Computational Analysis of Morphosyntactic Categories in Georgian

Sophiko Daraselia

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

University of Leeds
School of Languages, Cultures and Societies
April 2019

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

I would like to thank the International Educational Center (IEC) for giving me the opportunity of a lifetime by fully funding my PhD program at the University of Leeds. Without this financial support, this would not have been possible.

Of course, none of these would have been possible without my supervisory team, which has included Dr Serge Sharoff, Dr Diane Nelson and Dr Andrew Hardie. I would like to thank Dr Serge Sharoff, my first supervisor. I am very grateful to both him and my second supervisor, Dr Diane Nelson, for giving me their guidance throughout the years on my PhD project. Most importantly, they have always been supportive in my academic development beyond my PhD research. They have helped and supported me to lead the Kartvelian and Caucasian Studies research satellite at Language@Leeds. They have helped me in organizing the international annual workshop on Computational approaches to Morphologically Rich Languages (CAMRL2017 and CAMRL2018) at the University of Leeds.

I cannot thank enough and express how grateful I am to Dr Andrew Hardie, my external supervisor from Lancaster University. He has been so helpful, and I feel I have learnt a great deal from him. With Dr Hardie's supervision, I have managed to design a tagset for Georgian (KATAG), which is one of the main contributions in my PhD project.

I am very lucky to have my husband Thomas James Blackwell for supporting during my PhD journey. My loving thanks go also to each of the members of Thomas's family, who always made me feel very special.

Many thanks go to my family in Georgia, especially to my mother Nona Berishvili, who has always been there for me, showing her love and support every single day throughout the whole PhD project.

Alongside my PhD project, I was involved in a number of interesting projects with my Georgian friends and colleagues. I would like to thank Professor Tinatin Margalitadze, Professor Marine Beridze, Professor Nino Sharashenidze, Eto Churadze and Dr Rusudan Gersamia for their friendship, kindness and cooperation.

I also would like to thank the University of Leeds for accepting my PhD project and providing an incredibly professional and supportive research environment throughout my PhD study. I would like to thank my friends in Leeds with whom I shared the PhD experience and their love and support have given me strength during hard times, especially Faye Bennet, Polly Gallis, Aishah Mubaraki and Shifa Al Askari.

Abstract

This thesis describes the development of part-of-speech tagging resources for the Georgian language, consisting of i.) a new morphosyntactic language model for part-of-speech (POS) tagging purposes; ii.) tagging guidelines for tagging and post-editing; iii.) the KATAG tagset and iv.) the trained parameter files the probabilistic TreeTagger program needs to work on Georgian texts.

A new morphosyntactic model of Georgian for part-of-speech tagging purposes is described in the thesis. The thesis also describes a tagset (KATAG) defined in accordance with a new morphosyntactic model of the language and a set of design principles and tagging guidelines.

A stochastic methodology is used here to perform tagging in Georgian. Namely, the Treetagger - a probabilistic part-of-speech tagging program has been trained on Georgian texts. The justification for this choice is discussed.

I use two tokenisation approaches in part-of-speech tagging. An accuracy of 92.41% using an enclitic tokenisation approach and accuracy of 87.13% was achieved using a non-enclitic tokenisation approach, corroborating my hypothesis that treating enclitic elements separately from the host words results in better tagging performance.

To make the tagger program easily adaptable for a range of inputs (type, variety or genre of text), the performance of the probabilistic TreeTagger program was evaluated according to the obtained test set consisting of five different genres such as academic, informal, legal, fiction and news.

Table of Contents

List of Tables

List of Figures

Abbreviation

1O	1st person, Object	*POST*	Postposition
1S	1st person, Subject	*PRF*	Perfect
2O	2nd person, Object	*PRS*	Present
2S	2nd person, Subject	*PRV*	Preverb
3O	3rd person, Object	*PTCL*	Particle
3S	3rd person, Subject	*RES*	Resultative (tense)
ACC	Accusative case	*SBJV*	Subjunctive mood
ADV	Adverbial case	*SG*	Singular
AOR	Aorist	*SM*	Series marker
APPL	Applicative	*TAM*	Tense, Aspect, Mood
BEN	Benefactive	*THS*	Thematic suffix
COND	Conditional	*VOC*	Vocative case
COP	Copula		
DAT	Dative case		
DIM	Diminutive		
ERG	Ergative case		
FUT	Future		
GEN	Genitive case		
IMPERF	Imperfect		
INS	Instrumental case		
IPFV	Imperfective aspect		
MTE	Multext east		
NOM	Nominative case		
PL	Plural		

Chapter 1

Introduction

1.1 Motivation

This PhD thesis describes part-of-speech tagging in Georgian. Part-of-speech tagging is an established procedure in corpus linguistics. There are a wide range of applications of part-of-speech tagging software and tagged texts and corpora. These include information retrieval, machine translation, sentiment analysis, and the development of corpus-based grammars and dictionaries. It also supports additional layers of (automated) analysis, such as semantic annotation and discourse tagging (Leech, 1997; Leech and Smith, 1999; Hardie, 2004).

Thus, part-of-speech tagging is central to the field of corpus linguistics. Therefore, it is always a worthwhile task to develop part-of-speech tagging resources and extend part-of-speech tagging practices especially for under- or less-resourced languages such as Georgian.

It is worthwhile to mention that there have been a number of attempts of corpus annotation in Georgian. There are a handful of tagged Georgian corpora available. For example, the *Georgian analyser* is used to tag the Georgian Dialect Corpus (Lortkipanidze et al., 2013). The *Morphological Generator and Analyzer* is used to tag a corpus of Georgian literary language (Lobzhanidze, 2013), and a parser for Georgian using the Constraint Grammar (CG) framework (Meurer, 2015), which is used to tag the Georgian National Corpus (including Old, Middle and Modern Georgian) and the Georgian Reference corpus.

However, there are no tagsets or tagging guidelines available for these tagged corpora. Also, no tagger programs are available (with the exception of the Georgian parser). Furthermore, there is no information about the performance and/or accuracy of these tagging systems. Considering the state-of-the-art of corpus annotation in Georgian, developing part-of-speech tagging resources for the language and achieving a functional automated tagging is an undoubtedly novel task.

There are additional reasons why devising a part-of-speech tagging resources for Georgian can be even more interesting. First, Georgian is a member of the Kartvelian language family, for which no part-of-speech tagging has been done. As such, it may be hoped that the POS-tagging experience in Georgian may be of benefit to extend and develop part-of-speech tagging resources for other Kartvelian languages such as Megrelian, Laz and Svan. Secondly, Georgian is a morphologically complex language, meaning that it presents a number of interesting and possibly unique problems. For example, how to treat suffixaufnahme (double casing) case? How to tag argument agreement in verbs? How to treat numerous enclitic particles and postpositions? This gives an opportunity to solve such problems and makes the part-of-speech tagging process an interesting task.

1.2 Aims and objectives

The main aim of this thesis is to develop tagging resources and achieve functional automated part-of-speech tagging in Georgian. The other important aims of the thesis are as follows:

1. to devise a new morphosyntactic model of Georgian for POS-tagging purposes
2. to design a tagset for Georgian
3. to develop a set of tagging guidelines
4. to produce a set of parameter files for functional automated part-of-speech tagging in Georgian using the probabilistic TreeTagger (Schmid, 1994) program.

Additional aims of the thesis are as follows: 1) dealing with the complex Georgian morphology in POS tagging; 2) the part-of-speech tagging experience for Georgian may prove of benefit to later attempts to do the same for other Kartvelian languages.

Thus, the thesis describes the development of part-of-speech tagging resources for Georgian including a process of manual annotation of the training data, which is an essential prerequisite to achieve automated tagging.

1.3 Research questions

Apart from developing the part-of-speech tagging resources stated above and achieving functional automated part-of-speech tagging in Georgian, I will address the five research questions in the thesis, as follows:

1. **Is it possible to design a practically manageable hierarchical decomposable tagset for an agglutinative language, such as Georgian?**

By answering the first question, I will evaluate the practicality and manageability of the employed annotation schema. Georgian is an agglutinative language with complex morphology, meaning that it is hard to describe using the hierarchical-decomposable approach, for instance used by Hardie (2004), Khoja et al. (2001). This is because agglutinative languages have no finite paradigms and it is difficult to enumerate all conceivable combinations. However, the most problematic aspect of Georgian is the way in which it does not behave like an agglutinative language. For example, the verbal agreement paradigms are fusional (synthetic) with a high degree of syncretism.

In order to address this question, I will define possible hierarchies by going through category by category (see Chapter 4). Then I will put the proposed hierarchical decomposable tagset into practice by means of manual tagging (see Chapter 5) and finally I will evaluate the performance of the tagger based on the proposed hierarchical-decomposable KATAG tagset.

2. **Is a stochastic method an appropriate one in part-of-speech tagging of morphologically rich and complex languages such as Georgian?**

Selection of part-of-speech tagging methodology depends on many factors, such as the nature of the tagset, language typology etc. For example, Tapanainen and Voutilainen (1994) suggest that Markov model taggers operate better with small tagsets, whereas rule-based approaches perform better with large tagsets. It should be noted that this claim has been challenged and proved to be untrue. All taggers will perform better with fewer tags, as there are no fine-grained sub-categories in such tagsets (Eklund, 1993). Furthermore, Smith (1997, p.140) describes the comparison of the CLAWS system's performance with two English tagsets: C5, which was intended to be simple

(61 tags) and the larger C7 tagset (146 tags). Smith reports that larger tagset (C7) improves performance of the tagger (using Markov model).

The other important factors in selecting a tagging methodology include the typological features of a language. Morphologically rich languages have potentially freer word order and greater contextual ambiguity (Sánchez-León and Nieto-Serrano, 1997). These factors may suggest that a probabilistic model is unsuitable for Georgian. However, Hardie (2004, p. 296) points out that the free word order problems apply not only to Markov model taggers, but to rule-based approaches as well. Therefore, the probabilistic approach cannot be ruled out based on these factors.

Thus, I will evaluate the performance of the probabilistic TreeTagger program on Georgian texts. I will compare how different parameters, such as the size of lexicon, or context and affix lengths have effects on the tagger's performance.

3. **What are the challenges of the probabilistic TreeTagger program (with Markov model) when it is applied to Georgian?**

By answering this question, I will evaluate the overall performance of the employed probabilistic tagger for Georgian. I will identify the main challenges of the tagger program with regard to Georgian morphosyntax and provide solutions and suggestions for problematic areas (see chapter 6).

4. **What is the best approach in tokenisation when dealing with enclitics in Georgian?**

One of the preliminary tasks in part-of-speech tagging is tokenisation - dividing a text into tokens. It might seem that tokenisation is not a difficult task in Georgian, as there are clear word breaks by means of spaces. However, it is worthwhile to discuss the

clitic/affix distinction as it applies in POS tagging. In part-of-speech tagging an affix does not receive its own tag but may affect the grammatical features marked on the word; whereas a clitic receives its own tag. As Georgian is an agglutinative language it has numerous agglutinative postpositions and particles. There are two ways to treat such "enclitic" elements: 1) to tokenise into a unit of its own, or 2) to treat as a part of the word they are attached to.

It should be noted that there is no right or wrong choice regarding the "enclitics" in POS-tagging. Both enclitic and non-enclitic approaches are equally valid and have their advantages and disadvantages depending on the research question, end users etc. The main motivation for this question is to find out which approach is the best one for probabilistic part-of-speech tagging in Georgian. In order to do so, I will evaluate and compare the results of both approaches (see chapter 6).

5. **Which genres are most difficult in part-of-speech tagging in Georgian?**

The performance of the probabilistic TreeTagger program is evaluated on the obtained test set (see chapter 6) consisting of five different genres: academic, informal, fiction, news and legal. The main reason for this is to find out if the application of the tagger is limited because of the used resources (e.g. training set, lexicon) that have been trained for a particular variety or genre of text. In order to make the tagger program easily adaptable for a range of input (type, variety or genre of text), I will identify the genres, where the TreeTagger has a low performance level and provide possible solutions to improve the performance in these genres.

1.4 Thesis outline

There are several conventions used in this thesis. Georgian examples within the text are presented in bold type Georgian alphabet together with the Roman transliteration in brackets. The numbered Georgian examples in the thesis are glossed using the Leipzig[1] glossing rules. Italics are used for Georgian and English linguistic terms.

The thesis is organized as follows: Chapter 1 gives an overview of the topic and the motivations for the present research and outlines aims and objectives. Chapter 2 provides more detailed introductory discussions of Georgian, a language of which I am a native speaker. In it I discuss Georgian morphosyntax and claim that Georgian presents its unique and particular challenges for part-of-speech tagging. This chapter also discusses existing tagged corpora for Georgian.

Chapter 3 describes the necessary preliminaries of the design principles of the KATAG tagset. It also provides a brief overview of previous work in the field of tagset creation (for English). I will argue that the tagging standards, such as EAGLES recommendations for the morphosyntactic annotation of corpora[2] are not extensible and appropriate for a language like Georgian, which is a non-Indo-European language with complex morphology.

In chapter 4, I propose a new morphosyntactic model for Georgian for the purposes of part-of-speech tagging and accomplish the first main aim of the thesis by defining the KATAG tagset, by means of going through the proposed guidelines category by category.

[1] Revised version of February 2008 from https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf

[2] http://home.uni-leipzig.de/burr/Verb/htm/LinkedDocuments/annotate.pdf

Chapter 5 describes the process of manual tagging using the KATAG tagset. This allows me to assess whether or not the tagset is adequate to describe all the categories of Georgian. As a result of this, certain changes to the tagset are outlined and justified. Chapter 5 also describes the field of part-of-speech tagging methodology. I look at a number of different tagging methodologies, including rule-based method and probabilistic tagging using Markov models. I do this in order to be able to justify my choice of tagging approaches. This choice is made in the light of a number of factors, which are also discussed in this chapter.

Chapter 6 evaluates the performance of the trained parameter files of the probabilistic (TreeTagger) tagger program using the KATAG tagset with two different tokenisation approaches.

Chapter 7 is my conclusion and looks back across the preceding six chapters, considering the results of the study and possible future research.

Chapter 2

Background issues to the tagging of Georgian

In this chapter, I will discuss some background issues before I move on to describing the process of designing the part-of-speech tagging system. Firstly, section 2.1 describes the structure of Georgian. This was felt necessary because the language is not widely studied internationally and there are very few grammar books[3] on the language.

Secondly, as this PhD is about part-of-speech tagging in Georgian, I will briefly describe existing tagged Georgian corpora (sections from 2.3.1 to 2.3.4).

2.1 The Georgian Language

Georgian (ქართული - [**kartuli]**) belongs to the Kartvelian language family, which consists of four Kartvelian languages: Georgian, Laz, Megrelian and Svan.

Georgian is spoken by about 4 million people[4], mainly in Georgia as an official language. It is also spoken in Turkey, Iran, Azerbaijan and Russia. The history of the Georgian language has traditionally been divided into three main periods (Shanidze, 1976): Old Georgian (5th-11th c), Middle Georgian (11th-17/18th c) and Modern Georgian (from 18th c).

Modern Georgian is presented as the standard (literary) Georgian language and a wide range and variety of about 17 Georgian dialects, such as Imeretian (in Northwest

---

3 It should be noted that Georgian has a very rich literary tradition. There are some remarkable works on Georgian grammar (e.g. Shanidze, 1953) written in Georgian.

4 The information about the Georgian population is taken from the National Statistics Office of Georgia (GEOSTAT) as of 1 January, 2019. See more at: https://www.geostat.ge/ka/modules/categories/316/mosakhleoba-da-demografia.

Georgia), Gurian (in Southwest Georgia), Mtiuletian (in Northeast Georgia), Ingiloan (in Azerbaijan), Fereydanian (in Iran) etc.

Georgia has an ancient and rich literary tradition. The oldest literary text in Georgian (*The Passion of Saint Shushanik* by Iakob Tsurtaveli) dates back to the 5th century AD. The Georgian language has three unique alphabets - *Asomtavruli* (5th c.), *Nuskhuri* (9th c.), and *Mkhedruli* (from 10th c.) that are listed on the UNESCO's Representative List of Intangible Cultural Heritage[5].

Mkhedruli is the modern Georgian script. Thus, my PhD research utilizes this script: the Georgian web-corpus, tagset, manually tagged lexicon, training set and the test set are in Mkhedruli script.

The Mkhedruli alphabet originally consisted of 38 letters. Contemporary Georgian has 33 letters, as five letters became obsolete. The number of Georgian letters used in other Kartvelian languages varies. For instance, Megrelian uses 36 letters. Georgian has a high grapheme-to-phoneme and phoneme-to-grapheme correspondence. The Mkhedruli alphabet does not make a distinction between upper and lower cases. However, some Georgian fonts include "capitals", which are just larger versions of the letters. In June 2018, the obsolete *Asomtavruli* letters were added in Unicode version 11.0 to represent the capital letters in Georgian. They are capital letters with similar letterforms to *Mkhedruli*, but with descenders shifted above the baseline.

[5] https://ich.unesco.org/en/RL/living-culture-of-three-writing-systems-of-the-georgian-alphabet-01205

2.1.1 A brief overview of the structure of Georgian

In this section, I will sketch Georgian morphosyntax. It should be noted that here I will not describe commonly known and generally accepted linguistic facts. I will focus on the morphosyntactic features that are to some degree unique and particular to Georgian. My main sources where not otherwise specified, are Shanidze (1980), Gogolashvili (2011) and Melikishvili (2001, 2008, 2014).

2.1.1.1 Phonology

The following brief account of Georgian phonology is drawn from Shanidze's grammar (1980, pp. 7-23). Georgian has 5 vowels and 28 consonants. I have provided IPA notations alongside the Georgian symbols.

	Front	**Back**
Close	ი i	უ u
Mid	ე ɛ	ო ɔ
Open	ა ɑ	

Table 2. 1: Vowel phonemes (after Shanidze, 1980, p. 10).

		Labial	**Dental - Alveolar**	**Post-alveolar**	**Velar**	**Uvular**	**Glottal**
Nasal		მ m	ნ n				
Stop	**aspirated**	პ p^h	თ t^h		ქ k^h		
	voiced	ბ b	დ d		გ ɡ		
	ejective	პ’ p’	ტ’ t’		კ’ k’	ღ ɣ’	
Affricate	**aspirated**		ც $\widehat{ts}^h$	ჩ $\widehat{tʃ}^h$			
	voiced		ძ $\widehat{dz}$	ჯ $\widehat{dʒ}$			
	ejective		წ $\widehat{ts}$’	ჭ $\widehat{tʃ}$’			
Fricative	**voiceless**		ს s	შ ʃ	ხ x		ჰ h
	voiced	ვ v	ზ z	ჟ ʒ	ღ ɣ		
Tap/Flap			რ r				
Lateral			ლ l				

Table 2. 2: Consonant phonemes (after Shanidze, 1980, p. 13).

It should be noted that these vowels and consonants given Table 2.1 and Table 2.2 represent a standard literary Georgian language as having 33 letters and sounds. However, this is not true for Georgian dialects and their varieties. For example, unlike standard literary Georgian, Gurian dialect has additional two following approximants: [ჲ] [ĭ] and [ჳ][ŭ] (Gamkrelidze et al., 2006, p.14).

Standard literary Georgian has a wide range of ejective consonants in five places of articulation (labial, dental-alveolar, post-alveolar, velar and uvular). Some consonants show a strong affinity with certain other consonants. Shanidze (1980, p.23) describes these consonants as "harmonic groups /pairs". There are three so-called "harmonic groups" as follows:

1) [χ'] uvular ejective usually follows either of these ejectives: [p'], [t'], [t͡s'] and [t͡ʃ'] as in **ტყდომა [t'χ'doma]** "breaking, cracking";

2) [x] velar voiceless fricative follows these aspirated consonants: [pʰ], [tʰ], [t͡sʰ] and [t͡ʃʰ], as in **თხილი [tʰxili]** "hazel nut";

3) [ɣ] velar voiced fricative follows these voiced stops and affricates as follows: [b], [d], [d͡z] and [d͡ʒ], as in **დღე [dɣe]** "day".

One of the main characteristics of Georgian vowels is that there is no distinction in phonemic vowel length. However, it may exhibit sequences of identical vowel phonemes (vowel hiatus) that yield phonetically long vowels, such as **გააანალიზებ** [**gaaanalizeb**] "You will analyse it". Georgian does not use stress or pitch to give meaning to words.

2.1.1.2 Development of Georgian grammars

Prior to providing a short overview of the structure of the language, it is important to touch upon some background issues about the development of the Georgian grammars. There are relatively few grammars for the Georgian language.

It is worthwhile to mention that in the XVII-XVIII centuries Roman Catholic missionaries were well presented in Georgia. They founded schools in various regions of Georgia and started teaching Latin and Greek languages using Latin and Greek grammars and dictionaries (Tamarashvili, 1902, p.156).

Thus, early Georgian grammars from this period were influenced by Greek and Latin linguistic traditions. The first Georgian grammar was written by the Italian missionary Francisco-Maria Maggio in 1643. Earliest grammars of Georgian in the XVIII century were written by Zurab Shanshovani (1737) and Anton I Catholicos Patriarch of Georgia (1753, 1767)[6]. Gaioz Rektor's Georgian grammar written in 1789 (published in 1796)[7] was mainly based on Anton's grammar. These early works on Georgian grammar were influenced by Greek and Latin grammars (such as the "The Art of Grammar" [Tékhnē grammatikē] attributed to Dionysius Thrax[8] in 170-90 BC) and imported a Greek concept of grammar along with Greek terminology, which was inappropriate for the Georgian language. For example, they imposed a four-gender system (masculine, feminine, neuter and common) on Georgian declension despite the complete absence of grammatical gender in Georgian (or in any other Kartvelian

[6] For more information about Shanoshovani's and Anton's grammars See Potskhishvili (1981, pp. 22-52) and Babunashvili and Uturgaidze (1991).
[7] Gaioz Rektori's grammar were edited and published by Nikolaishvili (1970).
[8] For more detailed discussion see Karosanidze (2017).

languages). Moreover, Anton I Catholicos described Georgian as having prepositions despite the fact that Georgian is a postpositional language.

In the first half of the XIX century, there were a number of Georgian grammars written. These grammars include Eristavi (1802), Kartvelishvili (1809, 1815), Piralovi (1820), Bagrationi (1829), Dodaevi (1830), Brosset (1834, 1837) and Ioseliani (1840, 1851). It is worthwhile to mention that these grammars were influenced by Russian linguistic tradition (for more detailed discussion on this see Iluridze, 2006).

Akaki Shanidze was the first Georgian linguist to describe the language systematically in his *Fundamentals of the Georgian Language* published in 1953 (later reprinted in 1980 by his daughter Mzekala Shanidze). Since Shanidze's (1953) grammar, there have been few grammars written for Georgian. They are closely based on Shanidze's grammar (1953) with little novelty.

However, it should be mentioned that there is a great deal of work on each individual aspects of Georgian grammar (e.g. such as in morphology, syntax etc.) by Georgian and/or foreign linguists working on the Georgian language. Here I will not provide a detailed description of such works but will mention those authors that are relevant to this study. These includes Marr (1908, 1925), Chikobava (1928, 1968), Zorell (1930), Deeters (1930), Imnaishvili (1956, 1957), Topuria (1965), Gachechiladze (1979), Harris (1981), Sarjveladze (1984), Uturgaidze (1986), Hewitt (1995), Melikishvili (2001, 2008, 2014), Peikrishvili (2010), Gogolashvili (2011) and Sharashenidze (2014).

Thus, the works of the above mentioned authors have been used to some extent in this thesis. As for the systematic description of the Georgian language, Shanidze's (1953,

1980) work is most widely used and recognized as the traditional grammar until the present day.

One of the major problems in Shanidze's classification is that he uses semantic concepts and criteria to describe morphological categories. This causes a contradiction between form and meaning (Melikishvili, 2014) in Georgian morphosyntax. For example, the use of both semantic and formal criteria for grammatical functions - "subjects" and "objects" are in conflict. Furthermore, Harris (1981) argues that notions of "subject" and "direct object" are not appropriate for Georgian as there is no agreement between the three most obvious criteria for defining this concept: case, verb agreement, and semantic notion of subject.

Arnold Chikobava (1928, 1968) was one of the first Georgian linguists who identified and described the above problem of using semantic criteria to describe morphological categories. Later Melikishvili (2014) attempted to revise Shanidze's classification system by devising a new diatheses-based conjugation system of Georgian verbs. It is worthwhile to mention that grammar books at school and university levels are mainly based on Shanidze's traditional classification (Shanidze, 1953, 1980).

The main problem regarding this point that the language model (as in Shanidze, 1980) using semantic criterion to define morphological categories is not suitable for POS-tagging purposes. In designing the tagset, I have devised a new system of morphological categorisation, which focuses on purely morphological categories in Georgian. It should be highlighted that the proposed morphosyntactic model does not represent a new grammar of the language, but a simplified and practical approach for the purposes of part-of-speech tagging.

2.1.1.3 Morphology and Syntax

Georgian is a morphologically complex language, an agglutinative language with split ergativity. However, it is not purely agglutinative, as there are many examples of inflectional fusion as well. Georgian has no distinction of grammatical gender. While, there are some gender-specific words, such as დედა [**deda]** "mother" and მამა [**mama]** "father" etc. The kinship terms such as "niece" and "nephew" are gender neutral. However, parent of a "niece" or a "nephew" is gender specific. For example, ძმისშვილი [**dzmišvili]** can be translated as "niece" or "nephew" meaning "brother's child, either male or female", or დისშვილი [**dišvili]** "sister's child, either male or female".

There are no articles in Modern Georgian. However, Old Georgian used demonstrative pronouns as articles (Shanidze, 1980, pp. 618-620).

The agglutinative inflectional system is quite regular both for nominal declension and verb conjugation in Georgian. In the traditional case system (Shanidze, 1980), there are seven cases: Nominative, Ergative, Dative, Genitive, Instrumental, Adverbial and Vocative. Nominal modifiers may come either before or after the modified element. This affects the case and number agreement between the modifier and the head. For example, when the modifier appears before the noun it modifies, it does not agree in number, but in some cases, it agrees in case: it takes full case markers for nominative, ergative, and vocative. Optionally it takes "reduced" (as opposed to full marker) markers in genitive and instrumental - or takes no marker at all. However, this is only true when the modifier has a consonant-final root. When the modifier has a vowel final root and appears before the head, it does not agree in case and number. However, when

vowel-final modifiers appear after the head noun, they fully agree in case and number (see chapter 4).

Georgian has postpositions rather than prepositions. Few postpositions are independent words, for example, **შესახებ [šesaxeb]** "about"; most postpositions are always attached to a host, cliticised to the noun phrase. Each postposition governs a particular case and occurs after the case marker. Quite frequently, the case marker is deleted before the postposition due to the phonological rules in Georgian, such as the co-occurrence of two fricatives ([-s] and [-š]), as shown in the example below (from the KaWaC corpus.

(1)	saxl-i	saxl-(~~s~~)-ši
	house-NOM	house-(~~DAT~~)-POST
	"A house/home".	"In the house / at home".

In this example above, **სახლ** [**saxl]** "house / home" is a root form. In nominative, it adds the **[-i]** nominative case marker. Whereas in dative it adds the dative case marker **- [-s]**. When a postposition, such as the **[-ši]** "in" is added, the **[-s]** dative marker is deleted.

Another interesting phenomenon in the Georgian case system is *Suffixaufnahme* (suffix resumption), which is also known as case stacking. It is a genitive-based construction, where a genitive noun agrees with its head noun. It was first recognized in Old Georgian (Bopp, 1842) and is still actively used in Modern Georgian. This complex case system in Georgian is also characterized by two morphophonological phenomena: syncopation and apocopation. Syncope in phonology is the loss of one or more sounds in the middle of a word. Whereas an apocope is the deletion of one or more sounds from the end of a word. Syncopation usually occurs only in three cases,

Genitive, Instrumental, and Adverbial, where three vowels ([**a]**, [**e]** and [**o]**) are deleted. Apocope takes place in two cases, Genitive and Instrumental, and only two vowels are apocopated ([**a]** and [**e]**). Some words can be syncopated and apocopated at the same time.

(2) karkh**a**n**a** karkhn-is
factory.NOM factory-GEN
"A factory". "of a factory".

In this example, ქარხანა **[karkhana]** 'factory' undergoes both syncope and apocope at the same time. The middle vowel [**-a-**] and last vowel [**-a]** is deleted as a result of syncopation and apocopation in genitive and instrumental cases accordingly.

The morphology of the Georgian verb is very complex. Georgian traditional grammars describe the verb according to grammatical (argument agreement, number) and derivational (voice, aspect) categories. The Georgian verb can take up to three arguments, but only two arguments can be morphologically marked at the same time: 1) Subject (agent) and 2) either Direct Object (patient) or Indirect (oblique) Object.

(3) g-c'er-s
2O.SG-write.3S.SG.PRS
"S/he writes it to you".

It is a transitive verb, marked applicative. It has three arguments but agrees with only two of them: the interpretation of the third argument is recovered from the valency marking.

There are three types of case marking possible for the subject of the sentences according to a combination of morphological and case alignment criteria: Ergative, Nominative and Dative. However, this is conditioned by the Series (Tense, Aspect

and Mood) of the verb, as well as the voice category and transitivity. For example, the subject of the verb in the present tense has a nominative marking and in the past tense (aorist) an ergative marking.

In Georgian linguistics, the term ***screeve*** is used to express a system covering tense, aspect and mood (TAM). There are three TAM Series consisting of eleven screeves as follows:

Series	**Set**	**Screeve**
I	Present	Present
		Imperfect
		Present Subjunctive
	Future	Future
		Conditional
		Future Subjunctive
II	Past	Aorist
		Aorist Subjunctive (optative)
III	Perfect	Perfect – I Resultative (first evidential)
		Pluperfect – II Resultative (Second evidential)
		III subjunctive - (third evidential)

Table 2. 3: Series and screeves in Georgian.

Tense expresses time reference in Georgian, as in other languages. However, there is no single marker for each tense in Georgian; rather, various individual root forms mark the tense. Georgian verbs have so called "thematic suffixes" (from Greek *thema*), root forming suffixes, such as **[-ob]**, **[-av]**, **[-am]**, **[-ev]**, **[-en]**, **[-i]** and **[-op]**. Thematic suffixes are present and future stem formants. Thus, they appear in Series I (e.g. in present and future tenses) and also in Series III (e.g. in I resultative) since Series III verbs use stem formats from Series I (Shanidze, 1980, pp.387-388). Thematic suffixes are absent from Series II (e.g. in aorist).

(4) a) v-xat'-**av**
1S.SG-paint-THS.PRS
"I paint".

b) v-xat-e
1S.SG-paint-SM.AOR
"I painted".

(5) b) v-tamaš-**<u>ob</u>**
1S.SG-play-THS.PRS
"I play".

b) v-i-tamaš-e
1S.SG-BEN.APPL-play.SM.AOR
"I played".

Georgian verbs can encode a very complex information such as follows:

Preverb → verb root → thematic suffix → (APPL voice markers) → person / number agreement marker

(6) da-xat-av-s
PRV-paint-THS-3S.SG.FUT
"S/he will paint it"

Person agreement marker → verb root → thematic suffix→ (APPL voice markers) → number agreement markers

(7) v-xat-av-t
1S-paint-THS-PL.PRS
"We paint it".

Thus, a single Georgian verb may contain the following information: person and number features of subject, direct and indirect object, tense, aspect, voice, mood etc.

Another important characteristic of the Georgian verb is that some verb forms can have two or more readings. For example, the Present Tense root form can also express future tense, depending on the context.

(8) bržan-eb-s
order-THS-3S.SG.PRS/FUT
"S/he orders; s/he will order".

(9) asc'avlis
teach.3S.SG.PRS/FUT
“S/he teaches him/her it; s/he will teach him/her it”.

Furthermore, verbs in Georgian can also be described as having morphological syncretism in the agreement paradigm. For example, the verb form below can have two readings, as follows:

(10) gzrdit
raise.3S.SG.2O.PL.PRS
raise.1S.PL.2O.SG.PRS
“S/he raises you (PL)”; “We raise you (SG)”.

The category of aspect is derivational in modern Georgian. Prefixal morphemes (so called preverbs) that are cliticised to verbs and verbal nouns, mark perfective aspect.

(11) a) tex-av-s
break-THS-3S.SG.IMPERF
“S/he is breaking it”.
Imperfective aspect

b) ga-tex-av-s
PRV-break-THS-3S.SG.PERF
“S/he will break it”
Perfective aspect

There are about 22 preverbs (prefixal morphemes) in Modern Georgian. Together with marking aspect category, they also have other functions as follows:

1) Indicate location, direction and orientation of action and state in space. For instance, the [**še-**] preverb indicates the direction from outside to inside, for example, in the verb root **ვიდ [vid]**, **შევიდა** [**še**]**vida]** “S/he entered” and [**ga-**] preverb expresses the direction from inside to outside, e.g.: **გავიდა** [**ga**]**vida]** “S/he went out”;

2) Changes lexical meaning, for example in the verb root **გებ [geb]**, **გაგება [gageba]** "to understand", **მოგება [mogeba]** "to win", and **წაგება [cageba]** "to lose";

3) As mentioned above, they mark aspect and tense of the verb, for example, **აკეთებს [aketebs]** "S/he does, makes", Present tense, Imperfective aspect, **გააკეთებს [gaaketebs]** "S/he will do, make", Future Tense, Perfective aspect.

As demonstrated above, a single Georgian verb may encode a large number of morphosyntactic features. Thus, it is the most complex part-of-speech in Georgian, especially in terms of POS-tagging.

2.2 Previous work on corpus annotation in Georgian

In this section, I will discuss existing tagged corpora in Georgian. There are very few tagged corpora for Georgian, such as the KaWac corpus, Georgian Dialect corpus and the Georgian National corpus.

2.2.1 The KaWaC Corpus

The KaWaC[9] is a large web corpus of Georgian, which was created at the University of Leeds within my previous PhD project (Daraselia and Sharoff, 2014; Daraselia and Sharoff, 2015; Daraselia, 2015).

The KaWaC corpus was designed to be a large and diverse Georgian web-corpus representing a variety of internet genres on the web, such as press, news, fiction, personal blogs etc. The process started with identification of the more popular resources and crawling them from the internet using *wget*, with further processing by webpage cleaning and deduplication based on *BootCat* tools. The corpus texts were collected from 618,468 web pages from 697 websites. It contains over 180 million words.

The KaWaC corpus covers a wide range of text types, topics and regions. The text types are described using Functional Genre Dimensions, such as Argumentative, Instructional, Legalistic, etc. (Daraselia and Sharoff, 2014).

The KaWaC corpus was annotated using the MULTEXT-East Morphosyntactic Specifications Version 4[10] (Erjavec, 2012). The MULTEXT-East (MTE) language resources are a freely available large multilingual dataset for language engineering research and development. It focuses on harmonization of morphosyntactic

9 http://corpus.leeds.ac.uk
10 http://nl.ijs.si/ME/V4/msd/html/

specifications for sixteen languages, mainly from Central and Eastern Europe (Erjavec, 2012).

The MULTEXT-East Morphosyntactic Specifications define the main morphosyntactic categories and their attribute value pairs and describes morphosyntactic properties of words (called Morphosyntactic Descriptions - MSDs). For instance:

Verb, Type= indicative, Person = first, Number = singular, Tense = present

The specifications of the feature structure above correspond to a single MSD tag **Vi1sp**, which can be used in automatic morphological analysis and disambiguation (Santini et al., 2010).

The annotation scheme of the corpus uses a simplified approach based on the grammars of Shanidze (1980) and Gogolashvili (2011).

The tagset is designed according to MULTEXT-East Morphosyntactic Specifications. The MTE specifications of several corpora were directly taken from the MULTEXT-East resources. The new MSDs were created for specific Georgian cases.

The tagset contains 15 main categories: noun, verb, adjective, pronoun, adverb, adposition, conjunction, numeral, particle, interjection, masdar, participle, compound verb, abbreviation and residual. For each category the attributes and values appropriate for the category are given. These values are expressed as one-letter codes (Erjavec et al., 2003). There are in total 331 attribute-value pairs for Georgian appropriate to the main categories described above.

The probabilistic method was used to tag the KaWaC corpus. The performance of the probabilistic tagger program is below 70%, since it has been trained on a very small data (5,000 words) and without considering appropriate biases. Thus, the employed annotation schema has revealed a number of part-of-speech tagging errors, such as lexical gaps, disambiguation problems (Daraselia and Sharoff, 2014). I will critically engage with MULTEXT-EAST in next Chapter in section 3.3.

2.2.1.1 MULTEXT-East for another Georgian corpus

It should be noted that another Georgian corpus is tagged using the language model and the morphological lexicon developed according to the MULTEXT-East Morphosyntactic Specifications (Daraselia and Sharoff, 2014). The Georgian corpus of about 250 million words on the Aranea Corpora Portal (Benko, 2016) was tagged using the probabilistic TnT tagging software (Brants, 2000). The Arena portal consists of a Family of Comparable Gigaword Web Corpora[11] prepared by Benko (2018) within the framework of a joint Project of Department of Plurilingual and Intercultural Communication (Comenius University in Bratislava) and Ľ. Štúr Institute of Linguistics (Slovak Academy of Sciences). According to Benko (2018) the corpus coverage is low (75 % of all corpus tokens).

2.2.2 A parser for Georgian

Meurer (2007) describes a Georgian parser based on the Lexical Functional Grammar (LFG) framework (Kaplan and Bresnan, 1982). It uses the standard tool for morphological analysis with the XLE platform in the Xerox Finite State Tool (fst).

11 http://unesco.uniba.sk/guest/

Meurer uses the lexicon input to the Georgian morphological tranducer mainly from a digitized version of Georgian-German dictionary (Tschkeneli, 1964). The base form lexicon of the transducer comprises more than 74,000 nouns and adjectives and 3,800 verb roots (Meurer 2007). LFG analyses focus on two levels of syntactic structures. Constituent structure (c-structure) represents word order and phrasal groupings, and functional structure (f-structure) represents grammatical functions like subject and object. This annotation scheme is used to tag the Georgian National corpus[12] (GNC) including old, middle and modern Georgian texts and the Georgian reference corpus.

The list of grammatical features and codes ("tags") used in the CG parser for Georgian are available on the GNC website[13]. The "grammatical features" used in the CG parser are not POS-tags per se, as it accounts for syntactic and semantic information. For example, the <AuxTrans> is a grammatical feature, which is used in V (verbs), meaning that it is transitive auxiliary with non-human subject.

Eng. Code	Geo. Code	Variety	is POS?	Used with:	Explanation	
1	1	OG, NG		Pron	1st	1. პირი
2	2	OG, NG		Pron	2nd	2. პირი
3	3	OG, NG		Pron	3rd	3. პირი
<Advb>	<ვით>	OG, NG		Pp	Argument in Advb	არგუმენტი ვითარებითში
<AuxIntr>	<დამხმ-გარდაუ>	OG, NG		V	Intransitive auxiliary	გარდაუვალი დამხმარე ზმნა
<AuxTrans>	<დამხმ-გარდამ-ხულ>	OG, NG		V	Transitive auxiliary with non-Hum Subject	გარდამავალი დამხმარე ზმნა სულიერი სუბიექტით
<AuxTransHum>	<დამხმ-გარდამ>	OG, NG		V	Transitive auxiliary with Hum Subject	გარდამავალი დამხმარე ზმნა უსულო სუბიექტით
<DO:Dat>	<DO:მიც>	OG, NG		V	Dative direct object	პირდაპირი ობიექტი მიცემითში
<DO:Nom>	<DO:ხან>	OG, NG		V	Nominative direct object	პირდაპირი ობიექტი სახელობითში
<Dat/Gen>	<მიც/ნათ>	OG, NG		Pp	Argument in Dat or Gen	არგუმენტი მიცემითში ან ნათესაობითში
<Dat>	<მიც>	OG, NG		Pp	Argument in Dat	არგუმენტი მიცემითში
<Gen>	<ნათ>	OG, NG		Pp	Argument in Gen	არგუმენტი ნათესაობითში
<IO:Dat>	<IO:მიც>	OG, NG		V	Dative indirect object	ირიბი ობიექტი მიცემითში
<IO:Gen>	<IO:ნათ>	OG, NG		V	Genitive indirect object	ირიბი ობიექტი ნათესაობითში

Figure 2. 1: Grammatical features used in the Georgian parser

12 http://gnc.gov.ge
13 http://gnc.gov.ge/gnc/parse?session-id=247111275780348

The GNC corpus (including Old, middle and modern Georgian texts) and the Georgian reference corpus allows a number of filtered search options according to different metadata, such as author, genre, document, translator of the text etc. However, there is no search option according to specific part-of-speech tag or any of the given grammatical features. There are no guidelines available for the given set of grammatical features. Moreover, there is no information on the performance and accuracy of the parser.

The corpora developed within the GNC project was funded by the Volkswagen Foundation[14]. A number of significant corpus linguistic resources have been developed within this project including the Georgian corpora (as well as small sized corpora for Laz and Svan) and the Georgian parser with the CG framework. The parser is freely available on the GNC website, which indeed is a very useful tool to analyse Georgian texts.

2.2.3 Georgian morphological analyser

A group of Georgian linguists from the Georgian Technical University and Linguistics Institute are currently developing Georgian corpora and a morphological analyser (Lortkipanidze et al., 2013). The morphological analyser is mainly based on the Georgian monolingual dictionary (1950-1964). The analyser was first applied to the Georgian Dialect Corpus (GDC)[15]. The corpus includes the data (both spoken and written) of about 17 Georgian dialects.

[14] https://www.volkswagenstiftung.de/

[15] http://corpora.co/#/

The Georgian dialect corpus has grammatical markers, i.e. POS tags indicated. For example: **სახლში** [**saxlši**] "at home" receives the following tag: **N:Dat,Sg,Shi**. This can be interpreted as Noun, Dative, Singular, Postposition [ši] "in".

The same morphological analyser is also used to tag two specialized Georgian corpora: The Corpus of Otar Tchiladze [16], a Georgian writer and the Corpus of Akaki Shanidze[17], a Georgian linguist (Lortkipanidze et al., 2013). There are no tagset or the tagging guidelines available for this tagging scheme. The website of the Shanidze's specialized corpus notifies the users that the tagging process is not complete and thus, there might be some part-of-speech tagging errors occurring in the corpus.

2.2.4 Morphological Analyzer and Generator for Georgian

Lobazhanidze (2013) describes the Georgian Morphological Analyzer developed at Ilia State University in Tbilisi. The Morphological analyzer of Modern Georgian is developed using the Xerox Finite State Tools (Beesley and Karttunen, 2002). The system includes 13 "blocks" of the existing parts of speech of Modern Georgian including nouns, adjectives, numerals, pronouns, conjunctions, particles, adverbs, postpositions, verbs, verbal nouns and participles, as well as separate "blocks" for punctuation and abbreviations. The verbal paradigm is subdivided into additional 66 groups as described by D. Melikishvili (2001) and an additional group for irregular verbs (Lobzhanidze, 2013).

This morphological analyser is used to tag the Georgian corpora developed at the Ilya State university in Tbilisi. This includes the Georgian corpus (of literary texts) from

16 http://geocorpora.gtu.ge
17 http://textcorpora.tsu.ge

old, middle and modern Georgian [18] and bilingual corpora (Doborjginidze and Lobzhanidze, 2016), such as the bilingual corpus of *the Knight in the panther's skin*.

The query interface of the corpora enables simple search, as well as advanced search according to grammatical features. All morphological and semantic features associated with a given word appears as set of abbreviations of the linguistic terms. For example, კინო [**kino**] "movie/film" appears to be tagged as follows:

კინო+N+Com+Inanim+Sg+Nom

This reads as *Noun+common+inanimate+Singular+Nominative case*. The interesting thing is the appearance of the POS-tag combining not only morphological but semantic features (e.g. animacy). Like other annotation schemata in Georgian, there are no tagging guidelines available for this morphological analyser. Furthermore, there is no information about the performance and accuracy of this tagger.

2.3 Concluding Remarks

In this initial chapter, I have covered the preliminary issues around Georgian morphosyntax and described existing tagged Georgian corpora and tagging systems. As discussed above, all existing tagging systems in Georgian have three things in common: 1) there are no tagset documents and tagging guidelines; 2) there is no information about the performance and accuracy; 3) the application of tagger programs are limited to only these corpora and they are not available for other users.

[18] http://iliauni.edu.ge/ge/iliauni/institutebi-451/lingvistur-kvlevata-centri-467/qartuli-jesturi-enis-korpusi

Thus, developing part-of-speech tagging resources and achieving a functional automated part-of-speech tagging in Georgian is a novel task. The necessary first component to this part-of-speech tagging system is the tagset, which is the topic of the next chapter.

Chapter 3
Design principles of the Georgian Tagset

3.1 What is part-of-speech tagging?

Part-of-speech tagging is a type of corpus annotation. Leech (1997, p.2) defines corpus annotation as "the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written data". There are different types of corpus annotation, such as POS-tagging, semantic annotation, parsing etc.

The most common form of corpus annotation is part-of-speech tagging. Hardie (2004, p.40) defines part-of-speech tagging as "the process of assigning to each word in a running text a label which indicates the status of that word within some system of categorising the words of that language according to their morphological and/or syntactic properties". Tags are descriptive symbols and are called *part-of-speech tags*, since they indicate the parts of speech recognised by grammarians in the Latin/Greek tradition (Voutilainen, 1999, pp.3-4).

Corpora are now available for the majority of languages of the world and various forms of annotation are developed for languages other than English (Hardie, 2004, p.41). However, there are very few tagged corpora available for Georgian. Therefore, the use of corpora as a resource for linguistic study is not a common practice in Georgian.

3.2 Previous work on English part-of-speech tagsets

A tagset is a list of tags used for POS-tagging, representing a set of word categories (Garside et al., 1997). I will briefly describe the most important and influential works on the English tagsets.

The earliest work on the English tagsets started in the 1960s and early 1970s in the US. The most important tagsets of this earliest period are those of Klein and Simmons[19] (1963) and Greene and Rubin[20] (1971). The other tagging system at this time was developed at the University of Pennsylvania (Joshi and Hopely, 1997[21]). It is worthwhile to mention that these early works tended to stress the importance of part-of-speech tagging in parsing (Hardie 2004:47). For example, Klein and Simmons" (1963) tagging program was designed as a component of a parser. Likewise, the tagging software developed at the University of Pennsylvania was a finite-state parser.

Over the course of the 1980s and 1990s, a number of English tagsets have been devised at Lancaster University for use with the CLAWS (The Constituent Likelihood Automatic Word-tagging System) tagging software (Garside, 1987). There are a number of variations of the CLAWS tagset:

- The **CLAWS1** tagset, also known as the LOB tagset, was used in the tagging of the LOB corpus. It contained 132 tags. This tagset is similar to Brown corpus

[19] Klein and Simmons' CGC ("computational grammar coder") contains 30 tags. The authors reported (Klein and Simmons 1963:344) that they tagged several pages of children's encyclopedia with 90% accuracy.

[20] Green and Rubin's (1971) TAGGIT program was used for the linguistic annotation of the Brown University Corpus (Francis and Kučera 1967) containing 1.1 million words of American English representing 15 text genres. The Brown corpus tagset contains 77 tags. The later, refined Brown Corpus tagset contained 87 tags (Francis and Kučera 1982).

[21] It should be noted that they are reporting on a parser developed from the late 1950s.

tagset since these corpora were designed to be parallel in structure and the tags were also parallel.

- The **CLAWS2** tagset is finer-grained than the CLAWS1 tagset. It was the basis for the SUSANNE Wordtag Set (Sampson, 1995) and contains 166 tags.

The major subsequent development in the CLAWS tagset were the **C5** and **C7** tagsets. These tagsets were developed for the tagging of the BNC and the BNC sampler (Leech et al., 1994; Leech, 1997; Smith, 1997). The C7 tagset (146 tags) is more fine-grained than the C5 tagset and was used for the 2-million-word Sampler. The C5 tagset is a simplified version and it has 61 tags.

There are many other English tagsets. I will not discuss all of them in depth but will mention several influential tagsets in the field of corpus linguistics, such as TOSCA tagset, ICE tagset, Penn tagset, Lund tagset and EngCG tagset.

The **TOSCA**[22] tagset (Halteren and Oostdijk, 1993) makes many more distinctions of the syntactic function of the word than the CLAWS tagsets. It is made up of only 32-word class tags.

The **ICE**[23] tagset is an important development from the TOSCA tagset (Greenbaum and Yibin, 1996). It distinguishes 19-word classes but, like the TOSCA tagset, gives most words a feature list as well as a major word class tag.

[22] Tools for Syntactic Corpus Analysis.
[23] International Corpus of English.

The **Penn tagset** used in the Penn Treebank (Marcus et al., 1993) is based on the Brown Corpus tagset. The Penn tagset was modified in the direction of simplification. Thus, there are significantly fewer tags (36 tags).

The **Lund tagset** was designed for the annotation of the London-Lund Corpus of Spoken English (Svartvik, 1990). This tagset is significantly different from the Brown Corpus and CLAWS tagset tradition. It is more fine-grained and consists of over 200 tags. The Lund tagset was designed for spoken texts. Thus, it includes some discourse tags, such as swearing for example.

The tagset used by the EngCG tagger (Karlsson et al. 1995) is different from all tagsets reviewed above. It is described by (Heikkilä, 1995) as a "feature system" of "139 morphological or morphosyntactic features" rather than as a tagset per se.

Thus, this short account of the tagsets on the English language show that tagsets can vary in size and have different level of granularity. Some tagsets are large, fine-grained (e.g. C7), some tagsets are designed in the direction of simplification, having fewer tags (e.g. C5). The size and granularity of a tagset depend on many factors, such as size of the corpus, the kind of language data (e.g. spoken vs written) or the language typology may encourage differences in the annotations to be applied (Leech, 1997, p.7).

3.3 Design principles for a Georgian tagset

In this section, I will describe the annotation scheme for the proposed Georgian tagset. I will discuss the nature of the tagset, what information to include, the tagset appearance, tokenisation and disambiguation issues in Georgian.

According to Hardie "when POS tagging came to be applied to languages other than English, the need for the creation of standards became clear" (2004, p.55). The most recent standard on part-of-speech tagsets is the EAGLES[24] guidelines[25] (Leech and Wilson, 1999). The main disadvantage of the EAGLES guidelines is that it is a project of the European Union and it covers only English, Dutch, German, Danish, French, Spanish, Portuguese, Italian and Greek. Hence, it is not primarily designed for non-Indo-European languages such as Georgian in this case, which displays a complex agglutinative and inflectional nature different from the Indo-European languages. For example, there is no grammatical category of gender in Georgian, argument marking in verbs etc. It is worthwhile to mention that there are number of projects that use the EAGLES morphosyntactic framework for other languages than those mentioned above. For example, the MULTEXT project extends the tagset work to six languages of Central and Eastern Europe, including some non-Indo-European languages.

In general, there are advantages of compliance with standards (Hardie, 2004, p.68), such as the comparability of annotations in the same language or across languages. There are two main reasons why I will not comply with the EAGLES standards. First, Georgian is a member of the Kartvelian language family. The complex agglutinative and inflectional nature of the Georgian language makes it very distinct from the Indo-European languages (the main focus in the EAGLES guidelines). For example, there

[24] The Expert Advisory Group on Language Engineering Standards.
[25] EAGLES Recommendations for the Morphosyntactic Annotation of Corpora (1996).

is no grammatical category of gender in Georgian, which is one of the recommended subcategories in the EAGLES guidelines; the Georgian verb marks both subject and object agreement at the same time (a feature which is not covered in the EAGLES guidelines) etc. Secondly, there has been a previous attempt at compliance with the MULTEXT-east specifications (Daraselia and Sharoff, 2014) which is based on the EAGLES morphosyntactic framework. The previous attempt of adhering the standards did not prove to be appropriate in the construction of a Georgian tagset due to its distinct and complex morphosyntactic structure.

Thus, the main focus of annotation scheme of the proposed tagset is the Georgian language by retaining practicality and applicability of its complex morphosyntactic features.

3.3.1 Information to include

In this section, I will discuss what information a Georgian tagset should include and what information is excluded from the tagset.

Part-of-speech tags are categories as traditionally described in Latin/Greek grammars (Voutilainen, 1999). Under influence by the Latin/Greek tradition, 10 parts-of-speech have been proposed for Georgian. These will be considered as major word classes in the Georgian tagset. They are: noun, pronoun, adjective, numeral, verb, adverb, postposition, conjunction, particle and interjection. I will also include copula (see Section 3.3.12), punctuation (see Section 3.3.13) and "residual" (see Section 3.3.14) as three additional "categories". Thus, the proposed Georgian tagset will have 13 major classes.

Apart from the major word-classes, a Georgian tagset needs to include sub-categories and morphological features that are relevant to Georgian morphosyntax. Thus, a Georgian tagset will include three classes: major word class, sub-category and morphological features. Each of these will be given a single tag.

The Georgian tagset will not include derivational and etymological information, as this is marginal to morphosyntax (Hardie, 2004, p.73). It will not also consider syntactic information, such as syntactic roles, transitivity and applicative (benefactive, causative). Some tagsets, such as Brown corpus tagset[26] and C7[27] include semantic information in their morphosyntactic annotation. However, semantics is a separate field to morphosyntax, separate from part-of-speech tagging. Therefore, no semantic information will be included in the tagset.

3.3.2 Hierarchy and decomposability

Following Hardie (2004, p. 74), tagsets have become increasingly "hierarchical" and "decomposable" over the years and "these seem intuitively to be useful features for a tagset". Hardie (2004, p. 74) points out that it is easier for the user to memorise a small number of decomposable elements than a large number of tags. The other major advantage of decomposable tags is that it allows specific searches at "varying level of granularity" (Leech, 1997, p. 26).

A hierarchical tagset (aka feature hierarchy; Hardie, 2004) is a tree-like structure consisting of a number of categories. Cloeren (1999, pp. 39-40) suggests that major

26 For example, JJS is a semantically superlative adjective in the Brown corpus tagset.

27 C7 tagset includes some semantic features, the names of places as opposes to other proper noun.

word classes should be highest in the hierarchy, followed by subclassifications, and lastly morphological features. This is a common approach in hierarchical tagsets. For example, major word class (e.g. pronoun, verb) is the first category in the hierarchy, followed by a sub-category or sub-categories (e.g. personal / negative pronouns) and finally sub-sub-categories (morphological feature(s), such as number, case etc.).

A tag is considered "decomposable" if each tag consists of a string of concatenated elements and each of these elements represents a single feature in the definition of the category. It should be mentioned that language typology plays an important role when choosing a hierarchical-decomposable approach. Agglutinative languages are hard to describe using the hierarchical-decomposable approach, since they have no finite paradigms (Daraselia and Hardie, 2018). Thus, it is difficult to enumerate all conceivable combinations. The other approaches used for morphologically complex languages are for example, a feature-matrix (Sawalha and Atwell, 2013) for Arabic. Another possible solution is switching the task from part-of-speech tagging (per-word analysis) to morphological (per-morpheme) analysis (Hardie, 2017).

One of the main reasons of choosing a hierarchical-decomposable approach for Georgian is to ensure that it is reusable for as wide a range of end users as possible. For example, other annotators can expand it, design a more fine-grained tagset, or simplify the system of categories in the tagset considerably. The hierarchical-decomposable feature allows users to do so. Secondly, hierarchical-decomposable tagsets also allows the user to search for different sections of the paradigm via wildcard (*), for example, in the Georgian tagset:

- **V:*P:*** will look up for any plural verb
- **V:*:F** any future tense verb
- **V:1*** any verb, first-person of subject
- **NS*** any singular noun
- **N*E** any ergative noun

It is a widely accepted approach, easily understood and manageable, which is one of the main goals for annotators (Leech, 1997, pp. 6-7). Thus, the Georgian tagset will be fully decomposable and hierarchical.

3.3.3 Tokenisation

Dividing a text into tokens is not a difficult task in Georgian. The text makes clear word breaks by means of spaces. It is worthwhile to discuss the clitic/affix distinction as it applies in POS tagging. In part-of-speech tagging (as opposed to morphological annotation, for instance) an affix does not receive its own tag but may affect the grammatical features marked on the word of which it is part; whereas a clitic receives its own tag, for example the possessive "-'s" in English. In order to achieve this, it must be tokenised into a unit of its own, separate from the host word to which it is phonetically and/or orthographically attached, even if this involves splitting up what might be considered "one word".

There are two possible ways to treat encliticised words: 1) tag them as a one word or 2) split them from the host word and tag separately. It might seem more suitable for an agglutinative language to tag enclitic elements separately. However, I will consider both approaches in tokenisation and will introduce two terms accordingly: 1) *enclitic approach*, where enclitic elements are split from the host word and 2) *non-enclitic*

approach, where enclitic elements are treated as a single unit with the host word. In Chapter 5, I will demonstrate that splitting enclitics separately is the best approach in tokenisation for Georgian. I will evaluate (in chapter 6) the performance of both approaches and show that the enclitic approach improves the performance of the tagger.

There are two additional reasons to favour of the enclitic approach for Georgian. First, in the enclitic approach, the KATAG tagset has a finite size. Whereas in the non-enclitic approach the number of tags is infinite in the tagset, as it is impossible to conceive all possible combinations. Secondly, such an "infinite" tagset is difficult to manage and therefore, is very impractical for use in part-of-speech tagging.

3.3.4 Disambiguation

Van Halteren and Voutilainen (1999, p. 109) describe three main sub-tasks that an automatic tagging system involves:

- Tokenisation – segmentation of text into tokens
- Analysis: assignment of potential tags to tokens
- Disambiguation- figuring out the correct tags.

This section focuses on the third sub-task – disambiguation, which is the most problematic one in part-of-speech tagging. Cloeren (1999, p. 47) distinguishes several different senses of ambiguity, such as *grammatical homonymy*, where one wordform isolated from its context, belongs to more than one grammatical class. For example, the Georgian word დაწერა **[dac'era]** has two readings:

- Verbal noun ("to write")
- Verb ("S/he wrote it")

Another type of ambiguity is when a human annotator cannot decide on a single tag (Cloeren, 1999, p. 48). This is because the categories do not have clear boundaries, such as adjectives and adverbs in Georgian. The other thing is that linguists have different theoretical backgrounds and may have different opinions on the same data (Cloeren, 1999, p. 48).

Finally, Cloeren (1999, p. 48) describes *genuine* textual ambiguities, where text does not provide enough information for disambiguation. He discusses the exclamatory word "fire" as an example of this. It is unclear whether it is a verb or a noun.

The Georgian language has an additional level of ambiguity of *morphological syncretism*, when one wordform belongs to the same morphosyntactic category, but it is difficult to identify appropriate morphosyntactic features, such as tense and argument agreement in verbs. For example, the Georgian verb **გაძლევთ [gažlevt]** can have at least two readings:

- Verb, 3rd person of Subject singular and 2nd person of object Plural ("S/he gives you (PL) this)
- Verb, 1st person of Subject plural and 2nd person of object singular ("We give you this)

In the Georgian tagset, an ambiguous word will have two or more tags and it will be disambiguated at the POS tagging stage. For example, the word **და** [da] gets two tags as follows: **CC** – when it is a coordinating conjunction "and"; and **NSN** – when it is a singular noun, nominative, meaning "sister".

As for the words with no clear boundaries between the categories (nouns, adjectives, participles and verbal nouns, adverbs), there will be a lexicon for these categories. For

example, if a word appears in the adjective lexicon, but in the text, it functions as a noun, it still will be tagged as an adjective. This will be a consistent approach throughout the POS-tagging process.

For the fourth type of ambiguity of morphological syncretism, there will be appropriate tags provided in the lexicon for them and they will be manually disambiguated in the training corpus.

Thus, a Georgian tagset will include information on major word classes, subclassifications, and morphology. It will not include any derivational, etymological, syntactic or semantic information. The Georgian tagset will be fully decomposable and hierarchical. The tagset will tag by form rather than by function. Every word token will receive exactly one tag, with clitics tagged separately from the word they are attached to.

3.3.5 The tagset's appearance

The strings of the tags could be entirely arbitrary, but it is "preferable for the shape of the tag to reflect its meaning" (Hardie, 2004, p. 86). As Cloeren (1999) points out: "For reasons of readability there is a preference for mnemonic tags… Full-length names may be clearer individually but make the annotated text virtually unreadable."

For this reason, almost all tagsets have tags that are effectively abbreviations of the linguistic terms that describe their category. For instance, in CLAWS7 tagset, **AT** is a tag for articles; **NN** is the tag for common nouns, **VV0** is the tag for base form for lexical verbs. This is a practice that I shall follow. In order to retain some degree of comparability with the existing English tagsets and corpora, I will use the most

commonly encountered abbreviations (e.g. in CLAWS system's tagset) for the major word classes. For example, **N** is the tag for nouns, **V** for verbs, **J** for adjectives, **R** for adverbs and so on.

Some tagsets consists of upper-case letters only (e.g. CLAWS tagsets, Penn tagset), some tagsets consist of uppercase characters followed by lowercase characters (as in the MULTEXT tagset). The Georgian tagset will use uppercase letters only, as this is useful for the distinction between the tags and the actual words of the text (Erjavec, 2012).

To sum up, the forms of the tags in the Georgian tagset will follow these rules[28]:

- All tags will have mnemonic value as far as possible;
- Uppercase letters and the numeric symbols from 1 to 3 will be used, with the exception of: (colon) delimiter in verbs and enclitic elements (where enclitics are treated as a single word);
- The sequence of characters from left to right will represent a hierarchy of features ordered from the major word class to the morphosyntactic features.

[28] Cf. Hardie, 2004, pp. 88-89.

Chapter 4
Specification of the Tagset for Georgian

To create the categories of the tagset, it is necessary to have a model of the language to categorise. There are very few grammars for Georgian (see chapter 1, section 1.2). Shanidze's (1980) traditional grammar is most commonly used for Georgian. However, the language model as described by Shanidze proves inadequate for part-of-speech tagging purposes as it will be demonstrated in this chapter. Therefore, I will propose a new morphosyntactic categorisation to derive a language model for part-of-speech tagging.

Thus, the primary purpose of this chapter to devise a new morphosyntactic model and define a part-of-speech tagset for use in the tagging of Georgian, in compliance with the design principles described in chapter 3.

It should be noted that I will use the corpus evidence to develop a morphosyntactic scheme for part-of-speech tagging purposes. This will be a consistent approach throughout the tagset design process. All the examples used in my PhD thesis are from the KaWac corpus if not otherwise stated.

4.1 Noun (*arsebiti saxeli*)

The traditional categorisation of nouns (Shanidze, 1980) puts them into the following groups:

1) Animate and inanimate (*sulieri/usulo*)
2) Human and non-human (*vin/ra jgupis*)
3) Concrete and Abstract (*k'onk'ret'uli /abst'rak't'uli*)
4) Proper and common (*sak'utari/sazogado*)
5) Mass (*nivtierebata*)
6) Collective (*k'rebiti*)
7) Action (*mokmedebis*)

None of these categories is relevant for the morphosyntactic annotation scheme as they are not marked in the nominal morphology[29]. For example, the animate and inanimate binary has no place in a morphosyntactic tagset, as it is not marked in the morphology. This is true for all the other noun sub-categories listed above including concrete and abstract, mass and collective, human and non-human nouns.

I also will not categorise proper and common nouns separately in the tagset, as the distinction is not marked in Georgian orthography. There is no distinction between upper and lower cases and no articles are used in Georgian, one or other or both of those being the key formal characteristics of proper names in most of the languages of Europe. This lack of clear formal difference means there is both less need for, and lower feasibility of automatically accomplishing a morphosyntactic distinction between proper and common nouns. Thus, there will be no distinction between proper and common nouns in the tagset.

Unlike some other highly inflected languages including both Indo-European and Afro-asiatic languages, the category of Gender is not relevant for Georgian nominals.

The sub-categories for nouns that I will include in the tagset are 1) *Number* and 2) *Case*. These sub-categories of number and case are described below.

4.1.1 Number

In Georgian, nominals including nouns, adjectives, pronouns and numerals use three different suffixes to form their plural forms. Most plurals are formed by the pluralising suffix **-ებ** [**-eb]**, which is very productive in Modern Georgian. However, the usual formation for the plural in Old Georgian was the **-ნ** [**-n]** affix in Nominative and

29 However, they might be relevant to syntactic structure.

Vocative cases; and **-ო(ს)** [**-t(a)]** in Ergative, Dative and Genitive cases. The [**-t(a)]** is a fusional suffix indicating both case and number. Old pluralizing suffixes are still used in Modern Georgian, but they are less productive than the [**-eb]** suffix.

Thus, nouns and nominals in general, have three plural forming suffixes: [**-n], [-eb]** and bifunctional [**-t(a)]**, for example:

(1) k'ac-i
man-NOM.SG
"A man"

k'ac-**eb**-i
man-PL-NOM
"Men"

k'ac-**n**-i
man-PL-NOM
"Men".

As mentioned above, [**-t(a)]** is fusional suffix indicating both case and number and it is used in Ergative, Dative and Genitive Cases. For example, [**k'acta]** can be either of these three cases depending on the context. Thus, it will be difficult to automatically identify which cases [**-t(a)]** represents. In the tagset design, [**-t(a)]** will get three tags for Ergative, Dative and Genitive respectively and will be disambiguated at the POS-tagging stage.

For the purposes of POS-tagging, I will not make a distinction between old and regular plural forming suffixes. I will use just a single tag for both in the tagset as the main aim of a tagset is to abstract away from morphologically conditioned allomorphy and/or free variation for style.

4.1.2 Case System in Modern Georgian

The Georgian traditional case system is described in ways that are non-coherent in many grammars. Various ways of case descriptions exist, depending on vowel- or consonant-final roots, or syncopated or non-syncopated roots. Moreover, Shanidze (1980, pp. 73-77) discusses postpositional forms as case inflections[30]. There are many other problems in existing published descriptions of the Georgian case system, but they are beyond the scope of the present discussion.

In this section, I aim to simplify the model of the case system for the purposes of morphosyntactic annotation. In the traditional case system (Shanidze, 1980, pp. 44-108), it is considered that there are seven cases, as follows:

1. Nominative (*saxelobiti*)
2. Ergative (*motxrobiti*)
3. Dative (*micemiti*)
4. Genitive (*natesaobiti*)
5. Instrumental (*mokmedebiti*)
6. Adverbial (*vitarebiti*)
7. Vocative (*c'odebiti*)

Tallerman (2011, pp. 177-189) discusses ways of dividing and distinguishing three core arguments (S, A and O) by describing nominative-accusative, ergative-absolutive and split systems. This suggests that usually Nominative case is not expected in a language that displays ergative characteristics. Such languages are referred as having *ergative-absolutive* system. In ergative-absolutive languages, ergative case marks the

[30] Shanidze (1980, pp.73-77) names such cases as "local cases', since they indicate direction/orientation to/from a particular location.

subject of a transitive verb and absolutive case marks the subject of intransitive verbs and the direct object of transitive verbs. Whereas in *nominative-accusative* languages, nominative case marks the subject and accusative case marks the direct object of a transitive verb.

However, in Georgian the ergative case markers [**-m, -ma**] mark both subject of transitive, as well as subject of intransitive (unergatives) verbs. This is because Georgian displays characteristics of *split ergativity*, based on tense. Namely, the present tense (nonpast) has a nominative-accusative system, and in the past tense (aorist), an ergative-absolutive system.

Melikishvili (2008) describes Georgian as an active/ergative split Language. However, Amiridze (2006, p. 27) argues that Georgian is neither ergative nor split ergative language. According to Amiridze (2006, p. 29), Georgian shows split patterns between the nominative and active alignment as follows: the nominative alignment in TAM Series I and the active alignment in the TAM Series II and the TAM Series III. I will not go into further discussion of the alignment patterns of the case system in Georgian as this is beyond the scope of my PhD project.

To comply with the general concepts of *ergative-absolutive* and *nominative-accusative* and *split* language systems (Tallerman, 2011, pp. 177-189), it might seem reasonable to introduce the following terminology for the two cases as follows: *Nominative-absolutive* instead of *Nominative* and *Dative-accusative*, instead of *Dative*. Thus, hereafter I will consistently use the proposed terminology for these two cases in this thesis.

It should be mentioned that case markers and postpositions share certain properties. For example, both case markers and postpositions are suffixes which are cliticised to

nominals. I would like to discuss the criteria by which I decided that the above given cases are actually cases and not postpositions. There are several morphological phenomena that call for a distinction between case markers and postpositions. Firstly, postpositions govern a particular case, for instance, [**-ze]** "on" and [**-ši]** "in" govern the Dative-Accusative case, meaning that they can only be cliticised to a nominal in that case. Secondly, multiple postpositions and case markers cannot appear with nominals, with the specific exception of double case marking in a genitive construction.

(2) kal-**is**-**tvis**
woman-GEN-POST
"For a woman".

There are two additional cases, which fall outside the traditional case system. They are: 1) *Zero* (or null) and 2) *Suffixaufnahme* cases. I will briefly discuss these cases to justify my decision to include them in the tagset.

Zero (null) case was used in old Georgian in the V-XI cc. Marr (1908, 1925) was amongst one of the first Kartvelologists who classified the unmarked grammatical category as a zero case. This is debatable topic amongst Kartvelologists. Some grammarians including Shanidze (1934, p. 304; 1976, p. 31), Imnaishvili (1956, p. 59; 1957, p. 21), Zorell (1930) etc. recognize the unmarked form as a zero case. Some grammarians have different opinions on this matter. For example, Deeters (1930), Chikobava (1940, p. 13), Topuria (1965, p. 506), Sarjveladze (1984, p. 357), Uturgaidze (1986, p. 17) and Gogolashvili et al. (2011, p. 77) consider the zero case as a variation of a nominative case. However, they still differentiate a zero case from a nominative case and refer it as **გაუფორმებელი ფუძე** [**gauformebeli fuže**]

"unmarked root" (Chikobava, 1940), **უნიშნო სახელობითი** [**unišno saxelobiti**] "nominative without a marker" (Sarjveladze, 1984) or **არამარკირებული სახელდებითი ფორმა** [**aramarkirebuli saxeldebiti forma**] "unmarked nominative form" (Uturgaidze, 1986).

The zero case in old Georgian had its functions (Sarjveladze, 1984), for example, expressing a subject and a direct object (with certain types of verbs). Over the course of time, most functions of the zero case have been replaced by the marked nominative-absolutive case and hence, it was excluded from the traditional case system.

However, unmarked (zero case) form is still used in Modern Georgian. I will not go into a detailed discussion on this unmarked case, as it is not relevant to my PhD thesis. One of the main motivating reasons to include the zero case in the tagset (regardless of the discussion above if it is a case or not) is for clarity of analysis to count the unmarked form as a case. This can be very useful for linguistic research, for example, to look at the distribution of zero case in the corpus, analyse its functions and compare it to nominative-absolutive case.

I have also introduced an additional **Suffixaufnahme** case in the tagset. Suffixaufnahme (suffix resumption), is also known as **case stacking**. It is a linguistic phenomenon used in forming a genitive construction, where a genitive noun agrees with its head noun. For example:

(3) ded-isa-s
mother-GEN-DAT
"Of mother"

Suffixaufnahme was first recognized in Old Georgian (Bopp, 1842) and it was attested in all cases (see also Chapter 2, section 2.1.1.). Shanidze (1980, p. 92) discusses five cases with suffixaufnahme including nominative-absolutive, ergative, dative-accusative, adverbial and vocative cases. However, suffixaufnahme is more associated with Old Georgian than Modern Georgian. Contentiously, the KaWaC corpus data provides sufficient evidence for its existence in modern Georgian. According to the corpus data, suffixaufnahme occurs in four cases in Modern Georgian including ergative, dative-accusative, adverbial and vocative cases.

Table 4.1 demonstrated the observed frequency of suffixaufnahme cases in the KaWaC corpus.

Suffixaufnahme cases	**Observed freq.**
Genitive + Dative-Accusative	190,695
Genitive + Adverbial	75,658
Genitive + Ergative	1,037
Genitive + Vocative	114

Table 4. 1: The frequency of Suffixaufnahme case in the KaWaC.

Thus, suffixaufnahme case most frequently occurs with dative-accusative and adverbial cases. There are some rare examples of suffixaufnahme. For example, it can be used with an old plural in nominative-absolutive case.

(4) cql-isa-n-i

water-GEN-PL-NOM

“Of waters”

This example is a rare archaism, not part of the modern morphosyntax, and therefore it will receive the tag for genitive.

The corpus examples of the suffixaufnahme provides enough evidence to be included in the tagset. However, it should be noted that they are not as frequent as other cases. There are two main reasons for including suffixaufnahme in the tagset. First, it is part of the modern Georgian morphosyntax, since there is enough corpus evidence as shown in Table 4.1 above. Secondly, tagging suffixaufnahme can be very useful to extract the information about this phenomenon and analyse its syntactic or semantic features.

There are two possible ways to tag suffixaufnahme case. Firstly, it can get a tag for each case individually; for example, get separate tag for genitive and dative-accusative. Another possibility is a single tag for suffixaufnahme (e.g. for both genitive and dative-accusative). In the proposed tagset, the suffixaufnahme will receive a tag for each case individually. This will simply help the user to search or extract suffixaufnahme cases from the corpus more efficiently. For example, it will help the user to extract a set of individual pairs (genitive + dative-accusative or genitive + ergative), analyse and compare the frequency of their usage and distribution in the corpus.

Thus, I have introduced two additional cases together with the traditional case system. This results in total 12 cases for POS-tagging, as follows:

№	Case (English)	Case (Georgian)	Case Marker (Latin)	Case marker (Georgian)
1	Zero case	c'rfelobiti	Ø	Ø
2	Nominative-absolutive	saxelobiti	-i, -a, -o, -e, -u	-ი, -ა, -ო, -ე, -უ
3	Ergative	motxrobiti	-ma, -m	-მა, -მ
4	Dative-accusative	micemiti	-s, -sa	-ს, -სა
5	Genitive	natesaobiti	-is, -isa, -si	-ის, -ისა, -სი
6	Instrumental	mokmedebiti	-it, -ita, -ti	-ით, -ითა, -თი
7	Adverbial	vitarebiti	-ad, -d, -ada, -da	-ად, -დ, -ადა, -და
8	Vocative	c'odebiti	-o, -av	-ო, -ავ
9	Suffixaufnahme: Genitive + Ergative	natesaobiti + motxrobiti	-isam, -isama	-ისამ, -ისამა
10	Suffixaufnahme: Genitive + Dative-Accusative	natesaobiti + micemiti	-isas, -isasa	-ისას, -ისასა
11	Suffixaufnahme: Genitive + Adverbial	natesaobiti + vitarebiti	-isad	-ისად
12	Suffixaufnahme: Genitive + Vocative	natesaobiti + c'odebiti	-isav	-ისავ

Table 4. 2: The Case System in Georgian.

In addition to the set of cases, I have made a decision regarding each particular case marker. Nominative-absolutive as a rule is marked by the [**-i]** suffix. However, other vowels ([**a]**, [**o]**, [**e]** and [**u]**) also can function as nominative-absolutive case markers if a word ends in these vowels. Thus, in the tagset, I will consider these vowels as allomorphs for [**-i]** Nominative-absolutive marker.

In four cases, dative-accusative, genitive, instrumental and adverbial, nominals can add the [**-a**] element after the case marker. This [**-a**] element is the remnant of the article that was used in old Georgian. In modern Georgian, it can be affixed to the four cases including dative-accusative, instrumental and adverbial, especially before

enclitic postposition or particles. The use of the [**-a]** "affix" is optional in most cases. For example, it depends on the register or style of the language (e.g. in poetry etc.) or it simply depends on the choice of the user.

(5) a) ka'c-s
man-DAT
"To a man"

b) k'ac -s**a**
man-DAT
"To a man"

(6) a) ka'c-is-tvis
man-GEN-POST
"For a man".

b) k'ac -is**a**-tvis
man-GEN-POST
"For a man".

Both forms are grammatically correct and widely used. However, there are some preferences, where the [**-a]** element should be used, for example where there are two and more similar words conjoined (belonging to the same part-of-speech). In this case, the word just before the last word should add the [**-a]** affix. For example:

(7) lamaz-s amaq-s**a** da saxiers
beautiful-DAT proud-DAT and nice-DAT
"To someone beautiful, proud and nice."

I will not treat the [**-a]** affix separately. For the purposes of POS tagging, I will consider this as case allomorphy. For example, the dative-accusative case markers are [**-s]** and [**-sa]**, rather than just [**-s]**. The full list of all the case markers including the [**-a]** affix (as case allomorphy) is summarized in Table 4.3.

Case	**Original Case markers**	**[-a] allomorphy markers**
Dative-accusative	-s	-sa
Genitive	-is, -si	-isa
Instrumental	-it, -ti	-ita
Adverbial	-ad, -d	-ada, -da

Table 4. 3: The case marker "allomorphy" with the [-a] affix.

In general, the traditional classifications of the case system in Georgian are not very relevant for POS tagging purposes. They purely concern the specific forms taken by the different morphemes involved depending on different conditioning factors and are thus wholly matters of morphology rather than morphosyntax. The POS tags abstract away from all the above-discussed categories.

The consonant- and vowel-final roots have different declension paradigms – depending on the vowel or consonant-final root, the case markers change. This information can be very useful in POS-tagging.

The other interesting phenomena when dealing with case marking is syncope and apocope. As discussed in section 2.1.1.2, syncope is the loss of one or more sounds in the middle of a word and an apocope is the deletion of one or more sounds from the end of a word. In Georgian, syncope occurs in both nominal and verbal paradigms. In General, two or more syllable words undergo syncope if the final syllable consists of a vowel and sonorant (-VC). Syncopation usually occurs only in three cases: Genitive, Instrumental, and Adverbial, where three vowels syncopate. They are: [**a]**, [**e]** and [**o]**, when they form these syllables:

1. **[-al-], [-el-], [-ol-];**
2. **[-an-], [-en-], [-on-];**
3. **[-ar-], [-er-], [-or-];**
4. **[-am-], [-em-] [-om-]**.

As in:

(8) c'q'**al**i (NOM/ABS) → c'q'**l**is (GEN) “water”
berž**en**i (NOM/ABS) → berž**n**is (GEN) “Greek”
mx**ar**i (NOM/ABS) → mx**r**is (GEN) “side”
ir**em**i (NOM/ABS) → ir**m**is (GEN) “deer”
ob**ol**i (NOM/ABS) → ob**l**is (GEN) “an orphan”
mac**on**i (NOM/ABS) → mac**vn**is (GEN) “yoghurt”
mind**or**i (NOM/ABS) → mind**vr**is (GEN), “field”.

Thus, the three vowels – **[a]**, **[e]** and **[o]** are deleted when they form syllables with[**-l-], [-n-], [-r-]** and [**-m-]** consonants. However, in some cases, the [**-o-]** vowel can be reduced to [**-v-]** as in [**mindori**] "field" Nominative-absolutive to [**mindvrisa**] in Genitive.

It also should be noted that above given rules are not universal. Many nominals, however, end in such "syncopated" syllables, but they do not syncopate. These are known as non-syncopated nominals (Gogolashvili, 2011, pp.98-118). Thus, some nominals syncopate, and some do not. Prescriptive grammars simply provide lists of syncopated and non-syncopated nominals as they appear. For example, Gogolashvili (2011, pp. 98-118) discusses cases where syncopation occurs and gives a list of 375 non-syncopated and 349 syncopated words. However, the list is not corpus-based and there is no information regarding what sources have been used to identify non-syncopated and syncopated words.

I have used the corpus evidence to analyse vowel syncopation in Georgian. I have extracted over 5 million (more precisely 5,234,371) words with "syncopated syllables" in genitive, instrumental and adverbial cases from the KaWac corpus[31]. Based on the corpus examples, I have produced three types of lists. The first list includes the words that never syncopate. The second list covers the words that are always syncopated. The third list includes the words that can be found in both forms in the corpus: a) in some cases they are syncopated and b) sometimes they are not syncopated. These lists are as follows:

[31] The KaWac corpus (http://corpus.leeds.ac.uk/internet.html) has a limited search engine which does not allow advanced searches. Thus, I have used the Python programming language to analyse the vowel syncopation in the whole corpus. Namely, I have extracted the words with genitive, instrumental and adverbial case markers (by word endings) together with a corresponding POS tag and manually analysed them.

1) Non-syncopated words: **640 words**
2) Syncopated words: **335 words**
3) Words that are sometimes syncopated and non-syncopated: **50 words.**

This list is based on the corpus data and it may be useful to identify the patterns when the words are syncopated. The full list of vowel syncopation in Georgian is given in the Appendix B. Table 4.4 below illustrates the consonant final syncopated root in the case system as it appears in წყალი [**c'q'ali]** "water".

Case	**Singular**	**Plural**	**Old Plural**
Zero case	c'q'**al**	c'q'**l**eb	-
Nominative-absolutive	c'q'**al**i	c'q'**l**ebi	c'q'**al**ni
Ergative	c'q'**al**ma	c'q'**l**ebma	c'q'**al**t
Dative-accusative	c'q'**al**s	c'q'**l**ebs	c'q'**al**t
Genitive	c'q'**l**is	c'q'**l**ebis	c'q'**al**t
Instrumental	c'q'**l**it	c'q'**l**ebit	-
Adverbial	c'q'**l**ad	c'q'**l**ebad	-
Vocative	c'q'**al**o	c'q'**l**ebo	c'q'**al**no

Table 4. 4: Syncopation in [-al-] syllable.

Thus, in the singular forms [**-a-]** is syncopated in the Genitive, Instrumental and Adverbial Cases. In the plural [**-a-]** is syncopated in all cases, and in the Old plural, [**-a-]** is not syncopated at all.

In the consonant final non-syncopated type, the root never changes, and always has the same form regardless of what affixes are added to it. For example, **[k'aci]** "man", nominative-absolutive, [**k'acma]** "a man", Ergative.

In vowel final apocopated root, apocope takes place in two cases, in Genitive and Instrumental, and only two vowels are apocopated: [**a]** and [**e]**. For example, [**žma]** "brother", nominative-absolutive; **[žmis]** "brother", genitive. With a vowel final non-apocopated root, the root remains unchanged, but there are some changes in the case markers. For example, as in [**c'q'aro]** "river", nominative-absolutive to [**c'q'aroti]** in

genitive. Some vowel-final words are both syncopated and apocopated, for example, [**karxana]** “factory”, nominative-absolutive to [**karxnis]** in genitive.

4.1.3 Tags for Nouns

Thus, based on the discussions above, I have introduced two attribute values for nouns: *number* and *case*.

Value	**i) Number**	**ii) Case**
1	Singular	Zero case
2	Plural	Nominative-absolutive
3		Ergative
4		Dative-accusative
5		Genitive
6		Instrumental
7		Adverbial
8		Vocative
9		Suffixaufnahme: Genitive + Ergative
10		Suffixaufnahme: Genitive + Dative-accusative
11		Suffixaufnahme: Genitive + Adverbial
12		Suffixaufnahme: Genitive + Vocative

Table 4. 5: Attribute values for Nouns.

This gives 24 Tags for nouns. The full list of noun tags is given in the appendix A.

Description	**TAG**	**Examples (Latin)**	**Examples (Georgian)**
Noun Singular Zero Case	NSU	k'ac, saxl, kud	კაც, სახლ, ქუდ
Noun Singular Nominative - absolutive	NSN	k'aci, saxli, kudi	კაცი, სახლი, ქუდი
Noun Singular Ergative	NSE	k'acma, saxlma, kudma	კაცმა, სახლმა, ქუდმა

Table 4. 6: Sample tags for nouns.

4.2 Adjectives (*zedsartavi saxeli*)

Gogolashvili (2011, pp. 148-149) classifies adjectives according to their forms: 1) Primary Adjectives and 2) Derived Adjectives. Primary adjectives include, for example, [**didi]** "big", [**lamazi]** "beautiful", [**p^{h}arto]** "wide".

Derived adjectives are formed by derivational suffixes or prefix-suffix combinations (circumfixes), for example, [**-ian], [-ier]/[-iel], [-osan], [-ovan]** and [**-a]** suffixes. I will not consider derivational information further here, as it is not relevant for POS tagging.

Adjectives in Georgian can have degrees of comparisons as follows:

1. **Positive** - simply denotes a property, e.g. [**didi]** "big"; [**citeli]** "red"
2. **Attenuative**, is formed by [**mo-…-o]** circumfix: [**mo**dido**]** "slightly big", [**mo**citalo**]** "reddish".
3. **Superlative** is formed by **u-…-es** circumfix: [**u**dide**s**i**]** "biggest", [**u**citle**s**i**]** "reddest".

There are no specific morphemes that marks comparative degree of adjective. The method is via the addition of a functional element [**ufro]** "more". Thus, comparative degree is formed by **[ufro]** meaning "more", which precedes the adjective, for example [**ufro didi]** "bigger"". Alternatively, Superlative can also be formed by [**q'velaze]** meaning "most" preceding the adjective, for example, [**q'velaze didi]** "the biggest".

I will not consider degrees of comparison in POS-tagging. There are two reasons for this. First, degrees of comparison are derivational categories in Georgian. Since there are no morphological processes that signal comparative and superlative, there is no

need to include it in the tagset. Secondly, this will avoid another level of granularity and difficulty in POS-tagging.

Adjectives modify nouns and usually precede nouns, but they can also appear after nouns and with other elements intervening. Adjectives even may be used without nouns (function as nominal heads).

(9) gatenda lamaz-i dila
dawn.3S.SG.AOR beautiful-NOM morning.NOM
"A beautiful morning dawned."

(10) tval-eb-i-c lamaz-i gaqvs
eye-PL-NOM-PTCL beautiful-NOM have.2S.SG.PRS
"You have beautiful eyes too"

(11) damc'q'evla lamaz-ma
curse.3S.SG.AOR beautiful-ERG
"A beautiful one (woman) cursed me"

Adjectives decline like nouns depending on whether a given adjective appears before or after the noun it modifies. When an adjective appears after the noun it modifies, it takes all case markers like a noun, for example as in კაცი მართალი [**k'aci martali]** "true/honest man":

Case	**Singular**	**Plural**
Zero	k'ac martal	-
Nom./Abs.	k'aci martali	k'aceb**i** martaleb**i**
Erg.	k'ac**ma** martal**ma**	k'aceb**ma** martaleb**ma**
Dat./Acc.	k'ac**s** martal**s**	k'aceb**s** martaleb**s**
Gen.	k'ac**is** martl**is**	k'aceb**is** martleb**is**
Ins.	k'ac**it** martl**it**	k'aceb**it** martleb**it**
Adv.	k'ac**ad** martl**ad**	k'aceb**ad** martleb**ad**
Voc.	k'ac**o** martal**o**	k'aceb**o** martaleb**o**

Table 4. 7: Noun and adjective agreement.

As it is shown in Table 4.7 above, when the adjective modifies a Genitive noun with suffixaufnahme, it also copies the head noun case suffix, as in კაცისას მართლისას [**k'acisas martlisas]**, "true/honest man", suffixaufnahme, dative-accusative.

When an adjective appears before the noun it modifies, it takes the full case markers for three cases: nominative-absolutive, ergative, and vocative. Optionally it takes "partial" markers in two cases: genitive and instrumental - or takes no marker at all. However, this system applies only when the adjective has a consonant-final root (see Table 4.8).

Case	**Singular**	**Plural**
Zero	martal k'ac	martal k'aceb
Nom./Abs.	martal**i** k'ac**i**	martal**i** k'acebi
Erg.	martal**ma** k'ac**ma**	martal**ma** k'acebma
Dat./Acc.	martal k'ac**s**	martal k'acebs
Gen.	martal(**i**) k'ac**is**	martal(**i**) k'acebis
Ins.	martal(**i**) k'ac**it**	martal(**i**) k'acebit
Adv.	martal k'ac**ad**	martal k'acebad
Voc.	martal**o** k'ac**o**	martal**o** k'acebo
Suffix./Erg.	martal**i** k'ac**isam**	martal**i** k'aceb**isam**
Suffix./Dat.	martal**i** k'ac**isas**	martal**i** k'aceb**isas**

Table 4. 8: Noun and adjective agreement.

When an adjective appears after the noun it modifies, it agrees with the noun in case and number. However, an adjective preceding the noun partially agrees with the noun in case and not in number as demonstrated in the Table 4.8 above.

(12) c'el-s dedamic'a-s did-**ma** k'omet'a-**m** čaukrola
year-DAT earth-DAT big-ERG comet-ERG pass.3S.SG.AOR
"This year a big comet passed the earth"

(13) did-**i** c'armosaxv-**is** p'atroni xar
big-NOM imagination-GEN owner-NOM be.2S.SG.PRS
"You are the person of a big scope of imagination."

(14) axal did p'ort'-s ašenebs
new.ZER big.Ø port-DAT build.3S.SG.PRS
"(S/he) is building a new big port."

When the adjective has a vowel final root and appears before the noun, it takes no case marker at all regardless of the case of the noun, for example as in ყრუ კაცი [**q'ru k'aci]** "a deaf man" in Table 4.9 below.

Case	**Singular**	**Plural**
Zero	q'ru k'ac	q'ru k'aceb
Nom./Abs.	q'ru k'ac**i**	q'ru k'aceb**i**
Erg.	q'ru k'ac**ma**	q'ru k'aceb**ma**
Acc./Dat.	q'ru k'ac**s**	q'ru k'aceb**s**
Gen.	q'ru k'ac**is**	q'ru k'aceb**is**
Ins.	q'ru k'ac**it**	q'ru k'aceb**it**
Adv.	q'ru k'ac**ad**	q'ru k'aceb**ad**
Voc.	q'ru k'ac**o**	q'ru k'aceb**o**

Table 4. 9: Noun and adjective agreement.

Thus, we see that adjectives decline like nouns, but when used as an attribute they may or may not inflect for case. Thus, it is problematic to analyse an adjective which is used to modify a noun, but which has no case markers.

In addition to this, there is no clear difference between adjectives and nouns in Georgian. For example, adjectives can function as nominal heads (See example 11). However, this will not affect the tagging: it will be decided in the tagging lexicon whether a word is noun or adjective and so any given form will never have adjective/noun ambiguity.

4.2.1 Tags for Adjectives

Thus, like nouns, case and number categories are considered in the tagset design for adjectives. Before introducing POS tags for adjectives, I will briefly discuss the decision I have made to tag number and case. As mentioned above, the appearance of plural depends on whether an adjective appears before or after the noun. I follow the form and not the agreement: the adjective will be tagged as singular if it looks singular, even if it agrees with a plural head noun.

I have made several decisions concerning adjectives, which are used to modify a noun, but which have reduced or no case markers. Shanidze analyses (1980, pp. 81-85) such modifiers as having the same case as the head noun, even if there is no case marker on the adjective at all. It is worthwhile to mention that this is a right approach when analysing modifiers in Georgian. However, I will use a different approach for POS-tagging purposes to be consistent with the design principles that the tagset will tag by form rather than by function (see chapter 3).

I will discuss two cases to demonstrate the two possible ways of tagging an adjective when it has either a "partial" case marker, or no case marker**.**

Case 1: martal-Ø k'ac-is
Honest-Ø man-GEN
"Of honest man"

In Case 1, the adjective [**marta**l] "honest" could in theory be tagged in two different ways:

a) **martalØ**_JSU **k'acis**_NSG
b) **martalØ**_JSG **k'acis**_NSG

Thus, in case 1, [**martal]** could either be tagged according to its form, i.e. the base form, giving the tag **JSU** – Adjective_Singular_Zero Case. The second option it to tag

it according to its function. Hence, by the logic that it agrees with a genitive noun therefore it is genitive. This will give the following tag: **JSG** – Adjective_Singular_Genitive.

Case 2: martal-i(~~s~~) k'ac-is
Honest-~~GEN~~ man-GEN
"Of honest man"

In Case 2, the adjective [**marta**li] "honest" could also be tagged in two different ways:

a) **martal<u>i</u>**_JSN **k'ac<u>is</u>**_NSG
b) **martal<u>i</u>**_JSG **k'ac<u>is</u>**_NSG

In the second case, likewise [**martali]** could be tagged as **JSN** according to its form, since the **[-i]** is nominative-absolutive case marker or **JSG** according to its theoretical, unmarked agreement. I will use the first approach - **JSU** in the first case and similarly in the second case, **JSN**- Adjective_Singular_nominative-absolutive. Despite the fact that [**i**] in **martali** is etymologically part of the genitive, I will treat these forms as nominative-absolutive, as [**i**] is nominative-absolutive case marker and this approach will avoid a major problem of ambiguity of analysis everywhere in terms of POS-tagging. To conclude, adjectives will be tagged according to their morphological form and not unrealised grammatical features (position).

Thus, adjectives will be tagged according to two attribute values: number and Case. It gives 24 Tags for adjectives. The full list of adjective tags is given in the appendix A.

Description	**TAG**	**Examples (Latin)**	**Examples (Georgian)**
Adjective Singular Nominative-absolutive	JSN	cudi, martali,	ცუდი, მართალი
Adjective Singular Ergative	JSE	q'rum, martalma	ყრუმ, მართალმა

Table 4. 10: Sample tags for adjectives.

4.3 Pronouns (*nacvalsaxeli*)

Shanidze (1980, pp. 41-44) and Gogolashvili (2011, pp. 168-183) describe eleven types of pronouns including personal, reflexive, demonstrative, interrogative, possessive, interrogative-possessive, relative, reciprocal, intensive, indefinite and negative pronouns. I will discuss each of them in turn.

Personal Pronouns (*p'iris nacvalsaxeli*). Shanidze (1980, pp. 41-43) describes three personal pronouns: first, second and third personal pronouns. It is worthwhile to note that the third personal pronouns are demonstrative pronouns that function as third person pronouns. Thus, I will consider only two personal pronouns including first person and second person.

Singular	Plural	English
me	**čven**	I
šen	**tkven**	You

Table 4. 11: Personal pronouns.

(15) **me** alp'inist-i var
I alpinist-NOM be.1S.SG.PRS
"I am an alpinist".

(16) **šen** ašk'arad ničier-i p'oet-i xar
You obviously talented-NOM poet-NOM be.2S.SG.PRS
"You are obviously a talented poet."

Thus, the two personal pronouns are the first and second persons. Each have singular and plural forms. As mentioned above, a group of demonstrative pronouns function as third person pronouns, for example, ის **[is]** can mean "S/he; it"; ისინი **[isini]** "they" etc. Typologically, we would expect demonstrative to be the main function and third person pronoun to be the extra functions, since third person pronouns in many languages are frequently created by means of a process where demonstrative pronouns

are grammaticalized to third person pronouns over time (Haine and Song, 2011; Heine and Reh, 1984, p. 271; Diessel, 1997, 1999; Klausenburger, 2000). Therefore, I will discuss this group of pronouns as demonstratives only.

The personal pronouns as a rule have no case. However, there are exceptions regarding the second personal pronouns შენ [**šen]** and თქვენ [**tkven]**. These pronouns can have vocative case if they are used as modifiers. In particular, [**šen]** and [**tkven]** can get a proper vocative case marker [**-o]**, as in თქვენო აღმატებულებავ [**tkveno ağmat'ebulebav]** "your majesty". However, more commonly they do not get vocative case markers [**-o]** or [**-v]**, but instead drop the final [**n]** consonant.

(17) modi ak **še** mamažağl-o
come.2S.SG.AOR here you.VOC bitch-VOC
"Come here, you son of a bitch."

(18) rat'om damblok'et **tkve** ertujredian-eb-o
why block.2S.PL.1O.SG.AOR you.VOC.PL unicellular-PL-VOC
"Why did you block me, you unicellular (creatures)?"

Demonstrative pronouns (*čvenebiti nacvalsaxeli*). All Demonstrative pronouns in Georgian have deictic meaning (Gogolashvili, 2011, pp. 173-174). Some demonstratives can also function as 3rd personal pronouns. However, they will be referred as demonstrative pronouns for the purposes of part-of-speech tagging. This will avoid major disambiguation problem. These demonstrative pronouns are summed up in the Table 4.12 below.

Singular	Plural	English
es	**eseni**	"This" 1st person deixis
eg	**egeni**	"That" 2nd person deixis
is	**isini**	"That" 3rd person deixis
igi	**igini**	"That" 3rd person deixis

Table 4. 12: Demonstrative Pronouns.

(19) **eg** azr-i gakvs tav-ši
this idea-NOM have.2S.SG.PRS head-POST
"You have this idea in your head"

These demonstrative pronouns are different from the other demonstrative pronouns in many ways. Unlike other demonstrative pronouns, they use old plural forms and have irregular declension paradigms.

The irregularity of these four pronouns is that they show two different roots when declined: that is, they are suppletive roots. The root of the nominative-absolutive case occurs in singular and plural forms, but there is another root for the other cases in both singular and plural forms. Also, the four demonstratives do not have Vocative case.

These "secondary" roots are **ამა** [**ama], მაგა [maga], მა [ma]** and **იმა [ima]** and they are apocopated when declined. The difference here from the normal paradigm is that the ergative case marker is [**-n]**, instead of [**-ma/m]**.

Case	**es**		**eg**		**igi**		**is**	
	SG	**PL**	**SG**	**PL**	**SG**	**PL**	**SG**	**PL**
Nom. / Abs.	es(e)	eseni	eg(e)	egeni	igi	igini	is(i)	isini
Erg.	aman	**amatma**	magan	**magatma**	man	**matma**	iman	**imatma**
Dat./ Acc.	amas	amat(**s**)	magas	magat(**s**)	mas	mat(**s**)	imas	imat(**s**)
Gen.	amis	**amatis**	magis	**magatis**	mis	**matis**	imis	**imatis**
Ins.	amit	**amatit**	magit	**magatit**	mit	**matit**	imit	**imatit**
Adv.	amad	-	magad	-	-	-	ima	**imatad**

Table 4. 13: Declension of Demonstrative Pronouns.

According to Shanidze (1980) these demonstrative pronouns in plural have only two cases: nominative-absolutive and dative-accusative. However, there are possibly more than two cases if we take into account that these demonstrative pronouns form their old plural forms by the [**-t(a)]** fusional suffix, which marks both plurality and three

cases: ergative, dative-accusative and genitive (Shanidze, 1980, pp. 47-48; Gogolashvili, 2011, p. 140).

Suffixaufnahme cases with this fusional suffix are considered to be *possessive pronouns* (Shanidze, 1980). In table 4.13 above, I have introduced and highlighted examples, which is attested in the corpus data. It is quite obvious that they are not the third person possessive pronouns, but a genitive construction of old plural [**-t(a)]** + modern case markers.

As discussed above, we know that the [**-t(a)]** fusional suffix can represent either of these three ergative, dative-accusative and genitive cases. But in modern Georgian, it takes the regular case suffixes on top of the old one [**-t(a)**] suffix. In the Georgian Orthographic Dictionary[32], which prescribes the norms and rules, describes these forms including ([**amatma], [amats], [amatit]**) as incorrect and suggesting using them without the case markers. However, in the KaWaC corpus, there are many examples of such "incorrect forms", which I have analysed.

Wordform Georgian	**Wordform Latin**	**Observed freq. per 1000 lines**
მათმა	matma	873
მათს	mats	694
ამათმა	amatma	378
მათით	matit	252
მაგათმა	magatma	227
იმათმა	imatma	212
მათის	matis	106
მაგათის	magatis	23
იმათს	imats	22

Table 4. 14: Corpus frequency of demonstratives: Plural + case marker.

[32] http://ena.ge/

Most contexts that they are used in suggest standard Ergative or dative-accusative case. However, there are very few examples (see Table 4.14 above) where the extended forms are ambiguous – being used both for standard ergative and dative-accusative case and suffixaufnahme case. Thus, the decision is made in light of the corpus evidence and the discussion above.

To conclude, in these pronouns, the [**-t(a)]** fusional suffix as attested in the corpus are used as a plural marker only. Thus, the -[**t(a)]** suffix in this context is no longer fusional but a pluralizing suffix.

Another difference with these demonstratives is that they change the root depending on whether they function as modifiers (agree with the noun) or not. The root can be changed as follows: [**es] → [am]; [eg] → [ma], [igi] / [is] → [im]**. For example: [**es] / [eg]/ [is kaci]** "this/that man".

Case	**es**		**eg**		**is**	
	SG	**PL**	**SG**	**PL**	**SG**	**PL**
NOM/ ABS	es k'aci	es k'acebi	eg k'aci	eg k'acebi	is k'aci	is k'acebi
ERG	am k'acma	am k'acebma	mag k'acma	mag k'acebma	im k'acma	im k'acebma
DAT/ ACC	am k'acs	am k'acebs	mag k'acs	mag k'acebs	im k'acs	im k'acebs
GEN	am k'acis	am k'acebis	mag k'acis	mag k'acebis	im k'acis	im k'acebis
INS	am k'acit	am k'acebit	mag k'acit	mag k'acebit	im k'acit	im k'acebit
ADV	am k'acad	am k'acebad	mag k'acad	mag k'acebad	im k'acad	im k'acebad

Table 4. 15: Declension of demonstratives with the head noun.

Unlike other demonstrative pronouns, these pronouns do not agree with nouns in case and number. As discussed above, this type of demonstrative pronouns takes old plural forms. Other demonstrative pronouns form their plural by regular [**-eb]** pluralizing suffix.

Other demonstrative pronouns include for example, ასეთი [**aseti**] “such as”, “this kind of”. They inflect for case.

Singular	Plural	English
aseti	**asetebi**	“such as”, “this kind of”; 1st person deixis
amnairi	**amnairebi**	“this kind of”; 1st person deixis
magistana	**masgistanebi**	“this kind of”; 2nd person deixis

Table 4. 16: Demonstrative pronouns.

Interrogative Pronoun (*k'itxviti nacvalsaxeli*). The set of Interrogative Pronouns contains:

Singular	Plural	English
vin	vinebi	Who
ra	raebi, reebi	What
radara	-	What kind
raerti	-	how many/much
ramdeni	-	how much/many; so much/many
ranairi	ranairebi	what kind/sort of
rarigi	-	what sort, type, kind
rodindeli	rodindelebi	at/of/from what time
rogori	rogorebi	what sort/type/kind of
romeli	romlebi	which, who, what
sadauri	sadaurebi	From where

Table 4. 17: Interrogative pronouns.

(20) esen-i **vin** arian
these-NOM.PL who be.3S.PL.PRS
“Who are these (people)?”

(21) **ra** moxda
what happen.3S.SG.AOR
“What happened?”

Some interrogative pronouns inflect for case. However, the following two interrogative pronouns ვინ [**vin]** "who" and რა [**ra]** "what" show some irregularities when declined. They are defective, in particular, [**vin]** has only two forms in four cases: [**vin]** in nominative-absolutive and ergative case and **vis(a)** in dative-accusative and genitive case. In prescriptive grammars, [**vin]** and [**ra]** have no plural forms. However, the plural forms of the pronouns are attested in the corpus data. For example, the observed frequency for the wordform [**reebi**] in the corpus is 174 per 1000 corpus lines. Thus, based on the corpus evidence, the plural usage of these pronouns is quite common in modern Georgian.

(22) net'avi **vin-eb-i** igulisxmebian
wonder who-PL-NOM mean.3S.PL.PRS
"I wonder who (PL) are meant."

(23) arc ici **re-eb-i** vakete
not know.2S.SG.PRS what-PL-NOM do.1S.SG.AOR
"You don't know, what (things) I did."

Possessive Pronouns (*k'utvnilebiti nacvalsaxeli*). All possessive pronouns can be declined. The complete set of possessive pronouns are:

Singular	**Plural**
čemi "my"	**čveni** "our"
šeni "yours"	**tkveni** "your"

Table 4. 18: Possessive pronouns.

(24) nerviulobda **čem-i** žma
worry.3S.SG.IMPERF my-NOM brother.NOM
"My brother was worrying."

Thus, in traditional grammars there is a category of 1st and 2nd person possessive pronouns, but not for 3rd person possessive pronouns. According to Shanidze (1980,

p. 43; Gogolashvili, 2011, p. 175), demonstratives and reflexives function as third person possessive pronouns.

misi "his/her/its"	**mati** "their"
imisi "his/her/its"	**imati** "their"
tavisi "his/her/its"	**tavisebi** "their"
tavianti "his/her/its"	-
tvisi "his/her/its"	-

Table 4. 19: Third person possessive pronouns.

(25) **imat-i** gvar-eb-i aravin ar icis
their-NOM.PL surname-PL-NOM nobody not know.3S.SG.PRS
"Nobody knows theirs surnames."

In Table 4.19 above, მისი [**misi]**, იმისი [**imisi]**, მათი [**mati]** and იმათი [**imati]** are demonstratives in genitive or suffixaufnahme cases and they will be treated as such. As for თავისი [**tavisi]**, თავისები [**tavisebi]**, თვისი [**tvisi]** and თავიანთი [**tavianti],** they will get a special tag as reflexive possessives. The word **თავი [tavi]** in Georgian literally means "head". It will be treated as noun anywhere except these twelve forms: **თავისი [tavisi], თავისები [tavisebi], თვისი [tvisi], თავისმა [tavisma], თვისმა [tvisma], თავისად [tavisad], თავისას [tavisas], თვისას [tvisas], თვისით [tvisit], თავისით [tavisit], თავიანთ [taviant] and თავის [tavis].** These forms will be tagged as reflexive possessives.

It should be mentioned that some ungrammatical forms that are not discussed in the traditional grammar (Shanidze, 1980), are attested in the corpus data. For example, the wordform [**tavisebi**] occurs 11 times in the Georgian web-corpus.

Interrogative-Possessive Pronouns (*k'itxvit-k'utvnilebiti nacvalsaxeli*). Shanidze (1980, p. 42) and Gogolashvili (2011, p. 177) describe a separate category of

Interrogative-possessive pronouns in Georgian. There are three such pronouns: ვისი [**visi]** "whose"; რისა [**risa]** "which/whose" and რისი [**risi]** "which/whose".

(26) mašin **visi** bral-i-a
then whose fault-NOM-COP
"Whose fault is this then?"

(27) **risi** dablogva ğirs da **risi** ara
what blog.NOM worth.3S.SG.PRS and what no
"What is worth to write a blog about?"

Like third person possessive pronouns, these interrogative possessives are interrogative pronouns, with suffixaufnahme cases. Thus, they are not a separate category and will be treated as interrogative pronouns.

Relative Pronouns (*mimartebiti nacvalsaxeli*). This category of pronouns is formed by adding the [**c(a)]** particle to interrogative pronouns. The [**c(a)]** is treated as an enclitic particle with the particles. These pronouns usually function as conjunctions and can be declined. The set of relative pronouns are as follows:

Singular	**Plural**	**English Translation**
vinc	**vinebic**	"who", "whoever"
vinca	**-**	"who", "whoever"
visic	**visebic**	"whose"
rac	**raebic, reebic**	"what", "that"
raca	**-**	"what", "that"
ramdenic	**-**	"however many/much"
ranairic	**ranairebic**	"what"
risac	**-**	"what"
risic	**-**	"what"
rodindelic	**-**	"at/of/from what time"
rogoric	**rogorebic**	"what (sort/type/kind of):"
romelic	**romlebic**	"which" "that"
sadauric	**sadaurebic**	"from where"

Table 4. 20: Relative Pronouns.

The relative pronouns will not be treated separately, since [**c(a)]** is an enclitic particle and has a tag on its own. Thus, there is no need for a category of relative pronouns to be presented separately in the tagset.

Reciprocal Pronouns (*urtiertobiti nacvalsaxeli*). There are about four reciprocal pronouns in Georgian (Gogolashili, 2011, p. 181; Shanidze, p. 43) as follows: **ერთმანეთი** [**ertmanet**i], **ერთურთი** [**erturti]**, **ერთიმეორე** [**ertimeore] and ურთიერთი [urtierti]** "each other", "one another". Reciprocal pronouns vary for case. They can have all cases, except the vocative case.

(28) dğe-s vnaxet **ertmanet-i**
day-DAT see.1S.PL.AOR each other-NOM
"Today we saw each other."

(29) p'oliponi-it **erturts** at'k'bobdnen
polyphony-INS one another-DAT sweeten.3S.PL.IMPERF
"They were enjoying one another with polyphony."

Reflexive Pronoun (*uk'ukceviti nacvalsaxeli*). There is only one reflexive pronoun **თავი** [**tavi]** "self", which is a noun, meaning "head". It can function as both, noun and reflexive pronoun in the sentence.

(30) uar-is nišn-ad did-i **tav-i** gaaknia
refusal-GEN sign-ADV big-NOM head-NOM shake.3S.SG.AOR
"(S/he) shook (her/his) head as a sign of refusal."

(31) bela-m **tav-i** moik'la
Bela-ERG self-NOM kill.3S.SG.AOR
"Bella killed herself."

(32) sakutarØ **tav-sa-c** p'at'iv-s vcem
Own.Ø self-DAT-PTCL respect-DAT pay.1S.PRS
"I respect my own self too."

In order to avoid ambiguity in POS-tagging, I will be treating თავი [**tavi]** "head" as a noun. However, it will get special tags for reflexive possessive as discussed above.

Intensive Pronouns (*gansazğvrebiti nacvalsaxeli*). Some linguists use the term "emphatic" (Hewitt, 1995, pp. 84-85) to describe this type of pronouns. There are ten intensive pronouns in Georgian, out of which six can inflect for case. They are:

- თვითოელი **[tvitoeuli]**, თითოეული [**titoeuli**] "each single one";
- ყოველი **[q'oveli]** "very, any, each, all";
- ყველა **[q'vela]** "all, every; everything; everyone, everybody";
- სხვა **[sxva]** "other"; მავანი **[mavani]** "someone; some people; a certain (sb)".

The other four intensive pronouns cannot be declined (Gogolashvili, 2011, pp. 178-179) They are: თვით [**tvit]** "oneself, myself, yourself, itself"; თვითონ [**tviton]** "oneself"; თითონ [**titon]** "itself, oneself, myself, yourself"; თავად [**tavad]** "personally". Most intensive pronouns do not have number. However, one intensive pronoun სხვა [**sxva]** "other" can have plural number.

(33) žiur-is **titeul-i** c'evr-i damouk'idebel-i-a
jury-GEN each-NOM member-NOM independent-NOM-COP
"Each member of the jury is independent."

(34) **sxv-eb-i** ra-s it'q'vian
other-PL-NOM what-DAT say.3S.PL.FUT
"What will others say?"

Indefinite Pronouns (ganusaz*ğvrelobiti nacvalsaxeli*). Indefinite pronouns belong to a class of pronoun that indicates indefinite references. The indefinite pronouns can be declined. The full list of indefinite pronouns contains:

Singular	Plural	English
erti	**-**	one
vinme	**vinmeebi**	somebody, someone; anybody, anyone; some people
viğac(a)	**viğaceebi**	someone/ somebody, a certain person
zogi	**-**	some, a certain; one
zogierti	**zogiertebi**	one/several (of several/many); a certain person
rame	**rameebi**	Something, some
rağac(a)	**rağaceebi**	something, anything; some
rogoriğac(a)	**-**	something
romelime	**-**	somebody or other
romeliğac(a)	**-**	somebody; some, a certain (sb/sth)

Table 4. 21: Indefinite Pronouns.

Negative Pronouns (*uarq'opiti nacvalsaxeli*). There are about ten negative pronouns in Georgian, given in Table 4.21 below.

Georgian	English
aravin	Nobody
aeravin	Nobody/no one… can/may
vervin	Nobody/no one… can/may
nuvin	No one/nobody, don't anyone/anybody
nuravin	No one/nobody, don't anyone/anybody
nurvin	No one/nobody, don't anyone/ anybody
araperi	Nothing
veraperi	Nothing… can/may
nura	Nothing
nuraperi	Nothing, don't… anything

Table 4. 22: Negative Pronouns.

Unlike other pronouns, Negative Pronouns only have singular forms and some of them can be declined (Gogolashvili, 2011, pp. 181-183). The following Negative particles can be declined: არაფერი [**araperi]**; ვერაფერი [**veraperi]**; ნურა **[nura]** and ნურაფერი [**nuraperi]** "nothing".

4.3.1 Tags for Pronouns

Most, albeit not all, pronouns have irregular case inflections, and many pronouns lack plural forms. Thus, I will give attribute values for each type and then will define the tags for them.

Personal pronouns have the attribute values for number (suppletive), person and cases as follows:

Value	**i) type**	**ii) Person**	**iii) Number**	**iv) Case**
1	Personal	First	Singular	Zero
2		Second	Plural	Nominative -Absolutive
3				Dative-accusative
4				Vocative

Table 4. 23: Attribute values for personal pronouns.

Demonstratives have the attribute values for number and case as follows:

Value	i) type	ii) Number	iii) Case
1	Demonstrative	Singular	Zero
2		Plural	Nominative-Absolutive
3			Ergative
4			Dative-accusative
5			Genitive
6			Instrumental
7			Adverbial
8			Vocative
9			Suffixaufnahme-Ergative
10			Suffixaufnahme-Dative-accusative
11			Suffixaufnahme-Adverbial

Table 4. 24: Attribute values for Demonstrative Pronouns.

As discussed above, not all demonstrative pronouns inflect for case. Some of them have only one or two cases, some of them have the full case inflection. I will take into account all the exceptions in POS-tagging as each pronoun group is small enough to deal with.

The same approach can usefully be employed for other pronouns. For example, some interrogative pronouns are marked for zero case and some inflect for all the cases except vocative case. To sum up, attribute values of interrogative pronouns are as follows:

Value	i) type	ii) Number	iii) Case
1	Interrogative	Singular	Zero
2		Plural	Nominative-Absolutive
3			Ergative
4			Dative-accusative
5			Genitive
6			Instrumental
7			Adverbial
8			Suffixaufnahme-Ergative
9			Suffixaufnahme-Dative-accusative
10			Suffixaufnahme-Adverbial

Table 4. 25: Attribute values for Interrogative Pronouns.

The Possessive pronouns have the attribute values for person, number and case as follows:

Value	i) type	ii) Person	iii) Number	iv) Case
1	Possessive	First	Singular	Zero
2		Second	Plural	Nominative-Absolutive
3	Reflexive			Ergative
4				Dative-accusative
5				Genitive
6				Instrumental
7				Adverbial
8				Vocative
9				Suffixaufnahme-Ergative
10				Suffixaufnahme-Dative-accusative
11				Suffixaufnahme-Adverbial
12				Suffixaufnahme-Vocative

Table 4. 26: Attribute values for Possessive Pronouns.

Thus, possessive pronouns decline like nouns. However, some possessives do not inflect for all the cases (Gogolashvili, 2011, pp. 175-177). This information will be considered in part-of-speech tagging.

Reciprocal pronouns have attribute values for case only (Shanidze, 1980, pp. 98-99; Gogolashvili, 2011, p. 181):

Value	**i) type**	**ii) Case**
1	Reciprocal	Zero
2		Nominative-Absolutive
3		Ergative
4		Dative-accusative
5		Genitive
6		Instrumental
7		Adverbial
8		Suffixaufnahme-Ergative
9		Suffixaufnahme-Dative-accusative
10		Suffixaufnahme-Adverbial

Table 4. 27: Attribute values for Reciprocal Pronouns.

The intensive pronouns have attribute values for case inflection only, but **[sxva]** "other" can have singular and plural number. Thus, it will get tags for case and number. The rest of the intensive pronouns will get tags for case only.

Value	**i) type**	**ii) Case**
1	Empathic	Zero
2		Nominative-Absolutive
3		Ergative
4		Dative-accusative
5		Genitive
6		Instrumental
7		Adverbial
8		Vocative
9		Suffixaufnahme-Dative-accusative
10		Suffixaufnahme-Adverbial

Table 4. 28: Attribute values for Empathic Pronouns.

Indefinite Pronouns have attribute values for number and case.

Value	**i) type**	**ii) Number**	**iii) Case**
1	Indefinite	Singular	Zero
2		Plural	Nominative-Absolutive
3			Ergative
4			Dative-accusative
5			Genitive
6			Instrumental
7			Adverbial
8			Suffixaufnahme-Ergative
9			Suffixaufnahme-Dative-accusative
10			Suffixaufnahme-Adverbial

Table 4. 29: Attribute values for Indefinite Pronouns.

Negative Pronouns have number and case inflection. However, some negative pronouns are given in Zero case form and some of them in nominative-absolutive. Thus, the attribute values for Negative pronouns are as follows:

Value	i) type	ii) Case
1	Negative	Zero
2		Nominative-Absolutive
3		Ergative
4		Dative-accusative
5		Genitive
6		Instrumental
7		Adverbial
8		Suffixaufnahme-Dative-accusative

Table 4. 30: Attribute values for Negative Pronouns.

Overall, this gives 163 tags for pronouns. The full list of pronoun tags is given in the appendix A.

Description	TAG	Examples (Latin)	Examples (Georgian)
Pronoun Personal First Person Singular Nominative-Absolutive Case	PP1SN	me	მე
Pronoun Demonstrative Singular Ergative	PDSE	asetma, magnairma	ასეთმა, მაგნაირმა
Pronoun Negative Ergative	PNE	araferma, veraferma	არაფერმა, ვერაფერმა
Pronoun Negative Dative-accusative	PND	arafers, verafers	არაფერს, ვერაფერს

Table 4. 31: Sample tags for pronouns.

4.4 Numerals (*ricxviti saxeli*)

There are three types of numerals generally recognised in Georgian: **Cardinal, Ordinal,** and **Fraction**. I will introduce an additional type of **Diminutive** numeral.

The Diminutive numerals is formed by adding [**-ode]** suffix to cardinal numerals: the [**oriode]** "just two", [**xutiode]** "just five". The [**-ode]** suffix is usually considered to be a particle. However, it has very distinctive features from the rest of the particles in Georgian. It is used only with numerals, expresses the exact numbers, and has the sense of "not being sufficient".

(35) me-c ĉavurtav **or-i-ode** sit'q'va-s
Me-PTCL add-1S.SG.FUT two-NOM-DIM word-DAT
"I will also get a (two) word in".

Case markers appear after the [**-ode]** suffix. Thus, the [**-ode]** cannot be an enclitic particle - rather it must be either an inflectional or derivational affix, as enclitics are almost always expected to be further from the root than inflectional affixes.

The Cardinal numbers from one to ten are simple numerals, such as 1- [**erti]**, 2 – [**ori]**, 3 – [**sami]**, 4 – [**otxi]**, 5- [**xuti]**, 6 –[**ekvsi]**, 7 – [**ŝvidi]**, 8 – [**rva]**, 9 – [**cxra]**, 10 – [**ati]**. The numbers from eleven to nineteen are compound numerals with more than one root. For example, 11-[**tertmeti]**, 12 – [**tormeti]**, 13 – [**cameti]**, 14 – [**totxmeti]**, 15 – [**txutmeti]**, 16- [**tekvsmeti]**, 17 – [**čvidmeti]**, 18 – [**tvrameti]**, 19 – [**cxrameti]**.

The Ordinal numbers are formed by attaching the circumfix [**me…e]** to the root of the cardinal numerals, as in [**meore]** "2nd", [**mesame]** "3rd", [**meotxe]** "4th", [**mexute]** "5th" etc.

The Fraction numbers are formed by adding the circumfix **[me…edi]** to the cardinals (or ordinal+ suffix [**-di]**), for example: [**mesamedi]** "1/3", [**meotxedi]** "1/4".

Numerals decline like nouns to agree with the head noun they quantify. The numerals with consonant-final stem are non-syncopated; the only exception is an indefinite numeral, **mravali** "a lot/many" that is syncopated.

(36) biznes-i **xut-ma** adamian-**ma** davic'q'et
business-NOM five-ERG person-ERG start.1S.PL.AOR
"Five of us started a business".

In general, numerals do not have vocative case. However, there can be some exceptions and I will therefore include the vocative case in the tagset.

(37) žilinebisa nomer-o **or-o**
goodnight number-VOC. two-VOC
"Goodnight number two" (referring sb. who is second on the list).

As a rule, numerals use old plural forms, but there are cases where the [**-eb]** pluralising suffix is used. Like in nouns, the Old plurals and modern [**-eb]** plurals in numerals will receive the same tags.

(38) **xut-n-i** da-n-i viq'avit
five-PL-NOM sister-PL-NOM be.1S.PL.AOR
"We were five sisters."

(39) ar mecadineobda da mainc **xut-eb-i** hq'avda
Not study.3S.SG.IMPERF and despite five-PL-NOM have.3S.SG.IMPERF
"S/he was not studying, despite this, s/he had 5 (highest) marks."

Thus, the attribute values for numerals include the following: Four types of numerals (Cardinal simple, Cardinal Diminutive, Ordinal and Fraction) and two morphological

categories of case and number. However, no plural forms are available for diminutive numerals.

The following exceptions with regard to case should be taken into consideration: Zero case is given only for cardinal and approximative numerals in singular forms, because ordinal and fraction numerals are formed with the [**me-e/-edi]** circumfixes and the[**-e]** and [**-i]** suffix endings function as nominative-absolutive case markers. That means there is no "base" shorter than the nominative-absolutive form which could appear alone. As for the double genitive construction, it can only occur only in cardinal and ordinal numerals. Namely, it occurs in all four suffixaufnahme cases, and in three cases (ergative, dative-accusative, vocative) in ordinal numerals.

4.4.1 Tags for numerals

Thus, numerals will receive tags according to their type (cardinal, ordinal etc.) and the grammatical categories of number and case. The attribute values for numerals are summed up in the table 4.32 below:

Value	**i) type**	**ii) Number**	**iii) Case**
1	Cardinal Simple	Singular	Zero Case
2	Cardinal Approximative	Plural	Nominative-absolutive
3	Ordinal		Ergative
4	Fraction		Dative-accusative
5			Genitive
6			Instrumental
7			Adverbial
8			Vocative
9			Suffixaufnahme: Genitive + Ergative
10			Suffixaufnahme: Genitive + Dative-accusative
11			Suffixaufnahme: Genitive + Adverbial
12			Suffixaufnahme: Genitive + Vocative

Table 4. 32: Attribute values for numerals.

In total, this produces 58 tags. The full list of numeral tags is given in the appendix A.

Description	**TAG**	**Examples (Latin)**	**Examples (Georgian)**
Numeral Cardinal Singular Ergative	MCSE	samma, orma	სამმა, ორმა
Numeral Ordinal Singular Dative-accusative	MOSD	mesames, meores	მესამეს, მეორეს
Numeral Fraction Singular Genitive	MFSG	mesamedis, meoredis	მესამედის, მეორედის
Numeral Diminutive Singular Ergative	MDSE	samiodem, oriodem	სამიოდემ, ორიოდემ

Table 4. 33: Sample tags for numerals.

4.5 Adverbs (*zmnizeda* or *zmnisarti*)

Adverbs are words that mainly modify the meaning of verbs, but also adjectives and other adverbs. They can express manner, place, time or reason, aim and purpose (Shanidze, 1980, pp. 987-588). Like particles, adverbs cannot be inflected or declined. Adverbs in Georgian can be classified according to 1) their forms and 2) their functions.

According to their form, adverbs are classified into two major types of adverbs. They are **primary adverbs** and **derived adverbs (**Shanidze, 1980, pp. 587-594). There are very few primary adverbs that are originally adverbs, such as [**xval]** "tomorrow"; [**ak]** "here". The derived adverbs are formed by derivational adverb suffixes, which are: [**-gan]**, [**-iv]**, [**-re]**, [**-gzis]**, [**-jer]**, [**-da]**, [**-mo]**, [**-še]**, [**-ğam]**, [**-ma(rta**)], [**-mag]**, [**-kec]**, **[-xel]** and **[-xan]**. These derivational suffixes are usually used with numerals and sometimes with other nominals too. For example, [**xutjer]** "five times"; [**ganuc'q'vetliv]** "continuously", [**mravalgzis]** "many /multiple times".

There are other types of adverbs, which have exactly the same form as nominals with case inflections, namely: dative-accusative, genitive, instrumental and Aadverbial and also, zero case nominals.

Dative-accusative - nominal adverbs. These are nominals in dative-accusative case, but they can function as adverbs, for example:

[žiri] "bottom; base" in dative-accusative [**žirs**], functions as a noun:

(40)	xe	ğrm-ad	idgams	**žir-s**	mic'a-ši
	tree.NOM	deep-ADV	grow.3S.SG.PRS	root-DAT	soil-POST
	"Tree grows (roots) deeply in the soil".				

The same example can also functions as an adverb:

(41) **žir-s** vašl-eb-i eq'ara
Bottom-DAT apple-PL-NOM drop.3S.PL.IMPERF
"There were apples (dropped) at the bottom of (something)."

In the tagset, I will not classify dative-accusative nominal adverbs as adverbs, but they will be treated as nominals in dative-accusative case.

Genitive-nominal adverbs. In most cases, these adverbs involve multiply-marked nouns: the adverb then includes postpositions or two case markers, genitive and instrumental, for instance. Compare [**dğe]** "day/ daylight" and [**dğisit]** "during a day/ by day/ in daylight"; the latter has two case markers, [**-is]** for genitive and [**-it]** for instrumental.

(42) axla mxolod **dğ-is-it** mžinavs.
Now only day-GEN-INS. sleep.1S.SG.PRS
"Now I only sleep during a day."

In the tagset, I will not classify genitive-nominal adverbs as adverbs, but they will be treated as nominals in genitive and double cases will be treated as suffixaufnahme cases.

Instrumental-nominal adverbs. These are nominals in instrumental case, which function as adverbs.

(43) t'iroda im **ğam-it** anano.
cry.3S.SG.IMPERF that night-INS Anano.NOM.
"Anano was crying that night".

In the tagset, I will not classify instrumental-nominal adverbs as adverbs, but they will be treated as nominals in instrumental case.

Adverbial-nominal adverbs. Most case-marked adverbs are in adverbial case, not including the derivational-suffix adverbs. Sometimes adverbs take the full adverbial case marker [**-ad]**; sometimes the adverbial case marker is partially reduced to [**-a]**; and sometimes postpositions are added. Examples of full adverbial case are: [**almacerad]** "sideway", [**uecrad]** "suddenly" etc. The reduced case marker can be seen in words such as [**nela]** "slowly", the full form of which is [**nelad]**; or [**čkara]** "quickly", the full form is [**čkarad]**.

In the tagset, I will not classify adverbial-nominal adverbs as adverbs, but they will be treated as nominals in adverbial case.

Nominative-absolutive adverbs. Nominals in nominative-absolutive case can function as adverbs, for example: [**ğame]** "night / at night". In the tagset, I will not classify nominative-absolutive nominal adverbs as adverbs, but they will be treated as nominals in nominative-absolutive case.

There is also reduplication as another way to function as an adverb: [**t'q'e-t'q'e]** "throughout forest", [**nak'uc'-nak'uc']** "by little parts/pieces"; sometimes the reduplicated forms are conjoined by the conjunction **[da]** "and", e.g.: [**fexdafex]** "step by step", [**k'valdak'val]** "following someone's steps".

In the case of reduplication, if a nominal has a zero case, it will be tagged as an adverb. Otherwise it will be tagged as a nominal.

In order to make decisions on what types of adverbs to include in the tagset, I have analysed the types of adverbs according to their function as described by Shanidze (1980, pp.587-588) and Hewitt (1995, pp.65-69). Shanidze provides the following definition for adverbs (my translation of original Georgian):

> "An adverb is a word which is uninflectable and has its own lexical meaning. It modifies the (action) verb in terms temporal and spatial relationship, or shows in what conditions the action is happening, or how often and how many times it is happening, or what is the cause and purpose the action" (Shanidze, 1980, p. 587).

The types of adverbs according to their function does not match with the provided definition. Namely, Shanidze classifies adverbs into eight types according to their function, which are as follows:

1) Adverbs of place (*adgilis*)
2) Adverbs of time (*drois*)
3) Adverbs of manner (*vitarebis*)
4) Adverbs of measure (*zoma-c'onis*)
5) Adverbs of cause (*mizezis*)
6) Adverbs of purpose (*miznis*)
7) Interrogative adverbs (*k'itxviti*)
8) Relative adverbs (*mimartebiti*)

Only five types of the above given types of adverbs will be classified as adverbs in the tagset. They are adverbs of place, adverbs of time, adverbs of manner, adverbs of measure, adverbs of time and adverbs of cause.

Adverbs of place:

(44) **ak** xom sul gazapxul-i-a
here PTCL always spring-NOM-AUX.3S.PRS
"Here is always spring".

Adverbs of time:

(45) es q'velafer-i **gušin** ar dac'q'ebula
this all-NOM yesterday not start.3S.SG.PRF
"All of these have not started yesterday."

Adverbs of cause:

(46) me rom miq'varxar, **magitom** geubnebi
I that love.1S2S.PRS that's why tell.1S.SG.2O.SG.PRS
"I love you, that's why I am telling you (this)."

The adverbs of purpose do match with the adverb of definition. They do not form a special category but represent a special meaning of a category. In particular, they are demonstrative pronouns with enclitic postpositions, which are treated as enclitics in the proposed tagset. However, these are small number of adverbs and since it easy to deal with such small number of adverbs, they will not be tagged as clitics, but they will be tagged as adverbs.

(47) **am-is-tvis** q'vela-m unda vizrunot.
this-GEN-POST all-ERG must care.1S.PL.FUT
"We must all take care of this."

The interrogative Adverb category are the adverbs that can form interrogative sentences. It might seem that it is not a valid category and can be confused with interrogative pronouns. However, the differentiating criteria for this is case inflection. The ones that inflect for case will be treated as pronouns and those that do not inflect for case will be treated as adverbs. For example, **[sad]** "where" and **[rodis]** "when" cannot inflect for case and they will be tagged as adverbs. Whereas **[sadauri]** "from where" and **[rodindeli]** "at/of/from what time" can inflect for case and they will be tagged as pronouns.

(48) šen-s korc'il-ši **rodis** dagvatrob
your.SG-DAT wedding-POST when drink.2S.SG.1O.PL.FUT
"When will you get us drunk at your wedding?"

(49) ar vici **saidan** davic'q'o
not know.1S.SG.PRS where start.1S.SG.AOR.SBJV
"I don't know where I can start from."

Hewitt (1995, pp.65-69) adds another type of adverbs to this list - **Adverbs of Negation**. There is no definition provided for this category, but a paragraph lists the adverbs of negation. In fact, one part of the list are negation particles such as [**ar]** "not" and [**ver]** "not (potential)", which are accordingly classified as particles in the tagset. The rest of the adverbs are a combination of negation particles, interrogative pronouns and adverbs. For example, [**arsad]** "nowhere", [**ar]** is the negative particle and [**sad**] "where" adverb; [**nursed]** "nowhere" (prohibitional), [**nu]** is the negative particle and [**sad]** "where" adverb. In this instance, [**arsad]** will have the same tag as [**sad]**, i.e. they will be tagged as adverbs.

Some Georgian linguists (Gogolashvili et al., 2011; Shanidze, 1980) describe adverbs as if they have some limited declension system. For example, Table 4.34 (from Gogolashvili et al., 2011) demonstrates declension of **[dğe]** "day" as a noun and adverb:

Case	**Noun**	**Adverb**
Nom-Abs.	**dğe**	**dğe**
Erg.	**dğem**	-
Acc.-Dat.	**dğes**	**dğes**
Gen.	**dğis**	**dğeis**
Inst.	**dğit**	**dğeidan**
Adv.	**dğed**	**dğemde**
Voc.	**dğeo**	-

Table 4. 34: Declension of [dğe] "day" as a noun and adverb.

I will analyse these forms as nouns with inflections that can function as adverbs. I will not treat the so called “adverbs” separately. Therefore, I will deal with these as if there were inflections of the basic noun.

4.5.1 Tags for Adverbs

In this section, I will make the decision as to which adverb sub-categories to include in the tagset. The traditional classification provided by several authors (Shanidze, 1980) is not relevant for morphosyntactic tagging as it is based on both form and function. In POS-tagging it will be a difficult task to tag adverbs according to both their form and function. I have disregarded the classification by form where nominals in dative-accusative, genitive, instrumental, adverbial and nominative-absolutive cases function as adverbs. I will treat them as nominals. However, I will introduce a different approach for nominals that have lost nominal features and are only used as an adverb. There are a few examples of this, and I will treat them as adverbs.

I have also disregarded most sub-categories of function, but adverbs of negation and Interrogative adverbs will be considered in the tagset. Therefore, there are three tags for adverbs.

Description	**TAG**	**Examples (Latin)**	**Examples (Georgian)**
General Adverb	RR	ak, amağam, cin	აქ, ამაღამ, წინ
Adverbs of Negation	RN	arsad, arasodes	არსად, არასოდეს
Interrogative Adverb	RI	rogor, rodis, sad	როგორ, როდის, სად

Table 4. 35: Tags for Adverbs.

4.6 Conjunctions (*k'avširi*)

Conjunctions (k'avširi) in Georgian can be simple (*mart'ivi*) or compound (*rtuli*) according to their form. For example, simple conjunctions are: და [**da**]"and", თუ [**tu]** "if", ან [**an]** "or" etc. Compound conjunctions are made by joining two or more words (particles/pronouns/adverbs), e.g.: ვიდრე [**vidre]** (vid+re) "while", თუნდაც [**tundac]** (tu unda+c) "even if", თორემ [**torem**](tu ara+m) "otherwise" etc.

According to their function, conjunctions can be **Coordinating** (*maertebeli*) or **Subordinating** (*makvemdebarebeli*).

There are four types of coordinating conjunctions in Georgian. They are: 1) **Conjoining** (*majgupebeli*), 2) **Disjunctive** (*macalk'evebeli*), 3) **Adversative** (*map'irisp'irebeli*) and 4) **Illative / Resultative** (*maigivebeli*).

There are six types of subordinating conjunctions in Georgian: 1) **Locative** (*adgilis*), 2) **Temporal** (*drois*), 3) **Causal** (*mizezis*); 4) **Purposive** (*miznis*); 5) **Concessional** (*datmobis*) and 6) **Conditional** (*p'irobis*). Adverbs and relative pronouns ([**rodesac], [roca]** "when") often function as subordinating conjunctions.

4.6.1 Coordinating Conjunctions

Conjoining conjunctions. The following words can be conjoining conjunctions: და **[da]** "and", თუ [**tu]** "or". Conjoining conjunctions connect words, phrases, and clauses.

[da] “and” connects words, phrases and sentences:

(50) mze **da** mtvare
Sun.NOM **and** moon.NOM
“The Sun and the moon.”

(51) me q'ava davlie **da** t'elevizor-s vuq'ure
I coffee.NOM drink.1S.SG.AOR **and** TV-DAT watch. 1S.SG.AOR
“I drank coffee and watched the TV.”

The single word თუ **[tu]** meaning “or” can have several functions within subordinating conjunction. It can be categorized as a disjunctive conjunction or conditional conjunction. Sometimes it can have the same function as the conjoining conjunction [**da]** “and”. For example:

(52) q'vela movida: k'ac-i **tu** kal-i,
all.NOM come.3S.SG.AOR man-NOM **or** woman-NOM
“All came: man and woman.”

Disjunctive conjunctions. They are coordinating conjunctions that separate two or more mutually exclusive options presented in a sentence. The set of disjunctive conjunctions in Georgian contains the following: თუ [**tu]** “or”; ან [**an]** “or” and ხან [**xan]** “sometimes”.

(53) ğvino mogartva, **tu** c'q'al-i?
wine.NOM offer.1S.SG.2O.SG.COND **or** water-NOM
“What can I offer you, wine or water?”

(54) irc'mune om-i **an** mšvidoba
believe.2S.SG.AOR war-NOM **or** peace.NOM
“Believe in war or in peace.”

Adversative conjunctions. They are a type of coordinating conjunction which expresses comparisons and contrasts. Sometimes it is also known as a contrastive

conjunction. The set of adversative conjunctions in Georgian contains: **მაგრამ** [**magram]**, **მარა** [**mara]**, "but"; **ხოლო** [**xolo]** "and", **კი** [**k'i]** "but, however", and **თორემ** [**torem]** "otherwise".

(55) bevr-i vecade, **magram** veraper-i ševcvale.
lot-NOM try.1S.SG.AOR **but** nothing-NOM change. 1S.SG.AOR
"I tried a lot but I couldn't change anything."

(56) c'q'al-i gtxove, šen **k'i** ǧvino mogakvs.
water-NOM ask.1S.SG.2O.SG.AOR you **and** wine.NOM bring.2S.SG.PRS
"I asked you to give me some water and you are bringing me wine."

Illative conjunctions. They are coordinating conjunctions (also known as final conjunctions) that introduce clauses or phrases that draw inferences or conclusions from earlier ones. The set of Illative conjunctions contains: **ანუ** [**anu]** "thus, so", **ესე იგი** [**ese igi]**, **მაშასადამე** [**mašasadame]** "therefore".

(57) kartvel-i var da **mašasadame**, martlmadidebel-i.
Georgian-NOM be. 1S.SG.PRS and **therefore** orthodox-NOM
"I am Georgian, therefore, I am an orthodox Christian."

4.6.2 Subordinating Conjunctions

Locative conjunctions. They express a location relative to a main clause. Some examples of Locative conjunctions are: **სადაც** [**sadac]** "where", **საითაც** [**saitac]** "where to", **საიდანაც** [**saidanac]** "where from" etc. These words are interrogative adverbs following enclitic particles and thus, will be tagged as adverbs.

(58) šen-tan viknebi, **sadac** ar unda viq'o.
you-POST be.1S.SG.FUT **where** not shall be.1S.SG.PRS.SBJV
"I will be with you wherever I am."

(59) kargad vicodi, **saitac** mivdiodi.
Well know.1S.SG.IMPERF **where** go.1S.SG.IMPERF
"I knew well, where I was going."

Temporal conjunctions. They are used to express relations in time. Some Temporal conjunctions are: **როდესაც** [**rodesac]**, **როცა** [**roca], რაც [rac], რო [ro]** "when" and the general subordinator [**rom]** used in this meaning.

(60) dil-it **roca** iğviže**b**
morning-INS **when** wake.2S.SG.PRS
"When you wake up in the morning"

Causal conjunctions. They introduce a cause or result. Some examples of the Causal conjunctions are: **ვინაიდან** [**vinaidan]** "whilst", **რადგან** [**radgan], რადგანაც [radganac], რაკი [raki], რახან [raxan], ამიტომ [amit'om], ამიტომაც [amit'omac]**, აქაოდა [**akaoda], მაგიტომ [magit'om], მიტომ [mit'om]** "as, because".

(61) **raxan** davic'q'et, gavagrželot
since start.1S.PL.AOR. continue.1S.PL.COND
"Since we have already started (this), let us continue".

Purposive Conjunctions. As the term suggests, they indicate the purpose, "why" something has happened or has been done etc. Some Purposive Conjunctions are: **რომ** [**rom]**, **რათა** [**rata], ვითარმედ [vitarmed], რამეთუ [rametu]** "in order that", "that".

(62) dagežeb **rata** p'at'ieba gtxovo
look.1S.SG.2O.SG. **in order to** forgivness.NOM ask.1S.SG.2O.SG
“I am looking for you, in order to ask for your forgiveness”.

Concessional conjunctions. They express a fact or supposition in spite of which the assertion in the main clause is made. Some Concessional conjunctions in Georgian are: **თუმც [tumc], თუმცა [tumca], თუმცაღა [tumcağa], თუმცაღაკი [tumcağaki]** “although”, **თუნდ** [**tund], თუნდაც [tundac]** and **რომც** [**romc]** “even if”; **ოღონც [oğonc], ოღონდ [oğond], ოღონდაც [oğondac]** “only, except that”.

(63) bevr-sa-c it'irebs, **tumca** ar šemecodeba
Lot-DAT-PART cry.3S.SG.FUT **although** not feel-sorry.1S.SG.FUT
“(S/he) will cry a lot too, but I will not feel sorry for (him/her).

(64) c'aik'itxe, **oğond** aravi-s utxra
read.2S.SG.AOR **but** anyone-DAT tell.2S.SG.COND
“Read (this), but don’t tell anyone”.

Conditional conjunctions. They are dependent clauses which describe the conditions under which something may or may not happen. Some Conditional conjunctions in Georgian are: **თუ [tu]**, **თუკი** [**tuk'i]**, **უკეთუ** [**uk'etu]** “if” etc.

(65) mogiqvebit, **tuk'i** visurveb
tell.1S.SG.2O.PL.FUT. **if** wish.1S.SG.FUT
“I will tell you, if I wish to”.

4.6.3 Tags for Conjunctions

For the purposes of POS-tagging, I have disregarded the ten sub-categories of coordinating and subordinating conjunctions. The decision regarding what sub-categories should be included in the tagset has instead been based on the syntactic

behaviour of conjunctions. As discussed above, Conjunctions in Georgian have two main functions: to join two or more words/phrases and/or independent clauses, and to join one or more subordinate sentences with the main (independent) clause. This gives two sub-categories: coordinating and subordinating conjunctions. The two tags are as follows:

Description	**Tag**	**Examples (Latin)**	**Examples (Georgian)**
Coordinating Conjunction Simple	CC	da, magram	და, მაგრამ
Subordinating Conjunction	CS	oğond, rom	ოღონდ, რომ

Table 4. 36: Tags for Conjunctions.

4.7 Particles (*nacilak'i*)

In Shanidze's grammar (1980, pp. 607-616), the term "particle" is used for many different elements that do not necessarily form a coherent category. Some particles are used only with verbs; they occur before the verb, i.e. precede a verb, for example: **არ** [**ar], ვერ [ver], ნუ [nu]** "not". Some particles are only used with nominals (including noun, pronoun, adjective, numeral), such as **-ვე** [**-ve]**, **-ც(ა) [-c(a)**] "too" etc. They are merged with a word and cliticised.

(66) gogo-c
girl.NOM-PTCL
"A girl too/ even a girl".

Some particles are used with both nominals and verbs, such as **ო** [**-o]**, for example:

(67) mitxra gaicina lamaz-ma-o
tell.3S.SG.1O.SG.AOR smile.3S.SG.AOR beautiful-ERG-PTCL
"(S/he) told me that a beautiful (one) smiled".

(68) gepicebi araper-i utkvams-o
Swear.1S.SG.2S.SG.PRS nothing-NOM say-3S.SG.RES.PTCL
"(S/he said) I swear that s/he has said nothing".

Some particles can appear separately in the sentence as an independent word, such as **ხოლმე** [**xolme]** "usually", **ნუ** [**nu]** "don't". However, some particles are written with a hyphen joining them to the word they accompany, for example, **მოვალ-მეთქი** [moval-**metki]** "I said I will come". Some particles are cliticised, such as **-ცა** [**-c(a)]** "too"; **-ღა [-ğa]** "only" etc.

The functions and usage of particles are not well classified in Georgian. They are confusingly described and very often there is no clear difference between proposed

particle types/classes. There are cases where other parts-of-speech are discussed as particles; for example, [**igi], [ege], [ese]** are described as relative particles by Shanidze (1980, p. 609) and in the next chapter of the same book (Shanidze, 1980, pp. 616-621), they are described as articles. [**igi], [ege]** and **[ese]** can be inflected and their function in Old Georgian was to express definiteness and indefiniteness. Thus, they should be discussed within the article category. However, articles do not appear in Modern Georgian at all. There are other major and minor problems in description of Georgian grammar that I will not discuss here, as the main aim of my research is defining the tagset for Georgian and POS-tagging.

4.7.1 Tags for Particles

Before deciding which sub-categories to distinguish in the tagset, I will characterise the category of particle itself. Particles in Georgian have no lexical meaning and are uninflectable. Some particles are cliticised; others are used as independent words. I have disregarded most of the fine-grained distinctions among different types of particles described by several authors (Shanidze, 1980) as these distinctions are not relevant for the purpose of tagset design. The subcategory distinctions that I will be using are mapped according to syntactic behaviour. For instance, I will introduce a separate category if the particles in it behave in a specific way syntactically and will not split categories if there is no syntactic difference. This allows to make reference to particle categories when doing contextual disambiguation. Taking this into consideration, I have outlined the following categories for particles:

Interrogative Particles. There are four interrogative particles: [**gana], [nutu], [xom], [tu]**. Interrogative particles convert a statement into a rhetorical or yes-no question.

They are all used as separate words and are never cliticised. They can appear at the beginning, middle, or end of a sentence.

(69) **gana** p'atara var?
PTCL little.NOM be.1S.SG.PRS.
"Am I little?"

This question can be interpreted pragmatically as a rhetorical question: the meaning in context is "do you really think that I am little?".

(70) **nutu** martla dagĉ'irdi?
PTCL really need.2S.SG.1O.SG.AOR
"Do you really need me?"

(71) šen **xom** mimixvdi?
you PTCL understand.2S.SG.1O.SG.AOR
"You understand, don't you?"

(72) sik'vdil-is šemdeg **tu** arsebobs sicocxle?
death-GEN after PTCL exist.3S.SG.PRS life.NOM?
"Is there life after death?"

Speech Particles. "Speech Particle" is a term used by Hewitt (1995, p.89) and is a literal translation from the Georgian (*met'q'velebis nac'ilak'i*). I will instead use the term **Quotative Particle** as its main function is to mark a stretch of quoted speech within which the verbal tense and person/number agreement of the original utterance is preserved. The four Quotative particles are **მეთქი [metki], თქვა [tkva], თქო [tko]** and **-ო** [**-o]**. I also consider two rather informal variants of [**metki]** used in the Imeretian and Javakhian dialects: [**mevtkvi]** and [**metkvi].** [**metki], [mevtkvi]** and **[metkvi]** are used when a 1st person singular speaker repeats his/her own words, i.e. when the embedded clause subject is 1st person. [**tko]** and [**tkva]** (in the Imeretian

dialect) are used when a 1st person speaker addresses a 2nd person to pass his/her (1st person's) words to a 3rd person. [**-o]** is used when a 3rd person (either singular or plural) is the speaker. Unlike other quotative particles, [**-o]** is always encliticised to the main verb of the quoted material. As for [**metki]**, **[mevtkvi]**, **[metkvi]**, [**tko]** and [**tkva]**, they may be written with or without a hyphen, i.e. they may be cliticised but may also be used as separate words.

(73) vutxari gagik'eteb **metki**
say.1S.SG.AOR do.1S.SG.2O.SG.FUT PTCL
"I said to him/her: "I will do that for you."

(74) film-is gmir-s magoneb **tko**
film-GEN hero-DAT remind.1S.SG.2O.SG.PRS PTCL
"I said to him/her: you remind me of a movie hero."

(75) col-ad kartvel-i mindoda-**o**, mitxra.
wife-ADV Geo-NOM want.1S.SG.IMPERF-PTCL tell.3S.SG.1O.SG.AOR
"S/he told me: I always wanted a Georgian as a wife."

Prohibitive Particles. I will not use this term because not all the particles in question are specifically "prohibitive" in function; I will use the term **Particles of Negation** instead as they indicate negation including denial, refusal, or prohibition. The set of negation particles contains: [**ar], [ara], [ver], [vera], [nu],** and **[rodi]** "not, cannot". Also, by adding the [**-c(a**)] and [**-ğa]** particles, another set of negation particles are formed, such as the following: [**agar], [veğar], [veğarc], [nuğar], [nuğara], [arc], [arca], [nurc], [nurca], [ağarc], [araperic], [verc], [verca]**. They are all used as separate words and are never cliticised. They can appear at the beginning, middle, or end of a sentence. Taking into account the wider context of negation in Georgian, negation particles (alongside with the adverbs of negation and negative pronouns) are the primary way that sentences get negated in Georgian.

(76) gza uk've **agar** arsebobs
way.NOM already not exist.3S.SG.PRS
"Now, there is no way."

(77) gušin **ver** movicale
yesterday cannot free.1S.SG.AOR
"I couldn't get free yesterday".

(78) saertod **ar** maintereseb es politika
at all not interest.1S.SG.PRS this politics.NOM
"I am not interested in this politics at all."

(79) žalian gtxov, uar-s **nu** gvet'q'vi
very ask.1S.SG.2O.SG.PRS no-DAT don't say.2S.SG.1O.PL.FUT
"I am begging you very much, don't say no to us."

I have introduced additional three sub-categories for particles that are not covered within the traditional list. They are modal, nominal and general:

Modal Particles. In general, modality in Georgian is expressed by modal verbs and other words such as particles and adverbs that have modal functions (Sharashenidze, 2014, pp.80-90). For the purposes of POS-tagging, within modal particles, I have grouped those particles that indicate modality and are originally verbs or derived from verbs. Modal particles are usually used immediately before verbs, but also can appear after verbs or even can be split by some other word. Modal particles are uninflectable and do not cliticise. The set of modal particles contains:

1. **[unda]** "must" and its dialect variants [**un]** and [**una]**, which are, in particular, used in the Kakhetian, Meskhian and Javakhian dialects
2. **[šeižleba]**, **[šesažloa]** "can/may be" and the dialect variant [**šeileba]**, mainly used in the Gurian dialect, but also quite frequently used in spoken standard Georgian;

3. **[albat]** “may be”
4. **[gind], [ginda], [gindac], [gindat]** “want” and the dialect variants [**ginc]**, and [**gina]** used in Kiziqian which is a sub-dialect of Kakhetian;
5. **[vinicobaa]** “in case”
6. **[vinžlo]** “I bet”
7. **[mgoni], [mgonia]** “I think / suppose”
8. **[egeb]**, [**egeba]**, [**egebis]** “may be”; I also consider the dialect variants [**ageb]**, [**ageba]**, and [**agebis]**
9. **[ikneb], [ikneba]** “maybe”
10. **[tugind], [tuginda]** “let’s say”
11. **[lamis]** “almost”
12. **[titkmis]** “almost”
13. **[titkos]** “as if”
14. **[ragind]** “no matter…”
15. **[tund], [tunda], [tundac]** “even if”.

(80) **unda** icode ena-c
must know.2S.SG.AOR.SBJV language.NOM-PTCL
“You must know the language as well”

(81) tumca **šeižleba** vcdebode
however maybe mistake.1S.SG.PRS.SBJV
“However, I may be mistaken/wrong.”

Nominal Particles. This sub-category includes particles that are used only with nominals (including nouns, adjectives, pronouns and numerals) or with a particle itself (thus several particles conjoined). The three nominal particles are **-ც** [**-c(a)]**, **-ღ** [**-ğa]**, **-ვე** [**-ve]**. They are always post-nominal and enclitic. [**-c(a)]**, [**-ğa]**, [**-ve]** are

always the last enclitic on the word they are associated with (Shanidze, 1980, pp.71-72).

Nominal particles will be treated in POS-tagging as follows: 1) as enclitics when used on nouns, adjectives, pronouns and numerals and 2) as part of a word / not separated when used on other particles.

General Particles. In this sub-category, I classify the particles that may be used with both nominals and verbs. General particles are separate words; they are never cliticised and can appear before or after the word with which they are associated. General particles do not have a single function; they can, for instance, express a wish or a desire.

(82) gvianobamde vusmen **xolme** radio-s
late listen.1S.SG.PRS usually radio-DAT
"Usually, I listen to the radio till late."

(83) **maš** čven rağa vknat
so we what do.1S.PL.AOR.SBJV
"So, what else we should do."

(84) šedareba ar momec'ona **oğond**
comparison.NOM not like.1S.SG.AOR PTCL
"But I didn't like the comparison".

Therefore, the attribute values for particles I have classified are the following: Interrogative particles, Quotative particles, Particles of negation, Modal particles, Nominal particles and General particles.

This gives six tags for particles:

Description	Tag	Examples (Latin)	Examples (Georgian)
General Particle	XX	netav, diax	ნეტავ, დიახ
Interrogative Particle	XI	gana, xom	განა, ხომ
Quotative Particle	XQ	metki, tko	მეთქი, თქო
Negative Particle	XN	ar, veğar	არ, ვეღარ
Modal Particle	XM	vinžlo, titkmis	ვინძლო, თითქმის
Nominal Particle	XO	-ca, -ve	-ცა, -ვე

Table 4. 37: POS-tags for Particles.

4.8 Interjections (*šorisdebuli*)

Unlike other parts-of-speech, interjections are not part of the grammar of the clause. They usually occur at the beginning of a sentence or between clauses. However, they can also occur sentence-finally.

(85) **auu**, daviğale
Oh tire.1S.SG.AOR
"Oh, I got tired."

(86) es mizani ganvaxorciele, **vaša**
this aim-NOM fulfil.1S.SG.AOR **yay**
"I have fulfilled this aim, Yay!"

There are different thematic groups of interjections (denoting surprise, fear, displeasure etc.) described by Gachechiladze (1979, pp.138-224), Shanidze (1980, pp.621-628), Hewitt (1995, pp.99-100) and Peikrishvili (2010, pp.217-263). I will not analyse their functions and meanings here or introduce any subcategories in the tagset. Thus, there is a single tag for Interjection.

Description	**TAG**	**Examples (Latin)**	**Examples (Georgian)**
Interjection	UU	uime, eriha	უიმე, ერიჰა

Table 4. 38: Tags for Interjections.

4.9 Postpositions (*tandebuli*)

In Georgian, postpositions occur only with nominals selecting a particular case, for example, **-თან [tan]** "at" and **-ზე [ze]** "on" selects/governs dative-accusative case. In modern Georgian, postpositions may govern the following five cases: nominative-absolutive, dative-accusative, genitive, instrumental and adverbial cases.

4.9.1 Postpositions governing nominative-absolutive case

The **-ვით** [**-vit]** postposition usually governs dative-accusative Case. However, when the nominal root is consonant-final, the [**-vit]** "like" postposition governs nominative-absolutive case, but otherwise it governs dative-accusative case.

(87) p'irvel t'aks-s mxec-i-**vit** davet'ak'e
first.Ø taxi-DAT beast-NOM-POST attack.1S.SG.AOR
"I have attacked (grabbed) the first taxi like a beast."

(88) q'vela bat-i-**vit** iq'o dabneuli
all.NOM goose-NOM-POST be.3S.SG.AOR confused-NOM
"All were confused like a goose."

4.9.2 Postpositions governing dative-accusative case

The set of postpositions governing dative-accusative case contains: **-ვით** [**-vit]** "like", **-თან** [**-tan]** "at", **-ზე** [**-ze]** "on", **-ში** [**-ši]** "in", **შორის** [**šoris]** "between/among", **შუა** [**šua]** "between". [**-vit]**, [**-tan]**, [**-ze]**, [**-ši]** are cliticised to the preceding nominals; [**šoris]** and [**šua]** are used as separate words.

(89) čven davkarget k'avšir-i mic'a-s-**tan**
we lose.1S.PL.AOR connection-NOM soil-DAT-POST
"We lost the connection to the soil."

(90) bednier vasrsk'vlav-**ze** xart dababedul-i
Lukcy.Ø star-POST be.2S.PL.PRS born-NOM
"You are born under a lucky star."

4.9.3 Postpositions governing Genitive Case

The set of postpositions governing genitive case are: **-თვის** [**-tvis**] "for"; **-გან** [**-gan]** "from"; **-კენ** [**-k'en]** "to, towards", **-ებრ** [**-ebr]** "like"; **-თანავე** [**-tanave]** "as, immediately upon"; **-დამი** [**-dami]** "to", **-დმი** [**-dmi]** "to. They are all cliticised with nominals. Other postpositions governing genitive case which appear as separate words include: **მიერ** [**mier]** "by, with", **გამო** [**gamo]** "because", **მიმართ** [**mimart]** etc.

The [-a] affix may be attached to the cliticised postpositions governing the genitive case. This will be considered as allomorphy, for example: [**tvis]** and **[tvisa]** and [**k'en]** and **[k'ena]** and will receive the same tag. As discussed in section 4.1.2 of this chapter, there both forms coexist and are correct. For example:

(91) švil-i ded-is-**tvis** q'velaper-i-a
child-NOM mother-GEN-POST everything-NOM-COP
"A child is everything for a mother."

(92) sik'vdil-i mova da c-is-**k'en** aaxedebs
death-NOM come.3S.SG.FUT and sky-GEN-POST look.3S.SG.FUT
"The death will come and make him/her look up to the sky."

4.9.4 Postpositions governing instrumental case

The set of postpositions governing instrumental case are as follows: **-დან** [**-dan], -იდან [-idan], -დამ [-dam], -იდამ [-idam]** "from" and **-ურთ** [**-urt]** "(together) with".

(93) t'ekst'-eb-i targman-it-**urt** gamosca
text-PL-NOM translation-INS-POST publish.3S.SG.AOR
"(S/he) published texts with translations."

(94) k'viradğe-s saxli-**dan** gasvla ar miq'vars
Sunday-DAT house.INS-POST go.NOM not love.1S.SG.PRS
"I don't like going out on Sundays".

4.9.5 Postpositions governing Adverbial Case

The only postpositions governing adverbial case are **-მდე** [**-mde]** and **-მდის** [**-mdis]** "to", with its dialect variants **-მდინ** [**-mdin]**, **-მდისი** [**-mdisi]**, **-მდისინ** [**-mdisin]**.

(95) dili-dan sağamo-**mde** miq'ureb
morning-POST evening-POST look.2S.SG.1O.SG.PRS
"You look at me from morning till evening."

4.9.6 Tags for Postpositions

As described above, some postpositions are cliticised with nominals, and a smaller number appear as independent words. In total, there are about 36 postpositions, out of which 21 are always cliticised with nominals and 12 appear as independent words; 3 postpositions can be either enclitic or used as independent words. The enclitic postpositions need to be tagged as their own tokens, separately to the nominals. This will make the analysis of clitic and non-clitic postpositions more broadly equivalent.

It is not necessary to divide the postpositions up according to what case they govern. This will be obvious from the preceding case marker. In POS tagging, both types of postpositions will receive a single tag.

Thus, gives a single tag for Postpositions:

Description	**TAG**	**Examples (Latin)**	**Examples (Georgian)**
Postposition	II	mier, gamo, mimart, -ebr	მიერ, გამო, მიმართ, -ებრ

Table 4. 39: Tags for Postpositions.

4.10 Verbs (*zmna*)

In this section, I will describe verbs, one of the most complex parts-of-speech in the Georgian language. I will focus on the categories that are marked in morphology, i.e. the categories that are relevant for POS-tagging purposes. Therefore, I will describe the grammatical categories that I think should be included in the tagset.

In the traditional Georgian grammar of Shanidze (1980), morphological categories are not clearly defined. However, I will be only commenting on such issues where they are relevant for POS-tagging as this is the main aim of my PhD research.

The Georgian traditional grammars (Shanidze, 1980, pp.163-530; Gogolashvili, 2011, pp.266-634) describe verbs according to grammatical and derivational categories as follows:

Grammatical categories:

1) **arguments** (*p'iri*)
2) **Number of argument agreements** (*ricxvi*)
3) **Screeves** (*mc'k'rivi*) - **Tense**, **Mood**, **Iteration**, **Act**, **Accompaniment**.

Derivational categories:

1) **Direction** (*gezi*)
2) **Orientation** (*orient'acia*)
3) **Aspect** (*aspek't'i*)
4) **Voice** (*gvari*)
5) **Version** (*qceva*)
6) **Contact** (*k'ontakt'i*).

In the defined annotation scheme, I will disregard this traditional classification and rather focus on the following grammatical categories for verbs:

1) Person of Agreement (of Subject and Object)
2) Number of argument agreement
3) Screeves – covering Tense, Aspect and Mood.

4.10.1 Arguments and Number of Argument Agreement

The Georgian verb can have up to three arguments, but only two arguments can be morphologically marked at the same time: 1) Subject (Agent) and either Direct Object (Patient) or Indirect (Oblique) Object. There is a very simple rule to find out how many arguments the verb has. The verb is analysed without any context and it can give us the information about the number of arguments. As mentioned above, a verb can have up to three arguments, but morphologically only two arguments are marked (see Example 3 in section 2.1.12 of chapter 2).

There are two sets of markers in Georgian, for Subject and Object. I will introduce two terms: 1) **v-agreement** for subject markers and 2) **m-agreement** for object markers. The [**v-]** is usually subject marker (1st person of agreement), but it can be an object marker in certain types of verbs. The [**m-]** is usually an object marker (1st person agreement), but it can mark subject as well depending on the type of verb. This is a result of split ergative alignment. As a simplified approach in POS-tagging, **v-**agreement will be treated as subject and **m**-agreement as object.

Thus, taking into account this approach, the person of subject argument markers in Georgian are:

S/O	Subject and Object Argument agreement Markers	
	Singular	**Plural**
S1	v-	v-…-t
O1	m-	gv-
S2	Ø-, x-, h-, s-	Ø-, x-, h-, s-…-t
O2	g-	g-…-t
S3	-s, -a, -o	-en, -an, -nen, -n, -es
O3	h-, s-	h-, s-…(-t), Ø-

Table 4. 40: Subject and object Argument agreement Markers.

As illustrated in the Table 4.40 above, the person agreement markers are mostly prefixal and the number agreement markers are suffixal. However, there are a few exceptions regarding the 3rd person of subject (**S_3**): some suffixes ([-**en], [-an], [-n], [-es]**) can mark both, the person of agreement and its number.

The table below shows an example of the distribution of the arguments, as well as the number of arguments in the screeves. I am using the verb [**goraoba]** "to roll" as an example in three different voices: 1) active, intransitive: [**goravs]** "s/he rolls" (**S_3**); 2) active, transitive: [**agorebs]** "s/he rolls it" (**S_3O_3**); and 3) passive (reflexive), intransitive: [**gordeba]** "s/he rolls himself/herself" (**S_3**).

I Series			
Present Group			
	Present	**Imperfect**	**Present Subjunctive**
S1	**v**gorav, **v**agoreb, **v**gordebi	**v**goravdi, **v**agorebdi, **v**gordebodi	**v**goravde, **v**agorebde, **v**gordebode
S2	gorav, agoreb, gordebi	goravdi, agorebdi, gordebodi	goravde, agorebde, gordebode
S3	gorav**s**, agoreb**s**, gordeb**a**	goravd**a**, agorebd**a**, gordebod**a**	goravde**s**, agorebde**s**, gordebode**s**
Future Group			
	Future	**Conditional**	**Future Subjunctive**
S1	**v**igoreb, ga**v**agoreb, ga**v**gordebi	**v**igorebdi, ga**v**agorebdi, ga**v**gordebodi	**v**igorebde, ga**v**agorebde, ga**v**gordebode
S2	igoreb, gaagoreb, gagordebi	igorebdi, gaagorebdi, gagordebodi	igorebde, gaagorebde, gagordebode
S3	igoreb**s**, gaagoreb**s**, gagordeb**a**	igorebd**a**, gaagorebd**a**, gagordebod**a**	igorebd**a**, gaagorebde**s**, gagordebode**s**
II Series (Aorist)			
	Aorist		**Aorist Subjunctive**
S1	**v**igore, ga**v**agore, ga**v**gordi		**v**igoro, ga**v**agoro, ga**v**gorde
S2	igore, gaagore, gagordi		igoro, gaagoro, gagorde
S3	igor**a**, gaagor**a**, gagord**a**		igoro**s**, gaagoro**s**, gagorde**s**
III Series (Perfect)			
	I Resultative	**II Resultative**	**III Subjunctive**
S1	**m**igoria, ga**m**igorebia, ga**v**gorebulvar	**m**egora, ga**m**egorebina, ga**v**gorebuliqav(i)	**m**egoros, ga**m**egorebinos, ga**v**gorebuliqo
S2	gi**g**oria, ga**g**igorebia, gagorebulxar	ge**g**ora, ga**g**egorebina, gagorebuliqav(i)	ge**g**oros, ga**g**egorebinos, gagorebuliqo
S3	ugori**a**, gaugorebi**a**, gagorebul**a**	egor**a**, gaegorebin**a**, gagorebuliq**o**	egoro**s**, gaegorebino**s**, gagorebuliqo**s**

Table 4. 41: Argument Agreement across the screeve paradigm (from Melikishvili, 2014, p.101)

There are several possible combinations of person of agreement. To show all possible combinations of subject and object agreement, I will illustrate three examples[33] as follows:

Example 1: ვეზრდები [**vezrdebi]** "I am being raised (for him/her)";

Example 2: ვზრდი [**vzrdi**] "I raise him/her";

Example 3: ვუზრდი [**vuzrdi**] "I raise him/her for him/her".

The first example in Table 4.42 below is reflexive, transitive verb, and object oblique, and marked applicative (the beneficiary argument is expressed as an object, so the main object is the beneficiary, not the patient). The second example in Table 4.43 is an active, transitive verb. Whereas the third example in Table 4.44 is also an active, transitive verb and marked applicative. The beneficiary argument is expressed as an object.

Numbers from 1 to 3 represent the first, second and 3rd persons accordingly. Whereas the letter **S** here expresses a Subject and **O** an Object. There is no a special character for expressing singular forms. However, the plural forms are marked by letter **p**. For example, $\mathbf{S_2O_1^P}$ means that subject is the second person singular and object is the first person plural. The presence of syncretism is highlighted by using the same colour highlight.

[33] These examples are taken from Melikishvili (2014, pp.102)

O / S	O_1	O_2	O_3	$O_1{}^P$	$O_2{}^P$	$O_3{}^P$
S_1	-	**g**ezrdebi S_1O_2	**v**ezrdebi S_1O_3	-	**g**ezrdebi**t** $S_1O_2{}^P$	**v**ezrdebi $S_1O_3{}^P$
S_2	**m**ezrdebi S_2O_1	-	ezrdebi S_2O_3	**gv**ezrdebi $S_2O_1{}^P$	-	ezrdebi $S_2O_3{}^P$
S_3	**m**ezrdeb**a** S_3O_1	**g**ezrdeb**a** S_3O_2	ezrdeb**a** S_3O_3	**gv**ezrdeb**a** $S_3O_1{}^P$	**g**ezrdeba**t** $S_3O_2{}^P$	ezrdeb**a** $S_3O_3{}^P$
$S_1{}^P$	-	**g**ezrdebi**t** $S_1{}^PO_2$	**v**ezrdebi**t** $S_1{}^PO_3$	-	**g**ezrdebi**t** $S_1{}^PO_2{}^P$	**v**ezrdebi**t** $S_1{}^PO_3{}^P$
$S_2{}^P$	**m**ezrdebi**t** $S_2{}^PO_1$	-	ezrdebi**t** $S_2{}^PO_3$	**gv**ezrdebi**t** $S_2{}^PO_1{}^P$	-	ezrdebit $S_2{}^PO_3{}^P$
$S_3{}^P$	**m**ezrdebia**n** $S_3{}^PO_1$	**g**ezrdebi**an** $S_3{}^PO_2$	ezrdebia**n** $S_3{}^PO_3$	**gv**ezrdebi**an** $S_3{}^PO_1{}^P$	**g**ezrdebi**an** $S_3{}^PO_2{}^P$	ezrdebi**an** $S_3{}^PO_3{}^P$

Table 4. 42: Subject and object combinations, Example 1.

O / S	O_1	O_2	O_3	$O_1{}^P$	$O_2{}^P$	$O_3{}^P$
S_1	-	**g**zrdi S_1O_2	**v**zrdi S_1O_3	-	**g**zrdi**t** $S_1O_2{}^P$	**v**zrdi $S_1O_3{}^P$
S_2	**m**zrdi S_2O_1	-	zrdi S_2O_3	**gv**zrdi $S_2O_1{}^P$	-	zrdi $S_2O_3{}^P$
S_3	**m**zrdi**s** S_3O_1	**g**zrdi**s** S_3O_2	zrdi**s** S_3O_3	**gv**zrdi**s** $S_3O_1{}^P$	**g**zrdi**t** $S_3O_2{}^P$	zrdi**s** $S_3O_3{}^P$
$S_1{}^P$	-	**g**zrdi**t** $S_1{}^PO_2$	**v**zdri**t** $S_1{}^PO_3$	-	**g**zrdi**t** $S_1{}^PO_2{}^P$	**v**zrdi**t** $S_1{}^PO_3{}^P$
$S_2{}^P$	**m**zrdi**t** $S_2{}^PO_1$	-	zrdi**t** $S_2{}^PO_3$	**gv**zrdi**t** $S_2{}^PO_1{}^P$	-	zrdit $S_2{}^PO_3{}^P$
$S_3{}^P$	**m**zrdia**n** $S_3{}^PO_1$	**g**zrdi**an** $S_3{}^PO_2$	zrdia**n** $S_3{}^PO_3$	**gv**zrdi**an** $S_3{}^PO_1{}^P$	**g**zrdi**an** $S_3{}^PO_2{}^P$	zrdi**an** $S_3{}^PO_3{}^P$

Table 4. 43: Subject and object combinations, Example 2.

O / S	O_1	O_2	O_3	$O_1{}^P$	$O_2{}^P$	$O_3{}^P$
S_1	-	**g**izrdi S_1O_2	**v**uzrdi S_1O_3	-	**g**izrdi**t** $S_1O_2{}^P$	**v**uzrdi $S_1O_3{}^P$
S_2	**m**izrdi S_2O_1	-	uzrdi S_2O_3	**gv**izrdi $S_2O_1{}^P$	-	uzrdi $S_2O_3{}^P$
S_3	**m**izrdi**s** S_3O_1	**g**izrdi**s** S_3O_2	uzrdi**s** S_3O_3	**gv**izrdi**s** $S_3O_1{}^P$	**g**izrdi**t** $S_3O_2{}^P$	uzrdi**s** $S_3O_3{}^P$
$S_1{}^P$	-	**g**izrdi**t** $S_1{}^PO_2$	**v**uzdri**t** $S_1{}^PO_3$	-	**g**izrdi**t** $S_1{}^PO_2{}^P$	**v**uzrdi**t** $S_1{}^PO_3{}^P$
$S_2{}^P$	**m**izrdi**t** $S_2{}^PO_1$	-	uzrdi**t** $S_2{}^PO_3$	**gv**izrdi**t** $S_2{}^PO_1{}^P$	-	uzrdit $S_2{}^PO_3{}^P$
$S_3{}^P$	**m**izrdia**n** $S_3{}^PO_1$	**g**izrdi**an** $S_3{}^PO_2$	uzrdia**n** $S_3{}^PO_3$	**gv**izrdi**an** $S_3{}^PO_1{}^P$	**g**izrdi**an** $S_3{}^PO_2{}^P$	uzrdi**an** $S_3{}^PO_3{}^P$

Table 4. 44: Subject and object combinations, Example 3.

As discussed above, a single verb can have up to three arguments, but only two are marked in the agreement. In POS tagging, I will consider the two arguments that are marked in morphology. I will also make a decision which argument agreement combinations are relevant for POS-tagging. For these purposes, I have classified the Argument Combinations as follows:

1. **Impossible Combinations** - combinations that never occur. These combinations do not exist and thus cannot be considered in the Tagset.

2. **Possible Combinations**

   a) **Unique Combinations** - combinations that have unique forms (markers) and more or less are unambiguous;

   b) **Ambiguous Combinations** – when a single form expresses two or more different agreement combinations.

There are eight **impossible combinations**:

1) **S_1O_1**
2) **$S_1O_1^P$**
3) **S_2O_2**
4) **$S_2O_2^P$**
5) **$S_1^PO_1$**
6) **$S_1^PO_1^P$**
7) **$S_2^PO_2$**
8) **$S_2^PO_2^P$**.

Overall, there are **28 possible combinations** as follows:

S_2O_1	**S_1O_3**	**$S_2^PO_1^P$**	**$S_3O_3^P$**
S_3O_1	**S_2O_3**	**$S_3^PO_1^P$**	**$S_1^PO_3^P$**
$S_2^PO_1$	**S_3O_3**	**$S_1O_2^P$**	**$S_2^PO_3^P$**
$S_3^PO_1$	**$S_1^PO_3$**	**$S_3O_2^P$**	**$S_3^PO_3^P$**
S_1O_2	**$S_2^PO_3$**	**$S_1^PO_2$**	
S_3O_2	**$S_3^PO_3$**	**$S_3^PO_2^P$**	
$S_1^PO_2$	**$S_2O_1^P$**	**$S_1O_3^P$**	
$S_3^PO_2$	**$S_3O_1^P$**	**$S_2O_3^P$**	

Out of which, there are 11 **unique combinations** in Georgian:

1) **S_1O_2**
2) **S_2O_1**
3) **$S_2O_1^P$**
4) **S_3O_1**
5) **S_3O_2**
6) **$S_3O_1^P$**
7) **$S_3O_2^P$**
8) **$S_2^PO_1$**
9) **$S_2^PO_1^P$**
10) **$S_3^PO_1$**
11) **$S_3^PO_1^P$**

There are **17 ambiguous combinations**. The general pattern of the ambiguous combinations is that there is no distinction between the singular and plural object - **O_3** from **O_3^P**. The ambiguous combinations are as follows:

1) **S_1O_3** is ambiguous with **$S_1O_3^P$**
2) **S_2O_3** is ambiguous with **$S_2O_3^P$**
3) **S_3O_3** is ambiguous with **$S_3O_3^P$**
4) **$S_1O_2^P$** is ambiguous with **$S_1^PO_2$** and **$S_1^PO_2^P$**
5) **$S_1^PO_3$** is ambiguous with **$S_1^PO_3^P$**
6) **$S_2^PO_3$** is ambiguous with **$S_2^PO_3^P$**
7) **$S_3^PO_2$** is ambiguous with **$S_3^PO_2^P$**
8) **$S_3^PO_3$** is ambiguous with **$S_3^PO_3^P$**

In ambiguous combinations, singular and plural forms have the same form. The subject and object of these type of combinations can be either plural or singular depending on the context. Thus, it can be a difficult task to tag them automatically. I have made a decision to treat these pairs of combinations as a singular category in POS-tagging.

In addition, I have made decision to exclude 6 agreement combinations, they are: **S_1O_3**, **S_2O_3**, **S_3O_3**, **$S_1^PO_3$**, **$S_2^PO_3$** and **$S_3^PO_3$**. There is no explicit marker for the 3rd person object, and this is also the form used for an intransitive verb. Thus, in morphology, there is no difference between these forms and the following: **S_1**, **S_2**, **S_3**, **S_1^P**, **S_2^P**, **S_3^P**.

Taking into consideration the above given classification and description, I will consider the following 19 agreement combinations in POS-tagging:

N	S/O	Examples
1	S_1	**vzrdi** “I grow”
2	S_2	**zrdi** “You grow”
3	S_3	**zrdis** “S/he grows”
4	S_1^P	**vzrdit** “we grow”
5	S_2^P	**zrdit** “You grow”
6	S_3^P	**zrdian** “They grow”
7	$S_2O_1^P$	**gvzrdi** “you grow us”
8	S_1O_2	**gzrdi** “I grow you”
9	S_2O_1	**mzrdi** “You grow me”
10	S_3O_1	**mzrdis** “S/he grows me”
11	$S_3O_1^P$	**gvzrdis** “S/he grows us”
12	S_3O_2	**gzrdis** “S/he grows you”
13	$S_3O_2^P$	**gzrdit**, “S/he grows you”; **gezrdebat** “s/he grows you”
14	$S_2^PO_1$	**mzrdit** “You grow me”
15	$S_2^PO_1^P$	**gvzrdit** “You grow me”
16	$S_3^PO_1$	**mzrdian** “They grow me”
17	$S_3^PO_1^P$	**gvzrdian** “They grow us”
18	$S_1^PO_2$	**gzrdit** “We grow you”; **gezrdebit** “We grow for you”
19	$S_3^PO_2$	**gzrdian** “They grow for you”

Table 4. 45: Combinations of argument agreement included in the tagset.

4.10.2 Screeves

Screeves in Georgian represent a system covering Tense, Aspect and Mood (TAM). I will use the following terms: screeve and/or screeves (in plural) to describe the Georgian verb paradigm. The term “screeve” in Georgian **მწკრივი [mc'k'rivi]** literally means “row”, “line”. The decision to use this term is simply because to be consistent with the terminology used in the traditional conjugation system (Shanidze, 1980, pp. 214-224). The screeves represent the conjugation paradigms covering tense, aspect and mood.

The conjugation (screeves and series) system was first classified by Nicholas Marr (1908). Based on Marr, Shanidze developed a new conjugation system described in

his “Georgian *Grammar*” published 1930, later in “*Fundamentals of the Georgian Language*”, published in 1953.

After Shanidze, most work on the verb conjugation has been based on his classification and contains little novelty. The verb description as well as the terminology used is not very accurate in some cases. For example, according to Shanidze, the **Screeves** cover:

1) Tense (*dro*)
2) Mood (*k'ilo*)
3) Iteration (*gzisoba*)
4) Act (*ak'ti*)
5) Accompaniment (*tanamdevroba*).

Shanidze (1980) describes morphologically irrelevant (not marked) categories (so called *iteration, act* and *accompaniment* that represent the literal translation of Georgian terms) within Screeves, but not the aspect category, which is a part of the verb paradigm, despite the fact it is derivational. I will not describe further details but will focus on the categories that should be considered in POS-tagging.

Tense expresses time reference in Georgian, as in other languages. There are no single markers for each tense in Georgian, rather specific root forms that mark the tense. Some Georgian verbs can have two or more meanings; the present tense root form also can express future tense, depending on the context. For example, the verbs such as: **bržanebs** “(will) order”; **asc'avlis** “(will) teach”; **scems** “(will) beat”; **uqurebs** “(will) watch”; **ižienbs** “(will) sleep” etc.

The **Mood** category is a part of the screeve system. There are up to 8 moods in Georgian described by different authors, out of which only four are relevant here:

1) Indicative (*txrobiti*)
2) Subjunctive (*k'avširebiti*)
3) Conditional (*p'irobiti*)
4) Optative (*nat'vriti*)

Aspect is a category that is a part of the screeve system. In Old Georgian, Aspect was morphologically marked: I Series verbs were **Imperfective aspect** and II Series **Perfective aspect**. In Modern Georgian, the aspect category is derivational and uses preverbs, which are prefixal morphemes that are attached to verbs and verbal nouns (Shanidze, 1980, pp.262-272). In particular, preverbs mark Perfective aspect, and the absence of a preverb marks imperfective aspect. For example, [**t'exavs]** "S/he breaks", Imperfective aspect and [**gat'exavs]** "S/he will break", Perfective aspect. There are about 22 preverbs (prefixal morphemes) in Modern Georgian.

For POS-tagging purposes, I will use Shanidze's classification for TAM series. The Series in this classification represents a broader set/group of tenses, which are further divided into screeves - each individual paradigm. Overall, there are eleven screeves distributed across three sets (series). They are:

I Series

a) Present set

1) Present Tense (*ac'mq'o*)
2) Imperfect (*uc'q'vet'eli*)
3) Present Subjunctive (*ac'mq'os k'avširebiti*)

b) Future set

4) Future Tense (*mq'ofadi*)
5) Conditional (*xolmeobiti*)

6) Future Subjunctive (*mq'ofadis k'avširebiti*)

II Series

7) Aorist (*c'q'vet'ili*)
8) Optative (II subjunctive) (*II k'avširebiti*)

III Series

9) Perfect (first evidential, resultative) (*I turmeobiti*)
10) Pluperfect (Second evidential, resultative) (*II turmeobiti*)
11) III subjunctive (third evidential) (*III k'avširebiti*)

There is another classification, the Diatheses system Melikishvili (2014). Diatheses is a Greek word (διάθεσις) meaning grammatical voice. The conjugation system is based on grammatical category of voice, which is more or less similar to Shanidze's classification. Diatheses classification is based on 15,000 verbs (over 9,000 verb roots) from the Georgian Monolingual Dictionary in eight volumes (1950-1964).

Melikishvili introduces three diatheses that are further divided into three series and eleven screeves as in the traditional conjugation system. In this classification, the verbs are grouped into smaller classes according to what kinds of thematic suffixes the verb takes and the grammatical voice of the verb. Basically, they are classes of verbs conjugated across diatheses. The diatheses system, like the traditional system, does not capture all the features of the Georgian verb. The problem that both Shanidze and Melikishvili run into is that they are trying to describe the system as if it was a true inflectional system, but many of the categories in the Georgian verb are more derivational than they are inflectional. For instance, the exact meaning and function of preverbs are not systematic, but rather idiosyncratic in Georgian. For example, some verbs can have only certain preverbs. It is quite rare that all preverbs can occur with

every verb (Gogolashvili, 2011, pp.313-316). Some verbs can only have one preverb, for example as in **და**-ასახიჩრებს [**da**-asaxichrebs] “will mutilate”. Whereas, some verbs can two preverbs, as in **გა**-ნაღმავს [**ga**-nağmavs] “will demine something” and **და**-ნაღმავს [**da**-nağmavs] “will mine something”[34].

That is why Melikshivhili’s attempt to describe this all as an inflectional system results in many cases in a huge list of facts about individual verbs (or small classes of verbs) with long lists of exceptions. Therefore, this approach is not useful for POS-tagging purposes.

[34] These examples are taken from Gogolashvili (2011, pp. 313-316).

4.10.3 Tags for Verbs

In this section, I will define the tagset for verbs. For the purposes of POS-tagging, I have considered the morphologically marked features. They are: 1) argument agreement (as discussed above in sections 1.1.1) and number of argument agreement and 2) screeves. Thus, attribute value pairs for verbs are as follows:

Value	**i) Argument Agreement**	**ii) Screeves**
1	S_1	Present
2	S_2	Imperfect
3	S_3	Present Subjunctive
4	S_1^P	Future
5	S_2^P	Conditional
6	S_3^P	Future Subjunctive
7	S_1O_2	Aorist
8	S_2O_1	Aorist Subjunctive
9	S_3O_1	I Resultative
10	$S_3O_1^P$	II Resultative
11	S_3O_2	III Subjunctive
12	$S_3O_2^P$	
13	$S_2^PO_1$	
14	$S_2^PO_1^P$	
15	$S_3^PO_1$	
16	$S_3^PO_1^P$	
17	$S_1^PO_2$	
18	$S_3^PO_2$	
19	$S_2O_1^P$	

Table 4. 46: Attribute values for verbs.

This gives 209 tags for verbs. The tags are decomposable; I will use colons : to separate major category, person/number agreement and tense. For example, in **V:3S1P:F**, **V** = verb, which is followed by a colon and **3S1P -** argument and number agreement. Here, subject and object are represented in numbers (1,2,3) and their position (first or second) defines the role, namely the first element, in our case **3S** is a subject and **1P** is an object. **S** and **P** mark the number of agreement, singular or plural respectively. This

is then followed by colon again and **F** = Future tense. Thus, there are the following hierarchy: 1) major category (V), 2) argument and number agreement (**3S1P**) and 3) tense (**F**). The complete set of tags for verbs is given in the appendix A.

Description	**Tag**	**Examples (Latin)**	**Examples (Georgian)**
Verb **S_1** Singular, Present Tense	V:1S:P	vizrdebi, vtbebi	ვიზრდები, ვთბები
Verb **S_2** Singular, Future Tense	V:2S:F	gaizrdebi, gatbebi	გაიზრდები, გათბები
Verb **S_3** Singular, Present Subjunctive Tense	V:3S:B	izrdebodes, tbebodes	იზრდებოდეს, თბებოდეს
Verb **S_2O_1** Singular, Conditional Tense	V:2S1S:C	gamzrdidi, gamatbobdi	გამზრდიდი, გამათბობდი
Verb **S_3O_1** Singular, Imperfect Tense	V:3S1S:I	mzrdida, matbobda	მზრდიდა, მათბობდა
Verb **S_3O_2** Singular, Present Tense	V:3S2S:P	gzrdis, gatbobs	გზრდის, გათბობს
Verb **$S_2O_1^P$**, S singular / O Plural , I Resultative Tense	V:2S1P:R	gagizrdivart	გაგიზრდივართ
Verb **$S_2O_1^P$**, S singular / O Plural , II Resultative Tense	V:2S1P:G	gagezarde	გაგეზარდეთ
Verb **$S_2O_1^P$**, S singular / O Plural , III Subjunctive Tense	V:2S1P:S	gagezardo	გაგეზარდოთ

Table 4. 47: Sample tags for verbs.

4.11 Deverbal Adjectives and Nouns

4.11.1 Masdar (*masdari, sac'q'isi*)

The Georgian term for *verbal noun* is *sac'q'isi*, meaning "the beginning". Alternatively, the Arabic term **masdar** ("source") is widely used in Georgian. In the thesis, I will interchangeably use both terms as follows: *verbal noun* and *masdar*.

The *verbal noun* is derived from the verb. Unlike verbs, verbal nouns do not have argument agreement, and they cannot be conjugated. The verbal nouns are declined like nominals, but they do not have plural number.

(96) c'eril-eb-is **gzavna** šec'q'vita
letter-PL-GEN sending.NOM stop.3S.SG.AOR
"(S/he) has stopped sending the letters."

(97) daic'q'eba šek'itxv-eb-is **gamogzavna**
start.3S.SG.FUT question-PL-GEN sending.NOM
"There will start sending out of questions (from there)."

The marker for the verbal noun is the [**-a]** suffix, which is added to the thematic suffix of the I Series verbs (in Present or Future Tenses), and all the argument and voice markers are removed.

Even though verbal nouns are derived from verbs and can have some derivational verbal features, they cannot be conjugated or have argument agreement. The verbal nouns function like nouns and decline like nouns. The only difference between the noun and the verbal noun is that the verbal nouns do not have number except when verbal nouns have lost their verbal features; in this case, it can have plural form. For this reason, I have made a decision to treat verbal nouns under the noun category. Thus, there will not be separate tags introduced for verbal nouns.

4.11.2 Participle

The term for Participle in Georgian is ***mimğeoba*** meaning “derived from something”. Participles formed from the verb in Georgian mainly function as adjectives and decline like adjectives.

(98) c'aikitxa čveni **gzavnil-i** c'eril-i
read.3S.SG.AOR our sent-NOM letter-NOM
“S/he read our sent letter”

(99) c'eril-i anonom-is-gan iq'o **gamogzavnil-i**
letter-NOM anonym-GEN-POST be.3S.SG.AOR sent-NOM
“The letter was sent out by an anonymous (person).”

Like Masdars, participles are mainly used as an adjective and/or noun. Participles decline like nominals and can have number, when used as a noun. For this reason, I will treat participles as adjectives and nouns accordingly. Thus, there will not be a separate tag introduced for participles.

4.12 Prediction: Copular, Affixal

In Georgian, the auxiliary verb არის [**aris]** "is" is mostly commonly used copula (Gogolashvili, 2011, p. 771), which links the subject of a clause to the predicate. However, this copula can be affixed to complements. The term to describe this phenomenon in Georgian is *compound predicate* (*šedgenili šemasmeneli*). In particular, the auxiliary verb [**aris]** "is" is reduced to [**-a]** when it is affixed to a complement. Thus, the affix [-a] is a cliticised copula.

With nouns:

(100) is bavš-i-**a**
s/he child-NOM-COP
"S/he is a child."

With adjectives:

(101) gogo lamaz-i-**a**
girl.NOM beautiful-NOM-COP
"The girl is beautiful."

With numerals:

(102) čem-i nomer-i xut-i-**a**
my-NOM number-NOM five-NOM-COP
"My number is five."

With pronouns:

(103) saxl-i čem-i-**a**
house-NOM my-NOM-COP
"The house is mine."

With participles:

(104) sakitx-i gadac'q'vetil-i-**a**
issue-NOM solved-NOM-COP
“The issue is solved.”

The affixal copula will be treated as enclitic in POS-tagging, the [-**a]** suffix will get a single tag:

Description	**TAG**	**Examples (Latin)**	**Examples (Georgian)**
Copula	AUX	-a	-ა

Table 4. 48: Tags for copula.

4.13 Residual

The residual categories comprise various semi-linguistic and non-Georgian elements. There are five such tags. Sometimes, this element can be inflected as a verb or nominal, in this case it may be considered sufficiently a part of that category to be tagged as such. This particularly applies to foreign words, acronyms and abbreviations.

The tag for Foreign Word covers words from other languages such as Russian and English written in the Georgian alphabet. The unclassified category covers everything, particularly non-Georgian elements, such as foreign words written in Latin or Cyrillic alphabets.

Description	**TAG**	**Examples (Latin)**	**Examples (Georgian)**
Foreign Word	FF	news, job	-
Formula (e.g. Mathematical)	FO	2 × 2	-
Letter of the Alphabet	FZ	b, g, d	ბ, გ, დ
Abbreviation and Acronym: in Georgian	FG	šss, ašš	შსს, აშშ
Abbreviation and Acronym: English (other)	FE	LOL	-
Other unclassifiable non-Georgian element / transliteration variant of a foreign word	FU	cool	ქუულ

Table 4. 49: Tags for Residuals.

It is noteworthy that these residuals, such as abbreviations or acronyms can inflect for case. Thus, when the residuals inflect for case, they will be treated as nouns (or other nominals).

4.14 Punctuation

I will introduce four different labels for different categories of punctuation. They are:

1) Sentence final - punctuations that occur at the end of sentences
2) Sentence medial - Punctuations that occur in the middle of sentences
3) Quotations – opening, closing
4) Brackets – opening, closing

Thus, the Georgian tagset contains four tags for punctuation as follows:

Description	**TAG**	**Examples**
Sentence final	YF	. ? ! ?! …
Sentence medial	YM	, : ; - * ~
Quotations	YQ	" „ “ ”
Brackets	YB	() [] {} / \ < > «

Table 4. 50: Tags for Punctuation.

4.15 Concluding remarks

In this chapter, I have presented a new morphosyntactic language model by going through category by category.

The full list of the tagset is given as a separate document in the appendix A. Thus, I have met one of the main goals of corpus annotation. According to Leech (1997, p.6), the corpus user should have access to documentation including the annotation scheme- “a document describing and explaining the scheme of analysis employed for the annotations”.

Thus, in this chapter of the thesis, I have achieved my aim of defining a POS tagset for use in the tagging of Georgian, which is one of the major prerequisites of an automated part-of-speech tagging.

Chapter 5
Part-of-speech tagging methodologies

In this chapter, I discuss part-of-speech tagging methodologies and justify my choice of part-of-speech tagging program. This chapter also describes the process of manual annotation of the training data for the tagger program, which is an essential prerequisite for automated tagging.

5.1 A review part-of-speech tagging methodology

As discussed in chapter 3 (see section 3.2.4), the design of an automatic tagging system involves several sub-tasks, such as tokenisation, analysis and disambiguation. This section focuses on the disambiguation methodologies and techniques.

There is a wide range of techniques employed in part-of-speech disambiguation. However, "the contextual information analysed by a disambiguation algorithm is typically minimal… preceding or following words, or the tags that these words have, are the only information utilised to any great degree in disambiguation" (Hardie, 2004, p. 229).

Voutilainen (1999, p. 9) describes the linguistic approach and the data-driven approach in disambiguation. According to him, in the linguistic approach, the tagger uses written rules devised by grammarians. Whereas in the data-driven approach, "the language model is derived from automatically conducted statistical studies of large text samples" (Voutilainen, 1999, p. 9). In general, the data-driven approach accounts for a short word sequences and their frequencies and "the tagger selects from the alternatives the one with the highest probability".

Hardie (2004, p. 230) also groups models of language used in disambiguation in two ways, “Firstly, do the linguistic generalisations in the model derive from the grammatical knowledge of a linguist or from a corpus of texts? Secondly, are these linguistic generalisations expressed as rules or as probabilities? Combining these two classifications, four logically possible disambiguation methodologies exist”.

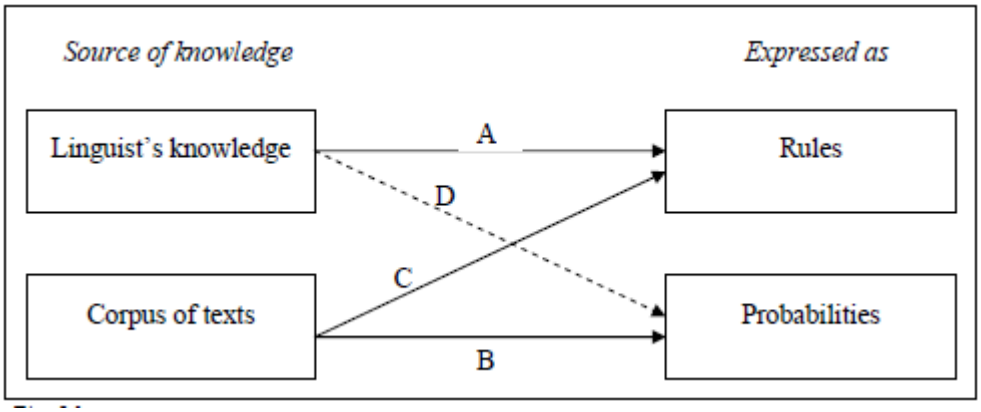


Figure 5. 1: Four logically possible disambiguation methodologies (Hardie 2004, p. 230)

Thus, Hardie (2004, p. 230) describes four possible disambiguation methodologies as follows:

- Type A: the linguistic knowledge is expressed as rules. These types of systems were the earliest to be developed in the 1960s and 1970s, although major advances were made in the 1990s.

- Type B: this method uses corpus-derived data to decide which of the possible tags given to a word is most likely given the surrounding tags, employing a statistical model such as a Markov model (Hardie, 2004). The probabilistic methods were the second to develop in the late 1970s and 1980s.

- Type C: in this method corpus-based rules are used. The methods were used from 1990s (Brill, 1992). Brill calls this approach "transformation-based error-driven learning".

- Type D: Hardie (2004, p. 230) indicates this method using a dashed line and noting that no such methodologies exist. According Hardie (2004, p. 230), no probabilistic model of linguistic knowledge in part-of-speech disambiguation has been based upon human-estimated probabilities.

Any given methodology can be combined that allows for more types of system (Hardie 2004, p. 230). For example, combination methodology, such as a rule-based and stochastic method is referred as *hybrid*[35] method (Garside et al., 1997).

5.1.1 Rule-based approaches

In rule-based approaches, a set of linguistic rules devised by a linguist or from grammars and dictionaries are used as the knowledge base. These linguistic rules are instructions describing a context where the rules should be applied (Hardie, 2004, p. 232). For example, a rule for a Georgian tagger might state that where one of the potential tags for a word is a *modal particle* or a *verb*, it should be tagged as a *modal particle* if it is followed by a word tagged as a *verb*. If the following word is not tagged as a *verb*, the preceding word should be tagged as a *verb* not as a *modal particle*.

It is worthwhile to mention that taking a "rule-based" approach to disambiguation in tagging does not imply using grammar rules as traditionally formulated by linguists. It typically makes use of short-range information (Hardie, 2004, p. 233).

[35] CLAWS4 is an example of a hybrid tagger.

The earliest works on rule-based tagging is associated with Klein and Simmons (1963) and Greene and Rubin (1971). Their work was the first attempt to solve the problem of automated part-of speech tagging disambiguation.

The more recent rule-based approach in 1990s is associated with Constraint Grammar (Karlsson et al., 1995). It should be noted that Constraint Grammar (CG) is not only a tagger but also a parser. CG uses a tokenizer, morphological analyser and a rule-based disambiguator. CG disambiguation rules, depending on rule type, select a correct reading or reject an illegitimate reading, on the basis of relevant words or tags in the left- or right-hand context. Together with the local context, CG rules can refer up to sentence boundaries (Voutilainen, 1999).

5.1.2 Probabilistic approaches

The probabilistic approaches use statistical information about the frequency of tags occurring in long stretches of running text. This information is used to deduce which of the different analyses is the correct one for an ambiguously tagged word.

Modern probabilistic taggers use a mathematical approach such as a Markov model (Charniak et al., 1993). Markov models allow the calculation of the probabilities of different tag sequences by combining different tag transition probabilities. According to Hardie (2004, p. 244) the most immediate advantage of a stochastic system over rule-based systems is that the linguist does not have to write rules to produce an effective part-of-speech tagging system.

“A Markov model estimates the probability of a chain of tags, given empirically-derived tag transition probabilities. By comparing the likelihoods of possible tag sequences for a sequence of ambiguous tokens, the likeliest, and hopefully correct,

sequence can be identified" (Hardie, 2004, p. 248). Thus, such a model uses more minimal contextual information than the rule-based approach. Early work on Markov models was undertaken by Bahl and Mercer (1976).

When tagging, a Markov model system knows what output symbols (words) were produced, but not what states (tags) produced them. For this reason, it is common for this type of Markov model to be called a "hidden Markov model" (HMM), since here state transitions are unobservable (Cutting et al., 1992).

An advantage of HMM taggers is that only a lexicon and some untagged text is needed for training a tagger (Voutilainen, 1999, p. 14).

An interesting property of HMM taggers is that they operate on long-distance information. In practice, however, the size of the contextual "window" is often limited to two or three words. Another attractive feature of HMM taggers is that linguistic information can be incorporated to some extent in the tagger coded biases (by manipulating the lexicon, the tagset and the initial tag probabilities (Voutilainen 1999, p.15).

The CLAWS1 tagging system, developed at the University of Lancaster in the 1980s, utilises a Markov model in its disambiguation module. This module consists of a program called CHAINPROBS, described by Marshall (1987).

A Markov model disambiguator such as CHAINPROBS resolves ambiguity in chains of ambiguously tagged words. This contrast with rule-based methods and the early probabilistic methods of Stolz et al. (1965), where only one word at a time is dealt with. Thus, Markov model parameters capture the probabilities of a word being associated with a given tag and of one tag following another tag.

5.2 Selecting a part-of-speech tagging method for Georgian

van Halteren (1999, p. 95) points out that the choice of the tagger program is determined by the language which is to be tagged and "all other factors must be weighed and, hence can be outweighed". He also points out that the selection is made simply on the basis of availability of the tagger. However, the prime consideration should always be that the tagger is suited for the job it is supposed to do (van Halteren, 1999, p. 96).

There are several factors that influenced my decision to use a stochastic method. The general factor of choosing a stochastic method over a rule-based approach was that rule-based approaches requires a set of generalized linguistic rules prior to tagging. This process can be very time consuming. The other factors that influenced my decision to choose a probabilistic TreeTagger program (Schmid, 1994) are as follows:

- First, the TreeTagger program uses a new probabilistic tagging method in estimating "transition probabilities from sparse data" (Schmid, 1994). This is the main problem for other Markov model based taggers. Most wordforms in any corpus occurs with a low frequency. Therefore, adequate statistics cannot be calculated for them individually. The TreeTagger program estimates transition probabilities using a decision tree and avoids the "sparse data" problem. The TreeTagger achieved 96.36% on Penn-Treebank data. Thus, it may be hoped that this probabilistic method will have reasonably good performance in Georgian.

- Secondly, the TreeTagger program is freely available
- It is very practical and easily manageable with clear instruction and guideline documentation

- Finally, the TreeTagger is easy to use both in training and tagging phases. Particularly, it is very user friendly for those without computational backgrounds, since there is a Graphical Interface for the Windows version of the TreeTagger developed (Ciarán Ó Duibhín, 2018).

Thus, I will use the probabilistic TreeTagger program to accomplish part-of-speech tagging in Georgian. This decision is made in light of a number of factors including availability, practicality and a tagging method used by the TreeTagger program as described above.

5.3 Manual tagging

Automated part-of-speech tagging includes manual tagging, which is needed as training data and is necessary for many computational part-of-speech tagging methods. Thus, tagged data is an essential prerequisite to implementing an automated tagger.

In addition to this, trying out a tagset manually may help to check that the categories actually reflect valid distinctions in the language. It also may help to identify those phenomena, which are difficult to categorise. For example, in Georgian the boundary between the categories of nouns and adjectives, nouns, verbal nouns and adjectives and participles; adjectives and adverbs, particularly adverbs in adverbial case, are often unclear. Words in these categories have a similar syntactic distribution (i.e. prior to or after noun; adjectives in adverbial case have adverb syntactic behaviour- they occur prior to verbs), and they have similar morphological marking. So, the division between the categories depends on semantic and sometimes on syntactic criteria. Therefore, it is arguable whether the word **[kargi]** "good" nominative-absolutive case and **[kargad]** "well" adverbial case (functions as adverb), belong in one category or the other. In this

case, the process of manual tagging allows such words to be identified and discussed, and a decision taken. In the case of adjectives in adverbial case, they were judged according to their form, not function. The problematic examples were assigned to one category or the other in the process of manual tagging.

I have used *enclitic* and *non-enclitic* approaches to tokenisation. In the non-enclitic approach, enclitics are treated as one word, where tags for these enclitic elements are "separated" by: (colon) delimiter.

(1) k'acisk'enacaa => NSG:II:XO:AUX
"Is directed toward a man too"

(1) is a noun, singular, genitive, postposition, particle and auxiliary, tagged as one word. In an enclitic approach, these enclitic elements are tagged separately:

(2) k'acis_NSG
k'ena_II
ca_XO
a_AUX
"Is directed toward a man too"

The tagging manual for the KATAG tagset is primarily designed for enclitic tokenisation. However, it can be used with non-enclitic tokenisation as well. Thus, the KATAG consists of the tagset definition document. The initial version of the former was based on the discussion of the tagset in chapter 4.

It should be noted that manual tagging was undertaken by myself as a native speaker of Georgian. At the first stage, I prepared a manually tagged lexicon of over **95,000 word-forms**, out of which about 13,000 enclitic (including some postpositions and particles) word forms have been removed and **82,851 word-forms** remained. At the next stage, the training set data - **7,425 sentences** (consisting of **90,872 word forms**)

were first randomly selected and then more data were added from the corpus, which was also manually annotated. In order to ensure accuracy and consistency of the tagging process, the manually tagged data (including the training set and the lexicon) was thoroughly revised three times. As a result of these revisions subsequent corrections were made. Thus, accuracy and consistency of the tagging process is ensured as far as possible.

The main sources of the initial 95,000-word lexicon are as follows: 1) KawaC corpus - 35,000 word-forms; 2) Georgian monolingual dictionary (1950-1964) – 40,000 words and 3) Georgian dialect dictionaries – 20,000 words. My intention was to annotate some spoken data, but this was not possible as there are no spoken data available for Georgian.

Chapter 6
Evaluation of the TreeTagger on Georgian texts

In this chapter, I will evaluate the parameter files of the TreeTagger, which is a probabilistic part-of-speech tagging program developed by Schmid (1994) and described in chapter 5.

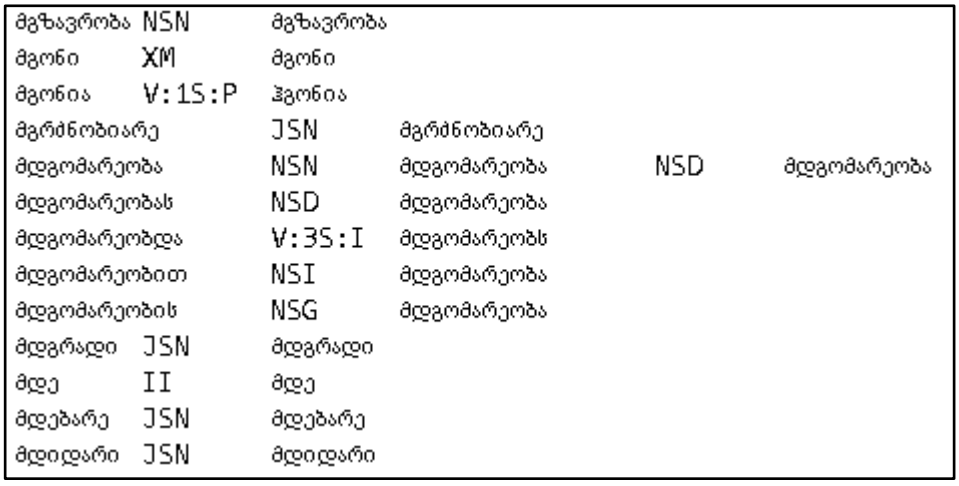
The main aim of this chapter is to measure the performance of the tagger program on Georgian texts. I will primarily consider the results with the enclitic tokenisation approach. Then I will compare the results with the non-enclitic approach. I will argue that the best approach for morphologically rich languages like Georgian, is to treat enclitic forms separately.

6.1 Evaluation of the Treetagger performance for Georgian

6.1.1 The lexicon

In this section, I will describe the performance of the TreeTagger program using the KATAG tagset with the enclitic tokenisation approach. The manually tagged lexicon and training set (described in chapter 5) were used to create a parameter file for an automatic part-of-speech tagging of Georgian texts using the training TreeTagger program. The TrainTreeTagger program requires the following datasets: a *fullform lexicon*, a *training set* and an *open class list*.

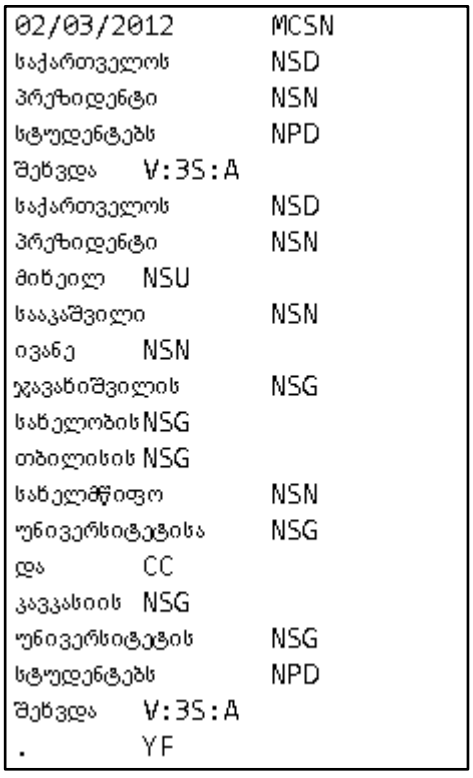
Each line of the lexicon corresponds to one word form and contains the word form itself followed by a Tab character and a sequence of tag-lemma pairs. The tags and lemmata are separated by whitespace.



მგზავრობა	NSN	მგზავრობა			
მგონი	XM	მგონი			
მგონია	V:1S:P	ჰგონია			
მგრძნობიარე		JSN	მგრძნობიარე		
მდგომარეობა		NSN	მდგომარეობა	NSD	მდგომარეობა
მდგომარეობას		NSD	მდგომარეობა		
მდგომარეობდა		V:3S:I	მდგომარეობს		
მდგომარეობით		NSI	მდგომარეობა		
მდგომარეობის		NSG	მდგომარეობა		
მდგრადი	JSN	მდგრადი			
მდე	II	მდე			
მდებარე	JSN	მდებარე			
მდიდარი	JSN	მდიდარი			

Figure 6. 1:Fullform lexicon

The training set file contains tagged training data (running text) in one-word-per-line format. This means that each line contains one token and one tag in that order separated by a tabulator.



02/03/2012	MCSN
საქართველოს	NSD
პრეზიდენტი	NSN
სტუდენტებს	NPD
შეხვდა	V:3S:A
საქართველოს	NSD
პრეზიდენტი	NSN
მიხეილ	NSU
სააკაშვილი	NSN
ივანე	NSN
ჯავახიშვილის	NSG
სახელობის	NSG
თბილისის	NSG
სახელმწიფო	NSN
უნივერსიტეტისა	NSG
და	CC
კავკასიის	NSG
უნივერსიტეტის	NSG
სტუდენტებს	NPD
შეხვდა	V:3S:A
.	YF

Figure 6. 2: Training set

The open class file contains a list of open class tags, i.e. possible tags of unknown word forms. This information is necessary to estimate likely tags of unknown words. The list covers six open class categories, such as adjectives, nouns, and verbs, but not postpositions, conjunctions or particles. The full list of these categories is given in Table 6.1 below. The open class file contains 133 tags in total.

Open class category	**No of tags**	**Example**
Verb	86	V:1P:A, V:1P:B, V:1P:C
Noun	17	NSE, NSG, NSI, NSN
Adjective	13	JSA, JSD, JSE, JSG
Numeral	11	MOSD, MOSE, MOSN, MOSU
Pronoun	1	PIPD
Residual	5	FE, FF, FG, FO, FU

Table 6. 1: Open class tags.

The tagger was trained on the disambiguated KaWac corpus. The data from the Georgian monolingual dictionary and dialect dictionaries contributed to its lexicon. First, the "fullform lexicon" of 95,000 word-forms were manually annotated. The words for the fullform lexicon were carefully selected from the following sources:

- The KawaC corpus (Daraselia and Sharoff, 2014) - 35,000 most frequently used word-forms in the corpus;
- The Georgian monolingual dictionary (1950-1964) - 40,000 words
- Georgian dialect dictionaries – 20,000 words representing a wide range of Georgian dialects.

At the tagset design stage, about 10,000 enclitic (some postpositions and particles) word-forms were removed[36] from the fullform lexicon since they are tagged separately from host words. After the enclitic forms were removed, the revised version of the

[36] They are treated as enclitics and tagged separately receiving their own tags.

fullform lexicon were reduced to about 85,000 word-forms. The training set was created from 7,425 sentences selected from the KaWaC corpus. It includes 90,872 word-forms, which were also manually annotated.

It should be noted that the size of the training corpus and the lexicon were decided based on practical reasons. This includes number of annotators and time limitations. Since, manual tagging was performed by a single person (myself), it was possible to annotate only 346,842[37] words considering the time limitations within this PhD project.

In order to assure consistency and quality of manual tagging process, tags were assigned according to the tagging guidelines defined in chapter 3 and chapter 4 and in section 5.3 of chapter 5.

During the process of training the TreeTagger, some corrections in the fullform lexicon became necessary. The TrainTreeTagger program automatically builds the suffix and prefix lexicon from the training set. However, the automatically derived suffix lexicon produced a number of major disambiguation errors in nominals.

Example 1:

(1) masala-**ze** vimušave
material-POST work.1S.SG.AOR.
"I worked on this material".

In this example, [**ze**] "on" is an enclitic postposition in [**masala<u>ze</u>**] meaning "on the material" and it is tagged separately. The problem here is that after decliticization the

[37] This includes manual tagging considering both enclitic and non-enclitic approaches as follows: 1) in enclitic approach 175,872 words were annotated - 90,872 words in the training set and the 85,000 in lexicon; 2) in non-enclitic approach 170,970 words were annotated – 83,753 in the training set and 87,217 in the lexicon.

remaining word-form [**masala**] "material" is ambiguous and can receive several tags as follows: **NSN** (Noun, singular, nominative-absolutive) or **NSD** (Noun, singular, dative-accusative) or **NSA** (Noun, singular, adverbial) depending on the postposition it follows. In this case, [**ze**] "on" enclitic postposition governs dative-accusative case in Georgian and it should receive an **NSD** (noun, singular, dative-accusative) tag as follows:

Tags for Example 1:

Word	**Tag**	**Tag Description**
masala	NSD	Noun, Singular, Dative-accusative
ze	II	Postposition
vimušave	V:1S:A	Verb, 1st subject, Singular, Aorist

Example 2:

saxl-**ši** movida
home-POST come.3S.SG.AOR.
"S/he came home"

In this example, [**ši**] "in" is an enclitic postposition in [**saxlši**] meaning "in the house/at home" and it is tagged separately. Like in the first example above, the remaining word-form [**saxl**] "house/home" is ambiguous with several possible tags: **NSU** (Noun, singular, zero case) or **NSD** (Noun, singular, dative-accusative) or **NSA** (Noun, singular, adverbial) depending on the postposition it follows. Here [**ši**] "in" postposition governs dative-accusative case in Georgian. Thus, it will get the NSD (noun, singular, dative-accusative) tag as follows:

Tags for Example 2:

Word	Tag	Tag Description
saxl	NSD	Noun, Singular, Dative-accusative
ši	II	Postposition
movida	V:3S:A	Verb, 3rd person Subject, Singular, Aorist

In the fullform lexicon, such ambiguous words have several possible tags, as in the example მასალა [**masala**] "material" and სახლ [**saxl**] "house/home" below:

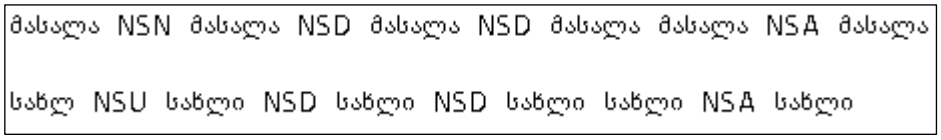


Figure 6. 3: Ambiguous word tagging

Thus, such ambiguous words receive several tags. The tag order is defined objectively based on the word-form and its case order as defined in the tagging guidelines (see chapter 4, section 4.1.2). For example, the word-form მასალა [**masala**] "material" have three potential tags as follows: **NSN** (noun, singular, nominative-absolutive), then it gets the second tag **NSD** when it is followed by a dative-accusative governing postposition, such as the **ze** "on" or **ši** "in"; or **NSA** tag if it is followed by an adverbial governing postposition such as the **mde** "till/until".

However, the TreeTagger cannot disambiguate such cases, so it assigns the most probable tags from the fullform lexicon. For example, the word-form [**masala]** "material" is tagged as **NSN** disregarding the postposition (dative-accusative or adverbial case governing) it follows.

This is because the number of occurrences of each noun/adjective (the same applies to other nominals, such as pronouns and numerals) followed by a postposition in the

training set is very low: the relative frequency (RF) of such occurrences (noun/adjective followed by a postposition) in the training set is **0.08%.** The RF of the first primary tag (e.g. **NSN**) is much higher in the fullform lexicon – **0.27%**. This means that the tags in the fullform lexicon "overrule" the disambiguation "rule" (in the training set) of noun/adjective followed by a postposition, and such word-forms always get the first probable tag as they appear in the fullform lexicon.

This problem was solved by normalizing the fullform lexicon, namely, by removing most nouns and adjectives of singular, nominative-absolutive cases (with **NSN**, **JSN** tags) from the fullform lexicon and approximating the relative frequency to the training set. The Table 6.2 below shows the process of normalizing the RF of **NSN**, **NSD** and **NSA** tags.

Category	**Tag**	**RF in the Training set**	**RF in the Fullform lexicon**	**RF in the normalized fullform lexicon**
Noun, Singular, Nominative-absolutive	NSN	0,08%	0,27%	0,15%
Noun, Singular, Dative-accusative	NSD	0,08%	0,03%	0,08%
Noun, Singular, Adverbial	NSA	0,002%	0,02%	0,003%

Table 6. 2: Normalization of the RF in the Fullform lexicon.

Thus, the relative frequency of the fullform lexicon was normalized. In the example above, the relative frequency of **NSN** tag in the fullform lexicon is 0,27%, which was normalized to 0,15%; and the RF of the NSD in the lexicon was normalized to 0,08% approximating the RF in the training set.

As a result of the normalization of the relative frequency of the nominal tags, about 76,500 word-forms were removed from the full-form lexicon. However, the removed

word-forms were used as an auxiliary lexicon in POS-tagging process. Thus, the number of items in the fullform lexicon was reduced to about 8,500 words. The normalized fullform lexicon improved the TreeTagger performance. It successfully disambiguated **98%** of the nominals followed by postposition cases.

Thus, the annotated data used to train the TreeTagger program are as follows:

Fullform lexicon	8,488 words
Training set	90,872 words (7,425 sentences, 7,500 unique word forms)
Open class tags	133 tags
Auxiliary lexicon	84,683 words

Table 6. 3: Lexicons and training set.

6.1.2 Underrepresentation of tags

The Georgian tagset (KATAG) contains **502 tags** (in theory) in total. However, more than half of the tags have not actually been used in POS-tagging. For example, during the tagset design period, four suffixaufnahme cases were introduced for nominals (nouns, adjectives, pronouns and numerals), i.e. for the categories that inflect for case. In general, suffixaufnahme is quite rare in Georgian and some categories such as numerals and pronouns do not usually get them.

This means that most numerals and pronouns with suffixaufnahme tags do not occur in the training set at all.

A large number of verb tags have also not been utilized in the POS-tagging. These are the verbs with two-person argument agreement (of subject and object). This can be

explained by the overrepresentation of news/press texts in the training set. In news/press language, most verbs encode agreement with one argument, whereas, literary texts or informal speech may be very rich in verbs that encode agreement with two arguments.

In total, **219 tags** out of 502 are actually used in POS-tagging. Whereas, **283** tags never appear in the training set. The full list of unused tags is given in the Table 6.4 below.

Categories	No of unused tags	Percentage of unused tags
Verbs	132	46.64%
Pronouns	86	30.38%
Numerals	44	15.54%
Adjectives	11	3.88%
Nouns	8	2.82%
Residuals	1	0.35%
Punctuation	1	0.35%

Table 6. 4: Unused tags from KATAG tagset.

6.2 The TreeTagger performance for Georgian texts

For the evaluation of the TreeTagger program I selected sample texts for tagging, hereafter referred as "test set". The test set consists of twelve different texts representing five different genres as follows: academic, informal, legal, fiction and news. Each genre consists of several text types. For example, the academic genre, consists of two sample collections from humanities and science fields, whereas fiction genres consist of two texts samples from two different authors.

Genres	Number of words
Academic, humanities	262
Academic, science	561
Informal, author 1	578
Informal, author 2	451
Legal, civil	487
Legal, criminal	380
Fiction, author 1	710
Fiction, author 2	656
News, hard news	121
News, press release	240
News, entertainment	186
News, tv program	249

Table 6. 5: Genres in sample texts.

The texts in the test set were tokenized using the inbuilt tokenizer of the TreeTagger that prints each token on a vertical line. In addition to this, I applied a "rule-based" tokenizer (the same as for training) which identifies token boundaries for enclitic elements, such as postpositions and particles. Thus, the text sample collection covers a range of genre varieties. In total, the test set includes about 5,000 words (including the punctuation and other symbols).

6.3 Results

The performance of the Treetagger was tested on the test set described above. Several variations of the Treetagger program were tested applying different parameters, such as n-gram length and length of the suffix lexicon.

Default values was used for smoothing (Schmid, 1994). For example, the minimum decision tree gain value is 0.7. This means that if the information gain at a leaf node of the decision tree is below this threshold (0.7), the node is deleted. Default value for equivalence class weight is 0.15. Equivalence class weight is used to get reasonable probability estimates for words. The influence of the beginning and ending of a word is calculated using the affix tree gain function. The default value is 1.2. Thus, the information gain at a leaf of an affix tree is below this threshold (1.2), it is deleted. The threshold probability for lexical entries is 0.1. It is a value, which is used to replace zero lexical frequencies. Zero frequencies occur when a word/tag pair appears in the lexicon but not in the training corpus.

Thus, the best results compared to different variations of the TreeTagger program for Georgian were obtained within the following default values of the parameters of the TreeTagger:

- Minimum decision tree gain: 0.7
- Equivalence class weight: 0.15
- Minimum affix tree gain: 1.2
- Threshold probability for lexical entries: 0.001

context	Prefix lexicon	Suffix lexicon	No of nodes	Max. pass length
Bigram	37 nodes	206 nodes	57	15
Trigram	37 nodes	206 nodes	85	16
Quatrogram	37 nodes	206 nodes	101	16

Table 6. 6: Number of n-grams, affix nodes and the depth of the tree.

In the first variation, zero frequencies are used and in the second variation, zero frequencies are replaced by 0.1 before the tag probabilities, to see how strong the influence of the choice of this parameter on the tagging accuracy is. However, changing the replacement value for zero frequencies in the decision tree from a very small value to 0.1 did not improve the accuracy. In both variations, the TreeTagger achieved an accuracy[38] of **88.45%**.

In another test, it was examined how much the tagging accuracy depends on the size of the lexicon, in particular, the auxiliary lexicons combined with different context (n-gram) and suffix lengths.

[38] Accuracy (also known as "correctness") here is defined as follows: percentage of correctly tagged tokens, divided by the total number of tokens (see van Halteren, 1999, p. 82).

6.3.1 Tests for improvement of the TreeTagger performance for Georgian

The inclusion of the auxiliary lexicon (85,000 words) initially dropped the accuracy of the TreeTagger to below **70%** (initial accuracy without auxiliary lexicon is **88.45%**). This is because the auxiliary lexicon was inconsistent with the predefined biases in the training set and the lexicon. This mainly includes the ambiguous categories after decliticization which were not initially considered in the auxiliary lexicon.

The auxiliary lexicon was then revised. Namely, missing ambiguous tags were added to the lexicon. For example, vowel-ending nominals ([-a], [-o], [-e] and [-u]) are ambiguous endings both for nominal and verbal paradigm in Georgian. In the original auxiliary lexicon, such words were presented with only one tag – **NSN** (Noun, singular, nominative-absolutive). In the revised auxiliary lexicon, two or more lines for ambiguous tags (NSD or NSA) were added. This improved the performance of the tagger as it successfully disambiguated such cases.

Finally, the influence of the pruning threshold on the accuracy of the trigram version and the quatrogram version of the TreeTagger was tested. As shown in Table 6.7 below, increasing the context to trigram and quatrograms did not result in any improvement.

Method	**Context**	**Accuracy**
TreeTagger	bigram	88.45%
TreeTagger (0.1)	bigram	88.45%
TreeTagger (auxiliary lexicon)	bigram	70%
TreeTagger (revised auxiliary lexicon)	bigram	92.41 %
TreeTagger (revised auxiliary lexicon)	trigram	92.41 %
TreeTagger (revised auxiliary lexicon)	quatrogram	92.41 %

Table 6. 7: Comparison of accuracy of the TreeTagger program.

After normalising the lexicon, the TreeTagger achieved an accuracy of **92.41 %**. The main contribution came from a better lexicon rather than longer contexts. Thus, the quality of the human expert input is very important. The main reason why the context length does not show any improvements is that these types of errors are not context related. More detailed discussion on the error analysis are given in section 6.4 below.

6.4 Error analysis of the trained TreeTagger on Georgian texts

The TreeTagger was tested on the text samples described in section 6.3 above and it achieved an accuracy of **92.41 %**. Tagging errors in part-of-speech categories varies greatly. Figure 6.4 illustrates the total count of errors in all categories.

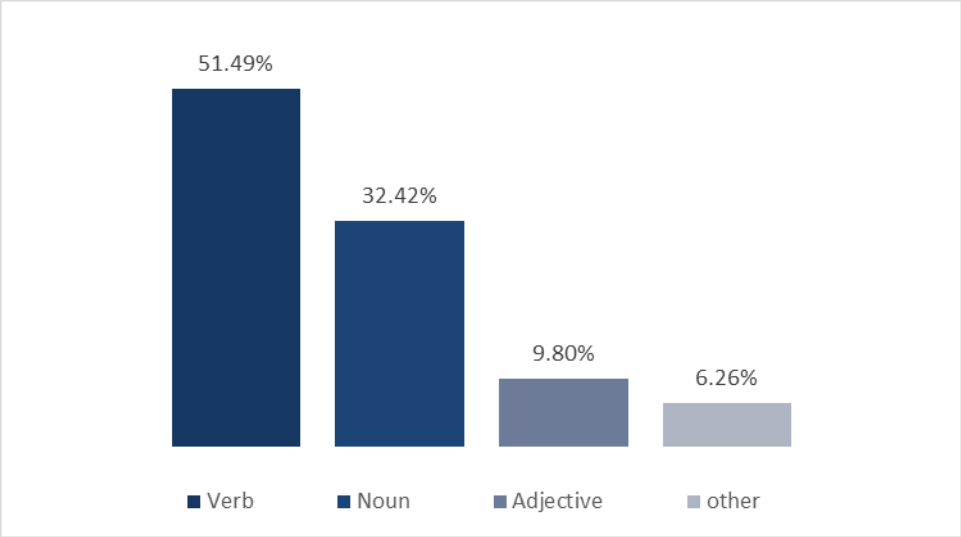


Figure 6. 4: Tagging errors by part-of-speech categories

Thus, verbs and nouns are the categories with the highest error rate. Half of the incorrectly assigned tags are for verbs - **51.49%**, followed by a noun – **32.42%**.

The types of errors produced by the tagger in each category are illustrated in Table 6.8.

Part-of-speech	Error rate	Relative error	Coverage
Verbs	3.89%	34.05%	11.43%
Nouns	2.43%	8.28%	29.44%
Adjectives	0.73%	6.71%	10.98%
Pronouns	0.14%	2.05%	6.96%
Numerals	0.1%	3.35%	3.05%
Residuals	0.2%	22.22%	0.94%

Table 6. 8: Incorrectly assigned POS tags.

The error rate in Table 6.8 refers to the total error count for this category covering both "known" and "unknown" words in the lexicon. The "known" words are the words that

are covered in the tagging lexicon, whereas “unknown” words do not appear in the lexicon.

The relative error reflects the amount of word forms within each category. In particular, the relative error rate reflects how difficult the category is for the tagger, e.g. a 34.05% rate for verbs means one out of 3 verbs gets a tag which is incorrect in at least one position and one out of 15 nouns (8.28%) gets a wrong tag (the noun is not recognized or the case is not assigned correctly), while one out of 50 pronouns (2.05%) gets a wrong tag (case is not assigned correctly). The coverage refers to the total amount of such POS tags in the test set. This indicates the relative importance of the category.

I have analysed the tagger performance for both known and unknown words separately. Overall, 19.03% of the words in the test-test are unknown words. The TreeTagger program assigns correct tags to 61.73% of the “unknown” words.

Part-of-speech	**Error rate for unknown**	**Error rate for known**	**Relative error for unknown**	**Relative error for known**
Verbs	3.87%	0.02%	33.87%	0.17%
Nouns	2.35%	0.08%	8%	0.27%
Adjectives	0.69%	0.04%	6.34%	0.37%
Pronouns	-	0.14%	-	2.05%
Numerals	0.04%	0.06%	1.34%	2.01%
Residuals	0.2%	-	22.22%	-

Table 6. 9: Error rate for known and unknown words.

Table 6.9 shows that the error rate for known words is much lower compared to the error rates for unknown words. For example, the error rate for unknown verbs is **3.87%** and for known verbs it is **0.02%**. Similarly, the error rate is much lower for known nouns and adjectives. However, the error rate for pronouns is related to only known words. This illustrates the disambiguation problem, where the tagger assigned

incorrect case tags. As for the residual category, all the words in this category are unknown words.

Thus, the evaluation of individual categories reveals that the most difficult category is the category of verb, followed by nominals, which includes nouns and adjectives, as well as pronouns and numerals.

It is important to analyse the performance of the TreeTagger across the different genres. The accuracy of the tagger varies for each genre.

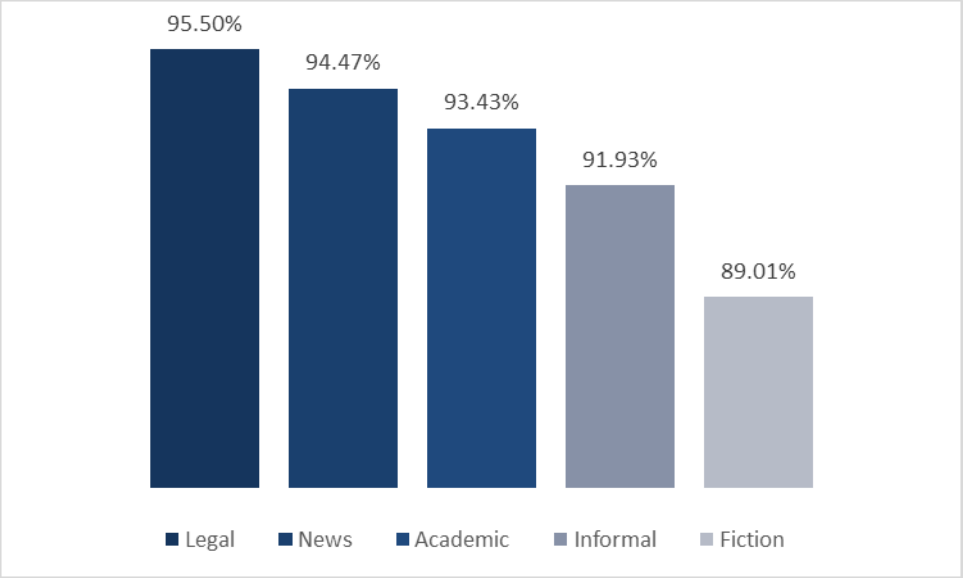


Figure 6. 5: Comparison of accuracy in genres

Figure 6.5 shows the accuracy of the TreeTagger in each genre. The highest performance of the TreeTagger is achieved in legal texts, which is **95.50**%, while the lowest accuracy appears in fiction and informal genre. This is because of the nature of the language used in this test set compared to the training set. For instance, the language (both style and structure) used in legal and news test sets are very similar to those used in the training set. This explains the high performance of the tagger in these genres.

The legal test set was compared to the training set. It revealed the similarities in style and structure of the language used. For example, so called "descriptive" [**ağc'eriti c'armoeba**][39] language is used in both legal and training sets. This "descriptive" language is characterised by passive verbs, such as კეთდება [**k'etdeba**] "is done, is made", გათვალისწინებულ იქნა [**gatvalisc'inebul ikna**] "(it) was considered". Thus, the type of verbs (e.g. argument agreement marking/tense, voice) were both similar in legal test set and in the training set. Table 6.10 shows the distribution of verbs, nouns and pronouns in the training set, legal and fiction test sets.

Part-of-speech	Training set texts	Legal texts	Fiction texts
Verbs	10.94%	9.22%	14.14%
Nouns	35.46%	33.79%	26.61%
Pronouns	12.47%	6.68%	9.38%
Adjectives	10.24%	14.99%	9.44%

Table 6. 10: Part-of-speech distribution in genres.

Compared to the training set and legal test set, the fiction test set has higher frequency of verbs. Table 6.11 below shows that a high number of errors in fiction and informal test sets are incorrectly assigned tags for verbs. This explains the low accuracy in these genres compared to other genres such as legal or news.

POS	Error rate according to genres				
	Legal	**News**	**Academic**	**Informal**	**Fiction**
Verbs	15.38%	26.19%	7.4%	**83.13%**	**67.78%**
Nouns	51.28%	59.52%	57.4%	8.43%	20.13%
Adjectives	33.3%	7.14%	12.9%	7.22%	8.05%
Pronouns	-	-	-	1.2%	4.02%
Numerals	-	59.52%	5.5%	-	-
Residuals	-	2.38%	16.6%	-	-

Table 6. 11: Error rate according to each genre.

[39] For more detailed discussion about the passive voice in Georgian see Melikishvili (2014, pp. 62-68).

Thus, taking into consideration that one out of three verbs gets an incorrect tag, this table explains the low accuracy in informal and fiction genres. These genres usually are rich in verbs, especially verbs which agree with two arguments, as opposed to news/press texts (the main genre in the training set), which are rich in nouns and adjectives, and verbs with a single argument agreement. Thus, most verb forms have incorrect tags in informal (83.13%) and fiction (67.78%) genres, which results in low performance of the Treetagger in these genres.

6.4.1 Types of POS-tagging errors

There are a number of types of tagging errors, including incorrect tags for the part-of-speech, incorrect number or case for nominals; incorrect tense, person/number of agreement in verbs etc. The full list of the type of errors is summarized in Table 6.12 below.

Type of errors	**Examples**
Adjectives tagged as nouns	ლინგვისტურ_NSD, იმპერიულ_NSD
Nouns tagged as adjectives	ხერხად_JSA, თარჯიმნად_JSA, დანაშაულად_JSA
Incorrect Tense in verbs	მოიაზრებოდა_V:3S:A, შეინიშნებოდა_V:3S:A
Incorrect Person/number argument agreement	მოგვეჩვენოს_V:3S:E, მომერგო_V:3S:A,
Incorrect tags for enclitics	მასალა_NSN, ენა_NSN, ხელოვნება_NSN
PL nouns tagged as SG	პარალელების_NSG, დარგებ_NSD
PL adjectives tagged as SG	ასეთებად_JSA, წამყვანები_JSN
SG nouns tagged as PL	მარის_NPG
Verbs tagged as nouns	ფლობდეს_NSD, შემომთავაზა_NSN, ავიღეთ_NSD,
Verbs tagged as adjectives	ჩამოვუყევი_JSN, დავარტყი_JSN, შევლასლასდი_JSN,
Adjectives tagged as verbs	მართლსაწინააღმდეგო_V:3S:A,
Nouns tagged as verbs	წუთებს_V:3S:F, ჭაობ_V:1S:P, გამოფენა-გაყიდვა_V:3S:A
Incorrect case in nouns	ლიტერატურასა_NSG, ნაშრომ_NSE, ზუგდიდსა_NSG
Incorrect case in adjectives	სულელმა_JSN, კარგადა_JSN,
Incorrect tags - residuals	პ_NSD, ჰ_NPG, შშმ_NSE, ე.წ_NSD
Ambiguous words	სასტუმრო_NSN, შექმნის_V:3S:F,
Incorrect tags for wrongly spelled words	რაღაცებს_NPD, რადგაბნ_NSG

Table 6. 12: Part-of-speech tagging errors.

Thus, most errors occur in verbs, followed by nouns and adjectives. To understand the type of errors and why such errors occur, I have analysed the errors for each part-of-speech and compared them to the training set.

There are overall 53 incorrectly assigned tags for verbs, out of which 9 tags do not appear in the training set or the lexicon. Thus, these are the tags that have not been utilized during the manual tagging of the training corpus, since these types of verbs never occurred in the training corpus. However, there are tags for these types of verbs in the KATAG tagset. They are:

TAG	Category
V:1S2S:C	Verb, 1st person SG, 2nd person SG, Conditional Tense
V:2S:C	Verb, 2nd person SG, Conditional Tense
V:2S1S:E	Verb, 2nd person SG, 1st person SG, Aorist Subjunctive Tense
V:3P1P:F	Verb, 3rd person PL, 1st person PL, Future Tense
V:3P2S:P	Verb, 3rd person PL, 2nd person SG, Present Tense
V:3S1S:E	Verb, 3rd person SG, 1st person SG, Aorist Subjunctive Tense
V:3S1S:F	Verb, 3rd person SG, 1st person SG, Future Tense
V:3S2S:A	Verb, 3rd person SG, 2nd person SG, Aorist Tense
V:3S2S:P	Verb, 3rd person SG, 2nd person SG, Present Tense

Table 6. 13: Missing verb tags in the training data.

The rest of the verbs (with incorrectly assigned tags) have low coverage in the training set. Table 6.14 reflects some verb examples and their coverage in the training set.

TAG	Category	Coverage in the Training set
V:3P1S:P	Verb, 3rd person PL, 1st person SG, Present Tense	0.001%
V:3S1P:F	Verb, 3rd person SG, 1st person PL, Future Tense	0.001%
V:3S1S:I	Verb, 3rd person SG, 1st person SG, Imperfect Tense	0.001%
V:3S2S:F	Verb, 3rd person SG, Future Tense	0.001%
V:2P1S:P	Verb, 2nd person PL, 1st person SG, Present Tense	0.002%
V:3S1S:P	Verb, 3rd person SG, 1st person SG, Present Tense	0.002%

Table 6. 14: Tagging errors in verbs.

Thus, these types of verbs (two person of agreement) are very rare, or do not occur in the training set. Hence, there is a very low frequency of such verbs in the training set and for some verbs (nine verbs), there are no tags in the training set. However, they are quite frequently used in informal and fiction texts. This explains the high rate of errors in verbs and low tagger accuracy in informal and fiction test sets.

Table 6.15 summarises type of errors and their error rate in verbs. **61.57%** of the errors in verbs are incorrectly assigned part-of-speech tags, where verbs are tagged as nouns or adjectives.

Type of errors in verbs	**Error rate**
Incorrect POS tag, e.g. verbs tagged as nouns or adjectives etc.	61.57%
Incorrect person and number agreement and tense	10.52%
Incorrect tense	21.05%
Incorrect person and number agreement	6.84%

Table 6. 15: Type of errors in verbs.

Thus, the most errors in verbs are incorrectly assigned tags. This is due to the ambiguous endings in verbs, which can be the same for nominal categories. The problem here is that "the suffix tree in the TreeTagger is searched during a lookup along the path, where the nodes are annotated with the letters of the word suffix in reversed order" (Schmid, 1994).

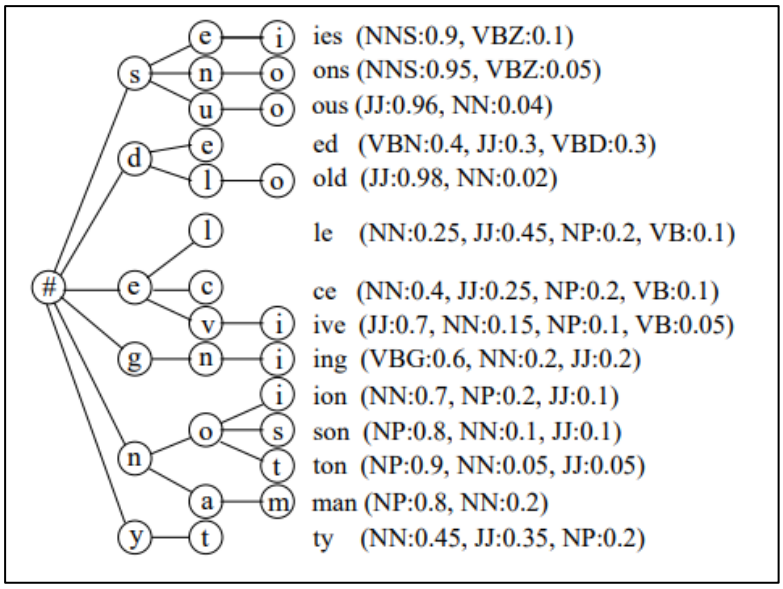


Figure 6. 6: A sample suffix Tree of length 3 (Schmid, 1995)

Thus, the same word endings between verbs and nominal categories are very problematic to disambiguate. As the error rate shows, this is the main reason for incorrectly assigned tags in verbs and in nominals as well. Table 6.16 shows most common examples of such ambiguous endings in verbs and nominals.

Verb ending and its usage	**Nominal ending**	**Error example in verbs**
[-bis] Present tense	**[-bis]** Genitive, singular or Plural	[dar**bis**] “S/he runs” [de**bis**] “of sisters”
[-odi] Conditional, aorist, future, imperfect and present tenses	**[-odi]** nominative-absolutive singular	[avdgeb**odi**] “I would get up” [kom**odi**] “shelf”
[-eba] ending in: Present, Future tenses	**[-i]** nominative-absolutive singular	[gibrund**eba**] “S/he/it is returning to you” [gaket**eba**] “to do”
[-bit] ending in: Aorist, present, aorist subjunctive, future and imperfect tenses	**[-bit]** Instrumental case, singular or plural	[vxde**bit**] “We become” [nabije**bit**] “by steps”
[-ebs] ending in: Future tense	**[-ebs]** Dative-accusative, singular or plural	[inan**ebs**] “S/he will regret this” [saxl**ebs**] “to houses”

Table 6. 16: Ambiguous endings in verbs.

Thus, many major class categories (verbs, nominals) can have the same ending, which can be very problematic in POS-tagging using the TreeTagger program. This contributes to most of the errors in verbs (61.57%).

The other types of errors are also related to the ambiguous endings. For example, incorrectly tagged tenses (**21.05%**), incorrect person and number and tense (**10.52%**) and incorrectly tagged person and number agreement (**6.84%**) are due to the ambiguous endings. The word endings for tenses in Georgian are not consistent, for example, the verb ending on the [**-it]** can be found in plural verbs in aorist, present, aorist subjunctive, future or imperfect tenses. On the other hand, the [**-it]** is the instrumental case marker for nominals. As for the markers for person of argument agreement, they are prefixal (for 1st and 2nd). However, like suffixes, prefixes are also ambiguous with nominals.

The other level of complexity is detecting verbs which agree with two arguments. Firstly, there are very few examples for two-person of argument marking in the training set and in the lexicon. Secondly, the markers for the two person of argument agreement are very difficult to detect in the verb form. In the example, [da**m**icer**e**] "You wrote for me", I have highlighted the person of argument markers, where the [**-m-]** is the marker for the 1st person object and **[-e]** is the marker for the 2nd person of subject in aorist.

Similar problems are encountered in nominals. The full list of incorrectly assigned tags for nominals are given in Table 6.17 below. It includes nouns and adjectives, as well as numerals and pronouns.

TAG	Category	Coverage in the training set
JPA	Adjective, Plural, Adverbial	0.001%
JSA	Adjective, Singular, Adverbial	0.96%
JSE	Adjective, Singular, Ergative	0.1%
JSN	Adjective, Singular, Nominative-absolutive	**7.56%**
JSU	Adjective, Singular, Zero	1.49%
MCSU	Numeral Cardinal Singular Zero	0.1%
MOSN	Numeral Ordinal Singular Nominative - absolutive	0.18%
NPD	Noun, Plural, dative-accusative	1.1%
NPG	Noun, Plural, Genitive	1.41%
NPI	Noun, Plural, Instrumental	0.03%
NPN	Noun, Plural, Nominative-absolutive	0.75%
NSA	Noun, Singular, Adverbial	0.22%
NSD	Noun, Singular, Dative-accusative	**8.49%**
NSFD	Noun, Singular, Suffixaufnahme: Genitive + Dative-accusative	0.16%
NSG	Noun, Singular, Genitive	**8.43%**
NSI	Noun, Singular, Instrumental	0.95%
NSN	Noun, Singular, Nominative-absolutive	**8.56%**
NSU	Noun, Singular, Zero	0.49%
PDPG	Pronoun Demonstrative Plural Genitive	0.08%
PND	Pronoun Negative Accusative –Dative-accusative	0.05%
PP1PU	Pronoun Personal First Person Plural Zero Case	0.1%

Table 6. 17: Tagging errors in nominals.

Table 6.18 summarises type of errors and their error rate in nominals. This includes nouns, adjectives, pronouns and numerals.

Type of errors in verbs	**Error rate**
Incorrect POS tag, e.g. nouns tagged as adjectives or verbs, and adjectives tagged as nouns etc.	37.22%
Plural nominals tagged as singulars	31.11%
Singular nominals tagged as plurals	1.11%
Incorrect case tags for nominals	28.8%
Incorrect case and number tags for nominals	2.7%

Table 6. 18: Type of errors in nominals.

As the table shows above, **37.22%** errors in nominals are incorrectly assigned part-of-speech tags, such as nouns tagged as adjectives or verbs; and adjectives tagged as nouns. **31.11%** of plural nominals are tagged as singulars, while only **1.11%** of singular nominals are tagged as plurals. A large number of errors also occurs in case tagging. About **28.8%** nominals have incorrectly assigned case tags. All these POS-tagging errors are due to the ambiguous endings. It is very difficult for the tagger to assign the correct tags in nominals, when they have the same endings (same case markers). The other major problem in nominals is distinguishing plurals from singulars. This is because the plural marker **[-eb-]** occurs before the case marker and cannot be captured within the suffix Tree length of 3. To capture the plural marker and the case markers in the suffix Tree it should have a length of at least 4 or 5. For example:

(2) ded-**eb-isa**
mother-PL-GEN
“Of mothers”.

Taking into consideration the example above, the suffix Tree should have a length of at least 4 or 5 to account for plural markers. The problem here is that increasing the suffix tree length dropped overall tagger performance.

To sum up, the stochastic TreeTagger program struggles to analyse the complex morphological features in Georgian for several obvious reasons. Firstly, it is difficult to tag the person of argument agreement in verbs. The main reason for this is the basic principle how the TreeTagger program works. Using the automatically generated suffix and prefix lexicon with different context lengths is not simply sufficient enough to disambiguate Georgian verbs, where the person and number of argument agreement are incorporated within the verb form, as in [da**m**icer**e**] “You wrote for me”.

Word with incorrect tags	**English Translation**	**Error rate**
ჩვენ [čven]	We /us *Zero, Genitive, Dative-accusative*	0.08%
ვეტყოდი [vet'q'odi]	I would say to him/her	0.08%
დანაშაულად [današaulad]	crime *Adverbial case*	0.08%
ვახო [vaxo]	Vaxo *Proper name for men Nominative-absolutive case*	0.08%
განზრახი [ganzraxi]	Intentional *Nominative-absolutive case*	0.06%

Table 6. 19: Most common incorrectly tagged words.

A more detailed look at the sources of errors presented in table 6.19 reveals the following problems:

i) Disambiguation problems specific to Georgian morphosyntax:

1. Distinguishing between the major word classes such as verbs, nouns and adjectives due to ambiguous endings;
2. Detecting person of argument agreement in verbs, especially the verbs with more than one person of argument marking;
3. Dealing with case marking in nouns, adjectives and pronouns, especially the postposition governed cases.

ii) Other disambiguation problems in Georgian:

4. Distinguishing between closely related POS classes, such as nouns and adjectives;
5. Distinguishing plural and singular cases in nouns and adjectives;
6. Distinguishing verb tenses.

In spite of the number of problems in statistical tagging, it demonstrated its reasonable performance. The overall accuracy of POS tagging achieved 92.41 %.

6.5 Comparison of enclitic and non-enclitic tokenization approaches

The POS-tagging using the KATAG tagset with the enclitic tokenisation approach achieved an accuracy of 92.41%. In this section I will compare the results with the non-enclitic tokenisation approach. It is worthwhile to mention that in the non-enclitic tokenisation approach, I used the same training set, lexicon and the test set that I have used with the enclitic tokenisation approach.

However, since the enclitic forms are treated as a single unit, the size of lexicon varies.

Fullform lexicon	87,217 words
Training set	83,753 words (7,200 sentences, 7,500 unique word forms)
Open class tags	198 tags

Table 6. 20: Lexicon and training set used with non-cliticised approach.

With the non-cliticised tokenisation approach, the KATAG tagset contains an infinite number of tags as it is impossible to encounter all possible variations. Therefore, the number of tags is unknown. In total **348 tags** are used in this approach (based on the training set).

Similarly, in POS-tagging, this approach uses the same TreeTagger parameters as it used for the enclitic approach.

context	Prefix lexicon	Suffix lexicon	No of nodes	Max. pass length
bigram	60 nodes	315 nodes	41	14
trigram	60 nodes	315 nodes	67	15
quatrogram	60 nodes	315 nodes	87	16

Table 6. 21: Number of n-grams, affix nodes and the depth of the tree, non-enclitic approach.

In the non-enclitic approach, the TreeTagger achieved an accuracy of **87.13%**, which is lower by 5.28% than the cliticised approach (92.41%). Like in the cliticised approach (see Table 6.7), increasing the context to trigram and quatrograms did not result in any improvement in this approach either.

Method	**Context**	**Accuracy**
TreeTagger	bigram	87.13%
TreeTagger	trigram	87.13%
TreeTagger	quatrogram	87.13%

Table 6. 22: Comparison of accuracy of the parameter files.

The types of POS-tagging errors are very similar to the errors described in the error analysis above for the cliticised approach. The accuracy of the TreeTagger program for unknown words in the enclitic approach is 61.73%, whereas in the non-enclitic approach it is 45.02%. This can be explained by high number of cliticised tokens that appears in the test-set. This is problematic since it is difficult to account for all possible clitics (postpositions or particles, or both) for each token in the training set or in the lexicon. Thus, the accuracy for unknown words in the non-clitic approach is much lower than in the enclitic approach. The accuracy of the tagger also varies in each genre.

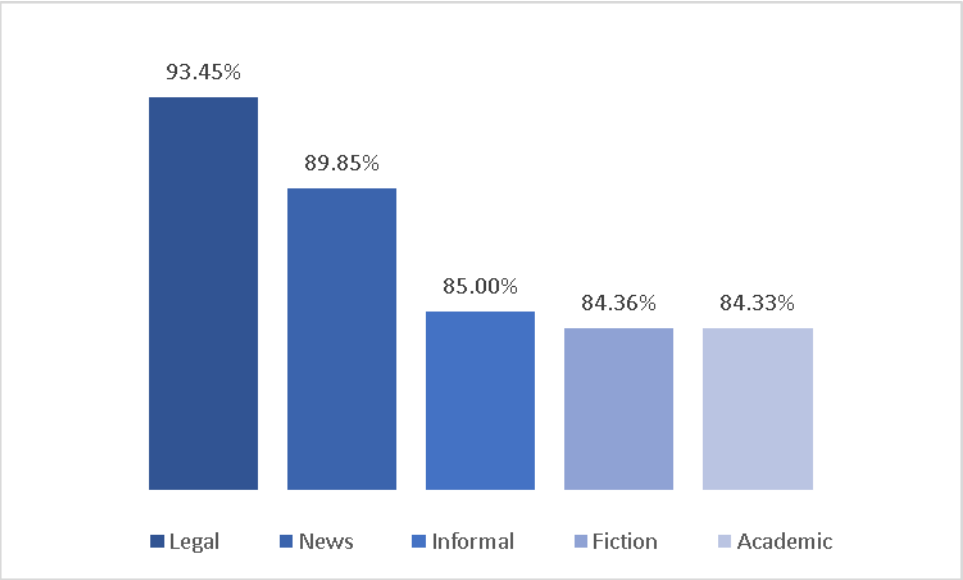


Figure 6. 7: Comparison of accuracy in genres, non-enclitic approach

Figure 6.7 shows the accuracy of the TreeTagger in each genre in the non- enclitic approach. The highest performance of the TreeTagger is achieved in legal texts, which is **93.45**%. The lowest accuracy appears in fiction and the academic genre. In enclitic approach (cf. Figure 6.5), similar results are achieved across the genres. For example, 95.5% accuracy in legal texts are shown in enclitic approach and 93.45% accuracy in non-enclitic approach. However, much worse results are shown in academic genre in non-enclitic approach. For example, 93.43% accuracy is achieved in enclitic approach and 84.33% in non-enclitic approach. This can be explained by variation in the use of enclitics across genres.

As mentioned above, types of errors encountered in both approaches are similar in a way that the probabilistic tagger finds it difficult to assign correct tags based on the suffix and prefix lexicon. It becomes even more problematic when enclitic forms are treated as a single word, as the tagger cannot deal with a long string of encliticized elements. Thus, the best results are obtained when enclitics are treated separately from the host words.

6.6 Comparison of the performance level of the trained TreeTagger program and the Georgian parser

As discussed in Chapter 1, there are no tagging guidelines or tagger programs available for Georgian for the wider academic community, with the exception of the Georgian parser (Meurer, 2007). I have analysed the test sets using the Georgian parser and compared it with the TreeTagger results. The performance and the accuracy of the Georgian parser is 83%, which is much lower than the TreeTagger results in both enclitic (92.41%) and non-enclitic (87.13%) approaches.

In addition to the low performance, the downside of the parser is that it leaves unknown words without tags.

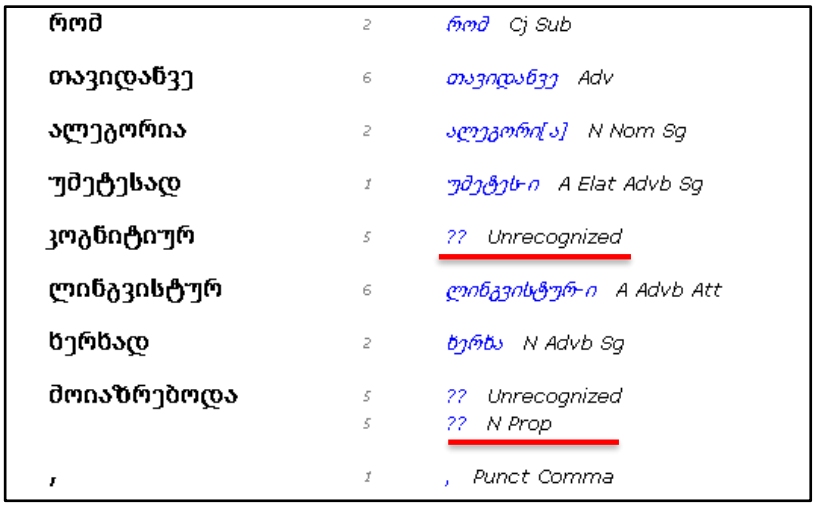


Figure 6. 8: Parsed Georgian text from the test set

Figure 6.8 shows above the "unrecognized" words, which are unknown words in the parser lexicon and which do not have grammatical features assigned to them. In some cases, grammatical features for unrecognized words are assigned incorrectly. For example, მოიაზრებოდა [**moiazreboda**] is an unrecognized word, a verb meaning "it was considered", with the possible grammatical features tag - "N prop", meaning proper noun.

6.7 Concluding remarks

In this section, I evaluated the performance of the probabilistic part-of-speech tagging program and analysed the POS-tagging errors. Thus, I used a stochastic methodology (TreeTagger; Schmid, 1994) taking two approaches: enclitic and non-enclitic approaches. An accuracy of 92.41% using an enclitic tokenisation approach and accuracy of 87.13% was achieved using a non-enclitic tokenisation approach, corroborating my hypothesis that treating enclitic elements separately from the host words results in better tagging performance.

To make the tagger program easily adaptable for a range of inputs (type, variety or genre of text), the performance of the probabilistic TreeTagger program was also evaluated according to five different genres: academic, informal, fiction, news and legal text samples.

In addition to this, I evaluated the performance of the TreeTagger and analysed the most commonly encountered part-of-speech tagging errors in Georgian. Obviously, the size of the training corpus, as well as the morphosyntactic complexity of Georgian, had some impact on the performance of the TreeTagger. Taking into consideration the morphosyntactic complexity of the language, the main challenges for the stochastic tagger on Georgian texts include: similar word endings, two argument agreement markings in verbs and morphological syncretism.

Chapter 7

Conclusions

7.1 Main contributions

The main contribution of my PhD thesis is achieving a functional automated part-of-speech tagging in Georgian using the probabilistic TreeTagger program with an accuracy of 92.41% using the enclitic approach.

The other major contributions of my PhD research are the new morphosyntactic model of Georgian for POS-tagging purposes and the part-of-speech tagging resources that I have developed including a tagset and set of tagging guidelines.

One of the major contributions that this study has made, as far as the structure of Georgian is concerned, is the new morphosyntactic model of the Georgian language for POS-tagging purposes. It is an adequate model for practical applications in Georgian language engineering. Knowing the applicability of this new morphosyntactic model to the field will allow future researchers in Georgian language engineering to make use of the model without uncertainty as to its suitability. Thus, it is an important output of this study.

Other important contributions include the research questions I have investigated. They are as follows:

1. **Is it possible to design a practically manageable hierarchical decomposable tagset for an agglutinative language, such as Georgian?**

I have designed a hierarchical decomposable KATAG tagset for Georgian, which is an agglutinative language with complex morphology. Agglutinative languages have

no finite paradigms and thus, it is difficult to encounter and describe all possible combinations hierarchically.

The KATAG tagset consists of 502 tags[40], which is four/five times larger than average English tagsets. The practicality and manageability of the KATAG tagset has been demonstrated at different stages of this research. First, possible hierarchies were defined going through by category-by-category (in Chapter 4). Then the proposed tagset was put into practice by means of manual tagging (in Chapter 5) of the training corpus representing the natural language data. Finally, the performance of the TreeTagger program using the KATAG tagset has been evaluated (in Chapter 6). Thus, the proposed hierarchical decomposable KATAG is practical and manageable despite the large number of tags it contains.

2. **Is a stochastic method an appropriate one in part-of-speech tagging of morphologically rich and complex language, such as Georgian?**

I have used a stochastic method in part-of-speech tagging in Georgian. Some researchers (Tapanainen and Voutilainen, 1994) suggest that Markov model taggers operate better with small tagsets and it is difficult to write good biases for the probabilistic tagger. Another disadvantage of stochastic methods (with Markov model taggers) when applied to morphologically rich languages is that these languages have potentially freer word order with greater contextual ambiguity (Sánchez-León and Nieto-Serrano, 1997) and thus might be unsuitable for such languages.

I have demonstrated that a probabilistic method is an appropriate approach in part-of-speech tagging for Georgian. According to Schmid (1994, p. 6) the TreeTagger was

[40] On this occasion, 219 tags have been utilized (out of 502 tags). These are the tags that appear in the training corpus. See discussion in section 5.1.2.

tested for English on the Penn-Treebank corpus (36 tags). 2 million words from the corpus were used for training and the TreeTagger achieved **96.36%** accuracy. For Georgian, the TreeTagger has been trained on a much small corpus (in total 90,872-word corpus including punctuation) and achieved an accuracy of **92.41%**. This suggests that with more and better lexicon, the performance of the TreeTagger for Georgian can be improved to achieve better results.

3. **What are the main challenges of the probabilistic TreeTagger program (with Markov model) when it is applied to Georgian?**

I have evaluated the performance level of the TreeTagger and analzyed the most commonly encountered errors of part-of-speech tagging in Georgian. Obviously, the size of the training corpus had some effects on the performance level of the TreeTagger. However, this research question addresses other problems, such as the morphosyntactic complexity of Georgian and what aspects of it are the most difficult for the TreeTagger program.

One of the main problems in tagging Georgian using the stochastic TreeTagger program is similar word endings in Georgian. Almost no word terminations in Georgian indicate exclusively a single category or even a small group of categories. Instead, a single morpheme may realise a large number of categories (for instance, *-a* which may indicate almost all the possible tags in the tagset (such as NSN, NSD, NSG, NSA, NSI, NSE, RR, JSN, V:2S:A etc.).

The other level of complexity is detecting two-person of argument marking in verbs. The markers for the two person of argument agreement are very difficult to detect in the verb form. For example, in დამიწერე **[damicere]** "You wrote for me", the marker

for the first person of object (“me”) is ‘infixal’ [**-m-**], whereas, the marker for the second person subject (“you”) is suffixal [**-e**]. It is difficult for the stochastic TreeTagger program to detect the subject and object agreement markers in Georgian verbs.

In addition to this, the other major problem in Georgian morphosyntax is *morphological syncretism,* when one wordform belongs to the same morphosyntactic category, but it is difficult to identify appropriate morphosyntactic features, such as tense and argument agreement in verbs. For example, the Georgian verb გაწუხებთ **[gac'uxebt]** can have at least two readings:

- Verb, 3rd person of Subject singular and 2nd person of object Plural (“S/he/it bothers you (PL))
- Verb, 1st person of Subject plural and 2nd person of object singular (“We bother you).

Thus, the main challenges for the stochastic tagger on Georgian texts include: the similar word endings, two argument agreement markings in verbs and the morphological syncretism as described above. Taking into consideration these challenges, better biases for the probabilistic tagger can be written by improving the lexicon to account for all problematic areas in Georgian stated above. Also, the most obvious means of improving the tagger is clearly to use a larger lexicon.

4. **What is the best approach in tokenisation when dealing with enclitics in Georgian?**

I have used two different approaches of tokenisation of “clitics” (as it applies in POS tagging). In the first approach, I have treated enclitic elements separately from the host

words and argued that a better performance level would be achieved using this approach. In the second approach, I have treated enclitics as a single word. Then I have compared the performance level of the TreeTagger program using both approaches.

I have demonstrated that the first tokenisation approach of splitting enclitics has advantage over the second approach. The encliticized tokenisation improves the performance of the tagger by 5.28%, from 87.13% to 92.41% of accuracy.

5. Which genres are most difficult in part-of-speech tagging in Georgian?

The performance of the probabilistic TreeTagger program is evaluated on the obtained test set consisting of five different genres: academic, informal, fiction, news and legal. The main reason for this is to find out if the application of the tagger is limited because the resources (e.g. training set, lexicon) used were trained for a particular variety or genre of text.

As expected, the accuracy of the tagger varies in each genre. The highest performance of the TreeTagger is achieved in legal and news texts in both enclitic and non-enclitic approaches. In legal texts, **95.50**% accuracy is shown (see Figure 6.5) in enclitic approach and **93.45%** accuracy in non-enclitic approach (see Figure 6.7). In news texts, **94.47%** accuracy is achieved in enclitic approach and **89.85%** in non-enclitic approach. The lowest accuracy appears in fiction (**67.78%**) and informal (**83.13%**) genres in enclitic approach. Whereas in non-enclitic approach, lowest accuracy is shown in fiction (**84.36%**) and academic (**84.33%**) genres. This is because of the style/register of the language used in these test set compared to the training set. For instance, the language (both style and structure) used in legal and news test sets are very similar to those used in the training set. This explains the higher performance level of the tagger in these genres.

On the other hand, the fiction and informal test set has a higher frequency of verbs. In the error analysis, a high number of errors in fiction and informal test sets are associated with incorrectly assigned tags for verbs. This explains the low accuracy in these genres compared to other genres such as legal or news.

To make the tagger program easily adaptable for a range of input (type, variety or genre of text), the training corpus should be expanded to include more fictional and informal texts proportionally.

7.2 Resources developed

The most important contributions of my PhD project are the part-of-speech tagging resources for Georgian that I have developed. They are:

1. KATAG tagset;
2. A set of tagging guidelines
3. Parameter files for functional automated part-of-speech tagging in Georgian using the probabilistic TreeTagger program;
4. Corpus based list of syncopated and non-syncopated words in Georgian
5. Manually annotated training corpus and lexicon.

The KATAG tagset obviously represents a major resource for Georgian corpus linguistics. This hierarchical decomposable tagset can be used for other stochastic taggers, or rule-based or hybrid tagger programs. Thus, its value as analysis scheme is independent of any particular application, and as such, it is a useful product of this study in its own right.

The trained parameter files of the TreeTagger program would be a valuable resource for the users, especially for those without a programming background. Thus, the parameter files of the probabilistic TreeTagger program trained on Georgian texts will be become publicly available. They are as follows:

TreeTagger parameter files	Method	Accuracy
kabigram-utf8.par	Bigram TreeTagger, enclitic tokenisation	92.41%
katrigram-utf8.par	Trigram TreeTagger, enclitic tokenisation	92.41%
kaquadrogram-utf8.par	Quadrogram TreeTagger, enclitic tokenisation	92.41%
geobigram-utf8.par	Bigram TreeTagger, non-enclitic tokenisation	87.13%
geotrigram-utf8.par	Trigram TreeTagger, non-enclitic tokenisation	87.13%
geoquadrogram-utf8.par	Quadrogram TreeTagger, non-enclitic tokenisation	87.13%

Table 7. 1: Trained TreeTagger parameter files.

Syncopation is an important part of the nominal declension in Georgian nominals. Thus, the information on which words undergo syncopation is very useful in part-of-speech tagging. Therefore, I have analysed over 5 million (5,234,371) words with "syncopated syllables" in Genitive, Instrumental and Adverbial Cases in the KaWac corpus. As a result of the corpus analyses, I have produced three types of lists, as follows:

Lists	No of words
List A: non-syncopated words	640
List B: syncopated words	335
List C: Both syncopated and non-syncopated	50

Table 7. 2:List of syncopated and non-syncopated words in Georgian.

The first list includes the words that are always syncopated. The second list covers the words with syncopated syllables, but they never syncopate. The third list includes the words that sometimes syncopate and sometimes remains unsyncopated. The full list of vowel syncopation in Georgian is given in the Appendix B.

The manually annotated training corpus and lexicon can also be considered as one of the most important part-of-speech tagging resources. This data may be of benefit of future research in part-of-speech tagging in Georgian.

Type of training Data	**Enclitic training data**	**Non-enclitic training data**
Training set	90,872 words (7,425 sentences, 7,500 unique word forms)	83,753 words (7,200 sentences, 7,500 unique word forms)
Fullform lexicon	8,488 words	87,217 words
Auxiliary lexicon	84,683 words	-

Table 7. 3: Manually Tagged training data.

7.3 Future works

The KATAG tagset represents the most important morphosyntactic features of Georgian. There is no claim that the presented tagset is optimal for Georgian. There is still room for improvement.

Some minor changes can be applied to the classification of adjectives. For example, the degrees of comparison for adjectives can be added to the tagset. The other important thing is to reconsider introducing the suffixaufnahme cases for numerals, as they have not actually been utilized for numerals. Another alteration can be made in relative pronouns, where enclitic particles can be considered as a part of the word form.

One of the main advantages of the proposed tagset is that it easily understandable and manageable. Therefore, if anyone wishes depending on her/his research interests, can alter the tagset to be more fine-grained or simplified.

There are a number of possible future research projects that can be carried out in the field of Georgian corpus linguistics using the KATAG tagset. For example, I have introduced two additional cases (outside the traditional case system, Shanidze 1980): 1) Zero (or null) and 2) Suffixaufnahme. These cases are not well studied in Georgian linguistics. Thus, the KATAG decomposable tagset will allow specific searches in the corpus to analyse the distribution patterns (syntactic behaviour for instance) and frequency of their usage.

The initial intention of this project involved modifying the Unitag - a rule-based tagger for Georgian. The Unitag program was originally developed to tag Urdu (Hardie, 2004) and then was used to tag Nepali (Hardie et al., 2011). It consists of a morphological and lexical analysis system and disambiguation modules, which is

based on hand-written rules and also uses a probabilistic system based on a Markov model. Much works on the rule-based disambiguation for the Unitag program has already been carried out. Thus, an obvious next step is a development of a rule-based Georgian tagger. It might be hoped that the rule-based tagger would improve the annotation accuracy.

Another important next step would be a development a semantic tagger in Georgian. A very first step toward the semantic tagging has been undertaken. In particular, I have enquired the possibility of expanding the USAS semantic tagger (Piao et al, 2015) for Georgian. The USAS's semantic lexicon has been used to automatically extract and map the translated Georgian dictionary entries from the English-Georgian Comprehensive Online dictionary[41]. The automatically derived semantic lexicon (for test sample letter A) for Georgian proved to be adequate with some manual post editing.

It may also be hoped that the experience of developing part-of-speech tagging resources to Georgian would support adaptation of the annotation scheme for other Kartvelian languages, such as Megrelian, Laz and Svan.

[41] https://dictionary.ge/

Appendix A

KATAG tagset and tagging guidelines

A1. Noun

There are two attribute values for nouns: number and case.

Value	i) Number	ii) Case
1	Singular	Zero case
2	Plural	Nominative-absolutive
3		Ergative
4		Dative-accusative
5		Genitive
6		Instrumental
7		Adverbial
8		Vocative
9		Suffixaufnahme: Genitive + Ergative
10		Suffixaufnahme: Genitive + Dative-accusative
11		Suffixaufnahme: Genitive + Adverbial
12		Suffixaufnahme: Genitive + Vocative

Table A1. 1: Attribute values for Nouns.

This gives 24 Tags for nouns as follows:

Description	TAG	Examples (Latin)	Examples (Georgian)
Noun Singular Zero Case	NSU	kac, saxl, kud	კაც, სახლ, ქუდ
Noun Singular Nominative-absolutive	NSN	kaci, saxli, kudi	კაცი, სახლი, ქუდი
Noun Singular Ergative	NSE	kacma, saxlma, kudma	კაცმა, სახლმა, ქუდმა
Noun Singular Dative-accusative	NSD	kacs, saxls, kuds	კაცს, სახლს, ქუდს
Noun Singular Genitive	NSG	kacis, saxlis kudis	კაცის, სახლის, ქუდის

Noun Singular Instrumental	NSI	kacit, saxlit, kudit	კაცით, სახლით, ქუდით
Noun Singular Adverbial	NSA	kacad, saxlad, kudad	კაცად, სახლად, ქუდად
Noun Singular Vocative	NSV	kaco, saxlo, kudo	კაცო, სახლო, ქუდო
Noun Singular, Suffixaufnahme: Genitive + Ergative	NSFE	kacisam, saxlisam	კაცისამ, სახლისამ
Noun, Singular, Suffixaufnahme: Genitive + Dative-accusative	NSFD	kacisas, saxlisas	კაცისას, სახლისას
Noun, Singular, Suffixaufnahme: Genitive + Adverbial	NSFA	kacisad, saxlisad	კაცისად, სახლისდ
Noun, Singular, Suffixaufnahme: Genitive + Vocative	NSFV	kacisav, saxlisav	კაცისავ, სახლისავ
Noun Plural Zero-case	NPU	kaceb, saxleb, kudeb	კაცებ, სახლებ, ქუდებ
Noun Plural Nominative-absolutive	NPN	kacebi, saxlni, kudebi	კაცები, სახლნი, ქუდები
Noun Plural Ergative	NPE	kacebma, saxlebma, kudebma	კაცებმა, სახლებმა, ქუდებმა
Noun Plural Dative-accusative	NPD	kacebs, saxlebs, kudebs	კაცებს, სახლებს, ქუდებს
Noun Plural Genitive	NPG	kacebis, saxlebis, kudebit	კაცების, სახლების, ქუდების
Noun Plural Instrumental	NPI	kacebit, saxlebit, kudebit	კაცებით, სახლების, ქუდებით
Noun Plural Adverbial	NPA	kacebad, saxlebad, kudebad	კაცებად, სახლებად, ქუდებად
Noun Plural Vocative	NPV	kacebo, saxlebo, kudebo	კაცებო, სახლებო, ქუდებო

Noun Plural, Suffixaufnahme: Genitive + Ergative	NPFE	kacebisam, saxlebisam	კაცებისამ, სახლებისამ
Noun, Plural, Suffixaufnahme: Genitive + Dative-accusative	NPFD	kacebisas, saxlebisas	კაცებისას, სახლებისას
Noun, Plural, Suffixaufnahme: Genitive + Adverbial	NPFA	kacebisad, saxlebisad	კაცებისად, სახლებისდ
Noun, Plural, Suffixaufnahme: Genitive + Vocative	NPFV	kacebisav, saxlebisav	კაცებისავ, სახლებისავ

Table A1. 2: Tags for nouns.

A2. Adjective

There are two attribute values for adjectives (like nouns): number and Case. It gives 24 Tags for adjectives as follows:

Description	**TAG**	**Examples (Latin)**	**Examples (Georgian)**
Adjective Singular Zero case	JSU	martal, did	მართალ, დიდ,
Adjective Singular Nominative-absolutive	JSN	cudi, martali,	ცუდი, მართალი
Adjective Singular Ergative	JSE	qrum, martalma	ყრუმ, მართალმა
Adjective Singular Dative-accusative	JSD	qrus, martals	ყრუს, მართალს
Adjective Singular Genitive	JSG	qrusi, martlis	ყრუსი, მართლის
Adjective Singular Instrumental	JSI	qruti, martlit	ყრუთი, მართლით
Adjective Singular Adverbial	JSA	qrud, martlad	ყრუდ, მართლად
Adjective Singular Vocative	JSV	qruv, martalo	ყრუვ, მართალო
Adjective Singular, Suffixaufnahme: Genitive + Ergative	JSFE	martlisam, cudisam	მართლისამ, ცუდისამ
Adjective, Singular, Suffixaufnahme: Genitive + Dative-accusative	JSFD	martlisas, cudisas	მართლისას, ცუდისას

Adjective, Singular, Suffixaufnahme: Genitive + Adverbial	JSFA	martlisad, cudisad	მართლისად, ცუდისად
Adjective, Singular, Suffixaufnahme: Genitive + Vocative	JSFV	cudisao, cudisao	მართლისაო, ცუდისაო
Adjective Plural Zero Case	JPU	qrueb, martleb	ყრუებ, მართლებ
Adjective Plural Nominative-absolutive	JPN	qruebi, martlebi	ყრუები, მართლები
Adjective Plural Ergative	JPE	qruebma, martlebma	ყრუებმა, მართლებმა
Adjective Plural Accusative- Dative-accusative	JPD	qruebs, martlebs	ყრუებს, მართლებს
Adjective Plural Genitive	JPG	qruebis, martlebis	ყრუების, მართლების
Adjective Plural Instrumental	JPI	qruebit, martlebit	ყრუებით, მართლებით
Adjective Plural Adverbial	JPA	qruebad, martlebad	ყრუებად, მართლებად
Adjective Plural Vocative	JPV	qruebo, martlebo	ყრუებო, მართლებო
Adjective Plural, Suffixaufnahme: Genitive + Ergative	JPFE	martlebisam, cudebsam	მართლებისამ, ცუდებისამ
Adjective, Plural, Suffixaufnahme: Genitive + Dative-accusative	JPFD	martlebisas, cudebisas	მართლებისას, ცუდებისას
Adjective, Plural, Suffixaufnahme: Genitive + Adverbial	JPFA	martlebisad, cudebisad	მართლებისად, ცუდებისად
Adjective, Plural, Suffixaufnahme: Genitive + Vocative	JPFV	martlebisao, cudebisao	მართლებისაო, ცუდებისაო

Table A2. 1: Tags for adjectives.

A3. Pronoun

Most, albeit not all, pronouns have irregular case inflections, and many pronouns lack plural forms. Thus, I will give attribute values for each type and then will give the full list of tags.

Value	i) type	ii) Person	iii) Number	iv) Case
1	Personal	First	Singular	Zero
2		Second	Plural	Nominative-Absolutive
3				Dative-accusative
4				Vocative

Table A3. 1: Attribute values for Personal Pronouns.

Value	i) type	ii) Number	iii) Case
1	Demonstrative	Singular	Zero
2		Plural	Nominative-Absolutive
3			Ergative
4			Dative-accusative
5			Genitive
6			Instrumental
7			Adverbial
8			Vocative
9			Suffixaufnahme-Ergative
10			Suffixaufnahme-Dative-accusative
11			Suffixaufnahme-Adverbial

Table A3. 2: Attribute values for Demonstrative Pronouns.

Value	i) type	ii) Number	iii) Case
1	Interrogative	Singular	Zero
2		Plural	Nominative-Absolutive
3			Ergative
4			Dative-accusative
5			Genitive
6			Instrumental
7			Adverbial
8			Suffixaufnahme-Ergative
9			Suffixaufnahme-Dative-accusative
10			Suffixaufnahme-Adverbial

Table A3. 3: Attribute values for Interrogative Pronouns.

Value	i) type	ii) Person	iii) Number	iv) Case
1	Possessive	First	Singular	Zero
2		Second	Plural	Nominative-Absolutive
3	Reflexive			Ergative
4				Dative-accusative
5				Genitive
6				Instrumental
7				Adverbial
8				Vocative
9				Suffixaufnahme-Ergative
10				Suffixaufnahme-Dative-accusative
11				Suffixaufnahme-Adverbial
12				Suffixaufnahme-Vocative

Table A3. 4: Attribute values for Possessive Pronouns.

Value	i) type	i) Case
1	Reciprocal	Zero
2		Nominative-Absolutive
3		Ergative
4		Dative-accusative
5		Genitive
6		Instrumental
7		Adverbial
8		Suffixaufnahme-Ergative
9		Suffixaufnahme-Dative-accusative
10		Suffixaufnahme-Adverbial

Table A3. 5: Attribute values for Reciprocal Pronouns.

Value	i) type	i) Case
1	Empathic	Zero
2		Nominative-Absolutive
3		Ergative
4		Dative-accusative
5		Genitive
6		Instrumental
7		Adverbial
8		Vocative
9		Suffixaufnahme-Dative-accusative
10		Suffixaufnahme-Adverbial

Table A3. 6: Attribute values for Empathic Pronouns.

Value	i) type	ii) Number	iii) Case
1	Indefinite	Singular	Zero
2		Plural	Nominative-Absolutive
3			Ergative
4			Dative-accusative
5			Genitive
6			Instrumental
7			Adverbial
8			Suffixaufnahme-Ergative
9			Suffixaufnahme-Accusative-Dative-accusative
10			Suffixaufnahme-Adverbial

Table A3. 7: Attribute values for Indefinite Pronouns.

Value	i) type	i) Case
1	Negative	Zero
2		Nominative-Absolutive
3		Ergative
4		Dative-accusative
5		Genitive
6		Instrumental
7		Adverbial
8		Suffixaufnahme-Dative-accusative

Table A3. 8: Attribute values for Negative Pronouns.

Overall, this gives 163 tags for pronouns:

Description	**TAG**	**Examples (Latin)**	**Examples (Georgian)**
Pronoun Personal First Person Singular Nominative-Absolutive Case	PP1SN	me	მე
Pronoun Personal Second person Singular Zero Case	PP2SU	šen	შენ
Pronoun Personal Second person Dative-accusative	PP2SD	šen	შენ
Pronoun Personal Second person Genitive	PP2SG	šen	შენ
Pronoun Personal First Person Plural Zero Case	PP1PU	čven	ჩვენ
Pronoun Personal First Person Plural Dative-accusative	PP1PD	čven	ჩვენ
Pronoun Personal First Person Plural Genitive	PP1PG	čven	ჩვენ
Pronoun Personal Second person Plural Zero	PP2PU	tkven	თქვენ
Pronoun Personal Second person Dative-accusative	PP2PD	tkven	თქვენ
Pronoun Personal Second person Genitive	PP2PG	tkven	თქვენ
Pronoun Personal Second person Singular Vocative	PP2SV	še, šeno	შე, შენო
Pronoun Personal Second person Plural Vocative	PP2PV	tkve, tkveno	თქვე, თქვენო
Pronoun Demonstrative Singular Zero Case	PDSU	es, eg	ეს, ეგ
Pronoun Demonstrative Singular Nominative - absolutive	PDSN	aseti, magnairi	ასეთი, მაგნაირი
Pronoun Demonstrative Singular Ergative	PDSE	asetma, magnairma	ასეთმა, მაგნაირმა

Pronoun Demonstrative Singular Dative-accusative	PDSD	asets, magnairs	ასეთს, მაგნაირს
Pronoun Demonstrative Singular Genitive	PDSG	asetis, magnairis	ასეთის, მაგნაირის
Pronoun Demonstrative Singular Instrumental	PDSI	asetit, magnairit	ასეთით, მაგნაირით
Pronoun Demonstrative Singular Adverbial	PDSA	asetad, magnairad	ასეთად, მაგნაირად
Pronoun Demonstrative Singular Vocative	PDSV	aseto, amnairo	ასეთო, ამნაირო
Pronoun Demonstrative Singular Suffixaufnahme-Ergative	PDSFE	amnairisam, amisam	ამნაირისამ, ამისამ
Pronoun Demonstrative Singular Suffixaufnahme-Dative-accusative	PDSFD	amnairisas, amisas	ამნაირისას, ამისას
Pronoun Demonstrative Singular Suffixaufnahme-Adverbial	PDSFA	amnairisad, amisad	ამნაირისად, ამისად
Pronoun Demonstrative Plural Nominative - absolutive	PDPN	asetebi, magnairebi	ასეთები, მაგნაირები
Pronoun Demonstrative Plural Ergative	PDPE	asetebma, magnairebma	ასეთებმა, მაგნაირებმა
Pronoun Demonstrative Plural Dative-accusative	PDPD	asetebs, magnairebs	ასეთებს, მაგნაირებს
Pronoun Demonstrative Plural Genitive	PDPG	asetebis, magnairebis	ასეთების, მაგნაირების
Pronoun Demonstrative Plural Instrumental	PDPI	asetebit, magnairebit	ასეთებით, მაგნაირებით
Pronoun Demonstrative Plural Adverbial	PDPA	asetebad, magnairebad	ასეთებად, მაგნაირებად
Pronoun Demonstrative Plural Vocative	PDPV	asetebo, amnairebo	ასეთებო, ამნაირებო
Pronoun Demonstrative Plural Suffixaufnahme-Ergative	PDPFE	amnairebisam, amebisam	ამნაირისამ, ამისამ
Pronoun Demonstrative Plural Suffixaufnahme-Dative-accusative	PDPFD	amnairebisas, amisas	ამნაირისას, ამისას

Pronoun Demonstrative Plural Suffixaufnahme-Adverbial	PDPFA	amnairebisa, amebisad	ამნაირისად, ამისად
Pronoun Interrogative Singular Zero	PTSU	vin, ra	ვინ, რა
Pronoun Interrogative Singular Nominative - absolutive	PTSN	romeli, rogori	რომელი, როგორი
Pronoun Interrogative Singular Ergative	PTSE	romelma, rogorma	რომელმა, როგორმა,
Pronoun Interrogative Singular Dative-accusative	PTSD	romels, rogors	რომელს, როგორს
Pronoun Interrogative Singular Genitive	PTSG	ramdenis, sadauris	რამდენის, სადაურის
Pronoun Interrogative Singular Instrumental	PTSI	ramdenit, sadaurit	რამდენით, სადაურით
Pronoun Interrogative Singular Adverbial	PTSA	ramdenad, sadaurad	რამდენად, სადაურად
Pronoun Interrogative Singular Suffixaufnahme-Ergative	PTSFE	sadaurisam, ramdenisam	სადაურისამ, რამდენისამ
Pronoun Interrogative Singular Suffixaufnahme-Dative-accusative	PTSFD	sadaurisas, ramdenisas	სადაურისას, რამდენისას
Pronoun Interrogative Singular Suffixaufnahme-Adverbial	PTSFA	sadaurisad, ramdenisad	სადაურისად, რამდენისად
Pronoun Interrogative Plural Nominative - absolutive	PTPN	romlebi, rogorebi	რომლები, როგორები
Pronoun Interrogative Plural Ergative	PTPE	romlebma, rogorebma	რომლებმა, როგორებმა
Pronoun Interrogative Plural Dative-accusative	PTPD	romlebs, rogoroebs	რომლებს, როგორებს
Pronoun Interrogative Plural Genitive	PTPG	sadaurebis, ranairebis	სადაურების, რანაირების

Pronoun Interrogative Plural Instrumental	PTPI	sadaurebit, ranairebit	სადაურებით, რანაირებით
Pronoun Interrogative Plural Adverbial	PTPA	sadaurebad, ranairebad	სადაურებად, რანაირებად
Pronoun Interrogative Plural Suffixaufnahme-Ergative	PTPFE	sadaurebisam, romlebisam	სადაურებისამ, რომლებისამ
Pronoun Interrogative Plural Suffixaufnahme- Dative-accusative	PTPFD	sadaurebisas, romlebisas	სადაურებისას, რომლებისას
Pronoun Interrogative Plural Suffixaufnahme-Adverbial	PTPFA	sadaurebisad, romlebisad	სადაურებისად, რომლებისად
Pronoun Possessive Singular First Person Nominative - absolutive	PV1SU	čem	ჩემ
Pronoun Possessive Singular First Person Nominative - absolutive	PV1SN	čemi	ჩემი
Pronoun Possessive Singular First Person Ergative	PV1SE	čemma	ჩემმა
Pronoun Possessive Singular First Person Dative-accusative	PV1SD	čems	ჩემს
Pronoun Possessive Singular First Person Genitive	PV1SG	čemis	ჩემის
Pronoun Possessive Singular First Person Instrumental	PV1SI	čemit	ჩემით
Pronoun Possessive Singular First Person	PV1SA	čemad	ჩემად

Adverbial			
Pronoun Possessive Singular First Person Vocative	PV1SV	čemo	ჩემო
Pronoun Possessive Singular First Person Suffixaufnahme-Ergative	PV1SFE	čemisam	ჩემისამ
Pronoun Possessive Singular First Person Suffixaufnahme- Dative-accusative	PV1SFD	čemisas	ჩემისას
Pronoun Possessive Singular First Person Suffixaufnahme-Adverbial	PV1SFA	čemisad	ჩემისად
Pronoun Possessive Singular First Person Suffixaufnahme-Vocative	PV1SFV	čemisav	ჩემისავ
Pronoun Possessive Singular Second Person Nominative - absolutive	PV2SN	šeni	შენი
Pronoun Possessive Singular Second Person Ergative	PV2SE	šenma	შენმა
Pronoun Possessive Singular Second Person Dative-accusative	PV2SD	šens	შენს
Pronoun Possessive Singular Second Person Genitive	PV2SG	šenis	შენის
Pronoun Possessive Singular Second Person Instrumental	PV22I	šenit	შენით
Pronoun Possessive Singular Second Person Adverbial	PV2SA	šenad	შენად
Pronoun Possessive Singular	PV2SFE	šenisam	შენისამ

Second Person Suffixaufnahme-Ergative			
Pronoun Possessive Singular Second Person Suffixaufnahme- Dative-accusative	PV2SFD	šenisas	შენისას
Pronoun Possessive Singular Second Person Suffixaufnahme-Adverbial	PV2SFA	šenisad	შენისად
Pronoun Possessive-reflexive Singular Third Person Zero	PRXU	tvis	თვის
Pronoun Possessive-Reflexive Singular Third Person Nominative - absolutive	PRXN	tavisi	თავისი
Pronoun Possessive-Reflexive Singular Third Person Ergative	PRXE	tavisma	თავისმა
Pronoun Possessive-Reflexive Singular Third Person Dative-accusative	PRXD	tavis	თავის
Pronoun Possessive-Reflexive Singular Third Person Genitive	PRXG	tavisis	თავისის
Pronoun Possessive-Reflexive Singular Third Person Instrumental	PRXI	tavisit	თავისით
Pronoun Possessive-Reflexive Singular Third Person Adverbial	PRXA	tavisad	თავისად
Pronoun Possessive Plural First Person Nominative - absolutive	PV1PN	čveni	ჩვენი
Pronoun Possessive	PV1PE	čvenma	ჩვენმა

Plural First Person Ergative			
Pronoun Possessive Plural First Person Dative-accusative	PV1PD	čvens	ჩვენს
Pronoun Possessive Plural First Person Genitive	PV1PG	čvenis	ჩვენის
Pronoun Possessive Plural First Person Instrumental	PV1PI	čvenit	ჩვენით
Pronoun Possessive Plural First Person Adverbial	PV1PA	čvenad	ჩვენად
Pronoun Possessive Plural First Person Vocative	PV1PV	čveno	ჩვენო
Pronoun Possessive Plural First Person Suffixaufnahme-Ergative	PV1FE	čvenisam	ჩვენისამ
Pronoun Possessive Plural First Person Suffixaufnahme-Dative- Accusative	PV1FD	čvenisas	ჩვენისას
Pronoun Possessive Plural First Person Suffixaufnahme-Adverbial	PV1FA	čvenisad	ჩვენისად
Pronoun Possessive Plural Second Person Nominative - absolutive	PV2PN	tkveni	თქვენი
Pronoun Possessive Plural Second Person Ergative	PV2PE	tkvenma	თქვენმა
Pronoun Possessive Plural Second Person Dative-accusative	PV2PD	tkvens	თქვენს
Pronoun Possessive Plural Second Person Genitive	PV2PG	tkvenis	თქვენის
Pronoun Possessive Plural Second Person Instrumental	PV2PI	tkvenit	თქვენით
Pronoun Possessive Plural Second Person Adverbial	PV2PA	tkvenad	თქვენით

Pronoun Possessive Plural Second Person Suffixaufnahme-Ergative	PV2FE	tkvenisam	თქვენისამ
Pronoun Possessive Plural Second Person Suffixaufnahme-Dative-Accusative	PV2FD	tkvenisas	თქვენისას
Pronoun Possessive Plural Second Person Suffixaufnahme-Adverbial	PV2FA	tkvenisad	თქვენისად
Pronoun Reciprocal Zero	PCU	ertmanet, erturt	ერთმანეთ, ერთურთ
Pronoun Reciprocal Nominative - absolutive	PCN	ertmaneti, erturti	ერთმანეთი, ერთურთი
Pronoun Reciprocal Ergative	PCE	ertmanetma, erturtma	ერთმანეთმა, ერთურთმა
Pronoun Reciprocal Dative-accusative	PCD	ertmanets, erturts	ერთმანეთს, ერთურთს
Pronoun Reciprocal Genitive	PCG	ertmanetis, erturtis	ერთმანეთის, ერთურთის
Pronoun Reciprocal Instrumental	PCI	ertmanetit, erturtit	ერთმანეთით, ერთურთით
Pronoun Reciprocal Adverbial	PCA	ertmanetad, ertimeored	ერთმანეთად, ერთიმეორედ
Pronoun Reciprocal Suffixaufnahme-Ergative	PCFE	ertmanetisam, ertimeorisam	ერთმანეთისამ, ერთიმეორისამ
Pronoun Reciprocal Suffixaufnahme- Dative-accusative	PCFD	ertmanetisas, ertimeorisas	ერთმანეთისას, ერთიმეორისას

Pronoun Reciprocal Suffixaufnahme-Adverbial	PCFA	ertmanetis ad, ertimeoris ad	ერთმანეთ ისად, ერთიმეორ ისად
Pronoun Intensive Zero Case	PFU	tvit, tviton	თვით, თვითონ
Pronoun Intensive Nominative - absolutive	PFN	titoeuli, qoveli	თითოეუ ლი, ყოველი
Pronoun Intensive Ergative	PFE	titoeulma, qovelma	თითოეუ ლმა, ყოველმა
Pronoun Intensive Dative-accusative	PFD	titoeuls, qovels	თითოეუ ლს, ყოველს
Pronoun Intensive Genitive	PFG	titoeulis, qovlis	თითოეუ ლის, ყოვლის
Pronoun Intensive Instrumental	PFI	titoeulit, sxvit	თითოეუ ლით, სხვით
Pronoun Intensive Adverbial	PFA	titoeulad, sxvad	თითოეუ ლად, სხვად
Pronoun Intensive Vocative	PFV	qovelo, titoeulo	ყოველო, თითოეუ ლო
Pronoun Intensive singular Nominative - absolutive	PFSN	sxva	სხვა
Pronoun Intensive singular Ergative	PFSE	sxvam	სხვამ
Pronoun Intensive singular Dative-accusative	PFSD	sxvas	სხვას
Pronoun Intensive singular Genitive	PFSG	sxvis	სხვის, სხვისი
Pronoun Intensive singular Instrumental	PFSI	sxvit	სხვით
Pronoun Intensive singular Adverbial	PFSA	sxvad	სხვად
Pronoun Intensive singular Vocative	PFSV	sxvav	სხვავ
Pronoun Intensive singular Suffixaufnahme-Dative-accusative	PFSFD	sxvisas	სხვისას

Pronoun Intensive Plural Suffixaufnahme-Adverbial	PFSFA	sxvisad	სხვისად
Pronoun Intensive Plural Nominative - absolutive	PFPN	sxvebi	სხვა
Pronoun Intensive Plural Ergative	PFPE	sxvebma	სხვამ
Pronoun Intensive Plural Dative-accusative	PFPD	sxvebs	სხვას
Pronoun Intensive Plural Genitive	PFPG	sxvebis	სხვის, სხვისი
Pronoun Intensive Plural Instrumental	PFPI	sxvebit	სხვით
Pronoun Intensive Plural Adverbial	PFPA	sxvebad	სხვად
Pronoun Intensive Plural Vocative	PFPV	sxvebo	სხვავ
Pronoun Intensive Plural Suffixaufnahme- Dative-accusative	PFPFD	sxvebisas	სხვისას
Pronoun Intensive Plural Suffixaufnahme-Adverbial	PFPFA	sxvebisad	სხვისად
Pronoun Indefinite Singular Zero	PISU	ert-ert	ერთ-ერთ
Pronoun Indefinite Singular Nominative - absolutive	PISN	viğac, zog	ვიღაც, ზოგი
Pronoun Indefinite Singular Nominative - absolutive	PISN	viğaca, zogi	ვიღაცა, ზოგი
Pronoun Indefinite Singular Ergative	PISE	viğacam, zogma	ვიღაცამ, ზოგმა
Pronoun Indefinite Singular Dative-accusative	PISD	viğacas, zogs	ვიღაცას, ზოგს
Pronoun Indefinite Singular Genitive	PISG	viğacis, zogis	ვიღაცის, ზოგის
Pronoun Indefinite Singular Instrumental	PISI	viğacit, zogit	ვიღაცით, ზოგით
Pronoun Indefinite Singular Adverbial	PISA	ramed, vinmed	რამედ, ვინმედ
Pronoun Indefinite Singular Suffixaufnahme-Ergative	PISFE	ramisam, zogisam	რამისამ, ზოგისამ
Pronoun Indefinite Singular Suffixaufnahme-Dative-Accusative	PISFD	ramisas, zogisas	რამისას, ზოგისას
Pronoun Indefinite Singular Suffixaufnahme-Adverbial	PISFA	ramisad	რამისად

Pronoun Indefinite Plural Nominative-absolutive	PIPN	zogiertebi, rameebi	ზოგიერთები, რამეები
Pronoun Indefinite Plural Ergative	PIPE	zogiertebma, rameebma	ზოგიერთებმა, რამეებმა
Pronoun Indefinite Plural Dative-accusative	PIPD	zogiertebs, rameebs	ზოგიერთებს, რამეებს
Pronoun Plural Singular Genitive	PIPG	vinmebis, rameebis	ვინმეების, რამეების
Pronoun Indefinite Plural Instrumental	PIPI	rameebit, ragaceebit	რამეებით, რაღაცეებით
Pronoun Indefinite Plural Adverbial	PIPA	zogiertebad, rameebad	ზოგიერთებად, რამეებად
Pronoun Indefinite Plural Suffixaufnahme-Ergative	PIPFE	rameebisam	რამეებისამ
Pronoun Indefinite Plural Suffixaufnahme-Dative-Accusative	PIPFD	rameebisas	რამეებისას
Pronoun Indefinite Plural Suffixaufnahme-Adverbial	PIPFA	rameebisad	რამეებისად
Pronoun Negative Zero case	PNU	aravin, nurvin	არავინ, ნურვინ
Pronoun Negative Nominative-absolutive	PNN	araferi, veraferi	არაფერი, ვერაფერი
Pronoun Negative Ergative	PNE	araferma, veraferma	არაფერმა, ვერაფერმა
Pronoun Negative Dative-accusative	PND	arafers, verafers	არაფერს, ვერაფერს
Pronoun Negative Genitive	PNG	araferis, veraferis	არაფრის, ვერაფრის
Pronoun Negative Instrumental	PNI	arafrit, verafrit	არაფრით, ვერაფრით
Pronoun Negative Adverbial	PNA	arafrad, verafrad	არაფრად, ვერფრად
Pronoun Negative Suffixaufnahme- Dative-accusative	PNFD	arafrisas, verafrisas	არაფრისას, ვერაფრისას

Table A3. 9: Tags Pronouns.

A4. Numeral

The attribute values for numerals are summed up in the table 4.1 below:

Value	i) type	ii) Number	iii) Case
1	Cardinal Simple	Singular	Zero Case
2	Cardinal Approximative	Plural	Nominative-absolutive
3	Ordinal		Ergative
4	Fraction		Dative-accusative
5			Genitive
6			Instrumental
7			Adverbial
8			Vocative
9			Suffixaufnahme: Genitive + Ergative
10			Suffixaufnahme: Genitive + Dative-accusative
11			Suffixaufnahme: Genitive + Adverbial
12			Suffixaufnahme: Genitive + Vocative

Table A4. 1: Attribute values for numerals.

In total, it gives 58 tags, as follows:

Description	TAG	Examples (Latin)	Examples (Georgian)
Numeral Cardinal Singular Zero	MCSU	sam, or	სამ, ორ
Numeral Cardinal Singular Nominative - absolutive	MCSN	sami, ori	სამი, ორი
Numeral Cardinal Singular Ergative	MCSE	samma, orma	სამმა, ორმა
Numeral Cardinal Singular Dative-accusative	MCSD	sams, ors	სამს, ორს
Numeral Cardinal Singular Genitive	MCSG	samis, samis	სამის, ორის
Numeral Cardinal Singular Instrumental	MCSI	samit, orit	სამით, ორით

Numeral Cardinal Singular Adverbial	MCSA	samad, orad	სამად, ორად
Numeral Cardinal Singular Vocative	MCSV	samo, oro	სამო, ორო
Numeral Cardinal Singular, Suffixaufnahme: Genitive + Ergative	MCSFE	samisam, orisam	სამისამ, ორისამ
Numeral Cardinal Singular, Suffixaufnahme: Genitive + Dative-accusative	MCSFD	samisas, orisas	სამისას, ორისას
Numeral Cardinal Singular, Suffixaufnahme: Genitive + Adverbial	MCSFA	samisad, orisad	სამისად, ორისად
Numeral Cardinal Singular, Suffixaufnahme: Genitive + Vocative	MCSFV	samisav, orisav	სამისავ, ორისავ
Numeral Ordinal Singular Zero	MOSU	pirvel	პირველ
Numeral Ordinal Singular Nominative - absolutive	MOSN	mesame, meore	მესამე, მეორე
Numeral Ordinal Singular Ergative	MOSE	mesamem, meorem	მესამემ, მეორემ
Numeral Ordinal Singular Dative-accusative	MOSD	mesames, meores	მესამეს, მეორეს
Numeral Ordinal Singular Genitive	MOSG	mesamis, meoris	მესამის, მეორის
Numeral Ordinal Singular Instrumental	MOSI	mesamit, meorit	მესამით, მეორით
Numeral Ordinal Singular Adverbial	MOSA	mesamed, meored	მესამედ, მეორედ
Numeral Ordinal Singular Vocative	MOSV	mesamev, meorev	მესამევ, მეორევ
Numeral Ordinal Singular, Suffixaufnahme: Genitive + Ergative	MOSFE	mesamisam, meorisam	მესამისამ, მეორისამ
Numeral Ordinal Singular, Suffixaufnahme: Genitive + Dative-accusative	MOSFD	mesamisas, meorisas	მესამისას, მეორისას
Numeral Ordinal Singular, Suffixaufnahme: Genitive + Vocative	MOSFV	mesamisv, meorisav	მესამისავ, მეორისავ
Numeral Fraction Singular Nominative -absolutive	MFSN	mesamedi, meoredi	მესამედი, მეორედი
Numeral Fraction Singular Ergative	MFSE	mesamedma, meoredma	მესამედმა, მეორედმა
Numeral Fraction Singular Dative-accusative	MGSD	mesameds, meoreds	მესამედს, მეორედს

Numeral Fraction Singular Genitive	MFSG	mesamedis, meoredis	მესამედის, მეორედის
Numeral Fraction Singular Instrumental	MFSI	mesamedit,meoredit	მესამედით, მეორედით
Numeral Fraction Singular Adverbial	MFSA	mesamedad, meoredad	მესამედად, მეორედად
Numeral Fraction Singular Vocative	MFSV	mesamedo, meoredo	მესამედო, მეორედო
Numeral Diminutive Singular Zero Case	MDSU	samiod, oriod	სამიოდ, ორიოდ
Numeral Diminutive Singular Nominative-absolutive	MDSN	samiode, oriode	სამიოდე, ორიოდე
Numeral Diminutive Singular Ergative	MDSE	samiodem, oriodem	სამიოდემ, ორიოდემ
Numeral Diminutive Singular Dative-accusative	MDSD	samiodes, oriodes	სამიოდეს, ორიოდეს
Numeral Diminutive Singular Genitive	MDSG	samiodis, oriodis	სამიოდის, ორიოდის
Numeral Diminutive Singular Instrumental	MDSI	samiodit, oriodit	სამიოდით, ორიოდით
Numeral Diminutive Singular Adverbial	MDSA	samioded, orioded	სამიოდედ, ორიოდედ
Numeral Cardinal Plural Nominative-absolutive	MCPN	samni, orebi	სამნი, ორები
Numeral Cardinal Plural Ergative	MCPE	samta, orta	სამთა, ორთა
Numeral Cardinal Plural Dative-accusative	MCPD	samebs, orebs	სამებს, ორებს
Numeral Cardinal Plural Genitive	MCPG	samebis, orebis	სამების, ორების
Numeral Cardinal Plural Instrumental	MCPI	samebit, orebit	სამებით, ორებით
Numeral Cardinal Plural Adverbial	MCPA	samebad, orebad	სამებად, ორებად
Numeral Cardinal Plural Vocative	MCPV	samebo, orebo	სამებო, ორებო
Numeral Ordinal Plural Nominative-absolutive	MOPN	mesameni, meoreni	მესამენი, მეორენი
Numeral Ordinal Plural Ergative	MOPE	mesameta, meoreta	მესამეთა, მეორეთა
Numeral Ordinal Plural Dative-accusative	MOPD	mesmeebs, meoreebs	მესამეებს, მეორეებს

Numeral Ordinal Plural Genitive	MOPG	mesameebis, meoreebis	მესამეების, მეორეების
Numeral Ordinal Plural Instrumental	MOPI	mesameebit, meoreebit	მესამეებით, მეორეებით
Numeral Ordinal Plural Adverbial	MOPA	mesameebad, meoreebad	მესამეებად, მეორეებად
Numeral Ordinal Plural Vocative	MOPV	mesameebo, meoreebo	მესამეებო, მეორეებო
Numeral Fraction Plural Nominative-absolutive	MFPN	mesamedni, meoreni	მესამედნი, მეორედნი
Numeral Fraction Plural Ergative	MFPE	mesamedta, meoreta	მესამედთა, მეორედთა
Numeral Fraction Plural Dative-accusative	MFPD	mesamedebs, meoredebs	მესამედებს, მეორედებს
Numeral Fraction Plural Genitive	MFPG	mesamedebis, meoredebis	მესამედების, მეორედების
Numeral Fraction Plural Instrumental	MFPI	mesamedebit, meoredebit	მესამედებით, მეორედებით
Numeral Fraction Plural Adverbial	MFPA	mesamedebad, meoredebad	მესამედებად, მეორედებად
Numeral Fraction Plural Vocative	MFPV	mesamedebo, meoredebo	მესამედებო, მეორედებო

Table A4. 2: Tags for Numerals.

A5. Adverb

There are three tags for adverbs.

Description	TAG	Examples (Latin)	Examples (Georgian)
General Adverb	RR	ak, amağam, cin	აქ, ამაღამ, წინ
Adverbs of Negation	RN	arsad, arasodes	არსად, არასოდეს
Interrogative Adverb	RI	rogor, rodis, sad	როგორ, როდის, სად

Table A5. 1: Tags for Adverbs.

A6. Conjunction

There are two tags for conjunctions.

Description	Tag	Examples (Latin)	Examples (Georgian)
Coordinating Conjunction Simple	CC	da, magram	და, მაგრამ
Subordinating Conjunction	CS	oğond, rom	ოღონდ, რომ

Table A6. 1: Tags for Conjunctions.

A7. Particle

There are six tags for particles.

Description	Tag	Examples (Latin)	Examples (Georgian)
General Particle	XX	netav, diax	ნეტავ, დიახ
Interrogative Particle	XI	gana, xom	განა, ხომ
Quotative Particle	XQ	metki, tko	მეთქი, თქო
Negative Particle	XN	ar, veğar	არ, ვეღარ
Modal Particle	XM	vinžlo, titkmis	ვინძლო, თითქმის
Nominal Particle	XO	-ca, -ve	-ცა, -ვე

Table A7. 1: POS-tags for Particles.

A8. Interjection

There is one tag for Interjection.

Description	TAG	Examples (Latin)	Examples (Georgian)
Interjection	UU	uime, eriha	უიმე, ერიჰა

Table A8. 1: Tags for Interjections.

A9. Postposition

There is one tag for Postposition.

Description	TAG	Examples (Latin)	Examples (Georgian)
Postposition	II	mier, gamo, -ši, -ze	მიერ, გამო, -ში, -ზე

Table A9. 1: Tags for Postpositions.

A10. Verb

There are two attribute value pairs for verbs.

Value	i) Argument Agreement	ii) Screeves
1	S_1	Present
2	S_2	Imperfect
3	S_3	Present Subjunctive
4	S_1^P	Future
5	S_2^P	Conditional
6	S_3^P	Future Subjunctive
7	S_1O_2	Aorist
8	S_2O_1	Aorist Subjunctive
9	S_3O_1	I Resultative
10	$S_3O_1^P$	II Resultative
11	S_3O_2	III Subjunctive
12	$S_3O_2^P$	
13	$S_2^PO_1$	
14	$S_2^PO_1^P$	
15	$S_3^PO_1$	
16	$S_3^PO_1^P$	
17	$S_1^PO_2$	
18	$S_3^PO_2$	
19	$S_2O_1^P$	

Table A10. 1: Attribute values for verbs.

This gives 209 tags for verbs.

Description	Tag	Examples (Latin)	Examples (Georgian)
Verb **S_1** Singular, Present Tense	V:1S:P	vizrdebi, vtbebi	ვიზრდები, ვთბები
Verb **S_1** Singular, Imperfect Tense	V:1S:I	vizrdebodi, vtbebodi	ვიზრდებოდი, ვთბებოდი
Verb **S_1** Singular, Present Subjunctive Tense	V:1S:B	vizrdebode, vtbebode	ვიზრდებოდე, ვთბებოდე

Verb **S₁** Singular, Future Tense	V:1S:F	gavizrdebi, gavtbebi	გავიზრდები, გავთბები
Verb **S₁** Singular, Conditional Tense	V:1S:C	gavizrdebodi, gavtbebodi	გავიზრდებოდი, გავთბებოდი
Verb **S₁** Singular, Future Subjunctive Tense	V:1S:D	gavizrdebode, gavtbebode	გავიზრდებოდე, გავთბებოდე
Verb **S₁** Singular, Aorist Tense	V:1S:A	gavizarde, gavtbi	გავიზარდე, გავთბი
Verb **S₁** Singular, Aorist Subjunctive Tense	V:1S:E	gavizardo, gavtbe	გავიზარდო, გავთბე
Verb **S₁** Singular, I Resultative Tense	V:1S:R	gavzrdilvar, gavmtbarvar	გავზრდილვარ, გავმთბარვარ
Verb **S₁** Singular, II Resultative Tense	V:1S:G	gavzrdiliqavi, gavmtbariqavi	გავზრდილიყავი, გავმთბარიყავი
Verb **S₁** Singular, III Subjunctive Tense	V:1S:S	gavzrdiliqo, gamtbarviqo	გავზრდილიყო, გავმთბარიყო
2			
Verb **S₁ᵖ** Plural, Present Tense	V:1P:P	vizrdebit, vtbebit	ვიზრდებით, ვთბებით
Verb **S₁ᵖ** Plural, Imperfect Tense	V:1P:I	vizrdebodit, vtbebodit	ვიზრდებოდით, ვთბებოდით
Verb **S₁ᵖ** Plural, Present Subjunctive Tense	V:1P:B	vizrdebodet, vtbebodet	ვიზრდებოდეთ, ვთბებოდეთ
Verb **S₁ᵖ** Plural, Future Tense	V:1P:F	gavizrdebit, gavtbebit	გავიზრდებით, გავთბებით
Verb **S₁ᵖ** Plural, Conditional Tense	V:1P:C	gavizrdebodit, gavtbebodit	გავიზრდებოდით, გავთბებოდით
Verb **S₁ᵖ** Plural, Future	V:1P:D	gavizrdebodet, gavtbebodet	გავიზრდებოდეთ, გავთბებოდეთ

Subjunctive Tense			
Verb **S₁ᵖ** Plural, Aorist Tense	V:1P:A	gavizardet, gavtbit	გავიზარდეთ, გავთბით
Verb **S₁ᵖ** Plural, Aorist Subjunctive Tense	V:1P:E	gavizardot, gavtbet	გავიზარდოთ, გავთბეთ
Verb **S₁ᵖ** Plural, I Resultative Tense	V:1P:R	gavzrdilvart, gavmtbarvart	გავზრდილვართ, გავმთბარვართ
Verb **S₁ᵖ** Plural, II Resultative Tense	V:1P:G	gavzrdiliqavit, gavmtbariqavit	გავზრდილიყავით, გავმთბარიყავით
Verb **S₁ᵖ** Plural, III Subjunctive Tense	V:1P:S	gavzrdiliqot, gamtbarviqot	გავზრდილიყო, გავმთბარიყოთ
3			
Verb **S₂** Singular, Present Tense	V:2S:P	izrdebi, tbebi	იზრდები, თბები
Verb **S₂** Singular, Imperfect Tense	V:2S:I	izrdebodi, tbebodi	იზრდებოდი, თბებოდი
Verb **S₂** Singular, Present Subjunctive Tense	V:2S:B	izrdebode, tbebode	იზრდებოდე, თბებოდე
Verb **S₂** Singular, Future Tense	V:2S:F	gaizrdebi, gatbebi	გაიზრდები, გათბები
Verb **S₂** Singular, Conditional Tense	V:2S:C	gaizrdebodi, gatbebodi	გაიზრდებოდი, გათბებოდი
Verb **S₂** Singular, Future Subjunctive Tense	V:2S:D	gaizrdebode, gatbebode	გაიზრდებოდე, გათბებოდე
Verb **S₂** Singular, Aorist Tense	V:2S:A	gaizarde, gatbi	გაიზარდე, გათბი
Verb **S₂** Singular, Aorist Subjunctive Tense	V:2S:E	gaizardo, gatbe	გაიზარდო, გათბე

Verb **S₂** Singular, I Resultative Tense	V:2S:R	gazrdilxar, gamtbarxar	გაზრდილხარ, გამთბარხარ
Verb **S₂** Singular, II Resultative Tense	V:2S:G	gazrdiliqavi, gamtbariqavi	გაზრდილიყავი, გამთბარიყავი
Verb **S₂** Singular, III Subjunctive Tense	V:2S:S	gazrdiliqo, gatbarviqo	გაზრდილიყო, გამთბარიყო
4			
Verb **S₂ᵖ** Plural, Present Tense	V:2P:P	izrdebit, tbebit	იზრდებით, თბებით
Verb **S₂ᵖ** Plural, Imperfect Tense	V:2P:I	izrdebodit, tbebodit	იზრდებოდით, თბებოდით
Verb **S₂ᵖ** Plural, Present Subjunctive Tense	V:2P:B	izrdebodet, tbebodet	იზრდებოდეთ, თბებოდეთ
Verb **S₂ᵖ** Plural, Future Tense	V:2P:F	gaizrdebit, gatbebit	გაიზრდებით, გათბებით
Verb **S₂ᵖ** Plural, Conditional Tense	V:2P:C	gaizrdebodit, gatbebodit	გაიზრდებოდით, გათბებოდით
Verb **S₂ᵖ** Plural, Future Subjunctive Tense	V:2P:D	gaizrdebodet, gatbebodet	გაიზრდებოდეთ, გათბებოდეთ
Verb **S₂ᵖ** Plural, Aorist Tense	V:2P:A	gaizardet, gatbit	გაიზარდეთ, გათბით
Verb **S₂ᵖ** Plural, Aorist Subjunctive Tense	V:2P:E	gaizardot, gatbet	გაიზარდოთ, გათბეთ
Verb **S₂ᵖ** Plural, I Resultative Tense	V:2P:R	gazrdilxart, gamtbarxart	გაზრდილხართ, გამთბარხართ
Verb **S₂ᵖ** Plural, II Resultative Tense	V:2P:G	gazrdiliqavit, gamtbariqavit	გაზრდილიყავით, გამთბარიყავით

Verb **S_{2P}** Plural, III Subjunctive Tense	V:2P:S	gazrdiliqot, gamtbariqot	გაზრდილიყოთ, გამთბარიყოთ
5			
Verb **S_3** Singular, Present Tense	V:3S:P	izrdeba, tbeba	იზრდება, თბება
Verb **S_3** Singular, Imperfect Tense	V:3S:I	izrdeboda, tbeboda	იზრდებოდა, თბებოდა
Verb **S_3** Singular, Present Subjunctive Tense	V:3S:B	izrdebodes, tbebodes	იზრდებოდეს, თბებოდეს
Verb **S_3** Singular, Future Tense	V:3S:F	gaizrdeba, gatbeba	გაიზრდება, გათბება
Verb **S_3** Singular, Conditional Tense	V:3S:C	gaizrdeboda, gatbeboda	გაიზრდებოდა, გათბებოდა
Verb **S_3** Singular, Future Subjunctive Tense	V:3S:D	gaizrdebodes, gatbebodes	გაიზრდებოდეს, გათბებოდეს
Verb **S_3** Singular, Aorist Tense	V:3S:A	gaizarda, gatba	გაიზარდა, გათბა
Verb **S_3** Singular, Aorist Subjunctive Tense	V:3S:E	gaizardos, gatbes	გაიზარდოს, გათბეს
Verb **S_3** Singular, I Resultative Tense	V:3S:R	gazrdila, gamtbara	გაზრდილა, გამთბარა
Verb **S_3** Singular, II Resultative Tense	V:3S:G	gazrdiliqo, gamtbariqo	გაზრდილიყო, გამთბარიყო
Verb **S_3** Singular, III Subjunctive Tense	V:3S:S	gazrdiliqos, gatbariqos	გაზრდილიყოს, გამთბარიყოს
6			
Verb **S_{3P}** Plural, Present Tense	V:3P:P	izrdebian, tbebian	იზრდებიან, თბებიან

Verb **S_3^p** Plural, Imperfect Tense	V:3P:I	izrdebodnen, tbebodnen	იზრდებოდნენ, თბებოდნენ
Verb **S_3^p** Plural, Present Subjunctive Tense	V:3P:B	izrdebodnen, tbebodnen	იზრდებოდნენ, თბებოდნენ
Verb **S_3^p** Plural, Future Tense	V:3P:F	gaizrdebian, gatbebian	გაიზრდებიან, გათბებიან
Verb **S_3^p** Plural, Conditional Tense	V:3P:C	gaizrdebodnen, gatbebodnen	გაიზრდებოდნენ, გათბებოდნენ
Verb **S_3^p** Plural, Future Subjunctive Tense	V:3P:D	gaizrdebodnen, gatbebodnen	გაიზრდებოდნენ, გათბებოდნენ
Verb **S_3^p** Plural, Aorist Tense	V:3P:A	gaizardnen, gatbnen	გაიზარდნენ, გათბნენ
Verb **S_3^p** Plural, Aorist Subjunctive Tense	V:3P:E	gaizardon, gatbnen	გაიზარდონ, გათბნენ
Verb **S_3^p** Plural, I Resultative Tense	V:3P:R	gazrdilan, gamtbaran	გაზრდილან, გამთბარან
Verb **S_3^p** Plural, II Resultative Tense	V:3P:G	gazrdiliqvnen, gamtbariqvnen	გაზრდილიყვნენ გამთბარიყავნენ
Verb **S_3^p** Plural, III Subjunctive Tense	V:3P:S	gazrdiliqon, gatbariqon	გაზრდილიყონ, გამთბარიყონ
7			
Verb **S_1O_2** Singular, Present Tense	V:1S2S:P	gzrdi, gatbob	გზრდი, გათბობ
Verb **S_1O_2** Singular, Imperfect Tense	V:1S2S:I	gzrdidi, gatbobdi	გზრდიდი, გათბობდი
Verb **S_1O_2** Singular, Present Subjunctive Tense	V:1S2S:B	gzrdide, gatbobde	გზრდიდე, გათბობდე

Verb **S_1O_2** Singular, Future Tense	V:1S2S:F	gagzrdi, gagatbob	გაგზრდი, გაგათბობ
Verb **S_1O_2** Singular, Conditional Tense	V:1S2S:C	gagzrdidi, gagatbobdi	გაგზრდიდი, გაგათბობდი
Verb **S_1O_2** Singular, Future Subjunctive Tense	V:1S2S:D	gagzrdide, gagatbobde	გაგზრდიდე, გაგათბობდე
Verb **S_1O_2** Singular, Aorist Tense	V:1S2S:A	gagzarde, gagatbe	გაგზარდე, გაგათბე
Verb **S_1O_2** Singular, Aorist Subjunctive Tense	V:1S2S:E	gagzardo, gagatbo	გაგზარდო, გაგათბო
Verb **S_1O_2** Singular, I Resultative Tense	V:1S2S:R	gamizrdixar, gamitbixar	გამიზრდიხარ, გამითბიხარ
Verb **S_1O_2** Singular, II Resultative Tense	V:1S2S:G	gamezarde, gametbe	გამეზარდე, გამეთბე
Verb **S_1O_2** Singular, III Subjunctive Tense	V:1S2S:S	gamezardo, gametbo	გამეზარდო, გამეთბო
8			
Verb **S_2O_1** Singular, Present Tense	V:2S1S:P	mzrdi, matbob	მზრდი, მათბობ
Verb **S_2O_1** Singular, Imperfect Tense	V:2S1S:I	mzrdidi, matbobdi	მზრდიდი, მათბობდი
Verb **S_2O_1** Singular, Present Subjunctive Tense	V:2S1S:B	mzrdide, matbobde	მზრდიდე, მათბობდე
Verb **S_2O_1** Singular, Future Tense	V:2S1S:F	gamzrdi, gamatbob	გამზრდი, გამათბობ
Verb **S_2O_1** Singular, Conditional Tense	V:2S1S:C	gamzrdidi, gamatbobdi	გამზრდიდი, გამათბობდი

Verb **S_2O_1** Singular, Future Subjunctive Tense	V:2S1S:D	gamzrdide, gamatbobde	გამზრდიდე, გამათბობდე
Verb **S_2O_1** Singular, Aorist Tense	V:2S1S:A	gamzarde, gamatbe	გამზარდე, გამათბე
Verb **S_2O_1** Singular, Aorist Subjunctive Tense	V:2S1S:E	gamzardo, gamatbo	გამზარდო, გამათბო
Verb **S_2O_1** Singular, I Resultative Tense	V:2S1S:R	gagizrdivar, gagitbivar	გაგიზრდივარ, გაგითბივარ
Verb **S_2O_1** Singular, II Resultative Tense	V:2S1S:G	gagezarde, gagetbe	გაგეზარდე, გაგეთბე
Verb **S_2O_1** Singular, III Subjunctive Tense	V:2S1S:S	gagezardo, gagetbo	გაგეზარდო, გაგეთბო
9			
Verb **S_3O_1** Singular, Present Tense	V:3S1S:P	mzrdis, matbobs	მზრდის, მათბობს
Verb **S_3O_1** Singular, Imperfect Tense	V:3S1S:I	mzrdida, matbobda	მზრდიდა, მათბობდა
Verb **S_3O_1** Singular, Present Subjunctive Tense	V:3S1S:B	mzrdides, matbobdes	მზრდიდეს, მათბობდეს
Verb **S_3O_1** Singular, Future Tense	V:3S1S:F	gamzrdis, gamatbobs	გამზრდის, გამათბობს
Verb **S_3O_1** Singular, Conditional Tense	V:3S1S:C	gamzrdida, gamatbobda	გამზრდიდა, გამათბობდა
Verb **S_3O_1** Singular, Future Subjunctive Tense	V:3S1S:D	gamzrdides, gamatbobdes	გამზრდიდეს, გამათბობდეს
Verb **S_3O_1** Singular, Aorist Tense	V:3S1S:A	gamzarda, gamatbo	გამზარდა, გამათბო

Verb **S_3O_1** Singular, Aorist Subjunctive Tense	V:3S1S:E	gamzardos, gamatbos	გამზარდოს, გამათბოს
Verb **S_3O_1** Singular, I Resultative Tense	V:3S1S:R	gavuzrdivar, gavutbivar	გავუზრდივარ, გავუთბივარ
Verb **S_3O_1** Singular, II Resultative Tense	V:3S1S:G	gavezarde, gavetbe	გავეზარდე, გავეთბე
Verb **S_3O_1** Singular, III Subjunctive Tense	V:3S1S:S	gavezardo, gavetbo	გავეზარდო, გავეთბო
10			
Verb **$S_3O_1^p$**, S Singular / O Plural, Present Tense	V:3S1P:P	gvzrdis, gvatbobs	გვზრდის, გვათბობს
Verb **$S_3O_1^p$** S Singular / O Plural, Imperfect Tense	V:3S1P:I	gvzrdida, gvatbobda	გვზრდიდა, გვათბობდა
Verb **$S_3O_1^p$** S Singular / O Plural, Present Subjunctive Tense	V:3S1P:B	gvzrdides, gvatbobdes	გვზრდიდეს, გვათბობდეს
Verb **$S_3O_1^p$** S Singular / O Plural, Future Tense	V:3S1P:F	gagvzrdis, gagvatbobs	გაგვზრდის, გაგვათბობს
Verb **$S_3O_1^p$** S Singular / O Plural, Conditional Tense	V:3S1P:C	gagvzrdida, gagvatbobda	გაგვზრდიდა, გაგვათბობდა
Verb **$S_3O_1^p$** S Singular / O Plural, Future Subjunctive Tense	V:3S1P:D	gagvzrdides, gagvatbobdes	გაგვზრდიდეს, გაგვათბობდეს
Verb **$S_3O_1^p$** S Singular / O Plural , Aorist Tense	V:3S1P:A	gagvzarda, gagvatbo	გაგვზარდა, გამათბო

Verb **$S_3O_1^p$** S Singular / O Plural, Aorist Subjunctive Tense	V:3S1P:E	gagvzardos, gagvatbos	გაგვზარდოს, გაგვათბოს
Verb **$S_3O_1^p$** S Singular / O Plural, I Resultative Tense	V:3S1P:R	gavuzrdivart, gavutbivart	გავუზრდივართ, გავუთბივართ
Verb **$S_3O_1^p$** S Singular / O Plural, II Resultative Tense	V:3S1P:G	gavezardet, gavetbet	გავეზარდეთ, გავეთბეთ
Verb **$S_3O_1^p$** S Singular / O Plural, III Subjunctive Tense	V:3S1P:S	gavezardot, gavetbot	გავეზარდოთ, გავეთბოთ
11			
Verb **S_3O_2** Singular, Present Tense	V:3S2S:P	gzrdis, gatbobs	გზრდის, გათბობს
Verb **S_3O_2** Singular, Imperfect Tense	V:3S2S:I	gzrdida, gatbobda	გზრდიდა, გათბობდა
Verb **S_3O_2** Singular, Present Subjunctive Tense	V:3S2S:B	gzrdides, gatbobdes	გზრდიდეს, გათბობდეს
Verb **S_3O_2** Singular, Future Tense	V:3S2S:F	gagzrdis, gagatbobs	გაგზრდის, გაგათბობს
Verb **S_3O_2** Singular, Conditional Tense	V:3S2S:C	gagzrdida, gagatbobda	გაგზრდიდა, გაგათბობდა
Verb **S_3O_2** Singular, Future Subjunctive Tense	V:3S2S:D	gagzrdides, gagatbobdes	გაგზრდიდეს, გაგათბობდეს
Verb **S_3O_2** Singular, Aorist Tense	V:3S2S:A	gagzarda, gagatbo	გაგზარდა, გაგათბო
Verb **S_3O_2** Singular, Aorist Subjunctive Tense	V:3S2S:E	gagzardos, gagatbos	გაგზარდოს, გაგათბოს

Verb **S_3O_2** Singular, I Resultative Tense	V:3S2S:R	gauzrdixar, gautbixar	გაუზრდიხარ, გაუთბიხარ
Verb **S_3O_2** Singular, II Resultative Tense	V:3S2S:G	gaezarde, gaetbe	გაეზარდე, გაეთბე
Verb **S_3O_2** Singular, III Subjunctive Tense	V:3S2S:S	gaezardo, gaetbo	გაეზარდო, გაეთბო
12			
Verb **$S_3O_2^P$**, S Singular / O Plural, Present Tense	V:3S2P:P	gzrdit, gatbobt	გზრდით, გათბობთ
Verb **$S_3O_2^P$,** S Singular / O Plural, Imperfect Tense	V:3S2P:I	gzrdidat, gatbobdat	გზრდიდათ, გათბობდათ
Verb **$S_3O_2^P$**, S Singular / O Plural, Present Subjunctive Tense	V:3S2P:B	gzrdidet, gatbobdet	გზრდიდეთ, გათბობდეთ
Verb **$S_3O_2^P$**, S Singular / O Plural, Future Tense	V:3S2P:F	gagzrdit, gagatbobt	გაგზრდით, გაგათბობთ
Verb **$S_3O_2^P$**, S Singular / O Plural, Conditional Tense	V:3S2P:C	gagzrdidat, gagatbobdat	გაგზრდიდათ, გაგათბობდათ
Verb **$S_3O_2^P$**, S Singular / O Plural, Future Subjunctive Tense	V:3S2P:D	gagzrdidet, gagatbobdet	გაგზრდიდეთ, გაგათბობდეთ
Verb **$S_3O_2^P$**, S Singular / O Plural , Aorist Tense	V:3S2P:A	gagzardat, gagatbot	გაგზარდათ, გაგათბოთ
Verb **$S_3O_2^P$**, S Singular / O Plural, Aorist Subjunctive Tense	V:3S2P:E	gagzardot, gagatbot	გაგზარდოთ, გაგათბოთ

Verb **$S_3O_2^P$**, S Singular / O Plural, I Resultative Tense	V:3S2P:R	gauzrdixart, gautbixart	გაუზრდიხართ, გაუთბიხართ
Verb **$S_3O_2^P$**, S Singular / O Plural, II Resultative Tense	V:3S2P:G	gaezardet, gaetbet	გაეზარდეთ, გაეთბეთ
Verb **$S_3O_2^P$**, S Singular / O Plural, III Subjunctive Tense	V:3S2P:S	gaezardot, gaetbot	გაეზარდოთ, გაეთბოთ
13			
Verb **$S_2^PO_1$,** S Plural, O Singular, Present Tense	V:2P1S:P	mzrdit, matbobt	მზრდით, მათბობთ
Verb **$S_2^PO_1$,** S Plural, O Singular, Imperfect Tense	V:2P1S:I	mzrdidit, matbobdit	მზრდიდით, მათბობდით
Verb **$S_2^PO_1$,** S Plural, O Singular, Present Subjunctive Tense	V:2P1S:B	mzrdidet, matbobdet	მზრდიდეთ, მათბობდეთ
Verb **$S_2^PO_1$,** S Plural, O Singular, Future Tense	V:2P1S:F	gamzrdit, gamatbobt	გამზრდით, გამათბობთ
Verb **$S_2^PO_1$,** S Plural, O Singular, Conditional Tense	V:2P1S:C	gamzrdidit, gamatbobdit	გამზრდიდით, გამათბობდით
Verb **$S_2^PO_1$,** S Plural, O Singular, Future Subjunctive Tense	V:2P1S:D	gamzrdidet, gamatbobdet	გამზრდიდეთ, გამათბობდეთ
Verb **$S_2^PO_1$,** S Plural, O Singular, Aorist Tense	V:2P1S:A	gamzardet, gamatbet	გამზარდეთ, გამათბეთ

Verb **$S_2^PO_1$,** S Plural, O Singular, Aorist Subjunctive Tense	V:2P1S:E	gamzardot, gamatbot	გამზარდოთ, გამათბოთ
Verb **$S_2^PO_1$,** S Plural, O Singular, I Resultative Tense	V:2P1S:R	gagizrdivart, gagitbivart	გაგიზრდივართ, გაგითბივართ
Verb **$S_2^PO_1$,** S Plural, O Singular, II Resultative Tense	V:2P1S:G	gagezardet, gagetbet	გაგეზარდეთ, გაგეთბეთ
Verb **$S_2^PO_1$,** S Plural, O Singular, III Subjunctive Tense	V:2P1S:S	gagezardot, gagetbot	გაგეზარდოთ, გაგეთბოთ
14			
Verb **$S_2^PO_1^P$,** Plural, Present Tense	V:2P1P:P	gvzrdit, gvatbobt	გვზრდის, გვათბობს
Verb **$S_2^PO_1^P$,** Plural, Imperfect Tense	V:2P1P:I	gvzrdidit, gvatbobdit	გვზრდიდით, გვათბობდით
Verb **$S_2^PO_1^P$,** Plural, Present Subjunctive Tense	V:2P1P:B	gvzrdidet, gvatbobdet	გვზრდიდეთ, გვათბობდეთ
Verb **$S_2^PO_1^P$,** Plural, Future Tense	V:2P1P:F	gagvzrdit, gagvatbobt	გაგვზრდით, გაგვათბობთ
Verb **$S_2^PO_1^P$,** Plural, Conditional Tense	V:2P1P:C	gagvzrdidit, gagvatbobdit	გაგვზრდიდით, გაგვათბობდით
Verb **$S_2^PO_1^P$,** Plural, Future Subjunctive Tense	V:2P1P:D	gagvzrdidet, gagvatbobdet	გაგვზრდიდეთ, გაგვათბობდეთ
Verb **$S_2^PO_1^P$,** Plural, Aorist Tense	V:2P1P:A	gagvzardet, gagvatbet	გაგვზარდეთ, გამათბეთ
Verb **$S_2^PO_1$, $S_2^PO_1^P$,** Plural, Aorist Subjunctive Tense	V:2P1P:E	gagvzardot, gagvatbot	გაგვზარდოთ, გაგვათბოთ

Verb **$S_2^PO_1^P$,** Plural, I Resultative Tense	V:2P1P:R	gagizrdivart, gagitbivart	გაგიზრდივართ, გაგითბივართ
Verb **$S_2^PO_1^P$,** Plural, II Resultative Tense	V:2P1P:G	gagezardet, gagetbet	გაგეზარდეთ, გაგეთბეთ
Verb **$S_2^PO_1^P$,** Plural, III Subjunctive Tense	V:2P1P:S	gagezardot, gagetbot	გაგეზარდოთ, გაგეთბოთ
15			
Verb **$S_3^PO_1$**, S Plural / O Singular, Present Tense	V:3P1S:P	mzrdian, matboben	მზრდიან, მათბობენ
Verb **$S_3^PO_1$**, S Plural / O Singular, Imperfect Tense	V:3P1S:I	mzrdidnen, matbobdnen	მზრდიდნენ, მათბობდნენ
Verb **$S_3^PO_1$**, S Plural / O Singular, Present Subjunctive Tense	V:3P1S:B	mzrdidnen, matbobdnen	მზრდიდნენ, მათბობდნენ
Verb **$S_3^PO_1$**, S Plural / O Singular, Future Tense	V:3P1S:F	gamzrdian, gamatboben	გამზრდიან, გამათბობენ
Verb **$S_3^PO_1$**, S Plural / O Singular, Conditional Tense	V:3P1S:C	gamzrdidnen, gamatbobdnen	გამზრდიდნენ, გამათბობდნენ
Verb **$S_3^PO_1$**, S Plural / O Singular, Future Subjunctive Tense	V:3P1S:D	gamzrdidnen, gamatbobdnen	გამზრდიდნენ, გამათბობდნენ
Verb **$S_3^PO_1$**, S Plural / O Singular, Aorist Tense	V:3P1S:A	gamzardes, gamatbes	გამზარდეს, გამათბეს
Verb **$S_3^PO_1$**, S Plural / O Singular, Aorist Subjunctive Tense	V:3P1S:E	gamzardon, gamatbon	გამზარდონ, გამათბონ

Verb **$S_3^PO_1$**, S Plural / O Singular, I Resultative Tense	V:3P1S:R	gavuzrdivart, gavutbivart	გავუზრდივართ, გავუთბივართ
Verb **$S_3^PO_1$**, S Plural / O Singular, II Resultative Tense	V:3P1S:G	gavezardet, gavetbet	გავეზარდეთ, გავეთბეთ
Verb **$S_3^PO_1$**, S Plural / O Singular, III Subjunctive Tense	V:3P1S:S	gavezardot, gavetbot	გავეზარდოთ, გავეთბოთ
16			
Verb **$S_3^PO_1^P$**, Plural, Present Tense	V:3P1P:P	gvzrdian, gvatboben	გვზრდიან, გვათბობენ
Verb **$S_3^PO_1^P$**, Plural, Imperfect Tense	V:3P1P:I	gvzrdidnen, gvatbobdnen	მგვზრდიდნენ, გვათბობდნენ
Verb **$S_3^PO_1^P$**, Plural, Present Subjunctive Tense	V:3P1P:B	gvzrdidnen, gvatbobdnen	გვზრდიდნენ, გვათბობდნენ
Verb **$S_3^PO_1^P$**, Plural, Future Tense	V:3P1P:F	gagvzrdian, gagvatboben	გაგვზრდიან, გაგვათბობენ
Verb **$S_3^PO_1^P$**, Plural, Conditional Tense	V:3P1P:C	gagvzrdidnen, gagvatbobdnen	გაგვზრდიდნენ, გაგვათბობდნენ
Verb **$S_3^PO_1^P$**, Plural, Future Subjunctive Tense	V:3P1P:D	gagvzrdidnen, gagvatbobdnen	გაგვზრდიდნენ, გაგვათბობდნენ
Verb **$S_3^PO_1^P$**, Plural, Aorist Tense	V:3P1P:A	gagvzardes, gagvatbes	გაგვზარდეს, გაგვათბეს
Verb **$S_3^PO_1^P$**, Plural, Aorist Subjunctive Tense	V:3P1P:E	gagvzardon, gagvatbon	გაგვზარდონ, გაგვათბონ
Verb **$S_3^PO_1^P$**, Plural, I Resultative Tense	V:3P1P:R	gavuzrdivart, gavutbivart	გავუზრდივართ, გავუთბივართ

Verb **$S_3^P O_1^P$**, Plural, II Resultative Tense	V:3P1P:G	gavezardet, gavetbet	გავეზარდეთ, გავეთბეთ
Verb **$S_3^P O_1^P$**, Plural, III Subjunctive Tense	V:3P1P:S	gavezardot, gavetbot	გავეზარდოთ, გავეთბოთ
17			
Verb **$S_1^P O_2$**, S Plural / O Singular, Present Tense	V:1P2S:P	gezrdebit	გეზრდებით
Verb **$S_1^P O_2$**, S Plural / O Singular, Imperfect Tense	V:1P2S:I	gezrdebodit	გეზრდებოდით
Verb **$S_1^P O_2$**, S Plural / O Singular, Present Subjunctive Tense	V:1P2S:B	gezrdebodet	გეზრდებოდეთ
Verb **$S_1^P O_2$**, S Plural / O Singular, Future Tense	V:1P2S:F	gagezrdebit	გაგეზრდებით
Verb **$S_1^P O_2$**, S Plural / O Singular, Conditional Tense	V:1P2S:C	gagezrdebodit	გაგეზრდებოდით
Verb **$S_1^P O_2$**, S Plural / O Singular, Future Subjunctive Tense	V:1P2S:D	gagezrdbodet	გაგეზრდებოდეთ
Verb **$S_1^P O_2$**, S Plural / O Singular, Aorist Tense	V:1P2S:A	gagezardet	გაგეზარდეთ
Verb **$S_1^P O_2$**, S Plural / O Singular, Aorist Subjunctive Tense	V:1P2S:E	gagezardot	გაგეზარდოთ
Verb **$S_1^P O_2$**, S Plural / O Singular, I	V:1P2S:R	gagizrdivart	გაგიზდივართ

Resultative Tense			
Verb **$S_1^PO_2$**, S Plural / O Singular, II Resultative Tense	V:1P2S:G	gagezardet	გაგეზარდეთ
Verb **$S_1^PO_2$**, S Plural / O Singular, III Subjunctive Tense	V:1P2S:S	gagezardot	გაგეზარდოთ
18			
Verb **$S_3^PO_2$**, S Plural / O Singular, Present Tense	V:3P2S:P	gezrdebian	გეზრდებიან
Verb **$S_3^PO_2$**, S Plural / O Singular, Imperfect Tense	V:3P2S:I	gezrdebodnen	გეზრდებოდნენ
Verb **$S_3^PO_2$**, S Plural / O Singular, Present Subjunctive Tense	V:3P2S:B	gezrdebodnen	გეზრდებოდნენ
Verb **$S_3^PO_2$**, S Plural / O Singular, Future Tense	V:3P2S:F	gagezrdebian	გაგეზრდებიან
Verb **$S_3^PO_2$**, S Plural / O Singular, Conditional Tense	V:3P2S:C	gaezrdebodit	გაეზრდებოდით
Verb **$S_3^PO_2$**, S Plural / O Singular, Future Subjunctive Tense	V:3P2S:D	gaezrdbodet	გაეზრდებოდეთ
Verb **$S_3^PO_2$**, S Plural / O Singular, Aorist Tense	V:3P2S:A	gaezardet	გაეზარდეთ
Verb **$S_3^PO_2$**, S Plural / O Singular, Aorist Subjunctive Tense	V:3P2S:E	gaezardot	გაეზარდოთ

Verb **$S_3^PO_2$**, S Plural / O Singular, I Resultative Tense	V:3P2S:R	miumsgavsebixart	მიუმსგავსებიხართ
Verb **$S_3^PO_2$**, S Plural / O Singular, II Resultative Tense	V:3P2S:G	miemsgavsebinet	მიემსგავსებინეთ
Verb **$S_3^PO_2$**, S Plural / O Singular, III Subjunctive Tense	V:3P2S:S	miemsgavsebinot	მიემსგავსებინოთ
19			
Verb **$S_2O_1^P$**, S singular / O Plural, Present Tense	V:2S1P:P	gvzrdi	გვზრდი
Verb **$S_2O_1^P$**, S singular / O Plural , Imperfect Tense	V:2S1P:I	gvzrdidi	გვზრდიდი
Verb **$S_2O_1^P$**, S singular / O Plural , Present Subjunctive Tense	V:2S1P:B	gvzrdide	გვზრდიდე
Verb **$S_2O_1^P$**, S singular / O Plural , Future Tense	V:2S1P:F	gagvzrdi	გაგვზრდი
Verb **$S_2O_1^P$**, S singular / O Plural , Conditional Tense	V:2S1P:C	gagvzrdidi	გაგვზრდიდი
Verb **$S_2O_1^P$**, S singular / O Plural , Future Subjunctive Tense	V:2S1P:D	gagvzridide	გაგვზრდიდე
Verb **$S_2O_1^P$**, S singular / O Plural , Aorist Tense	V:2S1P:A	gagvzarde	გაგვზარდე
Verb **$S_2O_1^P$**, S singular / O Plural , Aorist	V:2S1P:E	gagvzardo	გაგვზარდო

Subjunctive Tense			
Verb $\mathbf{S_2O_1^P}$, S singular / O Plural , I Resultative Tense	V:2S1P:R	gagizrdivart	გაგიზრდივართ
Verb $\mathbf{S_2O_1^P}$, S singular / O Plural , II Resultative Tense	V:2S1P:G	gagezarde	გაგეზარდეთ
Verb $\mathbf{S_2O_1^P}$, S singular / O Plural , III Subjunctive Tense	V:2S1P:S	gagezardo	გაგეზარდოთ

Table A10. 2: Tags for verbs.

A11. Copula

There is one tag for the [-a] affixal copula in Georgian.

Description	TAG	Examples (Latin)	Examples (Georgian)
Copula	AUX	-a	-ა

Table A11. 1: Tags for Copula

A12. Residual

There are six tags for residuals.

Description	TAG	Examples (Latin)	Examples (Georgian)
Foreign Word	FF	news, job	-
Formula (e.g. Mathematical)	FO	2 × 2	-
Letter of the Alphabet	FZ	b, g, d	ბ, გ, დ
Abbreviation and Acronym: in Georgian	FG	šss, ašš	შსს, აშშ
Abbreviation and Acronym: English (other)	FE	LOL	-
Other unclassifiable non-Georgian element / transliteration variant of a foreign word	FU	cool	ქუულ

Table A12. 1: Tags for Residuals.

A13. Punctuation

There are four tags for punctuation.

Description	TAG	Examples
Sentence final	YF	. ? ! ?! …
Sentence medial	YM	, : ; - * / \ <> ~ «
Quotations	YQ	" „ “ ”
Brackets	YB	() [] {}

Table A13. 1: Tags for Punctuation.

Appendix B

Corpus based wordlist of vowel syncopation in Georgian

Table B. 1: Corpus based list of non-syncopated words in Georgian

Word	Observed frequency
აბაზანა	442
აგრესორი	145
ადმინისტრატორი	200
ადმირალი	81
აეროდრომი	204
ავატარი	160
ავიალაინერი	94
ავტობანი	71
ავტომაგისტრალი	956
ავტომანქანა	3275
ავტორი	7589
ავღანელი	74
ავჭალა	523
ათონელი	171
აკლდამა	58
აკუმულატორი	101
ალბომი	1374
ალიგატორი	50
ალმოდოვარი	111
ამაზონი	179
ამაღლობელი	78
ამირანი	203
ამონაწერი	243
ამოცანა	1449
ანალიზატორი	74
ანაფორა	71
ანბანი	602
ანდერსენი	75
ანდერსონი	130
ანტენა	183
აპოლონი	124
არასრულწლოვანი	239
არსენალი	1151
არქიტექტორი	384
არჩევანი	6644
ასტრონომი	66
ატენა	344
აუდიოჩანაწერი	100
აუდიტორია	276
აქლემი	256
აქციონერი	71
აღდგომა	2105
აღმზრდელი	119
აღმოჩენა	1504
აღსაზრდელი	61
აღწერა	2062
აყალმაყალი	63
აცეტონი	84
ახალგორი	1129
ახტალა	124
ბაბილონი	698
ბადაგონი	114
ბაიკერი	60
ბაირონი	145
ბალონი	72
ბანერი	256
ბარათელი	60
ბარომეტრი	69
ბარსელონა	55
ბატონი	954
ბგერა	411
ბეთჰოვენი	222
ბეისბოლი	231
ბეკონი	138
ბესტსელერი	96
ბეტმენი	113
ბეტონი	1042
ბზარი	52
ბიბლიოთეკარი	63
ბიზნესმენი	2147
ბიზნესსექტორი	259
ბიულეტენი	199

ბიუსტჰალტერი	125
ბიძაჩემი	162
ბლოგერი	600
ბოდლერი	113
ბორჯომი	2059
ბოსტონი	530
ბოსფორი	168
ბოქვენი	76
ბრალი	57
ბრაუზერი	89
ბრიუსელი	476
ბროლა	17648
ბულბული	94
ბულვარი	424
ბუნკერი	78
გადადგომა	3550
გადამხდელი	547
გადასასვლელი	51
გადასახდელი	57
გადაცემა	13041
გავლენის	7273
გამგებელი	143
გამზრდელი	123
გამონაყარი	224
გამოსავალი	82
გამოფენა	1959
გამოცემა	5016
გამყიდველი	260
განაჩენი	2202
განსასჯელი	301
განსაცდელი	698
განცხრომა	51
გარგარი	186
გარნიზონი	101
გასასვლელი	100
გეგეჭკორი	108
გელოვანი	125
გენდერი	401
გენდირექტორი	92
გენერატორი	99
გერმანელი	110
გვარი	3688

გველი	1610
გიტარა	351
გოლკიპერი	53
გომბორი	267
გომორი	106
გორგასალი	71
გორდონი	99
გრიგოლი	328
გუბერნატორი	3177
გურმანი	54
დამცველი	8796
დანადგარი	610
დარეჯანი	69
დარიშხანი	91
დაუცველი	121
დედაჩემი	2053
დეკანი	611
დეტალი	494
დეტექტორი	74
დიაპაზონი	266
დიასპორი	6166
დიდგორი	739
დიზაინერი	1030
დილემა	398
დიპლომი	1190
დირექტორი	11955
დირიჟორი	180
დიქტატორი	483
დოლი	318
დონორი	434
დოქტორი	1665
დრაკონი	704
ეგზემპლარი	108
ეგზიუპერი	55
ევროზონა	724
ევროკომისარი	53
ევროპელი	95
ევროფესტივალი	60
ეთერი	1256
ეიფელი	384
ეკვადორი	336
ეკრანი	351

ელფერი	110
ესპანელი	75
ეშელონი	142
ვაგონი	271
ვალერი	163
ვატიკანი	269
ვაუჩერი	1301
ვაშინგტონი	1470
ვაშლოვანი	74
ვერტიკალი	112
ვერტმფრენი	85
ვესტმინსტერი	115
ვეტერანი	154
ვეტერინარი	125
ვექტორი	248
ვიდეოთვალი	139
ვიდეოკამერით	313
ვიდეომასალა	1000
ვიდეორგოლი	385
ვიდეოჩანაწერი	261
ვიცესპიკერი	53
ვიცე-სპიკერი	357
ვოკალი	255
ვულკანი	307
ზარი	1527
ზედამხედველი	85
ზვიგენი	176
ზოლი	1155
ზონა	3873
ზღვარი	524
თავდამსხმელი	440
თავჯდომარე	499
თანაგუნდელი	138
თანათავმჯდომარე	112
თანამგზავრი	451
თანამებრძოლი	122
თარგმანი	629
თბილისელი	151
თერჯოლა	570
თვალი	17552
თვითმხილველი	177
თინეიჯერი	146

იზოლატორი	295
იმპერატორი	1356
ინგლისელი	64
ინდიკატორი	59
ინვენტარი	750
ინვესტორი	808
ინიციატორი	148
ინკუბატორი	78
ინჟინერი	62
ინსპექტორი	357
ინსტრუქტორი	64
ინტერვალი	675
ინფორმატორი	69
იპოდრომი	178
იპოთეკარი	67
იუბილარი	59
იუმორი	1915
იუპიტერი	980
იჯარა	2790
კაბელი	94
კავალერი	56
კალათბურთელი	199
კალკულატორი	50
კამარა	339
კამერა	1915
კანონი	5289
კანტორა	161
კანცლერი	349
კაპელა	117
კარამელი	88
კარდინალი	77
კარნავალი	168
კარუსელი	117
კატალიზატორი	82
კვალი	1015
კითხვარი	195
კილოგრამი	494
კინორეჟისორი	136
კინოფესტივალი	887
კლანი	487
კლასელი	160
კოდორი	1344

კოლაგენი	101
კოლექტორი	90
კოლექციონერი	58
კოლონა	313
კომენტარი	4897
კომენტატორი	140
კომისარი	60
კომპიუტერი	3346
კომპიუტერი	3133
კომპოზიტორი	561
კონდიციონერი	112
კონსტანტინოპოლი	697
კონსტრუქტორი	65
კონტეინერი	107
კონტრაქტორი	51
კოორდინატორი	526
კორდონი	222
კორიდორი	246
კოშმარი	85
კრედიტორი	286
კრემი	596
კრიმინალი	826
კრისტალი	145
კურატორი	85
ლაზერით	318
ლაინერი	298
ლამაზმანი	114
ლეგიონერი	135
ლექსიკონი	561
ლექტორი	449
ლიბერალი	407
ლითონი	1475
ლიტერატორი	93
ლორწოვანი	231
ლუციფერი	118
მაგისტრალი	1052
მაგნიტოფონი	64
მადაგასკარი	62
მაიდანი	289
მაკარონი	192
მაკედონელი	95
მაკიაველი	59
მამაჩემი	1985
მანერა	318
მანქანით	13099
მაჟორიტარი	318
მარათონი	337
მარანი	92
მართლწერა	78
მარშალი	732
მარჩენალი	106
მაუწყებელი	100
მებრძოლი	347
მეგობარი	56
მედიატორი	317
მედპერსონალი	374
მეკარე	401
მემბრანა	90
მენეჯერი	1212
მენტორი	63
მეოთხედფინალი	143
მეპატრონე	389
მეტადონი	113
მეტალი	1078
მეტაფორა	88
მექანიზატორი	271
მეცხვარე	63
მეწყერი	287
მეჯვარე	106
მზერა	1199
მზრუნველი	90
მთელი	232
მთესველი	101
მთქმელი	63
მთხრობელი	76
მიკროზონა	54
მიკროფლორა	255
მიკროფონი	151
მილიარდელი	519
მილიონერი	160
მილსადენი	324
მიმოწერა	406
მინერალი	77
მირონი	253

მიქსერი	139
მკვლევარი	430
მკითხველი	8246
მკურნალი	219
მნახველი	64
მოდელი	5320
მოდერატორი	55
მოვალე	321
მოვლენა	79
მონაცემი	568
მონიტორი	504
მორალი	1139
მოსარჩელე	138
მოსაცდელი	62
მოტორი	52
მოცხარი	165
მოწამე	466
მოხელე	928
მრიცხველი	323
მსმენელი	576
მსურველი	111
მუქარა	362
მუშახელი	256
მუხროვანი	710
მფარველი	151
მფლობელი	1677
მყიდველი	315
მყინვარწვერი	99
მცენარე	1396
მცველი	558
მწეველი	123
მწერი	251
მწვანილი	176
მწვერვალი	221
მწვრთნელად	4254
მხარდამჭერი	281
ნაგავსაყრელი	506
ნავთობსადენი	196
ნათელი	68
ნაკერი	50
ნაკრძალი	332
ნამქერი	53

ნაპრალი	125
ნარკომანი	68
ნაქალაქარი	110
ნაღველი	69
ნაყენი	299
ნაშრომი	2791
ნაცვალსახელი	99
ნაწერი	379
ნაწილი	4594
ნახევარი	58
ნახევარმცველი	378
ნახევარფინალი	246
ნახველი	181
ნახტომი	597
ნიჟარა	106
ნიუტონი	237
ნობელი	1883
ნოდარი	182
ნოველა	224
ოზონი	241
ომბუდსმენი	1630
ონკანი	130
ოპერა	2493
ოპერატორი	788
ოპოზიციონერი	225
ორატორი	58
ორგანიზატორი	488
ორდენი	1581
ორდერი	129
ორიგინალი	280
ოსკარი	1273
ოფიცერი	52
პანელი	184
პაპაჩემი	125
პარალელი	171
პარკინსონი	183
პარლამენტარი	336
პაროლი	177
პასტერი	69
პასტორი	77
პატრონი	1028
პენსიონერი	141

პენტაგონი	536
პერსონა	705
პიონერი	119
პირველი	989
პლატონი	378
პლუტონი	342
პოკერი	328
პორტალი	556
პორტფელი	248
პოსტერი	62
პოტერი	589
პრაიმერი	977
პრესსპიკერი	456
პრეს-სპიკერი	113
პრინტერი	77
პრობაციონერი	65
პრობლემა	16732
პროგრამა	31071
პროდიუსერი	272
პროვაიდერი	82
პროვოკატორი	258
პროკურორი	8345
პრორექტორი	70
პროტოკოლი	375
პროფესიონალი	303
პროფესორი	1803
პროცესორი	96
ჟარგონი	59
ჟინვალი	171
ჟურნალი	5891
რადარი	196
რადიატორი	61
რადიოლოკატორი	59
რბოლა	561
რგოლი	1165
რეაქტორი	272
რეგისტრატორი	90
რედაქტორი	1625
რევოლვერი	75
რევოლუციონერი	75
რეინჯერი	116
რეკლამა	1867
რეკორდსმენი	411
რენტგენი	451
რეპერი	101
რეჟისორი	2933
რეფორმატორი	88
რექტორი	4524
რთველი	646
რიხტერი	95
რკალი	128
რკინაბეტონი	174
რომანი	2153
რქაწითელი	151
რწმენა	703
სავანე	195
საზომი	262
სათვალე	974
სათქმელი	364
სალონი	495
სამგორი	755
სარგებელი	115
სარჩელი	2275
სასტვენი	124
სასურველი	113
სასწორი	590
სასჯელი	5943
სატანა	384
საფარი	1107
საფუძველი	73
საყრდენი	153
საცოლე	186
საწოლი	996
სახდელი	237
სახელი	6102
სეზონი	8460
სელექციონერი	141
სემინარი	1286
სენატორი	279
სერვერი	127
სექტორი	9740
სიბერე	620
სიბნელე	536
სიგანე	134

სიგნალი	678
სილიკონი	359
სიმპტომი	116
სიმშრალე	128
სიმწარე	69
სინდრომი	1668
სირაქლემა	118
სისტემა	33820
სისხლდენა	565
სიჩქარე	748
სიძველე	338
სიძნელე	80
სკანდალი	1568
სკვერი	435
სლოგანი	273
სმარტფონი	222
სნაიპერი	348
სპიკერი	510
სპირალი	67
სპონსორი	282
სპორტსმენი	669
სტავროპოლი	151
სტენდალი	109
სტრიქონი	202
სულთანი	50
სუპერმენი	76
სურამი	372
სურნელი	673
სცენა	1353
სცენარი	2431
სხდომა	11423
ტამპონი	72
ტანდემი	208
ტელევიზორი	1486
ტელეფონი	7772
ტელეწამყვანი	134
ტენდერი	2001
ტერმინალი	560
ტესტოსტერონი	262
ტვიტერი	554
ტიპიკონი	58
ტოტალიზატორი	135

ტრაილერი	99
ტრანსფერი	736
ტრანსფორმატორი	52
ტრაპიზონი	241
ტრაქტორი	245
ტრეილერი	53
ტრენერი	122
ტრიბუნალი	266
ტყვარჩელი	112
უკანალი	260
უკანასკნელი	55
უკრაინელი	331
უნარი	2087
ურანი	506
უფლებადამცველი	185
ფაქტორი	2037
ფენომენი	538
ფეოდალი	71
ფერმერი	331
ფერწერა	499
ფესტივალი	6803
ფეხბურთელი	2596
ფინალი	1091
ფიროსმანი	416
ფიქალი	143
ფიჭვნარი	76
ფლაკონი	76
ფლომასტერი	52
ფლორა	234
ფოლკლორი	1309
ფოლკნერი	157
ფოსფორი	285
ფოტომასალა	139
ფრენა	1248
ფრინველი	722
ფსკერი	246
ფტორი	78
ფურგონი	53
ქალბატონი	2135
ქვეპროგრამი	173
ქვესკნელი	78
ქლორი	274

ქნარი	85
ქსელი	967
ღვეზელი	59
ღვთისმეტყველი	80
ღუმელი	51
ყელსაბამი	61
ყველი	255
ყოველი	100
ყუმბარა	235
შაშხანა	324
შევარდენი	455
შემოდგომა	2849
შემქმნელი	186
შესასვლელი	528
შეცდომა	2394
შპალერი	158
შრომა	15866
ჩამონათვალი	390
ჩანაწერი	1566
ჩანჩქერი	157
ჩოგბურთელი	86
ჩურჩხელა	100
ჩხავერი	242
ცაგერი	508
ცათამბჯენი	100
ციკლონი	64
ცირკულარი	162
ცისარტყელა	75
ცოცხალი	60
ცუნამი	413
ცხოველი	1911
ძველი	615

წამყვანი	416
წარმოდგენა	1270
წარწერს	1576
წაღვერი	58
წეროვანი	204
წვდომა	532
წვევამდელი	55
წვენი	857
წვერი	402
წინამძღოლი	253
წინასწარმეტყველი	363
წმინდანი	3938
წნორი	299
წყალსადენი	326
წყენა	253
ჭიპლარი	167
ხელნაწერი	390
ხელოვანი	401
ხელყუმბარა	89
ხელწერა	110
ხვანჭკარა	86
ხველა	367
ხსნარი	414
ხსნარით	94
ჯავშანმანქანა	112
ჯვრისწერა	431
ჰაკერი	55
ჰამქარი	60
ჰონორარი	290
ჰორმონი	680
ჰორორი	98
ჰოსპიტალი	296

Table B. 2: Corpus based list of syncopated words in Georgian

Word	Observed frequency
აბჯარი	56
ადგილსამყოფელი	96
ავლაბარი	1298
ავტოქარხანა	111
აივანი	458
ალაზანი	548
ალუბალი	767
ამომრჩეველი	2107
ანაბარი	481
არაჟანი	149
არასრულწლოვანი	254
არჩევანი	709
ატამი	604
აღმასრულებელი	140
აღმსარებელი	148
ახლობელი	601
ბადრიჯანი	113
ბაზარი	9886
ბალი	209
ბატკანი	130
ბებერი	98
ბორბალი	142
ბოსტანი	69
ბრალმდებელი	82
ბუღალტერი	107
ბუხარი	334
გალავანი	801
გამავრცელებელი	66
გამანადგურებელი	262
გამგებელი	7795
გამთამაშებელი	76
გამომგონებელი	135
გამომცემელი	86
გამომძიებელი	516
გამოსავალი	457
გამღიზიანებელი	91
განავალი	147
განმანათლებელი	137
განმცხადებელი	457
გარგარი	98
გარდაბანი	803
გასავალი	65
გენერალი	619
გენმდივანი	69
გორგასალი	377
გუთანი	58
გულშემატკივარი	419
დამაარსებელი	325
დამამზადებელი	113
დამთვალიერებელი	80
დამკვირვებელი	748
დამლაგებელი	165
დამპყრობელი	214
დამრიგებელი	179
დამსაქმებელი	375
დამფინანსებელი	67
დამფუძნებელი	1023
დამქირავებელი	133
დედანი	86
დედინაცვალი	88
დედოფალი	1523
დერეფანი	717
დიაკვანი	53
დუქანი	54
ევროკომისარი	145
ეკალი	127
ერევანი	520
ექთანი	195
ვაგზალი	327
ვაზისუბანი	327
ვაშლიჯვარი	121
ვერცხლისწყალი	119
ზამთარი	10906
ზარბაზანი	118
ზეწარი	50
ზღავრი	481
ზღაპარი	971
თავგადასავალი	114
თავმჯდომარე	55
თავშესაფარი	2059

თათარი	100
თანამშრომელი	2381
თარგმანი	241
თარჯიმანი	84
თაყვანისმცემელი	75
თვენახევარი	370
თიაქარი	193
თირკმელი	1448
ინჟინერი	142
ისანი	356
ისარი	346
კაკალი	826
კალამი	315
კალენდარი	988
კანონმდებელი	104
კაპიტანი	726
კაშხალი	386
კედელი	4804
კეისარი	333
კვარტალი	1419
კიდობანი	147
კისერი	1410
კომისარი	674
კოჯორი	399
კურდღელი	529
ლაშქარი	394
ლერწამი	129
ლიმონი	184
მადანი	169
მავნებელი	130
მათხოვარი	97
მაიდანი	216
მაკრატლით	162
მამალი	214
მანდილოსანი	68
მარანი	726
მართალი	179
მარცვალი	219
მასპინძელი	394
მასწავლებელი	3608
მატარებლით	1748
მაუწყებელი	8113

მაყარი	82
მაყვალი	122
მაყურებელი	2660
მაჩაბელი	172
მაჩვენებელი	2770
მაცივარი	252
მაცხოვარი	5217
მახასიათებელი	167
მბრძანებელი	158
მგელი	777
მდივანი	9971
მეგობარი	6797
მედალი	1967
მეზობელი	2481
მელანი	189
მეომარი	393
მერქანი	215
მერცხალი	128
მეწყერი	64
მთავარე	533
მთავარსარდალი	358
მიზნად	61906
მიმწოდებელი	240
მინანქარი	438
მკვდარი	191
მკვლევარი	102
მოედანი	2426
მოვაჭრე	75
მომავალი	7642
მომღერალი	2081
მომწოდებელი	59
მომხმარებელი	2917
მომხსენებელი	247
მონაზონი	114
მონასტერი	5174
მოსავალი	2164
მოსაკრებელი	797
მოყვარე	88
მოძღვარი	1041
მრავალი	99
მრჩეველი	572
მსესხებელი	159

მსხალი	309
მტაცებზწლი	66
მტევანი	291
მტვერით	1026
მტკვარი	1429
მტრად	5667
მუშტარი	71
მუცელი	5056
მღვდელი	916
მღვდელმთავარი	284
მშენებელი	73
მშვილდოსანი	242
მშობელი	1815
მწარმოებელი	582
მწერალი	3181
მწყერი	100
მხატვარი	2352
მხედარი	128
მხრით	621
ნათლით	486
ნათლისმცემელი	108
ნაკელი	68
ნამუშევრით	4468
ნამცხვარი	339
ნასოფლარი	83
ნატახტარი	72
ნაღველი	905
ნაცარი	524
ნაჭერი	432
ნახევარი	1386
ნეკერჩხალი	110
ნიშანი	14567
ნომრად	3031
ორთქლმავალი	84
ოფიცერი	795
პაემანი	248
პატარძალი	328
პატიმრად	5181
პირჯვარი	183
პრესმდივანი	253
რესტორანი	1258
რუსთაველი	134

რძალი	254
საბანი	200
საგალობელი	99
საგანი	3283
სავარძელი	297
საზღვარი	7808
საკანი	346
საკმეველი	107
საკურთხეველი	581
სამართალი	41346
სამართებელი	55
სამძიმარი	487
სანთელი	1425
სანიტარი	259
საპონი	883
სარგებელი	1200
სარდალი	327
სარკმელი	811
სარტყელი	168
სარქველი	120
სასმელი	354
სასუფეველი	268
სასწავლებელი	2984
საუბარი	11072
საფეთქელი	159
საფუძველი	1128
საქონელი	7808
საყდარი	285
საყვარელი	342
საჩივარი	1697
საჩუქარი	4903
საცვალი	86
საცხოვრებელი	167
საძინებელი	228
საძირკველი	164
საჭმელი	2335
სახამებელი	112
სახსარი	701
სვეტიცხოველი	849
სიზმრის	1044
სიმწარე	304
სოფელი	48644

სტუმარი	8011
სულთანი	255
ტანსაცმელი	2681
ტაძარი	11453
ტომარა	98
ტყემალი	311
უბანი	5599
უფალი	9329
უღელი	103
ფანქარი	293
ფანჯარა	1906
ფარავანი	122
ფარისეველი	114
ფეხსაცმელი	1840
ფიცარი	129
ფოთოლი	607
ფორთოხალი	1038
ფურცელი	733
ფუტკარი	663
ფუძემდებელი	158
ქამარი	286
ქარიშხალი	500
ქარხანა	3759
ქვეყნად	130488
ქვითარი	283
ქმარი	4157
ქოთანი	164
ღვთისმშობელი	7127
ღმრთისმშობელი	211
ღრუბელი	511
ყვარელი	693
ყველაფერი	185
ყოველი	256
ყურძენი	5323
შადრევანი	152
შარვალი	590
შაქარი	2878
შემოსავალი	4224
შემსრულებელი	1813
შესავალი	214
შველი	84
შორაპანი	54

შუამავალი	322
შუქნიშანი	98
ჩანგალი	108
ჩემოდანი	112
ჩირაღდანი	290
ჩოგანი	58
ცელოფანი	237
ცისკარი	315
ცხედარი	675
ცხენისწყალი	98
ცხვარი	1896
ძვალი	1301
ძმარი	234
წამლით	2592
წარმომადგენელი	6329
წელი	502657
წერეთელი	1341
წითელი	62
წილკანი	66
წინამძღვარი	243
წინანდალი	293
წინაპარი	334
წიფელი	122
წუთისოფელი	344
წყალი	29193
ჭარხალი	291
ჭინჭარი	437
ჭიშკარი	184
ჭურჭელი	1051
ხანი	43804
ხანძარი	2600
ხეივანი	232
ხელისუფალი	121
ხელოსანი	50
ხერხემალი	973
ხინკალი	359
ხმალი	535
ხორბალი	2002
ჯადოქარი	180
ჯავშანი	517
ჯვარი	6117
ჯირკვალი	1789

ჰამქარი	101
ჰექტარი	89

ჰოსპიტალი	193

Table B. 3: Corpus based list of syncopated and non-suncopated words in Georgian

Word	Observed frequency for syncopated word form	Observed frequency for non-syncopated word form
ადგილსამყოფელი	96	677
არასრულწლოვანი	254	239
არჩევანი	709	6644
ბაზარი	9886	55
გამგებელი	7795	143
გამოსავალი	457	82
ევროკომისარი	145	53
თავშესაფარი	2059	158
თანამშრომელი	2381	59
თირკმელი	1448	61
ინჟინერი	142	62
კვარტალი	1419	508
კომისარი	674	60
ლიმონი	184	1690
მარანი	726	92
მართალი	179	60
მეგობარი	6797	56
მეწყერი	64	287
მკვლევარი	102	430
მონასტერი	5174	55
მრჩეველი	572	176
ნათელი	486	68
ნაღველი	905	69
ნახევარი	1386	58
ოფიცერი	795	52
სამართალი	41346	118
სარგებელი	1200	115
სარტყელი	168	96
სასმელი	354	203
საფუძველი	1128	73
საცხოვრებელი	167	75
საძირკველი	164	70
საჭმელი	2335	68
ტანსაცმელი	2681	81
ტაძარი	11453	84

უკანასკნელი	55	1831
ფეხსაცმელი	1840	983
ქარიშხალი	500	75
ქვეყანა	130488	294
ქსელი	967	7683
ყველაფერი	185	307
ყოველი	256	100
შემოსავალი	4224	54
წარმომადგენელი	6329	393
წელი	502657	1340
წითელი	62	54
წყალი	29193	125
ჯვარი	6117	294
ჰამქარი	101	60
ჰოსპიტალი	193	296

References

Amiridze, N. 2006. *Reflexivization Strategies in Georgian*. LOT Dissertation Series 127, Utrecht Institute of Linguistics, Utrecht University, The Netherlands, 2006. ISBN-10: 90-76864-96-9, ISBN-13: 978-90-76864-96-9

Anton, the first. 1753. *kartuli ğrammat'ik'a*. Tbilisi. S/S “p'irveli st'amba”.

Anton, the first. 1767. *kartuli ğrammat'ik'a*. meore gamocema. Tbilisi. S/S “p'irveli st'amba”.

Babunashvili, E. and Uturgaidze, T. 1991. *ant'on p'irvelis “kartuli ğrammat'ik'a” da misi erovnul-ist'oriuli mnišvneloba*. Tb.: mecniereba (m/a st'.)

Bagrationi, I. 1829. *k'almasobiseuli kartuli gramat'ik'a*.

Bahl, LR. and Mercer, RL. 1976. Part of speech assignment by a statistical decision algorithm. In: *IEEE International Symposium on Information Theory*, 88-89. Ronneby.

Beesley, R. and Karttunen, L. 2002. *Finite-State Morphology: Xerox Tools and Techniques*. Center for the Study of Language and Information.

Benko, V. 2016. Two Years of Aranea: Increasing Counts and Tuning the Pipeline. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC2016*). Portorož, Slovenia.

Benko, V. 2018. Aranea Project Main NoSketch Engine Site, Georgian corpus. Available from: http://aranea.juls.savba.sk/

Bopp, F. 1842. *Analysis of the phenomenon in Georgian*. The Academy of Sciences of Berlin. Germany.

Brants, T. 2000. TnT: a statistical part-of-speech tagger. In: *Proceedings of the sixth Conference on Applied Natural Language Processing (ANLC’00)*, pp. 224-231. Seattle, Washington.

Brill, E. 1992. A simple rule-based part of speech tagger. In: *Proceedings of the Third Conference on Applied Natural Language Processing (ANLP'92)*. Trento.

Brosset, M. 1834. *xelovneba aznaurebiti gina qartulis enisa: tvit masc'avlebeli: [kartuli enis gramat'ik'a frangulad]*. Paris.

Brosset, M. 1837. Eléments de la langue géorgienne. Paris.

Charniak, E., Hendrickson, C., Jacobson, N. and Perkowitz, M. 1993. Equations for part of speech tagging. In: *Proceedings of the Eleventh National Conference on Artificial Intelligence*. Menlo Park: AAAI Press/MIT Press.

Chikobava A. 1940. *mesame p'iris užvelesi nišani kartvelur enebši*. animk'is moambe,V-IV.

Chikobava, A. 1968. *mart'ivi c'inadadebis p'roblema kartulši*. meore gamocema. Tbilisi. mecniereba.

Cloeren, J. 1999. Tagsets. In: van Halteren, H. ed. *Syntactic wordclass tagging*. Dordrecht: Kluwer Academic Publishers, pp. 37-54.

Cutting, D., Kupiec, J., Pederson, J. and Sibun, P. 1992. A practical part-of-speech tagger. In: *Proceedings of the third conference on Applied Natural Language Processing*, ACL.

Daraselia S. and Sharoff, S. 2014. Morphosyntactic Specifications for KaWaC, a Web Corpus for Georgian. International Conference - *Humanities in the Information Society II*. Batumi, Georgia, pp. 326-329.

Daraselia S. and Sharoff, S. 2015. The Main Steps of the Georgian Web-Corpus Construction. Tbilisi. Arnold Chikobava Institute of Linguistics, *Journal of Linguistics*, Volume XXXVIII, pp 52-62.

Daraselia, S. 2015. *Issues of Compilation of New Georgian-European Learner's Dictionary Using the Corpus Methodology*. Ph.D. thesis, Ivane Javakhishvili Tbilisi State University.

Daraselia, S., Hardie, A. 2018. A New Morphosyntactic Annotation Scheme in Georgian. *Batumi Summer School in Digital Humanities*. Batumi Shota Rustaveli State University.

Deeters, G. 1953. *Die Schtelung der Kartvelsprachen unter den kaukasichen sprechen: bedi Kartlisa*, №3.

Diessel, H. 1997. *The diachronic reanalysis of demonstratives in crosslinguistic perspective*. Chicago Linguistic Society (CLS) 33, 83-98.

Diessel, H. 1999. *Demonstratives: Form, function, and grammaticalization* (Typological Studies in Language). Amsterdam & Philadelphia: John Benjamins.

Doborjginidze, N. and Lobzhanidze, I. 2016. Corpus of the Georgian Language, Proceedings of the XVII EURALEX International Congress, *Ivane Javakhishvili Tbilisi University Press*, 328-335. Available from: https://euralex2016.tsu.ge/publication2016.pdf.

Dodaevi, S. 1830. *šemok'lebuli kartuli ğrammat'ik'a*. t'pilisi. gamomcemlis k'omit'et'is st'.

Eklund, R. 1993. A probabilistic tagging module based on surface pattern matching. In: Eklund, R (ed.) (1993) *NODALIDA '93 - Proceedings of 9:e Nordiska Datalingvistikdagarna*. Stockholm: Stockholm University.

Eristavi, V. 1802. *kratkaja gruzinskaja grammatika, sochinennaja na rossijskom jazyke*. Spb. : pri imp. Akademii nauk.

Erjavec, T., Krstev, C., Petkevič, V., Simov, K., Tadić, M., Vitas, D. 2003. The MULTEXT-East Morphosyntactic Specification for Slavic Languages. Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages. Association for Computational Linguistics. Budapest, Hungary. Pp. 25-32.

Erjavec, T. 2012. MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, 46/1, pp.

131-142. Available from: https://link.springer.com/article/10.1007%2Fs10579-011-9174-8.

Gachechiladze, O. 1979. *šorisdebuli axal salit'erat'uro kartulši: leksik'oniturt*. tbilisis universit'et'is gamomcemloba.

Gamkrelidze, N., Kotetishvili, Sh., Lezhava, J., Lortkipanidze, L., and Javakhidze, L. 2006. *Phonetic Analysis of Georgian Normative and Dialectal Speech*. gamomcemloba “nekeri”.

Garside, R. 1987. The CLAWS Word-tagging System. In: Garside, R., Leech, G. and Sampson, G. eds. *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Garside, R., Leech, G. and McEnery A. 1997. *Corpus annotation: linguistic information from computer text corpora*. London: Longman.

Gogolashvili, G., Arabuli, A., Sukhishvili, M., Manjgaladze, M. and Tchumburidze, N. 2011. *tanamedrove kartuli enis morpologia, salit'erat'uro ena*. Tbilisi. Meridiani.

Greenbaum, S. and Yibin, N. 1996. About the ICE Tagset. In: Greenbaum, S. ed. *Comparing English Worldwide: the International Corpus of English*. Oxford: Clarendon Press.

Greene, BB. and Rubin, GM. 1971. *Automatic grammatical tagging of English*. Providence, Rhode Island: Brown University Department of Linguistics.

Hardie, A. 2004. *The computational analysis of morphosyntactic categories in Urdu*. Ph.D. thesis, Lancaster University.

Hardie, A., Lohani, R. and Yadava, Y. 2011. Extending corpus annotation of Nepali: advances in tokenisation and lemmatisation. In: *Himalayan Linguistics*. 10, 1, p. 151–165.

Hardie, A. 2017. Morphosyntactic Annotation Schemata. *CAMRL2017: Workshop on Computational Approaches to Morphologically Rich Languages.* The University of Leeds.

Harris, A. 1981. *Georgian Syntax: A Study in Relational Grammar*. Cambridge: Cambridge University Press.

Heikkilä, J. 1995. A TWOL-based lexicon and feature system for English. In: Karlsson, F., Voutilainen, A., Heikkilä, J. and Anttila, A. eds. *Constraint Grammar: a language independent system for parsing unrestricted text*. Berlin: Mouton de Gruyter.

Heine, B., Reh, M. 1984. *Grammaticalization and reanalysis in African languages*. Hamburg: Buske.

Heine, B., Song, K. 2011. On the grammaticalization of personal pronouns. *Journal of Linguistics*, 47(3), 587-630.

Hewitt, G. 1995. *Georgian: A Structural Reference Grammar*. John Benjamins Publishing. Language Arts & Disciplines.

Iluridze, K. 2006. *saxelta bruneba XIX sauk'unis I naxevris kartuli enis gramat'ik'ebši*. Ph.D. thesis, arnold chikoabavas enatmecnierebis inst'itut'i.

Imnaishvili, Iv. 1956. *c'rfelobiti brunvis sak'itxi sak'utar saxelebši: saxelta brunebis ist'oriisatvis kartvelur enebši*. tbilisis universist'et'is gamomcemloba.

Imnaishvili, Iv. 1957. *saxelta bruneba da brunvata funkciebi žvel kartulši*. tbilisis universist'et'is gamomcemloba.

Ioseliani, P. 1840. *p'iruel-dac'q'ebitni k'anonni kartulis gramat'ik'isa*. tp., i. da d. arzanovta st'.

Ioseliani, P. 1851. *p'irueldac'q'ebitni k'anonni kartulis ǧrammat'ik'isa*. Tpilisi, guliancis c'igntsabechdavi.

Joshi, AK. and Hopely, P. 1997. A parser from antiquity. In: *Natural Language Engineering*, 2(4): 291-294.

Kaplan, R., Bresnan, J. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In: Bresnan, J. ed., *The Mental Representation of Grammatical Relations*. Chapter 4, The MIT Press, pp.173-281.

Karlsson, F., Voutilainen, A., Heikkilä, J. and Anttila, A. eds. 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text*. Berlin: Mouton de Gruyter.

Karosanidze, L. 2017. *ant'kur-bizant'iuri teoriebi enis šesaxeb da kartuli gramat'ik'uli azrovneba* (*Antique and Byzantine Theories about the Language and Georgian Grammatical Thought)*. Tbilisi. tsu gamomcemlobis st'amba.

kartuli enis ganmart'ebiti leksik'oni. 1950-1964. 8 tomeuli. Tbilisi. Available from: http://www.ice.ge/liv/liv/ganmartebiti.php.

kartuli enis ortografiuli leksik'oni. 2014. Available from: http://www.ena.ge/.

Kartvelishvili, I. 1801. *kartuli ğrammat'ik'a*. Tbilisi.

Kartvelishvili, I. 1815. *kartuli ğrammat'ik'a*. Tbilisi.

Khoja, S, Garside, R, and Knowles, G .2001. A tagset for the morphosyntactic tagging of Arabic. *Corpus Linguistics 2001 conference*, Lancaster.

Klausenburger, J. 2000. *Grammaticalization: Studies in Latin and Romance morphosyntax* (Amsterdam Studies in the Theory and History of Linguistic Science). Amsterdam & Philadelphia: John Benjamins.

Klein, S. and Simmons, RF. 1963. A computational approach to grammatical coding of English words. In: *Journal of the Association for Computing Machinery*, 10: 334-347.

Leech, G., Garside, R. and Bryant, M. 1994. CLAWS4: The tagging of the British National Corpus. In: *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)* Kyoto, Japan: 622-628.

Leech, G. 1997. Introducing corpus annotation. In: Garside, R., Leech, G. and McEnery T. eds. *Corpus annotation: linguistic information from computer text corpora*. London: Longman.

Leech, G. 1997. Grammatical tagging. In: Garside, R., Leech, G. and McEnery T. eds. *Corpus annotation: linguistic information from computer text corpora*. London: Longman.

Leech, G. and Wilson, A. 1999. Standards for tagsets. In: van Halteren, H. ed. *Syntactic wordclass tagging*. Dordrecht: Kluwer Academic Publishers, pp. 55-80.

Leech, G. and Smith, N. 1999. The use of tagging. In: van Halteren, H. ed. *Syntactic wordclass tagging*. Dordrecht: Kluwer Academic Publishers.

Lobzhanidze, I. 2013. Morphological Analyzer and Generator of Modern Georgian Language. In: *Georgian Language and Modern Technologies*. Tbilisi, pp. 82-83.

Lortkipanidze L., Beridze M. and Nadaraia D. 2013. Dialect Dictionaries with the Functions of Representativeness and Morphological Annotation in Georgian Dialect Corpus. *Theoretical Computer Science and General Issues. 10th International Tbilisi Symposium on Logic, Language, and Computation, TbiLLC 2013*, Gudauri, Georgia.

Marcus, M., Santorini, B. and Marcinkiewicz, MA. 1993. Building a large annotated corpus of English: the Penn Treebank. In: *Computational Linguistics*, 19(2): 313-330.

Marshall, I. 1987. Tag selection using probabilistic methods. In: Garside, R., Leech, G. and Sampson, G. eds. *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Marr, N. 1908. osnovnye tablicy k grammatike dreviegruzinskogo âzyka, S- Pb.. imp. akad. nauk.

Marr, N. 1925. grammatika drevneliteraturnogo gruzinskogo âzyka. materialy po yafeticheskomy yazykoznaniyu XII. Leningrad 1925.8.XXIV, 30, 216.

Melikishvili, D. 2001. *kartuli zmnis uğlebis sist'ema*. Logos Press. Tbilisi.

Melikishvili, D. 2008. *The Georgian Verb: A Morphosyntactic Analysis*. Washington, Dunwoody Press. ISBN: 978-1931546-51-5.

Melikishvili, D. 2014. *kartuli zmnis sist'emuri morposint'aksuri analizi*. ISBN 978-9941 437-74-8.

Melikishvili, I. 2008. Georgian as an Active/Ergative Split Language. *Bulletin of the Georgian National Academy of Sciences*, vol. 2, no. 2.

Meurer, P. 2007. A Computational Grammar for Georgian. *7th International Tbilisi Symposium on Logic, Language, and Computation, TbiLLC 2007*, Tbilisi, Georgia.

O Duibhín, C. 2018. *Windows Interface for Tree Tagger*. Available from: http://www.smo.uhi.ac.uk/

Peikrishvili, Zh. 2010. *kartuli enis morpologia*. meotxe gamocema. kutaisis universit'et'is gamomcemloba. ISBN 99928-79-08-4.

Piao, S., Bianchi, F., Dayrell, C., D'Egidio, A. and Rayson, P. 2015. Development of the multilingual semantic annotation system. *In proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015)*, Denver, Colorado, United States, pp. 1268-1274.

Piralovi, G. 1820. *tvit masc'avlebeli, romeli ip'q'robs tvis šoris ğramat'ik'asa, zneobis sc'avlasa, saubarsa da leksik'onsa rusulsa da qartulsa enasa zeda* (*samouchitel', soderzhashhij v sebe grammatiku, razgovory, nravouchenija i leksikon na Rossijskom i gruzinskom jazykah*). sanktpeterburği: st'anbasa šina iosen ioanesovisa.

Potskhishvili, A. 1981. *kartuli gramat'ikuli azris sataveebtan*. c'erilebi kartuli enisa da lit'erat'uris sak'itebze. tb.

Rektori, G., Nikolaishvili, E. ed. 1970. *kartuli ğrammat'ik'a: [Kartuli enis gramat'ik'is saxelmžğ*vanelo]. tb.: mecniereba.

Sampson, G. 1995. *English for the computer: the SUSANNE Corpus and Analytic Scheme*. Oxford: Clarendon Press.

Sánchez León, F. and Nieto Serrano, AF. 1997. Retargeting a tagger. In: Garside, R., Leech, G. and McEnery T. eds. *Corpus annotation: linguistic information from computer text corpora*. London: Longman.

Santini, M., Mehler, A. and Sharoff, S. 2010. Riding the rough waves of genre on the web. In Mehler, A., Sharoff, S., and Santini, M., eds. *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Sarjveladze, Z. 1984. *kartuli salit'erat'uro enis ist'oriis šesavali*. ganatleba.Tbilisi.

Sawalha, M. Atwell, ES. and Abushariah, M. 2013. SALMA: Standard Arabic Language Morphological Analysis. In Proceedings of *ICCSPA International Conference on Communications, Signal Processing, and their Applications*, pp. 1-6.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of *International Conference on New Methods in Language Processing*, Manchester, UK.

Schmid, H. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Shanidze, A. 1934. *žveli kartuli ena: žveli qartuli ena da lit'erat'ura*. saxelgamis sasc'. ped. sekt'oris ga-ma. Tbilisi.

Shanidze, A. 1953. *kartuli enis gramat'ik'is safužvlebi*. saxelmc'ipo universit'et'is gamomcemloba.

Shanidze, A. 1976. *žveli kartuli enis gramat'ik'a*. Tbilisi. saxelmc'ipo universit'et'is gamomcemloba.

Shanidze, A., Shanidze, M. ed. 1980. *kartuli enis gramat'ik'is safužvlebi*. Tbilisi.

Shanshovani, Z. 1737. *mok'le ğrammat'ik'a kartulisa enisa*. sank't' peterbugi. a. Tsagarelis gamocema.

Sharashenidze, N. 2014. *modalobis kat''egoria kartulši, misi sc'avlebis metodebi da st'rat'egiebi ucxoeltatvis*. Tbilisi.

Smith, N. 1997. Improving a tagger. In: Garside, Leech Garside, R., Leech, G. and McEnery T. eds. *Corpus annotation: linguistic information from computer text corpora*. London: Longman.

Stolz, WS., Tannenbaum, PH. and Carstensen, FV. 1965. A stochastic approach to the grammatical coding of English. In: *Communications of the ACM*, 8 (6): 399-405.

Svartvik, J. 1990. Tagging and parsing on the TESS project. In: Svartvik, J. ed. *The London-Lund Corpus of Spoken English: description and research. Lund*. Lund University Press.

Tallerman, M. 2011. *Understanding syntax*. Understanding language series, Routledge, 3rd ed.

Tamarashvili, M. 1902. *ist'oria k'atolik'obisa kartvelta šoris namdvilis sabutebis šemot'anita da ganmart'ebit XIII sauk'unidgan vidre XX sauk'unemde*. t'pilisi: avt'oris mier gamocemuli. elekt'r. sabetchdi st'amba kart. c'ig. gam. amx.

Tapanainen, P. and Voutilainen, A. 1994. Tagging accurately – don't guess if you know. In: *Proceedings of the Fourth Conference on Applied Natural Language Processing*. Stuttgart.

Tschenkeli, K. 1960. *Georgisch-Deutsches Wörterbuch*. Zürich: Amirani-Verlag. xxxviii, 2470 pp.

Topuria, V. 1965. *c'rfelobiti brunvisatvis žvel qartulši*. sal. mecn. akademiis moambe, XXXVIII, №2.

Uturgaidze, T. 1986. *kartuli enis saxelis morponologiuri analizi*. mecniereba. Tbilisi.

van Halteren, H. and Oostdijk, N. 1993. The TOSCA analysis system. In: Aarts, J., de Haan, P. and Oostdijk, N. English language corpora: design, analysis and exploitation. *Papers form the thirteenth International Conference on English Language Research on Computerised Corpora, Nijmegen 1992*. Amsterdam: Rodopi.

van Halteren, H. and Voutilainen, A. 1999. Automatic taggers: an introduction. In: van Halteren, H. ed. *Syntactic wordclass tagging*. Dordrecht: Kluwer Academic Publishers, pp. 109-115.

van Halteren, H. ed. 1999. *Syntactic wordclass tagging*. Dordrecht: Kluwer Academic Publishers.

Voutilainen, A. 1999. A short history of tagging. In: van Halteren, H. ed. *Syntactic wordclass tagging*. Dordrecht: Kluwer Academic Publishers, pp.3-4.

Zorell, F. 1930. *Grammatik zur altgeorgischen Bibelübersetzung mit Testproben und Wörtvereichnis,* Boma.