# Building a Universal Dependencies Treebank for Georgian

**Irina Lobzhanidze, Erekle Magradze, Svetlana Berikashvili, Anzor Gozalishvili, Tamar Jalaghonia**

Ilia State University
irina_lobzhanidze@iliauni.edu.ge, erekle.magradze@iliauni.edu.ge,
svetlana.berikashvili@iliauni.edu.ge,
anz2.gozalishvili@gmail.com, jalaghonia98@gmail.com

## Abstract

This paper presents the design and development of the Georgian Syntactic Treebank within the Universal Dependencies (UD) framework, addressing the unique morphosyntactic challenges of Georgian, a Kartvelian language. We describe the methodology for selecting and annotating 3,013 sentences from Wiki, mapping existing tagsets to the UD scheme, and converting data into the CoNLL-U format. The paper also details the training of a UDPipe model using this preliminary treebank.

## 1 Introduction

The development of syntactic treebanks is essential for advancing natural language processing (NLP) across diverse languages, enabling computational models to better understand and process linguistic structures. The Universal Dependencies (UD) (Nivre et al., 2017) framework provides a standardized approach to syntactic annotation that facilitates cross-linguistic consistency and data sharing. The data freely available on GitHub is generally used for training various models like UDPipe (Straka, 2016), UDify (Kondratyuk et al., 2019), Stanza (Qi et al., 2020) and others.

However, many languages, particularly those with complex morphosyntactic characteristics, remain underrepresented in these resources. Georgian, a Kartvelian language, is one such language that presents challenges due to its split-ergative structure, free word order, and rich inflectional morphology. This paper addresses the compilation of a Georgian Syntactic Treebank consisting of 151 utterances (2123 tokens) from the Georgian Language Corpus (GLC) and 3013 utterances (54116 tokens) from Wiki; totaling 3164 utterances (56239 tokens). This work contributes to the development of computational tools for under-resourced languages.

The paper consists of five sections. The first section provides a brief review of previous work concerning the Georgian language. The second section offers a detailed description of the data selection, annotation process, tagset mapping, and conversion to the CoNLL-U format. The third section includes information on the training of the UDPipe model, and the fourth section presents the results and their analysis. The fifth section summarizes the findings.

## 2 Background on Georgian Language Treebank

The development of treebanks for Kartvelian languages, a family characterized by its unique morphosyntactic structure and phonological properties, can be considered as new within the field of natural language processing (NLP). From this perspective the syntactic Treebank of the Laz language, another Kartvelian language, can be considered as the first attempt to create the Universal Dependencies Treebank and to make it available online (Turk et al. 2020). Common features shared by Georgian and other Kartvelian languages include the following:

- A relatively uniform sound system;
- A well-developed system of word inflection and derivation;
- Agglutinating and inflecting systems that make use not only of a large variety of grammatical affixes, but also of ablaut and other types of processes typical of internal stem inflection;



*Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 40 - 45

- The split-ergativity (Boeder 1979; Harris 1981, 1985; Hewitt 1983, 1987; Tuite 2017; Baker and Bobaljik 2017; Berikashvili 2024 and others).

All of these features pose unique difficulties at all levels of language processing and present interesting challenges for the compilation of robust language processing systems.

Prior to the efforts documented in this paper, Georgian had been largely underrepresented in major syntactic annotation initiatives such as the UD framework. While, various research groups (Datukishvili, 1997; Gurevich, 2006; Kapanadze, 2009 and others) have developed some tools for the processing of Modern Georgian morphology or for the creating of corpora (Gippert et al., 2011; Doborjginidze et al., 2012), the problem of syntax remained unsolved. Early attempts to create syntactic resources for Georgian included efforts to develop the ParGram Treebank within the Lexical Functional Grammar (LFG) framework (Sulger et al., 2013) and the GRUG treebank combining constituency-based and dependency-based structures (Kapanadze, 2017). But tagsets (Erjavec, 2004; Meurer, 2007 and others) and annotation schemes were not fully compatible with the UD framework, preventing their integration and wider use. Thus, it was important to adapt the existing tagsets and the mapping of Georgian linguistic features to the UD framework, to ensure that the syntactic annotation of Georgian could align with the UD, allowing the possibility of comparative linguistic studies.

As a result, the initial test version was limited in coverage and consisted of 151 utterances, that did not fully capture the linguistic characteristics of Georgian. Additionally, the tools available for syntactic parsing, such as the UDPipe model, had not been trained on sufficient Georgian data.

## 3 Methodology and Annotation Process

The development of the Georgian Syntactic Treebank followed a systematic approach to address the specific challenges posed by the Georgian language's complex morphosyntactic structure. The methodology encompassed several key strategies: determining syntactic functions and compiling annotation guidelines, improving the initial annotation scheme developed for the initial 151 utterances from the GLC by revising and standardizing the use of dependency relations, selecting and annotating data, and contributing to the UD GitHub repository. Additionally, the training of the UDPipe model using the annotated data is described in detail.

### 3.1 Data Selection

The Georgian Language Corpus (GLC) (Doborjginidze et al. 2012) served as the initial source for the treebank, offering a collection of texts of different genres and periods (6th-21st centuries). From this corpus, a total of 151 sentences reflecting Modern Georgian were selected. The selection criteria focused on ensuring a representative sample of Georgian syntax, including various sentence lengths, structures, and complexity levels. But these data were not enough to train the model and to complement the data from the GLC and introduce a more diverse linguistic style were also selected from Georgian Wikipedia. As a result, 3,013 sentences were selected from Wikipedia, covering 131 different scientific domains. The selection process prioritized sentences that demonstrate a variety of syntactic constructions, including simple, coordinated and subordinated complex clauses, as well as those that feature unique or less common linguistic phenomena. All these sentences were checked to include different morphosyntactic features.

### 3.2 Data annotation

The annotation process was preceded by the compilation of annotation guidelines and the development of the UD annotation scheme for Georgian. These guidelines were made available in the language-specific documentation section of the UD GitHub repository[1]. The development of the scheme for Georgian involved adapting the tags used in the Georgian morphological analyzer (Lobzhanidze 2022) to ensure compatibility with UD standards. After the mapping of tagsets, a special Python code was written to convert the analyzer's output into the CoNLL-U format and to provide additional tokenization. It was especially important to provide segmentation of multi-word tokens, which were not covered by the analyzer's

[1] https://universaldependencies.org/ka/index.html

output and to fill information on lemmas, part-of-speech (POS) tags, and morphosyntactic features. The main differences between the analyzer's output and the UD scheme like tokenization as well as different linguistic phenomena connected to split-ergativity and other features of Georgian can be summarized as follows: a) the main core dependency arguments, which are used in Georgian are nominal subject, direct and indirect objects. While in Indo-European languages, the verb generally agrees with the subject of the sentence, in Georgian the verb agrees not only with the subject, but with its objects (direct and indirect) as well. However, as a result of the strong Person Case Constraint (PCC) effect, the direct object is always the third person in ditransitive constructions, and the third person agreement is always null. Therefore, there are no cases where all three arguments agree simultaneously. As a result, Georgian verbs have core and peripheral arguments. A core argument agrees morphologically with the verb by means of person and number markers, while a peripheral argument does not. In Georgian, a nominal subject is a nominal that serves as the subject of the verbal predicate in ergative or nominative or dative cases; a direct object is a nominal or noun phrase that serves as the object of the verbal predicate in nominative or dative; the indirect object of a verb is a dative-marked complement. The Georgian treebank uses all the main non-core dependent's tags except of `expl` and `dislocated`. All nominal dependent tags are used except of classifier `(clf)`. As a result each sentence was annotated to capture syntactic dependencies, including subject, object, and modifier relationships. Taking into consideration the complexity of Georgian syntax - characterized by split-ergativity and free word order - special attention was paid to accurately representing the syntactic roles of words within sentences and to the case-marking of subject, direct and indirect objects.

### 3.3 UDPipe Model Training

To evaluate the quality of the annotations and provide a baseline for further development, a UDPipe model was trained using the annotated data. The training set consisted ka_glc-ud-dev.conllu (470 utterances), ka_glc-ud-test.conllu (481 utterances) and ka_glc-ud-train.conllu (2213 utterances) . The UDPipe model was trained on the Georgian data using the default parameters. Performance metrics, including tokenization accuracy, POS tagging accuracy, and parsing accuracy (both Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS)), were calculated to assess the model's effectiveness.

### 3.4 Validation and Corrections

Following the automatic annotation and model training, a manual validation process was implemented. This involved reviewing a sample of the annotated sentences to identify and correct errors in tokenization, POS tagging, and syntactic annotation. Corrections were made directly in the CoNLL-U files, and the model was retrained as necessary to incorporate these improvements.

### 3.5 Contribution to the UD repository

The validated treebank files, including ka_glc-ud-test.conllu and ka_glc-ud-train.conllu, were uploaded to the repository, along with related documentation files such as README.md. The treebank passed the UD validation process[2]. At this moment the treebank is available in the dev branch of the repository and will be unified with the master branch after the twenty-first release of annotated treebanks in Universal Dependencies, v2.15, to be implemented in November.

## 4 Model training

UDPipe (Straka et al. 2016) is a trainable pipeline for tokenization, tagging, lemmatization, and dependency parsing of CoNLL-U files. UDPipe is language-agnostic and can be trained given annotated data in CoNLL-U format. For the Georgian language case, we have used Version 1.3.1-dev. Data training has been implemented on 3164 utterances (sentences) consisting of 56239 tokens. We trained UDPipe models (tokenizer, tagger, parser) using training set. The method used for training was "morphodita_parsito" which is the only supported method in udpipe version 1.3. We used default parameters for each model in a pipeline. The training results are as follows:

- Tokenizer: Epoch 44, logprob: -1.6215e+03, training acc: 99.87%, heldout tokens:

[2] https://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/validation-report.pl?UD_Georgian-GLC

99.83%P/99.84%R/99.84%, sentences: 98.08%P/97.87%R/97.97%;

- Tagger: Iteration 20: done, accuracy 99.85%, heldout accuracy 89.49%t/91.80%l/85.38%b;

- Parser: Iteration 8: training logprob -2.0778e+04, heldout UAS 79.04%, LAS 74.75%

While the testing for accuracy on ka_glc-ud-test.conllu gives the following results:

- Tokenizer: Number of SpaceAfter=No features in gold data: 1523; Tokenizer tokens - system: 9288, gold: 9283, precision: 99.69%, recall: 99.74%, f1: 99.71%; Tokenizer multiword tokens - system: 742, gold: 751, precision: 97.71%, recall: 96.54%, f1: 97.12%; Tokenizer words - system: 10035, gold: 10039, precision: 99.15%, recall: 99.11%, f1: 99.13%; Tokenizer sentences - system: 497, gold: 481, precision: 92.35%, recall: 95.43%, f1: 93.87%

- Tagger: Tagging from gold tokenization - forms: 10039, upostag: 93.34%, xpostag: 93.34%, feats: 85.42%, alltags: 85.18%, lemmas: 89.89%

- Parser: Parsing from gold tokenization with gold tags - forms: 10039, UAS: 80.34%, LAS: 76.01%

Comparing the results some frequent misinterpretations were noted concerning the complex subordinate clauses. The gold standard files included more complex structures, while the parser tried to simplify them. For example, the parser sometimes had difficulties distinguishing the subject and object of sentences marked with `Case=Nom` or `Case=Dat`, which can be explained by the split-ergativity of Georgian. Additionally, it assigned the modifier relation differently depending on sentence context or positional emphasis, and showed discrepancies in the representation of clitics like postpositions and particles.

## 5 Results and discussion

The primary outcome of this project is the creation of an initial Georgian Syntactic Treebank, consisting of 3164 sentences (56239 tokens). This treebank was developed by mapping existing Georgian linguistic resources to the UD framework, ensuring compatibility with cross-linguistic standards. The treebank was validated and made available for use within the UD community, representing a significant milestone for the computational processing of the Georgian language. The main components of the treebank are the following:

- Universal POS Tags (UPOS): The mapping of Georgian part-of-speech tags to the UD's UPOS tags ensured the cross-linguistic consistency of the treebank. The main difference revealed is as follows: `NOUN` and `PROPN`. as opposed to `+Noun+Com` and `+Noun+Prop`;

- Morphological Features (FEATS): The detailed morphological features in the FEATS column allowed the representation of Georgian's morphosyntactic properties. We have added `AdpType`, `AdvType`, `PartType`, `NameType`, `VerbType`, `Subcat`, `PunctType` to Lexical Features; `NumForm` to Inflectional Features and `Person[subj]`, `Person[obj]`, `Person[io]`, `Number[subj]`, `Number[obj]`, `Number[io]` to Verbal Features;

- Syntactic Dependencies (DEPREL): The syntactic annotation, including the identification of heads and dependency relations, provided a structured representation of Georgian syntax. We have used all tags except `expl`, `dislocated`, `clf`, and `reparandum`.

At the same time, the implementation of the project revealed some areas for further improvement:

- Mapping and Compatibility: The mapping of Georgian morphosyntactic tagset to the UD revealed that some features and categories were not directly compatible with existing UD tags. For instance, some of the tags indicating voice and connected to the category of diathesis are not compatible with the UD framework (e.g. autoactive, inactive (inverse active);

- Annotation Accuracy: The treebank was validated through a series of automated and manual checks by two annotators, ensuring the accuracy of the syntactic annotations. The reliance on existing tools like the morphological analyzer and the UD validator can be considered as effective, but the manual correction highlighted the importance to add some additional syntactic dependencies like `flat:foreign`, `flat:name` etc.;

- Challenges in Complex Structures: The analysis identified particular difficulties in accurately annotating sentences with complex syntactic structures, such as those involving multiple clauses, valency-changing operations, and free word order. The Georgian verb reflects relations between two or three arguments and provides a mapping between morphology and syntactic features such as the roles of participants. Especially, impersonal verbs do not have a subject at all, intransitive verbs take a subject only, indirect transitive verbs take two arguments: a subject and an indirect object; transitive verbs take two arguments: a subject and a direct object and, ditransitive verbs take three arguments: a subject and a direct and indirect object. As a result, the subject can be marked by the nominative, ergative or dative cases, while the objects are marked by the nominative or dative case with or without a postposition. All these affected the correct marking of arguments at the level of syntactic dependencies.

## 6 Conclusions

By this study we tried to represent an advancement in the development of linguistic resources for the Georgian language, particularly through the creation of a syntactic treebank within the Universal Dependencies (UD) framework. The implementation of this project has provided a resource for the computational processing of Georgian, addressing main challenges related to the complex morphosyntactic structure and contributing to the broader field of natural language processing (NLP) for under-resourced languages. Expanding the treebank with the complete GLC data, updating the UDPipe model can be considered as important future steps to improve the accuracy of Georgian NLP tools.

## Acknowledgments

This study was funded by the Shota Rustaveli National Science Foundation (No FR-22-20496) and the JESH (Joint Excellence in Science and Humanities) grant, financed by the Austrian Academy of Sciences. The authors would like to thank the anonymous reviewers for their helpful comments and feedback.

## References

Mark Baker and Jonathan Bobaljik. 2017. On inherent and dependent theories of ergative case. *In J. Coon, D. Massam, & L. Travis, The Oxford Handbook of Ergativity*. Oxford: Oxford University Press, pages 111–134. https://doi.org/10.1093/oxfordhb/9780198739371.013.5.

Svetlana Berikashvili. 2024. *Differential Subject Marking in Georgian.* Göttingen: Georg-August-Universität Göttingen.

Winfried Boeder. 1979. Ergative syntax and morphology in language change: the South Caucasian languages. *In F. Plank, Ergativity: towards a theory of grammatical relations*. Orlando: Academic Press, pages 435–480.

Ketevan Datukishvili. 1997. Some questions of computer synthesis of verb in Georgian. *In Proceedings of the Second Tbilisi Symposium on Language, Logic and Computation.* Tbilisi: t'bilisis saxelmcip'o universiteti (Tbilisi State University), pages 83-85.

Nino Doborjginidze, Irina Lobzhanidze, and Irakli Gunia. 2012. *Georgian Language Corpus (GLC)*. http://corpora.iliauni.edu.ge/ (Accessed: 16 August 2024).

Tomaž Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04),* Lisbon, Portugal: European Language Resources Association (ELRA).

Jost Gippert, Paul Meurer, Manana Tandashvili. 2011. *The Georgian National Corpus (GNC).* http://gnc.gov.ge/ (Accessed: 16 August 2024).

Olga Gurevich. 2006. A Finite-State Model of Georgian Verbal Morphology. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL.* New York: Association for Computational Linguistics, Pages 45-48.

Alice Harris. 1981. *Georgian Syntax: a study in Relational Grammar*. Cambridge: Cambridge University Press.

Alice Harris. 1985. *Diachronic syntax: the Kartvelian Case*. New York: Academic Press.

Oleg Kapanadze. 2009. "Describing Georgian Morphology with a Finite-State System." *In Proceedings of the 8th international conference on Finite-State Methods and Natural Language Processing.* Pretoria: Springer. 114-122.

Oleg Kapanadze. 2017. *Multilingual GRUG Parallel TreeBank — Ideas and Methods.* Saarbrücken: LAMBERT Academic Publisher.

Dan Kondratyuk, Milan Straka. 2019. 75 Languages, 1 Model: Parsing Universal Dependencies Universally. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Hong Kong: China, pages 2779–2795.

Irina Lobzhanidze. 2022. *Finite-State Computational Morphology: An Analyzer and Generator for Georgian*. Cham: Springer.

Paul Meurer. 2007. A Computational Grammar for Georgian. *In Lecture Notes in Computer Science.* Berlin: Springer, pages 1-15.

Joakim Nivre, Daniel Zeman, Filip Ginter, Francis Tyers. 2017. Universal Dependencies. *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Online: Association for Computational Linguistics. 101–108.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).* Portorož, Slovenia: European Language Resources Association (ELRA), pages 4290–4297.

Sebastian Sulger, Miriam Butt, Tracy King, Paul Meurer, Tibor Laczko, Gyorgy Rákosi, Cheikh Dione, Helge Dyvik, Victoria Rosen, Koenraad De Smedt, Agnieszka Patejuk, Ozlem Çetinoğlu, I Wayan Arka, Meladel Mistica. 2013. ParGramBank: The ParGram Parallel Treebank. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Sofia, pages 550–560.

Kevin Tuite. 2017. Alignment and orientation in Kartvelian. *In J. Coon, D. Massam, & L. Travis, The Oxford Handbook of Ergativity*. Oxford: Oxford University Press, pages 1114–1138.

Utku Turk, Kaan Bayar, Aysegul Dilara Ozercan, Gorkem Yigit Ozturk, and Saziye Betul Ozates. 2020. First Steps towards Universal Dependencies for Laz. *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020).* Barcelona, Spain (Online): Association for Computational Linguistics, pages 189–194.