

# Parsing the Less-configurational Georgian Language with a Context-Free Grammar

**Oleg Kapanadze**

Tbilisi State University

1, Chavchavadze Ave.,

0179 Tbilisi

Georgia

okapanadze@uni-potsdam.de

## Abstract

A large part of the methodology for Natural Language Processing has been developed for languages with a strong syntactic configuration. At the other end of the configurational spectrum there are languages with rich derivational and inflectional morphology. These languages for *morphologically rich and less-configurational* features are referred to as **MR&LC**. In our study we have addressed Georgian - a language with less-configurational constraints, though, with a rich inflectional morphology and a very little fixed structure on the sentence level, and therefore, the most syntax-level information for the Georgian language is conveyed by its productive morphology.

This paper features issues concerned with development of a crucial NLP resource for the Georgian language - a Context-Free Syntactic Parser.

**Keywords:** Georgian Language Processing, Morphologically rich and less-configurational languages, Context-Free Syntactic Parser

## Résumé

ბუნებრივი ენების ტექნოლოგიისათვის განკუთვნილი მეთოდოლოგიის უდიდესი ნაწილი შემუშავებულია სინტაქსური კონფიგურაციაზე მკაცრი შეზღუდვების მქონე ენებისთვის. კონფიგურაციული სპექტრის საპირისპირო მხარეს წარმოდგენილია ენები, რომლებიც ნაყოფიერი დერივაციული და ფლექსიური მორფოლოგიით გამოირჩევიან. ასეთი ტიპის ენებს მათი ფართო მორფოლოგიური შესაძლებლობებისა და სინტაქსური თვალსაზრისით ნაკლებად შეზღუდულობის გამო *მდიდარი მორფოლოგიისა და ნაკლებად კონფიგურირებულ - მმ&ნკ* ენებს უწოდებენ. ამ კუთხით ჩვენი კვლევის საგანია ქართული - ენა, რომელს სტრუქტურა მცირე კონფიგურაციული შეზღუდვებით გამოირჩევა და მდიდარი ფლექსიური მორფოლოგიის წყალობით წინადადების დონეზე ნაკლებად ფიქსირებული სინტაქსური სტრუქტურებით არის წამოდგენილი. ამავდროულად ქართული ენის წინადადების სინტაქსური სტრუქტურის შესახებ ინფორმაცია ზმნის ვალენტობისა და მისი შესაბამისი მორფოლოგიური მარკერების საშუალებით არის ხელმისაწვდომი.

წინამდებარე სტატიაში განიხილულია საკითხები, რომლებიც ეხება ქართული ენის ტექნოლოგიისათვის უმნიშვნელოვანესი რესურსის - კონტექსტისაგან დამოუკიდებელი სინტაქსური პარსერის შემუშავებას.

## 1. Introduction

A large part of the methodology for natural language processing (NLP) has been developed for English which is known as a strongly configurational language. Hence, nearly all the syntactic information needed by any NLP application for English can be obtained by configurational analysis. At the other end of the configurational spectrum are the languages with rich derivational and inflectional morphology, such as Georgian that has very little fixed structure on the sentence level. These languages for *morphologically rich and less-configurational* features are referred to as **MR&LC** (Fraser et al., 2001). All of them are thriving to get a place in the modern digital world and in order to profit of the new opportunities offered by the Internet and digital devices must be modeled for using in high-quality computing systems. The long-term viability of languages not specifically supported by Human Language Technology is therefore put at risk and they can seriously face digital extinction.

There are a multitude of academic grammars and dictionaries developed for the Georgian language. However, this does not mean that there is a sufficient support

for computational applications involving Georgian, as these resources are not suited for NLP needs.

The proposed presentation will feature issues concerned with the development of a crucial NLP resource — a syntactic parser for the Georgian language. To this end we used a methodology that will extract a FST grammar and a consequent lexicon from a monolingual Georgian TreeBank. The compiled language resources will be utilized for the Georgian text syntactic annotation which terminal nodes are saturated with rich morphologic features.

## 2. Treebanking in NLP

A monolingual *TreeBank* is a parsed corpus in which sentences are annotated with syntactic structure. They are skeletal parses showing syntactic information – a *bank* of linguistic *trees*. Syntactic structure is commonly represented as *a tree structure* (in Mathematical terms – *an Oriented Graph*), hence the name *TreeBank*.

TreeBanks have become valuable resources as repositories for linguistic research, since corpus-based methods became useful in multilingual lexicography playing an important role in empirical language studies. They can be used in *languages contrastive studies* and *translation science*, in *corpus linguistics* for studying syntactic phenomena, in computational linguistics as evaluation corpora for different Human Language Technology systems or for training and testing *parsers* and as a database for *Translation Memory* systems.

*TreeBanks* can be created completely manually or semi-automatically, where a parser assigns some syntactic structure to a text that is then checked by linguists and, if necessary, corrected. Treebanks are often created on top of a corpus that has already been annotated with part-of-speech tags. The annotation can vary from constituent to dependency or tecto-grammatical structures. Additionally, treebanks are sometimes enhanced with semantic or other linguistic information.

Some TreeBanks follow a specific linguistic theory (e.g. the Bulgarian language follows HPSG), but most try to be less theory-specific. However, two main groups can be distinguished: treebanks that annotate *phrase structure* (the *Penn TreeBank* for Arabic, English and Chinese) and those that annotate *dependency structure* (the *Prague Dependency TreeBank* for the Czech language).

A significant part of modern treebanking literature is devoted to creation of large TreeBanks for the languages with a relatively simple morphology and the fixed word order. Data-driven treebanking is now at the state where naturally occurring text in the news domain can be automatically annotated with high accuracy according to standard parsing evaluation measures. However, when moving from languages with relatively fixed word order to languages with richer morphologies and less-rigid word orders, the standard issues for annotation TreeBanks developed for languages with fixed word order exhibit a large drop in accuracy.

### 3. Creating a Georgian Treebank and a Vanilla CFG

There are constituent TreeBanks for several languages in existence, along with a very limited number of parsing reports on them. The main challenge of constituent parsing for morphologically rich languages is in the handling of the huge number of word forms. According to the reports, the size of the preterminal set in the standard context-free grammar environment is crucial. If we use only the main part-of-speech (POS) tags as preterminals (as is the case with the strongly configurational languages), a considerable amount of information, encoded in the morphological description of the tokens, will be lost. Nevertheless, using the full morphological description as preterminal labels yields a set of over a thousand preterminals, resulting in data sparsity and performance problems (Szántó et al., 2014).

With this in mind, in order to manually construct the Georgian syntactically annotated trees, we had to perform the following text processing procedures:

- tokenization
- morphological analysis
- POS tagging and syntactic annotation.

Tokenization and morphological analysis were done by the Finite-State Transducer for Georgian (Kapanadze, 2010).

Before starting syntactic annotation procedures for the Georgian text, we made an overview of experience in building parallel TreeBanks for languages with different structures (Megyesi and Dahlqvist, 2007, Grimes et al., 2011, Rios et al., 2009, Samuelsson and Volk, 2005).

In a Quechua-Spanish parallel TreeBank, due to strong agglutinative features of the Quechua language, the monolingual Quechua TreeBank was annotated on morphemes rather than words. This allowed to link morpho-syntactic information precisely to its source. Besides, according to the authors, building phrase structure trees over Quechua sentences does not capture the characteristics of the language. Therefore, for its description a Role and Reference Grammar has been opted that allowed by using nodes, edges and secondary edges to represent the most important aspects of Role and Reference syntax for Quechua sentences (Rios et al., 2009).

Georgian is also an agglutinative language that uses for a wordform building both, suffixing and prefixing, though, there is no need to annotate the Georgian TreeBank on morphemes. Therefore, morphological analysis is one of the basic issues for agglutinating languages, since it provides useful clues for resolving syntactic ambiguity, and the parsing model should have a way of utilizing these hints. A lexicon-based parse engine has been oriented to capture the specifics of the Georgian morphology manifesting rich syntactic clues (among others the syntactic valency) encapsulated in the finite verb forms.

Syntactic annotation procedures were carried out manually using the *Synpathy* tool (Synpathy: Syntax Editor, 2006). It drew on an adapted version of the TIGER-XML encoding scheme (Brants and Hansen, 2002) that employs a SyntaxViewer developed for the TIGER-Research project (Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart). The POS tags pursue STTS (Stuttgart-Tübinger Tagset) guidelines with the necessary changes relevant to the Georgian grammar formal description and has been tested in the CLARIN-D project for the GRUG TreeBank repository building (Kapanadze, 2017).

In Georgian, as in many other languages, word order is much more flexible (for example, the subject may appear

either before or after a verb, etc.) as a result of its rich and productive morphology. In languages with flexible word

order the meaning of the sentence is realized using other structural elements, like word inflections or markers, which reflect morphological information.

A preferred basic word order without a Theme/Rheme bias for Georgian is SOV. The most notable feature in a syntactic description model for the Georgian clause is a phenomenon classified as a mutual government and agreement relations between verb-predicate and its actants (resp. NP), which number may reach up to three in a single clause. This anticipates control of the noun declension case markers by verbs, whereas, in its turn, the verb formants for person and number are governed by nouns presented in the clause. As a consequence of the verb-predicate capability to reflect morphologically the agreement relations with actants - Subject (SB), Direct Object (DO), Indirect Object (IO) as pronouns - can be omitted in the word order without a consequence for the clause meaning comprehension. The “reduced” clauses are equally “eligible” as their source ones in terms of the clause meaning representation.

In Figure 1 a syntactic tree of a Georgian complex sentence (\*) as an outcome of the CFG parse procedure is depicted.

(\*) თუ ღმერთი გწამთ, არ მითხრათ ახლა, რომ შავი თეთრია.

(Lit. “If you believe in god (=For god’s sake), do not tell me now that black is white”).

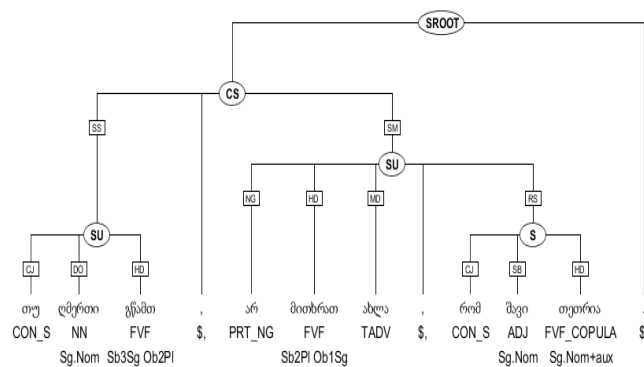


Figure 1: An adapted TIGER-XML scheme for a Georgian sentence.

The sentence in Figure 1 visualizes a hybrid approach to the syntactic annotation issue as the tree-like graphs and integrates annotation according to the constituency representations and functional relations. In a tree structure the node labels are phrasal categories.

The Complex Sentence (CS) in Figure 1 has two clauses as constituents:

– A clause without a Subject (SU) on the left side. As a daughter node it is linked by SS (Subordinate Sentence) Relation marked on an edge label.

– A simple sentence (S) on the right side. It is linked as a second daughter node by SM (Main Sentence) Relation that is marked on the consequent edge label.

– In its turn, S on the right side enjoys a simple sentence (S) as a daughter node linked by the RC (Relational Construction) as a constituent.

The edge labels for terminal nodes display the following syntactic functions: Conjunction (CJ), Subject (SB), Head (HD), Modifier (MD), Direct Object (DO).

The tokens in terminal nodes are annotated with POS tags such as Subordinating Conjunction (S\_CNJ), Normal Noun (NN), Adjective (ADJ), Finite Verb (FVF), Negation Particle (N\_PRT) and Adverb (ADV). They are saturated with morphological features of number (Sg) and case (Nom) for Normal Nouns. The Finite Verbs are annotated with features for person and number of subject and object (Sb3Sg Ob3Pl) (Sb2pl Ob1Sg), though, the Subject in the left constituent (SU) and both - the Subject and the Object in right (SU) one - are omitted in the word order. Thus, the monolingual treebanks converted into TIGER-XML format are a powerful database-oriented representation for graph structures in which each leaf (= token) and each node (= linguistic constituent) has a unique identifier.

Further, drawing on the sketched principle, we had manually built around 300 high quality morphologically and syntactically annotated trees. This repository had been used as training data for extracting a vanilla Context-Free Grammar and a lexicon for the Georgian language. The number of rules extracted from the syntactically annotated sentences has exceeded 1000. However, the rules are extracted with respect just to POS without morphological features as it is adopted in general while developing CFG parsers.

#### 4. Conclusion and Future Plans

In the future we intend to implement a mixed syntactic parsing method for the Georgian text that will utilize a traditional CFG approach combined with a morphological feature commonly known as syntactic valency of a verb-predicate. Morphological information of valency value will be extracted from verb which normally is the head (HD) of a clause.

E.g. In Figure 1:

- FVF - გწამთ - in SU (the left-hand constituent)
- FVF - მითხრათ - in SU (the right-hand constituent)
- FVF\_COPULA - თეთრია - in S (the secondary node in the right-hand constituent)

The verb syntactic valency feature will be used for determining syntactic structure of a clause in syntactic trees. To this end in the meantime we are developing a

new version of a Finite-State Morphoparser that will provide the Georgian verb parse output (alongside the POS tag) with the valency data.

For building a full-scale Georgian syntactic parser, we also intend to make use of the developed vanilla CFG that was extracted from the monolingual Georgian treebank. It will be utilized for finding optimal morphological features/preterminals for implementation in a Probabilistic Context-Free Grammar (PCFG) parser. The reason for such decision is the advantage of a deterministic part-of-speech tagger that can produce a morphologically annotated Georgian corpus achieving almost 100% accuracy after manual disambiguation (Kapanadze, 2010) and providing the tokens with POS saturated also with morphological information using features such as case, number for nouns and adjectives, and person, tense, syntactic valency for verbs.

In parallel we will extend a monolingual lexicon extracted from the Georgian TreeBank by adding all possible case forms for Nouns in singular and plural for each lexicon entry (14 forms for modern and 5 forms for old Georgian plural). For the mentioned procedures a Georgian FST morphological generator is intended to utilize.

According to the reports, the most successful supervised constituent parsers at the first stage apply a PCFG to extract possible parses. The  $n$ -best list parsers keep just the 50-100 best parses according to the PCFG. These feature templates exploit atomic morphological features and achieve improvements over the standard feature set. These methods use a large feature set — usually a few million features — and are engineered for English (Szántó and Farkas, 2014).

The innovative aspect of the proposed approach is a unique procedure for finding the optimal set of preterminals by merging morphological feature values.

The main advantage of this methodology over previous undertakings is the performance speed — it operates inside a PCFG instead of using a parser as a black box with retraining for every evaluation of a feature combination — and it can investigate particular morphological feature values instead of removing a feature with all of its values (Szántó and Farkas, 2014).

## 5. Bibliographical References

Fraser, A., Schmid, H., Farkas, R., Wang, R. and Schütze, H. (2013). Knowledge sources for constituent parsing of German, a morphologically rich and less-configurational language. In *Computational Linguistics*, Volume 39, Issue 1. MIT Press Cambridge, Ma, USA.

Kapanadze, O. (2010). Describing Georgian Morphology with a Finite-State System. In A. Yli-Jura et al. (Eds.): *Finite-State Methods and Natural Language Processing*

2009, *Lecture Notes in Artificial Intelligence*, Volume 6062, pp.114-122, Springer-Verlag, Berlin Heidelberg.

Szántó, Z. and Farkas, R. (2014). Special Techniques for Constituent Parsing of Morphologically Rich Languages. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden.

Synphaty: Syntax Editor. (2006). Manual – Nijmegen: Max Planck Institute for Psycholinguistics. The Netherlands.

Brants, S. and Hansen, S. (2000). Developments in the TIGER Annotation Scheme and their Realization in the Corpus. In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, pp. 1643–1649.

Kapanadze, O. (2017). *Multilingual GRUG Parallel TreeBank — Ideas and Methods*. LAMBERT Academic Publisher. 52 p. ISBN-13: 978-3-330-34810-3. EAN: 9783330348103.

Megyesi, B. and Dahlqvist, B. (2007). A Turkish-Swedish Parallel Corpus and Tools for its Creation. In *Proceedings of Nordiska Datalingvistdagarna (NoDaL- iDa 2007)*.

Grimes, S, Li, X., Bies, A., Kulick, S., Ma, X. And Strassel, S. (2011). Creating Arabic-English Parallel Word-Aligned Treebank Corpora at LDC. In *Proceedings of the Second Workshop on Annotation and Exploitation of Parallel Corpora. The 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011)*. Hissar, Bulgaria.

Rios, A., Göhring, A. and Volk, M. (2009). Quechua-Spanish Parallel Treebank. In *7th Conference on Treebanks and Linguistic Theories*, Groningen. The Netherlands.

Megyesi, B., Hein Sågval, A., Csató E.A. and Johanson, E. (2006). Building a Swedish-Turkish Parallel Corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa. Italy.

A multilingual German – Russian – Ukrainian - Georgian Parallel Treebank.  
<http://fedora.clarin-d.uni-saarland.de/grug/>

Samuelsson, Y. and Volk, M. (2005). Presentation and Representation of Parallel Treebanks. In *Proceedings of the Treebank-Workshop at Nodalida*. Joensuu, Finland.