Irina Lobzhanidze, Erekle Magradze, Svetlana Berikashvili, Anzor Gozalishvili, Tamar Jalagonia
Ilia State University, KakutsaCholokashvili Ave 3/5, Tbilisi 0162, Georgia

Syntactic Annotation of Georgian in the UD Schemes

Section 1. Introduction

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)[1], UniMorph[2], PARSEME[3] etc.). One of the most important resources from this perspective is the Universal Dependencies (UD) Initiative, which provides cross-linguistically consistent grammatical annotation for different languages and the possibility to share syntactic Treebanks online. The above-mentioned resources lack data on Kartvelian languages; even Georgian is not represented at the appropriate level and lacks appropriately annotated Georgian Treebank with sentences structured into syntactic trees. The above-mentioned issues are caused by two facts: firstly, Georgian suffers from data scarceness, i.e. the amount of data to train NLP tools is not sufficient and, secondly, the tools developed for other languages cannot be easily adopted in the case of Georgian due to the differences between morphosyntactic annotation schemes. The annotation tools developed for Georgian like tokenizer splitting Georgian text into sentences and tokens, morphological analyzer and generator (Lobzhanidze 2022) assigning information on lemmas, PoS-es, and morphosyntactic tags do not provide disambiguation and implementation of syntactic analysis; the data from the Georgian Language Corpus (GLC)[4] and the KartNLP[5] tool developed on the basis of the analyzer mentioned above also are not enough from the perspective of syntactic analysis.

Developing syntactic annotation schemes for the Georgian language is the basis for syntactic parsing and for improving the processing of languages with complex morphology, particularly those that are low-resourced. The present paper describes some of the issues associated with the development of the syntactic annotation schemes for Georgian implemented as a part of the project on the Georgian morphosyntactic computational analysis and tools for the annotation of universal syntactic dependencies (Projects No NEAR/TBS/2021/EARP/0086 and No FR-22-20496).

The paper consists of four sections. The introduction briefly describes the resources available for the annotation and tagging purposes of Georgian (XPOS level), existing corpora and levels of their annotation, and the importance of syntactic annotation. The second section focuses on the tools used for the mapping of the existing tagsets for Georgian to the UD format at the levels of UPOS and FEATS and the issues associated with mapping including those features which are omitted at UPOS and FEATS level, while the third section provides a thorough description of the syntactic annotation of Georgian with regard to the initial syntactic Treebank (Georgian-ud-test.conllu, README.md etc.) and language-specific documentation files (introduction.md and index.md) already uploaded to the GitHub repository[6]. The fourth section summarizes the work done and describes the future stages of the development of the syntactic TreeBank.

Section 2 Methods used, mapping between the existing tagsets of Georgian and the UD tagset

The theoretical prerequisites for the research on syntactic universals generally follow two theoretical approaches i.e. the functional-typological (Greenberg 1966, etc.) and the formal-generative (Chomsky 1976, etc.). The NLP realization of the above-mentioned theoretical approaches is generally focused on a unified framework, which provides cross-linguistic language description with regards to the Part-of-Speech (PoS) tagging, constituency, and dependency parsing and a corpus-based approach to Georgian. The development of the annotation scheme for Georgian includes, on the theoretical side, the description of language universals and underlying syntactic structures, while on the practical side, the use of language resources suitable for syntactic analysis. These prerequisites are used for the achievement of the following objectives: a) Determination of the syntactic functions and dependencies of the Georgian language; b) Compilation of the annotation guidelines for Georgian; c) Development of the UD annotation scheme for the separate PoS-es in accordance with the type of clauses: simple, subordinate complex, compound coordinate and, d) Compilation of a test Treebank. And the achievement of the above-mentioned objectives is closely connected to the dependency grammar methods used for the sentences of the Georgian dataset. Georgian, like other languages, has a hierarchical structure (Kvachadze, 1996)) containing nominals, modifiers, clauses, etc.

The tagsets developed for Georgian (Lobzhanidze, 2021-2022) have been compared with the universal UD tagsets and two lists of UD tags applicable for Georgian have been compiled:

---

[1] See, https://universaldependencies.org/, last accessed Oct. 12, 2022
[2] See, https://unimorph.github.io/, last accessed Oct. 12, 2022
[3] See, https://typo.uni-konstanz.de/parseme/, last accessed Oct. 12, 2022
[4] See, http://corpora.iliauni.edu.ge/, last accessed Oct. 12, 2022
[5] See, https://qartnlp.iliauni.edu.ge/, last accessed Oct. 12, 2022
[6] See, https://github.com/UniversalDependencies/UD_Georgian-GLC/tree/dev/ and https://github.com/UniversalDependencies/docs/blob/pages-source/_ka/ , last accessed Jul. 2, 2023

- A list of UPOS tags is used to mark the core part-of-speech categories. In addition to the UPOS tags, the CoNLL-U Format includes language-specific part-of-speech categories called XPOS tags. To demonstrate the direct correspondence between the existing tagsets, a list of XPOS tags was added to the test .conllu file. Several issues have been identified regarding the differences between the tagsets. For example, according to the Universal Dependencies (UD) approach, common nouns are marked as NOUN, while proper nouns are marked as PROPN. However, in the initial tagsets of Georgian (2021-2022), common nouns are represented as +Noun or Nc, and proper nouns are represented as +Noun+Prop or Np.
- A list of FEATS tags subdivided into lexical features e.g. PronType, NumType etc. and inflectional features for Nominal and Verbal paradigms e.g. Number, Case etc. Separate sub-lists for Nominal and Verbal inflections allow differentiating features used exclusively with nominals or verbs in Georgian. Some language-specific features are revealed as well, e.g. approximate numerals, different types of voice etc.

Syntactic tags applicable to Georgian have been represented in the form of a list of UD relations tags and compared to a list of Universal Dependency Relations available https://universaldependencies.org/u/dep/index.html. Language-specific information like tokenization and segmentation, information on Georgian morphology and syntax with information about the above-mentioned tags and features has been compiled and uploaded to the GitHub repository in the form of an initial introduction.md and index.md[7].

Section 3. Findings and Hypothesis

The Georgian Language Corpus (Doborjginidze et al. 2012) and annotation tools developed for Georgian: tokenizer, morphological analyzer, and generator (Lobzhanidze 2022) are used to partially fill the gap with regards to the lemmatization and morphosyntactic annotation of separate tokens. These resources were appended with regard to the labeling of syntactic functions. Therefore, the proposed paper is focused on the theoretical and computational approaches to the determination and assignment of clause boundaries and syntactic dependencies and, especially,

- Determination of the syntactic functions and dependencies of the Georgian language, which encompasses data description and analysis with regards to the phrase structure, functional relationships between constituents in a clause, and dependencies of the Georgian language;
- Compilation of the syntactic annotation guidelines for Georgian including information on nominals (e.g. modifier and function words dependents), simple (e.g. transitive and intransitive clauses, valency changing operation, etc.), and complex (e.g. predicate, coordinated and subordinated clauses) clauses and other constructions (if necessary);
- Development of the UD annotation scheme for the separate PoS-es in accordance with the type of clauses: simple, subordinate complex, compound coordinate includes documentation of tags, features, and syntactic relations and development of the UD annotation scheme in accordance with the annotation guidelines defined preliminary paying special attention to the dependency relations hold primarily between content words, the status of function words and the taxonomy of typed dependencies.

The initial Georgian-ud-test.conllu compiled for Georgian includes 151 utterances (sentences) or 2123 tokens. The number of sentences was determined as a minimum requirement for model training by means of UDPIPE[8]. The sentences have been randomly selected from the Georgian Language Corpus (Doborjginidze et al. 2012-2019) and consist of 10 columns including:

- ID: "Sentence" segmented according to dependency tree, "tokenization" from original GLC annotation supplemented with additional automatic tokenizaton of multiword tokens. Sentence numbering starting at 00001;
- FORM: word form or punctuation symbol from the original GLC annotation;
- LEMMA: lemma of word form determined from the original GLC annotation;
- UPOS: Universal part-of-speech tags Mapped from the original GLC annotation;
- XPOS: Language-specific part-of-speech tagset from Doborjginidze et al. (2012-2019) and Lobzhanidze (2022);
- FEATS: List of morphological features Mapped from the original GLC annotation;
- HEAD: Head of the current word determined and converted automatically with manual corrections;
- DEPREL: Universal dependency relation to the HEAD determined and converted automatically with manual corrections;
- DEPS: Enhanced dependency graph in the form of a list of head-deprel pairs determined and converted automatically with manual corrections;
- MISC: includes information on segmentation and transliteration converted automatically.

Section 4. Conclusions

[7] See, https://github.com/UniversalDependencies/docs/blob/pages-source/_ka/ , last accessed Jul. 2, 2023
[8] See, https://ufal.mff.cuni.cz/udpipe/2/models, last accessed Jul. 2, 2023

The UD tools provide cross-linguistically consistent Treebank annotations to develop multilingual parsers and cross-linguistic learning from a language typology perspective. And as there is no Georgian Treebank, the research on the syntactic model of universal dependencies is very important from different perspectives. Thus, the syntactic annotation scheme for Georgian represented in the paper can be considered as a first attempt to ensure the cross-linguistic compatibility of Georgian data and to compile the Treebank for Georgian. At this moment some features have to be added to language-specific documentation and a new update is due at the end of July. The future challenges are closely connected to the training of the Georgian linguistic model by means of the UDPIPE with the purpose to compile highly representative syntactic TreeBank for Georgian using the data already available at GLC.

Keywords: syntactic annotation, universal dependencies, Georgian language

References

Chomsky, N. (1976). *Reflections on language.* London: Temple Smith.

de Marneffe, Marie-Catherine, Bill MacCartney, Christopher D. Manning. (2021). Generating typed dependency parses from phrase structure parses. *Computational Linguistics, Vol. 47, Issue 2*, 255-308.

Doborjginidze, N., Lobzhanidze, I. (2012-2019). *Georgian language corpus.* Tbilisi: http://corpora.iliauni.edu.ge/.

Greenberg, J. (1966). *Universals of Language.* Cambridge: MIT Press.

Kvachadze, L. (1996). *t'anamedrove k'art'uli enis sintak'si (Syntax of Modern Georgian Language).* Tbilisi: Rubikoni.

Lobzhanidze, I. (2021, August 20). *MULTEXT-East Morphosyntactic Specifications, revised Version 6: Georgian Specifications*. Retrieved from MULTEXT-East Morphosyntactic Specifications: http://nl.ijs.si/ME/V6/msd/html/msd-ka.html

Lobzhanidze, I. (2022). *Finite-State Computational Morphology: An Analyzer and Generator for Georgian.* Cham: Springer.

Nivre, Joakim, Bandmann Megyesi, Beáta. (2007). Bootstrapping a Swedish Treebank Using Cross-Corpus Harmonization and Annotation Projection. *NEALT Proceedings Series, Vol. 1* (pp. 97-102). Bergen: Northern European Association for Language Technology (NEALT).