Irina Lobzhanidze, Erekle Magradze, Svetlana Berikashvili,
Anzor Gozalishvili, Tamar Jalaghonia
Ilia State University

Georgian Syntactic Treebank and UDPipe model

One of the most crucial Natural Language Processing (NLP) tasks is associated with the universality-driven development of language resources for different languages (e.g., Universal Dependencies (UD), UniMorph, PARSEME, etc.). The research describes the possibility of creating a Syntactic TreeBank for Georgian and consists of four sections: 1. Linguistic Resources and Syntactic Annotation; 2. Tools and Mapping to UD Format; 3. Principles of Syntactic Annotation and CoNLL-U Format; 4. Summary and UDPipe Model Development.

1. Linguistic Resources and Syntactic Annotation

The first section highlights the necessity of syntactic annotation and the compilation of the syntactic treebank for Georgian. Special attention is given to the Universal Dependencies (UD) repository, which lacks substantial data on Kartvelian languages, with the exception of the Laz language and the Georgian language. The current Laz treebank consists of 576 sentences and
2,306 tokens, annotated according to UD guidelines. The current Georgian treebank consists of 3164 sentences and 56239 tokens. By this moment the lack of data for Georgian and other Kartvelian languages was primarily due to two factors:

- Data Scarcity: The available data was and is insufficient to effectively train NLP tools for these languages.
- Tool Incompatibility: Tools developed for other languages cannot be easily adapted for Kartvelian languages.

Given these challenges, the description of available linguistic resources for the Georgian language focuses firstly on tools that can be used for annotating selected sentences, and secondly, on converting these sentences into the UD format. Additionally, the results of UDPipe models trained on Georgian data are discussed. UDPipe, a trainable pipeline for tokenization, tagging, lemmatization, and dependency parsing, can be utilized to enhance the annotation process for Georgian and to enrich the available datasets.

This structured approach aims to address the current gaps in linguistic data for Georgian and contribute to the broader goal of enriching the UD repository with comprehensive and accurate syntactic annotations for Kartvelian languages.

2. Tools and Mapping to UD Format

The second section delves into the methodologies and tools utilized for adapting existing Georgian linguistic resources to the Universal Dependencies (UD) format. Special attention is given to the annotation tools developed for Georgian, including the tokenizer for splitting Georgian text into sentences and tokens, and the morphological analyzer and

generator developed by Lobzhanidze (2022), which assigns information on lemmas, parts of speech (PoS), and morphosyntactic tags. However, these tools do not provide syntactic analysis capabilities. Furthermore, the data from the Georgian Language Corpus (GLC) and the KartNLP tool, developed based on finite state tools, are insufficient for comprehensive syntactic analysis. The existing tagsets, as per the MULTEX-EAST specification (Lobzhanidze
2021; Erjavec, 2004), also fall short in this regard.

3. Principles of Syntactic Annotation and CoNLL-U Format

The third section provides an in-depth description of the syntactic annotation principles applied in developing the Georgian Syntactic TreeBank. It represents the theoretical and practical guidelines followed to annotate syntactic structures consistently and accurately. Additionally, this section details the CoNLL-U format, a standard format used for annotating and sharing syntactic information in a machine-readable manner. Special attention is paid to the annotation guidelines for Georgian, which encompasses: a) simple clauses – providing annotation of transitive and intransitive clauses, including valency-changing operations and, b) complex clauses - addressing the annotation of predicates, coordinated and subordinated clauses, and other necessary constructions.

4. Summary and UDPipe Model Development

The fourth section offers a comprehensive summary of the project, encompasses the methodologies used during the creation of the Syntactic TreeBank for the Georgian language, and describes a significant outcome of the project, especially, the development of the UDPipe Model for Georgian. UDPipe is a tool designed to facilitate the automatic parsing and annotation of texts. UDPipe is language-agnostic, meaning it can be trained on annotated data provided in CoNLL-U format. For the Georgian language, Version 1.3.1-dev of UDPipe has been utilized and the training for the Georgian UDPipe model has been implemented on a dataset comprising 3,164 sentences and 45,874 tokens. The results of this training include various metrics across different components of the pipeline, especially, tokenizer's results, tagger's results and parser's results.

Keywords: *universal dependencies, Georgian syntactic treebank, UDPipe model.*

This work was supported by Shota Rustaveli National Science Foundation of Georgia (SRNSFG) [FR-22-20496]

References:

Doborjginidze, Nino, Lobzhanidze, Irina. (2016). Corpus of the Georgian Language. *Proceedings of the XVII EURALEX International Congress* (pp. 328-335). Tbilisi: Ivane Javakhishvili Tbilisi University Press.

Doborjginidze, Nino, Lobzhanidze, Irina, and Gunia, Irakli. (2012, December 19). *Georgian Language Corpus*. Retrieved October 30, 2019, from http://corpora.iliauni.edu.ge/

Erjavec, T. (2004, May 10). *MULTEXT-East Morphosyntactic Specifications*. Retrieved from Version 3.0: http://nl.ijs.si/ME/Vault/V3/msd/html/

Lobzhanidze, I. (2021, August 20). *MULTEXT-East Morphosyntactic Specifications, revised Version Georgian Specifications*. Retrieved from MULTEXT-East Morphosyntactic Specifications: http://nl.ijs.si/ME/V6/msd/html/msd-ka.html

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman. (2021).
Universal Dependancies. *Computational Linguistics 47 (2)*, 255–308.

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

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