notes on the phonetics of the Georgian word. *Bulletin of the School of Oriental and African Studies*, 15, 55–72.

Scannell, K. P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In C. Fairon, H. Naets, A. Kilgarriff, & G.-M. de Schryver (Eds.), *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop (WAC3)* (pp. 5–15). Louvain-la-Neuve, Belgium.

Tiberius, C., & Schoonheim, T. (2014). *A frequency dictionary of Dutch: Core vocabulary for learners*. Routledge.

Yarowsky, D., & Wicentowski, R. (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (pp. 207–216). Association for Computational Linguistics. https://doi.org/10.3115/1075218.1075245