A Finite-State Model of Georgian Verbal Morphology

Olga Gurevich
Department of Linguistics
University of California, Berkeley
3834 23rd Street
San Francisco, CA 94114
olya.gurevich@gmail.com

Abstract

Georgian is a less commonly studied language with complex, non-concatenative verbal morphology. We present a computational model for generation and recognition of Georgian verb conjugations, relying on the analysis of Georgian verb structure as a word-level template. The model combines a set of finite-state transducers with a default inheritance mechanism.[1]

1 Introduction

Georgian morphology is largely synthetic, with complex verb forms that can often express the meaning of a whole sentence. Descriptions of Georgian verbal morphology emphasize the large number of inflectional categories; the large number of elements that a verb form can contain; the inter-dependencies in the occurrence of various elements; and the large number of regular, semi-regular, and irregular patterns of formation of verb inflections (cf. Hewitt 1995). All of these factors make computational modeling of Georgian morphology a rather daunting task.

In this paper, we propose a computational model for parsing and generation of a subset of Georgian verbs that relies on a templatic, word-based analysis of the verbal system rather than assuming compositional rules for combining individual morphemes. We argue that such a model is viable, extensible, and capable of capturing the generalizations inherent in the Georgian verbal system at various levels of regularity. To our knowledge, this is the only computational model of the Georgian verb currently in active development and available to the non-Georgian academic community[2].

2 Georgian Verbal Morphology

The Georgian verb forms are made up of several kinds of morphological elements that recur in different formations. These elements can be formally identified in a fairly straightforward fashion; however, their function and distribution defy a simple compositional analysis but instead are determined by the larger morphosyntactic and semantic contexts in which the verbs appear (usually tense, aspect, and mood) and the lexical properties of the verbs themselves.

2.1 Verb Structure

Georgian verbs are often divided into four conjugation classes, based mostly on valency (cf. Harris 1981). In this brief report, we will concentrate on transitive verbs, although our model can accommodate all four conjugation types. Verbs inflect in tense/mood/aspect (TAM) paradigms (simplified here as tenses). There are a total of 10 actively used tenses in Modern Georgian, grouped into TAM series as in Table 1. Knowing the series and tense of a verb form is essential for being able to conjugate it.

The structure of the verb can be described using the following (simplified) template.

[1] This work was in part supported by the Berkeley Language Center. I'd like to thank Lauri Karttunen for introducing me to finite-state morphology and providing an updated version of the software, and Shorena Kurtsikidze and Vakhtang Chikovani for help with the Georgian data. All errors are my own.

[2] See Tandashvili (1999) for an earlier model. Unfortunately, the information in the available publications does not allow for a meaningful comparison with the present model.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 45–48,
New York, June 2006. ©2006 Association for Computational Linguistics

Series	Tense	2SGSUBJ:3SGOBJ
I	PRESENT	***xat’***-*av*
	IMPERFECT	***xat’***-*av-di*
	PRES. SUBJ.	***xat’***-*av-de*
	FUTURE	*da*-***xat’***-*av*
	CONDITIONAL	*da*-***xat’***-*av-di*
	FUT. SUBJ.	*da*-***xat’***-*av-de*
II	AORIST	*da*-***xat’***-*e*
	AOR. SUBJ.	*da*-***xat’***-*o*
III	PERFECT	*da-gi*-***xat’***-*av-s*
	PLUPERFECT	*da-ge*-***xat’***-*a*

Table 1: Tenses of the verb ‘to paint’. Root is in bold.

(Preverb)-(agreement1)-(version)-**root**-(thematic suffix)-(tense)-(agreement)

The functions of some of the elements are discussed below. As an illustration, note the formation of the verb xat’va ‘paint’ in Table 1.

2.2 Lexical and Semi-Regular Patterns

The complexity of the distribution of morphological elements in Georgian is illustrated by preverbs, thematic suffixes, and tense endings. The preverbs (a closed class of about 8) indicate perfective aspect and lexical derivations from roots, similar to verb prefixes in Slavic or German. The association of a verb with a particular preverb is lexical and must be memorized. A preverb appears on forms from the Future subgroup of series I, and on all forms of series II and III in transitive verbs. Table 2 demonstrates some of the lexically-dependent morphological elements, including several different preverbs (row ‘Future’).

Similarly, thematic suffixes form a closed class and are lexically associated with verb roots. They function as stem formants and distinguish inflectional classes. In transitive verbs, thematic suffixes appear in all series I forms. Their behavior in other series differs by individual suffix: in series II, most suffixes disappear, though some seem to leave partial “traces” (rows ‘Present’ and ‘Perfect’ in Table 2).

The next source of semi-regular patterns comes from the inflectional endings in the individual tenses and the corresponding changes in some verb roots (row ‘Aorist’ in Table 2).

Finally, another verb form relevant for learners is the masdar, or verbal noun. The masdar may or may not include the preverb and/or some variation of the thematic suffix (last row in Table 2).

	‘Bring’	‘Paint’	‘Eat’
Present	i-**gh**-*eb*-s	**xat’**-*av*-s	**ch’am**-*ø*-s
Future	*c’amo*-i-**gh**-eb-s	*da*-**xat’**-av-s	*she*-**ch’am**-s
Aorist	c’amo-i-**gh**-*o*	da-**xat’**-*a*	she-**ch’am**-*a*
Perfect	c’amo-u-**gh**-*ia*	da-u-**xat’**-*av-s*	she-u-**ch’am**-*ia*
Masdar	c’amo-**gh**-eb-a	da-**xat’**-v-a	**ch’**-am-a

Table 2: Lexical Variation. Roots are in bold; lexically variable affixes are in italics.

SUBJ	OBJ				
	1SG	1PL	2SG	2PL	3
1SG	—	—	g—ø	g—t	v—ø
1PL	—	—	g—t	g—t	v—t
2SG	m—ø	gv—ø	—	—	ø—ø
2PL	m—t	gv—t	—	—	—t
3SG	m—*	gv—*	g—*	g—t	—*
3PL	m—**	gv—**	g—**	g—**	—**

Table 3: Subject/Object agreement. The 3sg and 3pl suffixes, marked by * and **, are tense-dependent.

2.3 Regular Patterns

Verb agreement in Georgian is a completely regular yet not entirely compositional phenomenon. A verb can mark agreement with both the subject and the object via a combination of prefixal and suffixal agreement markers, as in Table 3.

The distribution and order of attachment of agreement affixes has been the subject of much discussion in theoretical morphological literature. To simplify matters for the computational model, we assume here that the prefixal and suffixal markers attach to the verb stem at the same time, as a sort of circumfix, and indicate the combined subject and object properties of a paradigm cell.

Despite the amount of lexical variation, tense formation in some instances is also quite regular. So, the Imperfect and First Subjunctive tenses are regularly formed from the Present. Similarly, the Conditional and Future Subjunctive are formed from the Future. And for most (though not all) transitive verbs, the Future is formed from the Present via the addition of a preverb.

Additionally, the number of possible combinations of inflectional endings and other irregularities is also finite, and some choices tend to predict other choices in the paradigm of a given verb. Georgian verbs can be classified according to several example paradigms, or inflectional (lexical) classes, similar to the distinctions made in Standard European languages; the major difference is that the number of classes is much greater in Georgian. For instance, Melikishvili (2001) distinguishes over 60 classes, of which 17 are transitive. While the exact number of inflectional classes is still in question, the general example-based approach seems the only one viable for Georgian.

3 Computational Model

3.1 Overview

Finite-state networks are currently one of the most popular methods in computational morphology. Many approaches are implemented as two-way finite-state transducers (FST) in which each arc corresponds to a mapping of two elements, for example a phoneme and its phonetic realization or a morpheme and its meaning. As a result, FST morphologies often assume morpheme-level compositionality. As demonstrated in the previous section, such assumptions do not serve well to describe the verbal morphology of Georgian. Instead, it can be described as a series of patterns at various levels of regularity. However, compositionality is not a necessary assumption: finite-state models are well-suited for representing mappings from strings of meaning elements to strings of form elements without necessarily pairing them one-to-one.

Our model was implemented using the *xfst* program included in (Beesley and Karttunen 2003). The core of the model consists of several levels of finite-state transducer (FST) networks such that the result of compiling a lower-level network serves as input to a higher-level network. The levels correspond to the division of templatic patterns into completely lexical (Level 1) and semi-regular (Level 2). Level 3 contains completely regular patterns that apply to the results of both Level 1 and Level 2. The regular-expression patterns at each level are essentially constraints on the templatic structure of verb forms at various levels of generality. The FST model can be used both for the generation of verbal inflections and for recognition of complete forms.

The input to the model is a set of hand-written regular expressions (written as FST patterns) which identify the lexically specific information for a representative of each verb class, as well as the more regular rules of tense formation. In addition to dividing verb formation patterns into lexical and regular, our model also provides a mechanism for specifying defaults and overrides in inflectional markers. Many of the tense-formation patterns mentioned above can be described as defaults with some lexical exceptions. In order to minimize the amount of manual entry, we specify the exceptional features at the first level and use the later levels to apply default rules in all other cases.

3.2 Level 1: The Lexicon

The first level of the FST model contains lexically specific information stored as several complete word forms for each verb. In addition to the information that is always lexical (such as the root and preverb), this network also contains forms which are exceptional. For the most regular verbs, these are: Present, Future, Aorist 2SgSubj, Aorist 3SgSubj, and Perfect.

The inflected forms are represented as two-level finite-state arcs, with the verb stem and morphosyntactic properties on the upper side, and the inflected word on the lower side.

The forms at Level 1 contain a place holder "+Agr1" for the prefixal agreement marker, which is replaced by the appropriate marker in the later levels (necessary because the prefixal agreement is between the preverb and the root).

3.3 Level 2: Semi-regular Patterns

The purpose of Level 2 is to compile inflectional forms that are dependent on other forms (introduced in Level 1), and to provide default inflections for regular tense formation patterns.

An example of the first case is the Conditional tense, formed predictably from the Future tense. The FST algorithm is as follows:

- Compile a network consisting of Future forms.
- Add the appropriate inflectional suffixes.
- Replace the tense property "+Fut" with "+Cond".
- Add the inflectional properties where needed.

An example of the second case is the Present 3PlSubj suffix, which is *-en* for most transitive verbs, but *-ian* for a few others (see Fig. 1). Xfst provides a simplified feature unification mechanism called *flag*

Lev. 1	*paint*+Pres *xat’-**av***	*paint*+Aor ***da**-xat’-**a***	*open*+PresPl *xsn-**ian***
Lev. 2	*paint*+Past+3Sg *xat’-av-**da***	*paint*+Pres+3Pl *xat’-av-**en***	default overridden
Lev. 3	*paint*+3PlSubj+1SgObj ***m**-xat’-av-en*	*open*+3PlSubj+1SgObj ***m**-xsn-ian*	

Figure 1: Verbs ‘paint’ and ‘open’ at three levels of the model. New information contributed by each form is in bold.

diacritics. Using these flags, we specify exceptional forms in Level 1, so that default inflections do not apply to them in Level 2.

The patterns defined at Level 2 are compiled into a single network, which serves as input to Level 3.

3.4 Level 3: Regular Patterns

The purpose of Level 3 is to affix regular inflection: object and non-3rd person subject agreement. As described in section 2, agreement in Georgian is expressed via a combination of a pre-stem affix and a suffix, which are best thought of as attaching simultaneously and working in tandem to express both subject and object agreement. Thus the compilation of Level 3 consists of several steps, each of which corresponds to a paradigm cell.

The operation of the model is partially illustrated on forms of the verbs ‘paint’ and ‘open’ in Figure 1.

3.5 Treatment of Lexical Classes

The input to Level 1 contains a representative for each lexical class, supplied with a diacritic feature indicating the class number. Other verbs that belong to those classes could, in principle, be inputted along with the class number, and the FST model could substitute the appropriate roots in the process of compiling the networks. However, there are several challenges to this straightforward implementation. Verbs belonging to the same class may have different preverbs, thus complicating the substitution. For many verbs, tense formation involves stem alternations such as syncope or vowel epenthesis, again complicating straightforward substitution. Suppletion is also quite common in Georgian, requiring completely different stems for different tenses.

As a result, even for a verb whose lexical class is known, several pieces of information must be supplied to infer the complete inflectional paradigm. The FST substitution mechanisms are fairly restricted, and so the compilation of new verbs is done in Java. The scripts make non-example verbs look like example verbs in Level 1 of the FST network by creating the necessary inflected forms, but the human input to the scripts need only include the information necessary to identify the lexical class of the verb.

4 Evaluation and Future Work

At the initial stages of modeling, we have concentrated on regular transitive verbs and frequent irregular verbs. The model currently contains several verbs from each of the 17 transitive verb classes mentioned in (Melikishvili 2001), and a growing number of frequent irregular verbs from different conjugation classes. Regular unaccusative, unergative, and indirect verbs will be added in the near future, with the goal of providing full inflections for 200 most frequent Georgian verbs.

The model serves as the basis for an online learner’s reference for Georgian conjugations (Gurevich 2005), which is the only such reference currently available.

A drawback of most finite-state models is their inability to generalize to novel items the way a human could. However, the output of our finite-state model could potentially be used to generate training sets for connectionist or statistical models.

References

Beesley, Kenneth and Lauri Karttunen. 2003. *Finite-State Morphology*. Cambridge University Press.

Gurevich, Olga. 2005. Computing non-concatenative morphology: The case of georgian. In *LULCL 2006*. Bolzano, Italy.

Harris, Alice C. 1981. *Georgian syntax: a study in relational grammar*. Cambridge University Press.

Hewitt, B. G. 1995. *Georgian: a structural reference grammar*. John Benjamins.

Melikishvili, Damana. 2001. *Kartuli zmnis ughlebis sist’ema [Conjugation system of the Georgian verb]*. Logos presi.

Tandashvili, M. 1999. *Main Principles of Computer-Aided Modeling, http://titus.uni-frankfurt.de/personal/manana/refeng.htm*. Tbilisi Habilitation.