ESUKA – JEFUL 2022, 13–2: 47–78

MORPHOLOGICAL INFLECTIONAL RULES FOR VERBS IN KARELIAN PROPER

Natalia Krizhanovskaya, Irina Novak, Andrew Krizhanovsky, Nataliya Pellinen

nataly@krc.karelia.ru, novak@krc.karelia.ru, andrew.krizhanovsky@gmail.com, nataliapellinen@gmail.com

Abstract. A methodology for the development and implementation of inflectional rules for verbs in Karelian Proper is presented. The materials for this study were lemmas and word forms from the Open corpus of Veps and Karelian languages (VepKar) and the electronic version of the Karelian language Dictionary. The system of rules for automatic verb inflection for the Karelian Proper supradialect is of both practical and theoretical scientific interest. The new rules have already enabled entering 141 000 Karelian Proper word forms in the VepKar dictionary. The new program for word form generation has significantly reduced the time for adding the full inflectional paradigm of any Karelian Proper verb to the VepKar dictionary. One only needs to fill in several template parameters instead of 125 word forms.

Keywords: Karelian language, Karelian Proper, Corpus linguistics, verbal inflection, inflectional paradigm, generating word forms

DOI: https://doi.org/10.12697/jeful.2022.13.2.02

1. Introduction

The electronic information and reference system Open corpus of Veps and Karelian languages (VepKar) is being developed jointly at the Institute of Language, Literature and History and the Institute of Applied Mathematical Research of the Karelian Research Centre RAS for over ten years. The VepKar corpus includes texts and dictionaries, as well as a computer program “corpus manager”, which provides an interface, search capabilities and processing of resource materials. This corpus manager is written in the PHP programming language in the Laravel framework. The data is stored in a MySQL database. The VepKar corpus and dictionaries data are available online (VepKar 2021).

The VepKar corpus is a continuation of the Veps language corpus created in 2009 (project leader Nina Zaitseva) (Zaiceva 2019: 39–40). In 2016, the corpus was extended to include the Karelian language.

The Karelian language belongs to the Baltic-Finnish group of the Uralic language family; it is the closest language to the Izhorian, Finnish and Veps languages. The total number of those who speak the Karelian language in Russia in 2010 amounted to 25.6 thousand people (Russian Census 2010). The number of the Karelian-speaking population of Finland, according to unofficial data, can reach 30 thousand people (Sarhimaa 2017: 111–115).

The Karelian language in the VepKar corpus is represented by three subcorpora corresponding to three dialects: Karelian Proper (northern and middle Karelia, Tver, Leningrad, Murmansk regions), Livvi (Olonets and Pryazhinsky districts of Karelia) and Ludian (Pryazhinsky, Prionezhsky, Kondopozhsky and Olonets districts of Karelia) (Krizhanovsky, Krizhanovsky & Novak 2019). This division dates back to the works of Dmitry Bubrikh (Bubrikh 1947, 1948). However, Finnish linguists believe that Ludian dialects are an independent Baltic-Finnish language (Pahomov 2017: 9, 33, 108, 285).

Karelian Proper is a direct descendant of the Old Karelian language, whereas the Livvi and Ludian supradialects formed through interactions between the Old Karelian and the Old Veps languages. Significant differences between Karelian language supradialects are seen at all levels of language, including morphological (for example, differences in case systems, peculiarities of the forms of personal pronouns, the use of the full temporal paradigm of the conditional in Livvi and Ludian dialects, as well as the presence of reflexive conjugation) (Novak et al. 2019: 21–27, Koivisto 2018: 57–58). The texts have been grouped into separate sub-corpuses for this reason.

In the UNESCO Atlas of the World's Languages in Danger (2016), the Karelian language (Karelian and Olonetsian) is assigned to the group "definitely endangered", and its Ludian dialect is assigned to the group "severely endangered" (Moseley 2010). The first attempts to standardize the Karelian language were undertaken in the 1930s (new written varieties of the Karelian language). The standardization work for two varieties of the new written Karelian language, developed on the basis of the Livvi and Karelian Proper (its North Karelian sub-dialects) supradialects has been going on in the Republic of Karelia for three

decades. At the same time, in the Tver region, a new written norm is developing, based on the local Karelian dialects of the Karelian Proper (Nagurnaya 2019: 75–77). The new written varieties of the Karelian language based on (1) the Ludian dialects in the Republic of Karelia and (2) the South Karelian dialects of the Karelian Proper supradialect of Border Karelia in Finland have been actively developed in the last decade. Spelling norms and grammar rules have been defined for each of the varieties.

The VepKar corpus contains all four varieties of the new written Karelian language that have been developed in Russia: North Karelian, Tver, Livvi and Ludian. The VepKar corpus is an electronic reference information system with journalistic and fiction texts (in new written language varieties) and dialectal texts (recorded written and oral texts since the second half of the 20th century) in Karelian and Veps. Some of the texts are supplied with a Russian translation aligned with the source text sentence by sentence. Karelian- and Veps-language texts are segmented into sentences, and sentences into tokens. The numerical characteristics of the dictionary (the number of lemmas, word forms) and the text corpus VepKar (the number of texts; the number of unique tokens in the language; tokens linked to the dictionary entries) are shown in Figure 1.

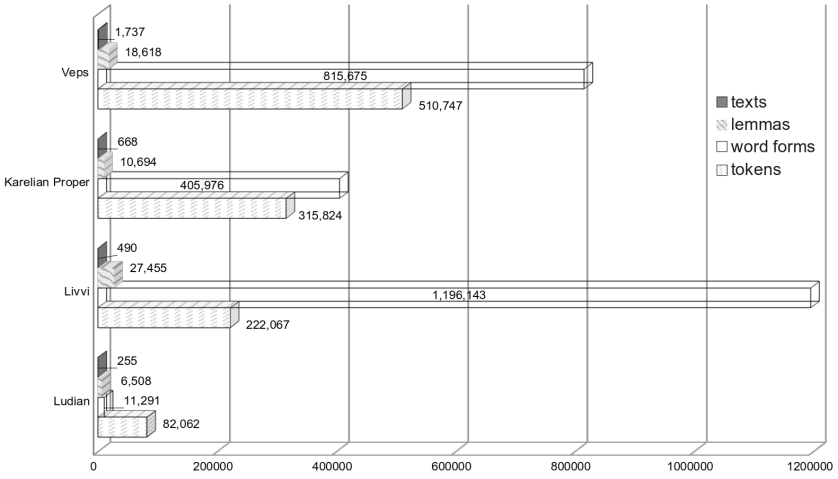

Figure 1. Volume of VepKar dictionaries and subcorpora for each language and supradialect (as of 16.09.2020).

The difference in data volume between subcorpuses is shown in Fig. 1. The Vep subcorpus contains the largest number of texts (1737), since the Veps corpus was created first (2009), and after that the Karelian language was included and the VepKar corpus was created (2016). The Livvi subcorpus is the leader in terms of the number of lemmas (27455), since all available newly-written dictionaries were included in the corpus. The Livvi subcorpus has reached a leading position in the number of word forms due to the fact that the rules for generating word forms were developed (Krizhanovskaya et al. 2021). The rules for generating word forms for the Ludian dialect have not yet been developed, therefore the number of word forms is negligible in the Ludian subcorpus.

A massive task for the future is to enable machine translation from Karelian and Veps into Russian and vice versa, so one of the routines for the corpus editors is semantic tagging of the texts. A key component of the system is the dictionary. It gets filled in rather slowly, because of a limited number of specialists involved on the one hand, and because word forms, which are essential for tagging, are entered manually. To wit, it takes some 20 minutes on average to fill in a verb inflection paradigm, which is made up of 125 (Karelian Proper) / 149 (Livvi, Ludian)[1] / 115 (Veps) word forms per lemma. This step, however, is unavoidable, since only a minor percentage of lemmas occur in texts in their root (dictionary) forms. For instance, after all word forms of the Karelian Proper verb *antua* 'give' had been entered, the number of tagged examples where the lemma was used in the texts increased from 32 to 549.

The process of text tagging and adding entries to the dictionary in the Karelian subcorpora can be accelerated by developing a program for automatic word form generation. In 2019, the VepKar corpus was supplied with a function of generating the full paradigm from basic word forms for the Veps language (Krizhanovskaya & Krizhanovsky 2019). It helped notably augment the share of not only Veps word forms, but also tagged tokens (word-uses in the text for which matches were spotted in the dictionary). The Veps subcorpus is now 83% tagged. The

1 The paradigm of the verbal inflection of Livvi and Ludian's adverbs differs from the paradigm of the Karelian Proper inflection in that the Livvi and Ludian contain four tense conjugation forms of verbs in the conditional (there are two tenses in the Karelian Proper).

same year, a similar program was produced for the Tver newly written variant of the Karelian language (one of the standardized varieties of the Karelian Proper supradialect developing in the Tver Region). Its application resulted in the addition of 130 000 word forms to the dictionary, and a 45% increase in tagging coverage. For this program to fill in the verb inflection paradigm, the editor needs to enter seven pseudo-endings for each lemma. In this research effort, the plan is to extend this program to cover the remaining Karelian supradialects, at the same time reducing the grammatical data input per lemma to a minimum.

The peculiarities of the development of a program for the automatic generation of word forms for the Karelian Proper supradialect of the Karelian language and the first results of the work of this program are described in this publication. The successful creation of a word form generator has important theoretical and practical implications. The program for automatic generation of word forms will serve as the basis for creating an automatic morphological analyzer of words, which will significantly increase the proportion of automatic marking of corpus texts. In turn, this will provide a large morphologically tagged corpus of texts, which is extremely necessary for the study of the under-researched question of the grammatical structure of the Karelian language. This article will present the process of designing the generation program in application to the analysis of the verb inflection system of the newly written variant of the Karelian Proper supradialect, namely:

- theoretical works used in the course of the research are presented in the article;
- materials and research methods are described;
- the methodology for building automatic generation (preceding the creation of the corresponding module for the corpus manager) is presented in the form of four stages, accompanied by detailed tables.

2. Study of the Karelian verb inflection system

The baseline sources for studying the Karelian verb inflection system were Karelian Proper grammar books by P. M. Zaikov “The grammar of Karelian” (Zaikov 2002) and “The grammar of Viena Karelian” (Zaikov 2013). These grammar books, separated by a fair time interval, are mutually complementary, considering that the rules and norms of

the newly written varieties of the Karelian language have been modified over this ten-year period, on the one hand, and that the earlier editions offer more detailed descriptions of some aspects of particular importance when working out word-form generation rules, on the other.

The grammar books of the newly written Karelian language varieties introduce the main inflection types and paradigms of nominals and verbs. In the process, however, we discovered that a few aspects have been covered insufficiently. One example is the lack of clear rules for ablaut in the imperfect indicative stem (alternations a : Ø, a / ä : o). These shortfalls hinder the formulation of word-form generation rules, calling for extra research, search for patterns and introduction of additional rules. The nature of the corpus itself (the presence of dialect texts in the corpus) also demands some departures from the standards accepted in the grammar books.

The theoretical sources for the study were P. M. Zaikov's "The verb in the Karelian language" (Zaikov 2000), which analyses the grammatical categories of person-number, tense, and mood in dialects of all the three Karelian supradialects in synchrony and diachrony, and E. L. Adel's "Verb inflection in the Karelian language" (Adel' 1998), with a description of the verb inflection system in South Karelian dialects of the Karelian Proper supradialect. Another source of material was the reference book "Karelian in Grammars" (Novak et al. 2019), reporting the results of a comparative study of the Karelian phonetic and morphological systems based on hand-written and published sources covering a century and a half. A whole section in this book is given to the description of the verb inflection system.

The morphological systems of the closely cognate Karelian and Finnish languages, descending from a common Balto-Finnic proto-language, have not changed so profoundly in the course of their separate evolution as have, for instance, their phonetic systems. It is therefore possible to use the valuable background produced by authors of The Descriptive grammar of Finnish (Hakulinen at al. 2004) in working with Karelian material.

3. Material and methods

This study is chiefly based on the VepKar corpus dictionary material (lemmas and word forms) in the Karelian Proper (ca. 9 000 lemmas and 129 000 word forms) and Livvi (ca. 21 000 lemmas and 557 000 word forms) newly written varieties. In 2018, 18 000 lemmas with major word forms were manually added to VepKar using material from the "Unabridged Karelian-Russian Dictionary (Livvi supradialect)" (Boiko 2018). In particular, the entries for verbs covered 1st and 3rd person singular in the present indicative, and 3rd person plural in the present indicative. Early in 2020, this list for verbs was augmented by adding 3rd person singular and plural in the imperfect indicative, enabling experiments to be run, revealing the laws of stem formation in the imperfect tense in Karelian.[2]

The electronic version of the Karelian language dictionary (Torikka 2009) comprising some 88 000 entries, amply illustrated with examples in different dialects of the Karelian Proper and Livvi supradialects of the first half of the 20th century was used for the analysis of the Karelian verb inflection system in this work.

The study relied upon inflectional analysis, comparative, and statistical methods. These methods in combination with experiments have enabled a determination of patterns in the Karelian verb inflection system. The actual development of the generator made use of the experience of creating a morphological analyzer for Kven[3] (Trosterud et al. 2017), and electronic parallel dictionaries of Uralic languages (e.g. Ludian-Russian-Finnish dictionary) (Rueter & Hämäläinen 2017).

At this time most Uralic languages still lack full-fledged morphological analyzers and large corpora (Pirinen et al. 2016). The work (Pirinen 2019b) describes the following infrastructure and technology for developing linguistic tools for the low-resource Karelian language:

2 The experimental results can be found online in the VepKar corpus section "Search for regularities in vowel gradation" at http://dictorpus.krc.karelia.ru/ru/experiments/vowel_gradation/verb_imp_3sg

3 The Kven language (kväänin kieli) is a Finnic language spoken by the Kven people in Northern Norway.

1. Machine translation is rule-based and implemented in the Apertium system.
2. The morphological analyzer-generator works on the basis of the Helsinki Finite-State Technology (HFST).
3. The corpus annotation is made in Universal Dependencies notation.

One of the important applications of the VepKar corpus will be the automatic analysis of the structure of sentences and the creation of syntactically annotated corpus (treebank). In (Pirinen 2019a), several texts of the VepKar corpus were prepared and marked in accordance with the requirements of the Universal Dependencies project. Thus, 125 sentences of VepKar in the Livvi and 228 sentences in the Karelian Proper supradialects were tagged (Pirinen 2019a). The fact that VepKar data is tagged and included in the international project Universal Dependencies is an important step in the research and preservation of languages.

Further studies plan to employ the available developments in morphological analysis and morphological generation programs, including those for low-resource languages, trained (using neural systems) on thoroughly tagged Bible texts in English and their translations into hundreds of other languages (Garrett et al. 2020). With Bible texts available in Veps and Karelian, morphological tools for these languages can hopefully be created.

4. Methodology for formulating rules for automatic word-form generation for verbs in Karelian

4.1. Stage 1: constructing the verb inflection paradigm

The first step towards creating a program for automatic generation of verb word forms is to derive the inflectional paradigm of the verb in Karelian. As well as other Baltic-Finnic languages, Karelian is agglutinative language with some elements of inflection. Words are inflected by attaching affixes with a grammatical meaning to the lexical stem. In doing so, only one inflectional marker of one grammatical category can be present in a single word form (Hakulinen et al. 2004: § 53, Novak et al. 2019: 162).

The verb inflection system in Karelian has active and passive voice forms to express how the action is related to its subject and object. Morphologically, the initial form is the active form, indicating the presence of the actor. The active form is contraposed to the morphologically derivative passive form, which denotes an action whose subject is unknown or is a generalized person. Passive voice forms in Karelian dialects (the equivalent of Finnish or Estonian passive marker) are used to form 3rd person plural forms of the verb (Novak et al. 2019: 268).

Verb conjugation in Karelian is differentiated into conjugated (finite) and non-conjugated (infinite, inflective) verb forms. Commonly conjugated verb forms have the following set of grammatical categories: person-number, tense, and mood (Zaikov 2013: 146–147). Another type of verb conjugation is generation of negative forms using the negative verb *ei*. The final meaning of a word form in Karelian is the sum of the meanings of its constituent parts, wherefore paradigmatic series comprise not only synthetic but also analytic forms (Adel' 1998: 4), such as complex tense constructs (e.g. *olen ottan* 1st person singular in the perfect indicative form, *olisit ottan* 2nd person singular in the pluperfect conditioned form from the word *ottua* 'take'). Non-conjugated forms (infinitives, participles), being inflected for number and case, combine verb traits with grammatical traits of nominals (Zaikov 2013: 189).

The basic, or dictionary, form of the verb is the 1st infinitive (A infinitive in the *Descriptive grammar of Finnish*). All verb forms are produced from the lexical verb stem by attaching inflectional markers.

A verb word form can consist of several components: 1) one lexical verb stem; 2) two components, e.g. verb stem + person-number ending; 3) several components, e.g. verb stem + passive voice marker + mood marker + person-number ending or verb stem + tense marker + person-number ending. The position nearest to the stem in finite verbs is occupied by the passive voice marker used in 3rd person plural forms, which is followed by the suffix for tense or mood, and then by the person-number marker, e.g., *hyö tul-t-i-h* 'they came'/ 'they have come', where *tul-* is the consonant stem of the verb *tulla* 'to come', *-ta* is the passive voice marker, *-i* is the imperfect tense suffix, *-h* is the person-number (3PL) ending. A single word form cannot contain both tense and mood markers at the same time. In infinite verbs, the stem is directly followed by an infinitive or a participle marker, and then come the number marker and the case ending, e.g. *otta-j-i-lla*, where *otta-* is the strong vowel

stem of the verb *ottua* ‘to take’, *-ja* is participle I active marker, *-i-* is the plural number marker, *-lla* is the adessive-allative case ending.

The inflectional paradigms in Table 1 demonstrate that Karelian is not perfectly agglutinative, as it has developed some flective inflection features such as stem-end vowel gradation, or consonant alternation, which is the basis for classifying stems into weak- and strong-grade ones (*anna-n* 1SG PRS IND – *anno-i-n* 1SG IMPF IND – *anta-u* 3 SG PRS IND, *rupie-n* 1SG PRS IND – *rupe-i-n* 1SG IMPF IND – *ruven-nut* PTCP II ACT). Another important feature to be taken into account is that the language has single-stem verbs, which have a vowel stem, and dual-stem verbs, which have a vowel stem and a consonant stem (*anta-u* 3 SG PRS IND – *anta-nut* PTCP II ACT, *rupie-u* 3 SG PRS IND – *ruven-nut* PTCP II ACT). The number of word forms required to identify the inflectional types of verbs is determined keeping these features in mind. In addition to the **dictionary form** (1st infinitive), they are the **1st person singular in the present indicative** with a weak vowel stem in single-stem verbs and a strong vowel stem in dual-stem verbs, **participle II active (contracted)** form with a strong vowel stem in single-stem verbs and a consonant stem in dual-stem verbs, and the **1st person singular in the imperfect indicative** form exhibiting a stem-vowel gradation.

Table 1. Verb inflection paradigm.[4]

Word form		Single-stem verb ***antua*** ‘to give’		Dual-stem verb ***ruveta*** ‘to begin’	
		Conjugated (finite) forms			
		affirmative	**negative**	**affirmative**	**negative**
Indicative	1SG PRS	**anna**-n	en **anna**	rupie-n	en rupie
	2SG PRS	**anna**-t	et **anna**	rupie-t	et rupie
	3SG PRS	anta-u	ei **anna**	rupie-u	ei rupie
	1PL PRS	**anna**-mma	emmä **anna**	rupie-mma	emmä rupie
	2PL PRS	**anna**-tta	että **anna**	rupie-tta	että rupie
	3PL PRS	**anne**-ta-h	ei **anne**-ta	***ruvet***-a-h	ei ***ruvet***-a
	1SG IMPF	**anno**-i-n	en anta-n	rupe-i-n / rupe-si-n	en ***ruven***-nun
	2SG IMPF	**anno**-i-t	et anta-n	rupe-i-t / rupe-si-t	et ***ruven***-nun
	3SG IMPF	anto	ei anta-n	rupe-i /rupe-si	ei ***ruven***-nun
	1PL IMPF	anto-ma	emmä anta-n	rupe-i-ma / rupe-si-ma	emmä ***ruven***-nun
	2PL IMPF	anto-ja	että anta-n	rupe-i-ta / rupe-si-ja	että ***ruven***-nun
	3PL IMPF	**ann**e-tt-i-h	ei ***anne***-ttu	***ruvet***-t-i-h	ei ***ruvet***-tu
	1 / 2 / 3SG / 1 / 2PL PERF	(olen / olet / on / olemma / oletta) anta-n	(en / et / ei / emmä / että) ole anta-n	(olen / olet / on / olemma / oletta) ***ruven***-nun	(en / et / ei / emmä / että) ole ***ruven***-nun
	3PL PERF	on **anne**-ttu	ei ole **anne**-ttu	on ***ruvet***-tu	ei ole ***ruvet***-tu
	1 / 2 / 3SG / 1 / 2PL PLUP	(olin / olit / oli / olima / olija) anta-n	(en / et / ei / emmä / että) ollun anta-n	(olin / olit / oli / olima / olija) ***ruven***-nun	(en / et / ei / emmä / että) ollun ***ruven***-nun
	3PL PLUP	oli **anne**-ttu	ei oltu **anne**-ttu	oli ***ruvet***-tu	ei oltu ***ruvet***-tu
Imperative	2SG PRS	**anna**	elä **anna**	rupie	elä rupie
	3SG PRS	anta-kkah	elkäh anta-kkah	***ruvek***-kah	elkäh ***ruvek***-kah
	1PL PRS	anta-kka	elkä anta-kka	***ruvek***-ka	elkä ***ruvek***-ka
	2PL PRS	anta-kkua	elkyä anta-kkua	***ruvek***-kua	elkyä ***ruvek***-kua
	3SG & PL	anta-kkah	elkäh anta-kkah	***ruvek***-kah	elkäh ***ruvek***-kah

4 Normal font in the table is used for a strong stem, **bold font** for a weak stem, vowel stems are not accentuated, *consonant stems* are in italics, and changes in the stem vowel are underlined.

Word form		Single-stem verb ***antua*** ‘to give’		Dual-stem verb ***ruveta*** ‘to begin’	
		Conjugated (finite) forms			
		affirmative	**negative**	**affirmative**	**negative**
Conditional	1SG IMPF	anta-si-n	en anta-is[5]	rupie-si-n	en rupie-is
	2SG IMPF	anta-si-t	et anta-is	rupie-si-t	et rupie-is
	3SG IMPF	anta-is	ei anta-is	rupie-is	ei rupie-is
	1PL IMPF	anta-si-ma	emmä anta-is	rupie-si-ma	emmä rupie-is
	2PL IMPF	anta-si-ja	että anta-is	rupie-si-ja	että rupie-is
	3PL IMPF	**anne**-tta-is	ei **anne**-tta-is	***ruvet***-ta-is	ei ***ruvet***-ta-is
	1 / 2 / 3SG / 1 / 2PL PLUP	(olisin / olisit / olis / olisima / olisija) anta-n	(en / et / ei / emmä / että) olis anta-n	(olisin / olisit / olis / olisima / olisija) ***ruven***-nun	(en / et / ei / emmä / että) olis ***ruven***-nun
	3PL PLUP	olis **anne**-ttu	ei olis **anne**-ttu	olis ***ruvet***-tu	ei olis ***ruvet***-tu
Potential	1SG PRS	anta-ne-n	en anta-ne	***ruven***-ne-n	en ***ruven***-ne
	2SG PRS	anta-ne-t	et anta-ne	***ruven***-ne-t	et ***ruven***-ne
	3SG PRS	anta-no-u	ei anta-ne	***ruven***-no-u	ei ***ruven***-ne
	1PL PRS	anta-ne-mma	emmä anta-ne	***ruven***-ne-mma	emmä ***ruven***-ne
	2PL PRS	anta-ne-tta	että anta-ne	***ruven***-ne-tta	että ***ruven***-ne
	3PL PRS	**anne**-tta-ne-h	ei **anne**-tta-ne	***ruvet***-ta-ne-h	ei ***ruvet***-ta-ne
	1 / 2 / 3SG / 1 / 2PL PERF	(lienen / lienet / lienöy / lienemmä / lienettä) anta-n	(en / et / ei / emmä / että) liene anta-n	(lienen / lienet / lienöy / lienemmä / lienettä) ***ruven***-nun	(en / et / ei / emmä / että) liene ***ruve***n-nun
	3PL PERF	lienöy **anne**-ttu	ei line **anne**-ttu	lienöy ***ruvet***-tu	ei line ***ruvet***-tu

5 According to P. M. Zaikov’s Grammar, the end consonant *-s* in 3rd person conditional is palatalized (e.g., *antais’* ‘he would give’ (would have given), *olis’ juonun* ‘he would drink’ (would have drunk)) (Zaikov 2013: 178). This norm, however, did not become mainstream in periodicals, fiction, language instruction, etc. The variant with non-palatalized conditional mood marker in 3rd person verb forms is adopted in this table and the rules below. In fact, P. M. Zaikov himself concurred to removing this norm in the process of editing the “Russian-Karelian Dictionary (north-Karelian subdialects)” (Zaikov et al. 2015).

Word form		Single-stem verb ***antua*** ‘to give’		Dual-stem verb ***ruveta*** ‘to begin’	
		Conjugated (finite) forms			
		affirmative	**negative**	**affirmative**	**negative**
Inflective (infinite) forms					
Infinitives	I	antu-a		***ruvet***-a	
	II INE	antua-s’s’a		***ruvet***-e-šša	
	II INSTR	anta-en		***ruvet***-e-n	
	III ADE-ALL	anta-ma-lla		rupie-ma-lla	
	III ELA	anta-ma-šta		rupie-ma-šta	
	III ILL	anta-ma-h		rupie-ma-h	
	III ABE	anta-ma-tta		rupie-ma-tta	
	III INE	anta-ma-šša		rupie-ma-šša	
Participles	I ACT	anta-ja		rupie-ja	
	II ACT CONTR	anta-n		***ruven***-nun	
	II ACT full	anta-nut		***ruven***-nut	
	I PASS	**anne**-ttava		***ruvet***-tava	
	II PASS	**anne**-ttu		***ruvet***-tu	

4.2. Stage 2: identifying the types of lexical verb stems

Karelian verb stems are classified into lexical and inflectional ones, which incorporate markers of tense, mood, or passive voice. The basis for identifying the types of lexical verb stems in Karelian is the presence of single- and dual-stem verbs in Baltic-Finnic languages (Zaikov 2002: 71–73, Zaikov 2013: 148–150). Table 2 gives examples of the main inflectional types of verbs from VepKar corpus dictionaries.

Table 2. Types of lexical verb stems.[6]

D.f.[7] ending in	**1st infinitive**	**1SG PRS IND**	**1SG IMPF IND**	**PTCP II ACT CONTR**
Single-stem verbs				
uo *yö*	kučču**o** ‘to call’ yhty**ö** ‘to get united’	ku**č**u-n y**h**y-n	ku**č**u-i-n y**h**y-i-n	kučču-n yhty-n
uo	kačču**o** ‘to look/watch’	ka**č**o-n	ka**č**o-i-n	kačč**o**-n
ie	ečč**ie** ‘to search’	e**č**i-n	e**č[Ø]**-i-n	ečči-n
ie	it**k**ie ‘to cry’	i**t**e-n	i**t[Ø]**-i-n	itke-n
ua *yä*	antua ‘to give’ ottua ‘to take’ kiäntyä ‘to turn’	a**nn**a-n o**t**a-n kiä**nn**ä-n	a**nno**-i-n o**t[Ø]**-i-n kiä**nn[Ø]**-i-n	anta-n otta-n kiäntä-n
ha / *hä* *ja* / *jä* *va* / *vä*	šuaha ‘to get’ viijä ‘to carry’ juuvva ‘to drink’	šua-n vie-n juo-n	š**a**-i-n v**e**-i-n j**o**-i-n	šua-nun vie-nyn juo-nun
ja / *jä*	kapaloij**a** ‘to swaddle’	kapaloi**č**e-n	kapaloi**č[Ø]**-i-n	kapaloi-nun
Dual-stem verbs				
la / *lä* *na* / *nä* *ra* / *rä* *sa* / *sä* *ša* / *šä*	tulla ‘to come’ männä ‘to depart’ purra ‘to gnaw’ paissa ‘говорить’ peššä ‘to wash’	tule-n mäne-n pure-n pakaja-n peše-n	tul**[Ø]**-i-n män**[Ø]**-i-n pur**[Ø]**-i-n paka-s-i-n pes**[Ø]**-i-n	tul-lun män-nyn pur-run pais-sun peš-šyn
ta / *tä*	ru**v**eta ‘to begin’ ke**r**itä ‘to manage/to be on time’	rupie-n kerkie-n	rup**e**-i-n / rupe-s-i-n kerk**e**-i-n / kerki-s-i-n	ru**v**en-nun ke**r**in-nyn
ta / *tä*	tarita ‘to suggest’	tariče-n	tarič**[Ø]**-i-n	tarin-nun
ta / *tä*	lyhetä ‘to shorten’	lyhene-n	lyhen**[Ø]**-i-n	lyhen-nyn
ta / *tä*	varata ‘to fear’	varaja-n	vara-s-i-n	varan-nun

6 Bold type in the table marks an alternation of vowel and consonant (weak-grade variant) phonemes.

7 D.w. and d.f. stand for dictionary (lexicalized) word and dictionary form. To wit, the dictionary word for the template *it|kie [e]* is *it|kie*, and the dictionary form (i.e. the form derived according to the rule) is *itkie*.

4.3. Stage 3: identifying the stems for generating verb word forms

Proceeding from the data obtained in the first and second stages, the next step is to determine the set of stems needed for generating word forms in the inflectional paradigm of verbs of any type, and the template for the lemma form to be entered into the generator by the editor. We strive to minimize the grammatical data input per lexeme. If, for instance, the editor only needs to enter three stems instead of seven to enable generation of all word forms, this will save the editor a lot of effort.

Fairly often, the dictionary form of the verb does not unambiguously indicate the form of its vowel stem (*ečči*e – *eči*-, *itkie* – *ite*-; *ruveta* – *rupie*-, *lyhetä* – *lyhene*-), so its autofill is not possible, therefore, a prerequisite is the input of two word forms manually. Importantly, the vowel stem of single-stem verbs should be given in the weak grade (if any).

Table 3 illustrates the rules for identifying verb stems, auxiliary stems (S) and pseudo-stems (PS) of verbs. The auxiliary stem and pseudo-stem are automatically generated by the program, and the difference between them is that the auxiliary stem coincides with one of the verb's stems, while the pseudo-stem is produced by the generator itself.

Table 3. Identification of stems for word-form generation.[8]

Stem	Stem generation rules	Example
d.f.	– | from the dictionary word. *D.w. in which the changeable part is separated from the unchangeable part by the symbol* |.	it|kie [e]→itkie, an|tua [na]→antua, ju|uvva [o]→juuvva, kapaloi|ja [če]→kapaloija, tul|la [e]→tulla, ru|veta [pie]→ruveta, tari|ta [če]→tarita

8 Notations for Tables 3 and 4: | denotes the border between between the unchangeable and the changeable parts of the word, – (minus sign) is the command to delete subsequent symbols, + is the command to add such symbols, = stands for equality, > is a replacement, / separates variants of the marker, ~ separates the front variant of the marker from the back variant, **V** is a vowel, **C** is a consonant, → is the transition to the next action, *italicized text* in the rules is the algorithm for deriving the lemma template.

Stem	Stem generation rules	Example
S1 (weak vowel stem)	D.w. part up to | + the first from brackets. Pseudo-ending of the stem is enclosed in brackets.	it|kie [e]→ite, an|tua [na]→anna, ju|uvva [o]→juo, kapaloi|ja [če]→kapaloiče, tul|la [e]→tule, ru|veta [pie]→rupie, tari|ta [če]→tariče
S2 (AUX strong vowel stem)	**A.** If d.f. ends in VV, then d.f. – terminal VV → + terminal V from S1. **B.** If d.f. ends in CV, then S2 = S1, if S1 ends in če, then *če* should be replaced by *čče* (*če* > *čče*).	**A.** itkie – ie → itk + e (S1=it**e**) = itke, antua – ua → ant + a (S1=ann**a**) = anta **B.** juuvva→juo, kapaloija→kapaloiče > kapaloičče, tulla→tule, ruveta→rupie, tarita→tariče > taričče
S3 (AUX strong vowel / consonant stem)	**A.** If d.f. ends in *va*/*vä*, *ha*/*hä*, *ja*/*jä* (d.f. has 2 syllables) OR d.f. ends in VV (any number of syllables) then S2 **B.** If d.f. ends in CV (other cases), then d.f. – terminal CV → **1)** if the resultant form ends in VV (except V*i*), CV, then + *t*	**A.** juuvva→juo, itkie→itke, antua→anta **B.** kapaloija – ja = kapaloi, tulla – la = tul, **B.1)** ruveta – ta → ruve + t = ruvet, tarita – ta → tari + t = tarit
PS4 (AUX strong vowel stem IMPF)	S2, and here, if S2 ends in **1)** C*e*, then *e* > **i**; **2)** *ja*/*jä,* and d.f. ends in CV, then – *ja*/*jä* → + **si**; **3)** C*ä,* and d.f. ends in VV, then *ä* > **i**; **4)** C*a,* and d.f. ends in VV, and S2 has 2 syllables of which the first one has the vowel *a* (*a*, *au*, *ai*, *ua*), then *a* > **o**; **5)** C*a*, and d.f. ends in VV, and S2 has three syllables or two syllables of which the first one has vowels *o*, *u*, then *a* > **i**; **6)** VV, and d.f. ends in *va*/*vä*, *ha*/*hä*, *ja*/*jä*, then – the first V from S2 → + **i**; **7)** VV, and d.f. ends in *ta*/*tä*, then – the first V from S2 → + **i / si**	**1)** itke > itki, kapaloičče > kapaloičči, tule > tuli, tariče > tariči **2)** varaja (d.f.=varata) – ja + si = varasi **3)** kiäntä (d.f.=kiäntyä) > kiänti **4)** anta (d.f.=antua) > anto, ruata (d.f.=ruatua) > ruato **5)** otta (d.f.=ottua) > otti **6)** juo (d.f.=juuvva) – u → jo + i = joi, **7)** rupie (d.f.=ruveta) – i → rupe + i / si = rupei / rupesi[9]

9 In the Karelian Proper the contracted verbs with *ta/tä* are characterized by the presence of two variants of the inflectional stem of the imperfect.

Stem	Stem generation rules	Example
PS5 (AUX weak vowel stem IMPF)	If S1 ends in **1)** C*u*, C*y*, C*o*, C*ö*, then + **i**; **2)** C*e*, then *e* > **i**; **3)** *ja*/*jä*, and d.f. ends in CV, then – *ja*/*jä* → + **si**; **4)** C*ä*, and d.f. ends in VV, then *ä* > **i**; **5)** C*a*, and d.f. ends in VV, and S1 has two syllables of which the first one has the vowel *a* (*a*, *au*, *ai*, *ua*), then *a* > **oi**; **6)** C*a*, and d.f. ends in VV, and S1 has three syllables or two syllables of which the first one has vowels *o*, *u*, then *a* > **i**; **7)** VV, and d.f. ends in VV, then S1 > PS4 → – terminal CV → + ***voi***; **8)** VV and d.f. ends in *va*/*vä*, *ha*/*hä*, *ja*/*jä*, then – the first V from S1 + **i**; **9)** VV and d.f. ends in *ta*/*tä*, then – the first V from S1+ **i / si**	**1)** kačo + i = kačoi **2)** ite > iti, kapaloiče > kapaloiči, tule > tuli, tariče > tariči **3)** varaja (d.f.=varata) – ja + si = varasi **4)** kiännä (d.f.=kiäntyä) > kiänni **5)** anna (d.f.=antua) > annoi, **6)** ota (d.f.=ottua) > oti **7)** rua (d.f.=ruatua) → ruato (PS4) – to → rua + voi = ruavoi **8)** juo (d.f.=juuvva) – u → jo + i = joi **9)** rupie (d.f.=ruveta) – I → rupe + i / si = rupei / rupesi
PS6 (AUX, 3PL, IMPF, IND)	**A**. If d.f. ends in VV, then d.f. > S1 → + **tt**, and here, if S1 ends in C*a*/C*ä*, then *a* > *e*, *ä* > *e* **B**. If d.f. ends in CV, then d.f. > S3 → + **t**	**A.** itkie > ite(S1) + tt = itett, antua > anna(S1) > anne + tt = annett **B.** juuvva > juo(S3) + t = juot, kapaloija > kapaloi(S3) + t = kapaloit, tulla > tul(S3) + t = tult, ruveta > ruvet(S3) + t = ruvett, tarita > tarit(S3) + t = taritt

4.4. Stage 4: deriving rules for word-form generation

For the stems, auxiliary stems and pseudo-stems obtained in stage 3 generation rules are developed for each grammatical form of the verb from Table 1, taking into account all possible inflectional types from Table 2.

Before looking into the rules, there is a feature of the Karelian phonetic system to be considered, i.e. vowel harmony. It plays an important role in the inflection process, as many affixes can appear with either front or back vowels (*mma~mmä*, *nun~nyn*, *ja~jä*). Hence, the choice of the grammatical marker's variant is programmed in the generator by introducing the following additional rule: 1) if d.f. has the vowels *a*, *o*, *u*, then the marker variant left of ~ is applied (e.g., for the 1st infini-

tive *antua* with affixes *mma~mmä* the output is *anna-mma*); 2) if, on the other hand, d.f. contains no vowels *a*, *o*, *u*, then the marker variant right of ~ is used (e.g., for the 1st infinitive *itkie* with affixes *mma~mmä* the output is *ite-mmä*). Another important factor for the generation of some verb forms is the number of syllables in the stem, which requires another specifying rule: a stem has as many syllables as it has indissoluble vowel component combinations (solo vowels, diphthongs, triphthongs) separated by consonants (*ka-pa-loi-če*, *ru-pieu*).

Table 4 suggests the rules for the full inflection paradigm, but the forms produced within the same mood and tense from the same stems, or analytical constructs of the same type are combined within the same cell and examples in a row are given only for the first one.

Let us take a sample line in Table 4 to explain how these rules are applied to get the word form. Line "INSTR II" in the table reads:

1) if **d.f.** ends in VV (except ie), then = S2 + **en**
2) else = **d.f.** (Ca/Cä > Ce) + **n**

This means that to get the instructive 2nd infinitive form, we check whether the dictionary form ends in two successive vowels (except for *ie*):

1) if yes, then the ending *en* is added to stem 2;
2) otherwise, *n* is added to the dictionary form, and in this case, if the dictionary form ends in a consonant + *a/ä*, then *a/ä* is first replaced with *e*, and then *n* is added.

Table 4. Rules for word-form generation in verb inflection in the newly written Karelian Proper supradialect.

Form	Rule (for affirmative word forms)	Example (affirmative)	Rule (for negative word forms)	Example (negative)
Indicative				
1SG PRS	**S1 + n** Similarly: *2SG*: + **t** *1PL*: + **mma~mmä** *2PL*: + **tta~ttä**	ite+n anna+n juo+n kapaloiče+n tule+n rupie+n tariče+n	**en S1** Similarly: *2SG*: **et** *1PL*: **emmä** *2PL*: **että**	en ite en anna en juo en kapaloiče en tule en rupie en tariče
3SG PRS	**S2** (terminal C*e* > C*o*~C*ö*) + **u~y**	itke > itkö+y anta+u juo+u kapaloičče > kapaloiččo+u tule > tulo+u rupie+u taričče > tariččo+u	**ei S1**	ei ite ei anna ei juo ei kapaloiče ei tule ei rupie ei tariče
3PL PRS	If **d.f.** ends in 1) VV, **d.f.** > **S1** (C*a*/C*ä* in S2 *a* > *e*, *ä* > *e*) + **tah~täh**; 2) CV, then + **h**	ite+täh anne+tah juuvva+h kapaloija+h tulla+h ruveta+h tarita+h	**ei** If **d.f.** ends in 1) VV, **d.f.** > **S1** (C*a*/C*ä* in S2 *a* > *e*, *ä* > *e*) + **ta~tä**; 2) CV, then = **d.f.**	ei ite+tä ei anne+ta ei juuvva ei kapaloija ei tulla ei ruveta ei tarita
1SG IMPF	**PS5** + **n** Similarly: *2SG*: + **t**	iti+n annoi+n joi+n kapaloiči+n tuli+n rupei+n / rupesi+n tariči+n	**en *PTCP II ACT CONTR*** Similarly: *2SG*: **et**	en itken en antan en juonun en kapaloinun en tullun en ruvennun en tarinnun

Form	Rule (for affirmative word forms)	Example (affirmative)	Rule (for negative word forms)	Example (negative)
3SG IMPF	**PS4** Similarly: *1PL*: + **ma~mä** *2PL*: if PS4 ends in 1) VV, then + **ta~tä** 2) CV, then + **ja~jä**	itki anto joi kapaloičči tuli rupei / rupesi taričči	**ei *PTCP II ACT CONTR*** Similarly: *1PL*: **emmä** *2PL*: **että**	ei itken ei antan ei juonun ei kapaloinun ei tullun ei ruvennun ei tarinnun
3PL IMPF	**PS6** + **ih**	itett-ih annett-ih juot-ih kapaloit-ih tult-ih ruvett-ih taritt-ih	**ei PTCP II PASS**	ei itetty ei annettu ei juotu ei kapaloitu ei tultu ei ruvettu ei tarittu
1SG PERF	**olen *PTCP II ACT CONTR*** Similarly: *2SG*: **olet** *3SG*: **on** *1PL*: **olemma** *2PL*: **oletta** PLUP: *1SG*: **olin** *2SG*: **olit** *3SG*: **oli** *1PL*: **olima** *2PL*: **olija**	olen itken olen antan olen juonun olen kapaloinun olen tullun olen ruvennun olen tarinnun	**en ole *PTCP II ACT CONTR*** Similarly: *2SG*: **et ole** *3SG*: **ei ole** *1PL*: **emmä ole** *2PL*: **että ole** PLUP: *1SG*: **en ollun** *2SG*: **et ollun** *3SG*: **ei ollun** *1PL*: **emmä ollun** *2PL*: **että ollun**	en ole itken en ole antan en ole juonun en ole kapaloinun en ole tullun en ole ruvennun en ole tarinnun
3PL PERF	**on *PTCP II PASS*** Similarly for PLUP: *3PL*: **oli**	on itetty on annettu on juotu on kapaloitu on tultu on ruvettu on tarittu	**ei ole *PTCP II PASS*** Similarly for PLUP: *3PL*: **ei oltu**	ei ole itetty ei ole annettu ei ole juotu ei ole kapaloitu ei ole tultu ei ole ruvettu ei ole tarittu

Form	Rule (for affirmative word forms)	Example (affirmative)	Rule (for negative word forms)	Example (negative)
Imperative				
2SG	**S1**	ite anna juo kapaloiče tule rupie tariče	**elä S1**	elä ite elä anna elä juo elä kapaloiče elä tule elä rupie elä tariče
3SG & PL	If **S3** (terminal *t* > ***k***) ends in 1) CV, then + **kkah~kkäh** 2) VV, C, then + **kah~käh** Similarly: *1PL*: + **kka~kkä** / **ka~kä** *2PL*: + **kkua~kkyä** / **kua~kyä**	itke+kkäh anta+kkah juo+kah kapaloi+kah tul+kah ruvet > ruvek+kah tarit > tarik+kah	**elkäh *POSIT form*** Similarly: *1PL.*: **elkä** *2PL*: **elkyä**	elkäh itkekkäh elkäh antakkah elkäh juokah elkäh kapaloikah elkäh tulkah elkäh ruvekkah elkäh tarikkah
Conditional				
1SG IMPF	If **S2** (C*e* > C***i***) ends in 1) CV (except C*i*), and S2 has 3 syllables, then + **isin**; 2) VV, and d.f. ends in *va*/*vä*, *ha*/*hä*, *ja*/*jä*, then – the first V of VV → + **isin**; 3) else, + **sin**; Similarly: *2SG*: + **sit / isit** *1PL.*: **+ sima~simä / isima~isimä** *2PL*: **+ sija~sijä / isija~isijä**	itke > itki+sin anta+sin juo > joisin kapaloičče > kapaloičči+sin tule > tuli+sin rupie+sin tariče > tariči+sin	**en *3 SG PRS COND AFF*** Similarly: *2SG*: **et** *1PL*: **emmä** *2PL*: **että**	en itkis en antais en jois en kapaloiččis en tulis en rupieis en tariččis

Form	Rule (for affirmative word forms)	Example (affirmative)	Rule (for negative word forms)	Example (negative)
3SG IMPF	If **S2** (C*e* > C***i***) ends in 1) C*i*, then + **s** 2) CV (except C*i*), then + **is** 3) VV, and d.f. ends in *ta*/*tä*, then + **is** 4) VV, and d.f. ends in *va*/*vä*, *ha*/*hä*, *ja*/*jä*, then – the first V of VV → + **is**	itke > itki+s anta+is juo > jo+is kapaloičče > kapaloičči+s tule > tuli+s tariče > tariči+s rupie+is	***ei 3 SG PRS COND AFF***	ei itkis ei antais ei jois ei kapaloiččis ei tulis ei rupieis ei tariičis
3PL IMPF	**PS6** + **ais~äis**	itett+äis annett+ais juot+ais kapaloit+ais tult+ais ruvett+ais taritt+ais	**ei PS6** + **ais~äis**	ei itett+äis ei annett+ais ei juot+ais ei kapaloit+ais ei tult+ais ei ruvett+ais ei taritt+ais
1SG PLUP	**olisin *PTCP II ACT CONTR*** Similarly: *2SG*: **olisit** *3SG*: **olis** *1PL.*: **olisima** *2PL.*: **olisija**	olisin itken olisin antan olisin juonun olisin kapaloinun olisin tullun olisin ruvennun olisin tarinnun	**en olis *PTCP II ACT CONTR*** Similarly: *2SG*: **et olis** *3SG*: **ei olis** *1PL*: **emmä olis** *2PL*: **että olis**	en olis itken en olis antan en olis juonun en olis kapaloinun en olis tullun en olis ruvennun en olis tarinnun
3PL PLUP	***olis PTCP II PASS***	olis itetty olis annettu olis juotu olis kapaloitu olis tultu olis ruvettu olis tarittu	***ei olis PTCP II PASS***	ei olis itetty ei olis annettu ei olis juotu ei olis kapaloitu ei olis tultu ei olis ruvettu ei olis tarittu

Form	Rule (for affirmative word forms)	Example (affirmative)	Rule (for negative word forms)	Example (negative)
Potential				
1SG PRS	If **S3** (*t* > ***n***) ends in 1) V, *n, h* + **nen**; 2) *l*, then + **len**; 3) *r*, then + **ren**; 4) *š*, then + **šen**; 5) *s*, then + **sen** Similarly: *2SG*: **+ net / let / ret / šet / set** *3SG*: **+ nou~nöy / lou~löy / rou~röy / šou~šöy / sou~söy** *1PL*: **+ nemma~nemmä / lemma~lemmä / remma~remmä / šemma~šemmä / semma~semmä** *2PL*: **+ netta~nettä / letta~lettä / retta~rettä / šetta~šettä / setta~settä**	itke+nen anta+nen juo+nen kapaloi+nen tul+len ruvet > ruven+nen tarit > tarin+nen	**en** If **S3** (*t* > ***n***) ends in 1) V, *n*, *h* + **ne**; 2) *l*, then + **le**; 3) *r*, then + **re**; 4) *š*, then + **še**; 5) *s*, then + **se** Similarly: *2SG*: **et** *3SG*: **ei** *1PL*: **emmä** *2PL*: **että**	en itkene en antane en juone en kapaloine en tulle en ruvenne en tarinne
3PL PRS	**PS6** + **aneh~äneh**	itett+äneh annett+aneh juot+aneh kapaloit+aneh tult+aneh ruvett+aneh taritt+aneh	**ei PS6** + **ane~äne**	ei itett+äne ei annett+ane ei juot+ane ei kapaloit+ane ei tult+ane ei ruvett+ane ei taritt+ane
1SG PERF	**lienen *PTCP II ACT CONTR*** Similarly: *2SG*: **lienet** *3SG*: **lienöy** *1PL*: **lienemmä** *2PL*: **lienettä**	lienen itken lienen antan lienen juonun lienen kapaloinun lienen tullun lienen ruvennun lienen tarinnun	**en liene *PTCP II ACT CONTR*** Similarly: *2SG*: **et liene** *3SG*: **ei liene** *1PL*: **emmä liene** *2PL*: **että liene**	en liene itken en liene antan en liene juonun en liene kapaloinun en liene tullun en liene ruvennun en liene tarinnun

Form	Rule (for affirmative word forms)	Example (affirmative)	Rule (for negative word forms)	Example (negative)
3PL PERF	**lienöy *PTCP II PASS***	lienöy itetty lienöy annettu lienöy juotu lienöy kapaloitu lienöy tultu lienöy ruvettu lienöy tarittu	**ei liene *PTCP II PASS***	ei liene itetty ei liene annettu ei liene juotu ei liene kapaloitu ei liene tultu ei liene ruvettu ei liene tarittu
Infinitives				
I	**d.f.**		itkie antua juuvva kapaloija tulla ruveta tarita	
II INE	If **d.f.** (Ca, Cä > Ce) ends in 1) C*e*, then + **šša~ššä**; 2) *ua, uo*, then + **s's'a**; 3) *yä, yö, ie*, then + **šša~ššä**		itkie+ššä antua+s's'a juuvva > juuvve+šša kapaloija > kapaloije+šša tulla > tulle+šša ruveta > ruvete+šša tarita > tarite+šša	
II INSTR	1) if **d.f.** ends in VV (except *ie*), then = S2 + *en* 2) else = **d.f.** (Ca/Cä > Ce) + **n**		itkie+n antua > anta+en juuvva > juuvve+n kapaloija > kapaloije+n tulla > tulle+n ruveta > ruvete+n tarita > tarite+n	
III ADE-ALL	**S2** (C*e* > C*o*~C*ö*) + **malla~mällä** Similarly: *ELA.*: **mašta~mäštä** *ILL*: **mah~mäh** *ABE*: **matta~mättä** *INE*: **mašša~mäššä**		itke > itkö+mällä anta+malla juo+malla kapaloičče > kapaloiččo+malla tule > tulo+malla rupie+malla taričče > tariččo+malla	

Form	Rule (for affirmative word forms)	Example (affirmative)	Rule (for negative word forms)	Example (negative)
Participles				
I ACT	**S2** (C*e* > C*i*) + **ja~jä**		itke > itki+jä anta+ja juo+ja kapaloičče > kapaloičči+ja tule > tuli+ja rupie+ja taričče > taričči+ja	
II ACT CONTR	If **S3** (*t* > *n*) ends in 1) CV, then + **n**; 2) VV, *n*, *h*, then + **nun~nyn**; 3) *l*, then + **lun~lyn**; 4) *r*, then + **run~ryn**; 5) *š*, then + **šun~šyn**; 6) *s*, then + **sun~syn**		itke+n anta+n juo+nun kapaloi+nun tul+lun ruvet > ruven+nun tarit > tarin+nun	
II ACT full	If **S3** (*t* > *n*) ends in 1) CV, then + **nut~nyt**; 2) VV, *n*, *h*, then + **nut~nyt**; 3) *l*, then + **lut~lyt**; 4) *r*, then + **rut~ryt**; 5) *š*, then + **šut~šyt**; 6) *s*, then + **sut~syt**		itke+nyt anta+nut juo+nut kapaloi+nut tul+lut ruvet > ruven+nut tarit > tarin+nut	
II PASS	**PS6** + **u~y** Similarly: *PTCP I PASS*: + **ava~ävä**		itett+y annett+u juot+u kapaloit+u tult+u ruvett+u taritt+u	

The full version of Tables 3 and 4 is available online (Krizhanovskaya, Novak & Pellinen 2021).

5. Design of generator software for the verb inflection paradigm

The rules thus derived serve as the basis for automatic word-form generation from a minimized template, implemented as functions in Dictorpus software for the Karelian Proper supradialect of the Karelian language. VepKar project editor fills in the template (sample templates: *it|kie [e], an|tua [na], ju|uvva [o], kapaloi|ja [če], tul|la [e], ru|veta [pie], tari|ta [če]*) in the lemma field (Fig. 2), and the system automatically generates 125 word forms of the verb paradigm. For impersonal verbs[10] (corresponding box ticked by the editor), 1st and 2nd person singular and plural forms, and the 3rd person plural form are not filled in the program.

The program code was written in PHP using the Laravel platform and the code is available online at GitHub.[11]

10 Verbs that express actions or states that occur without the participation of the actor.

11 See https://github.com/componavt/dictorpus

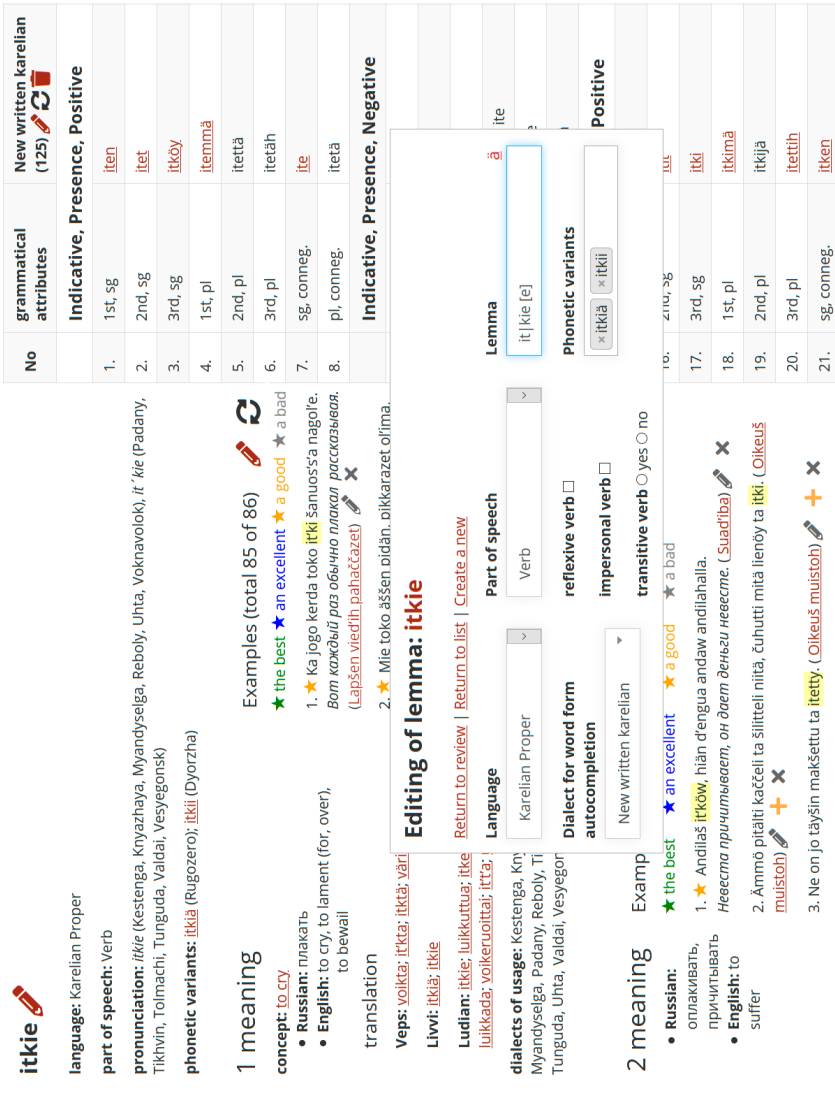

Figure 2. Adding word forms via the lemma template. The window in the center shows a fragment of the lemma editing form. The Lemma field is filled with the template *it|kie [e]*. The right-hand part of the figure is a fragment of the list of 125 word forms generated from this template. The part on the left is a fragment of the *itkie* dictionary entry with examples (quotations) from the VepKar corpus.

6. Results

Having studied the verb inflection system of newly written varieties of the Karelian language with view to creating an automatic word-form generator, we managed to identify and unify the generation rules for all grammatical forms of verbs of major inflectional types. The minimal amount of grammatical information that needs to be manually filled in the VepKar corpus database to enable automatic completion of the entire verb infection paradigm was reduced to one pseudo-ending. This was made possible by developing rules for generating imperfect forms for single-stem verbs with the stem ending in *a-, ä-* since such rules were insufficiently described in existing grammar books. Fairly clear-cut regularities in the generation of such forms were revealed by running some experiments based on material from the Livvi subcorpus of VepKar. The rules identified for the Livvi supradialect are valid in the Karelian Proper supradialect, as well.

It was found during the study that there is no uniformity in the generation of analytical 3rd person plural forms using participle II passive: perfect indicative – *on anne-ttu*, *ei ole anne-ttu*, pluperfect indicative *oli anne-ttu*, *ei oltu anne-ttu*. In the perfect tense the auxiliary verb *olla* appears in the 3rd person singular, but in the pluperfect the affirmative form of the auxiliary verb is the 3rd person singular, while the negative form is the 3rd person plural. This lack of uniformity suggests rules for the standard variant of the language can be refined further.

Full automation of the process of verb inflectional paradigm generation, i.e. training the program to automatically recognize the verb's inflectional type, so far seems impossible since indication of the vowel stem type is not always contained within the appearance of the dictionary word (basic form of the word). Still, the verb word-form generator in its present form has considerably reduced the time needed to complete the inflection paradigm per verb: down to 10 seconds, i.e. 120-fold less, on average.

Thus, the generator of word-forms for the Veps language, the Karelian Proper and Livvi supradialects of the Karelian language is part of the VepKar corpus. Note that the morphological analyzer and word form generator of the Livvi dialect was previously implemented in the Giellatekno framework based on the HFST program (Rueter 2014) and

is available online.[12] We believe that cooperation with the developers of Giellatekno (for example, with the aim of interchanging dictionary data and comparing the work of paradigm generators) would be promising.

7. Conclusions

This article is based on a solid background of preceding work to analyze the inflectional system of standardized varieties of the Karelian language, identify patterns in the generation of certain word forms, and fill in rule gaps. The latest developments (based on experiments with corpus materials) help make corrections and additions to existing grammar rules, and immediate effects of that will be changes to the practices of Karelian language instruction in the Republic of Karelia, and submission of recommendations to editorial departments of minority-language media outlets.

The verb inflectional paradigm generator was created for the Karelian Proper as well as for the Livvi subcorpora of VepKar, and a similar program has been previously designed for their nominal inflection systems. The new rules have already enabled entering 275 000 Karelian Proper word forms in the VepKar dictionary, improving the tagging coverage of the Karelian Proper subcorpus by 7%. With all that, we approach the creation of a Karelian spell checker application, and take a large step towards designing a morphological analyzer and enabling machine translation.

The morphologically marked corpus will become the largest base for theoretical research, including the analysis of the little-studied syntactic system of the language, as well as various statistical studies. Mobile applications based on the corpus data will be developed further, this is very important for maintaining the vital potential of the language in the modern world.

12 See https://giellatekno.uit.no/cgi/p-olo.eng.html

Abbreviations

1, 2, 3 – 1st, 2nd, 3rd person, ABE – abessive, ADE – adessive, ACT – active voice, ADE-ALL – adessive-allative, AFF – affirmative form, AUX – auxiliary, COND – conditional, CONTR – contracted form, d.f. – dictionary form, d.w. – dictionary word, ELA – elative, full – full form, ILL – illative, IMPF – imperfect, INE – inessive, INSTR – instructive, PASS – passive, PERF – perfect, PL – plural, PLUP – pluperfect, PRS – present, PS – pseudo-stem, PTCP – participle, SG – singular.

References

Adel', Elena L. 1998. *Glagol'noe slovoizmenenie v karel'skom jazyke (padanskij dialekt).* Petrozavodsk.

Boiko, Tatyana P. 2018. *Bol'shoy karel'sko-russkiy slovar' (livvikovskoye narechiye)*. Petrozavodsk: Periodika.

Bubrikh, Dmitry. 1947. Proiskhozhdenie karel′skogo naroda : povest′ o soi͡uznike i druge russkogo naroda na severe. Petrozavodsk: Karelo-Finnish SSR. http://knk.karelia.ru/site/bub/bubrih_new/index.html.

Bubrikh, Dmitry. 1948. Istoricheskoye proshloye karel'skogo naroda v svete lingvisticheskikh dannykh // Newsletter of the Karelian-Finnish Research Facility of the USSR Academy of Science. Petrozavodsk.

Garrett, Nicolai, L. Dylan, A. D. McCarthy, A. Mueller, W. Wu & D. Yarowsky. 2020. Fine-grained Morphosyntactic Analysis and Generation Tools for More Than One Thousand Languages. *Proceedings of The 12th Language Resources and Evaluation Conference*. 3963–3972. https://www.aclweb.org/anthology/2020.lrec-1.488.

Hakulinen, Auli, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen & Irja Alho. 2004. *Iso suomen kielioppi.* (Editions of Finnish Literature Society 950). Helsinki: Finnish Literature Society.

Koivisto, Vesa. 2018. Border Karelian dialects – a diffuse variety of Karelian. In Marjatta Palander, Helka Riionheimo & Vesa Koivisto (eds.), *On the border of language and dialect*, 56–84. Helsinki: Suomalaisen Kirjallisuuden Seura. http://library.oapen.org/handle/20.500.12657/29742.

Krizhanovskaya, Natalia & Andrew Krizhanovsky. 2019. Semi-automatic methods for adding words to the dictionary of VepKar corpus based on inflectional rules extracted from Wiktionary. *Corpus linguistics – 2019*, 211–217. Saint Petersburg: SPbU. https://events.spbu.ru/eventsContent/events/2019/corpora/corp_sborn.pdf#page=211.

Krizhanovsky, Andrew, Natalia Krizhanovsky & Irina Novak. 2019. Dialects in the Open corpus of Veps and Karelian languages (VepKar). *Corpus linguistics – 2019*, 288–295. Saint Petersburg: SPbU. https://events.spbu.ru/eventsContent/events/2019/corpora/corp_sborn.pdf#page=288.

Krizhanovskaya, Natalia, Irina Novak, Natalia Pellinen & Tatyana Boiko. 2021. Rules of generation of nominal word forms using a minimized template for newly written variants of the Karelian Proper and Livvi dialects (in Russian) // Figshare. 2021. Preprint. https://doi.org/10.6084/m9.figshare.14241833.v1.

Krizhanovskaya, Natalia, Irina Novak & Nataliya Pellinen. 2021. *Pravila generatsii glagol'nogo slova po minimizirovannomu shablonu dlya novmennogo severnokarel'skogo varianta karel'skogo yazyka*. Figshare. Preprint. https://doi.org/10.6084/m9.figshare.14237843.v7.

Moseley, Christopher (ed.). 2010. *Atlas of the World's Languages in Danger*. 3rd edn. Paris: UNESCO Publishing.

Nagurnaya, Svetlana. 2019. Karel'skaya pis'mennost'. *Narody Karelii: istoriko-etnograficheskie ocherki*, 65–77. Petrozavodsk: Periodika.

Novak, Irina, M. Penttonen, A. Ruuskanen & L. Siilin. 2019. *Karel'skiy yazyk v grammatikakh*. Petrozavodsk: KarRC RAS.

Pahomov, Miikul. 2017. Lyydiläiskysymys: Kansa vai heimo, kieli vai murre? Helsinki: Helsingin yliopisto.

Pirinen, T. A., T. Trosterud, F. M. Tyers, V. Vincze, E. Simon & J. Rueter. 2016. Foreword to the Special Issue on Uralic Languages. *Northern European Journal of Language Technology* 4(1). 1–9. https://doi.org/10.3384/nejlt.2000-1533.1641.

Pirinen, Tommi, 2019a. Building minority dependency treebanks, dictionaries and computational grammars at the same time – an experiment in Karelian treebanking, *Proceedings of the Third Workshop on Universal Dependencies*, 132–136. UDW, SyntaxFest. https://doi.org/10.18653/v1/W19-8016.

Pirinen, Tommi. 2019b. Workflows for kickstarting RBMT in virtually No-Resource Situation, *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, 11–16.

Rueter, Jack. 2014. The Livonian-Estonian-Latvian Dictionary as a threshold to the era of language technological applications. *Journal of Estonian and Finno-Ugric Linguistics* 5(1). 251–259. http://doi.org/10.12697/jeful.2014.5.1.14.

Rueter, Jack & Mika Hämäläinen. 2017. Synchronized Mediawiki based analyzer dictionary development. *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, 1–7. St. Petersburg. https://doi.org/10.18653/v1/W17-0601.

Russian Census. 2010. *Volumes of the official publication of the results of the Russian Census 2010*. http://www.gks.ru/free_doc/new_site/perepis2010/croc/perepis_itogi1612-tom4.htm (27 October, 2021).

Sarhimaa, Anneli. 2017. *Vaietut ja vaiennetut, Karjalankieliset karjalaiset Suomessa*. Tietolipas 256. Helsinki: Suomalaisen Kirjallisuuden Seura.

Torikka, Marja (ed.). 2009. *Karjalan kielen verkkosanakirja*. Kotimaisten kielten tutkimuskeskuksen verkkojulkaisuja. Helsinki. URL: http://kaino.kotus.fi/kks (27 October, 2021).

Trosterud, Sindre Reino, Trond Trosterud, Anna-Kaisa Räisänen, Leena Niiranen, Mervi Haavisto & Kaisa Maliniemi. 2017. A morphological analyser for Kven. *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, 76–88. https://doi.org/10.18653/v1/W17-0608.

VepKar. 2021. *Open corpus of Veps and Karelian languages*. Karelian Research Centre of RAS. http://dictorpus.krc.karelia.ru (27 October, 2021).
Zaiceva, Nina. 2019. Veps language heritage in Karelia. *Multi lingual Finnic*. Helsinki: Suomalais-Ugrilainen Seura, 379–400. https://doi.org/10.33341/uh.85043.
Zaikov, Pekka. 2000. *Glagol v karel'skom jazyke*. Petrozavodsk: PetrSU.
Zaikov, Pekka. 2002. *Karjalan kielioppi*. Petroskoi: Periodika.
Zaikov, Pekka. 2013. *Vienankarjalan kielioppi*. Helsinki: Karjalan Sivistysseura.
Zaikov, P. M., V. I. Karakina, M. A. Spitsyna, N. N. Arkhipova, T. I. Medvedeva, N. A. Pellinen, I. P. Novak, G. E. Lettieva. 2015. *Russian-Karelian Dictionary (north-Karelian subdialects)*. Petrozavodsk: Periodika.

Kokkuvõte. Natalia Krizhanovskaya, Irina Novak, Andrew Krizhanovsky, Nataliya Pellinen: Morfoloogilised muutereeglid päriskarjala verbide jaoks. Artiklis esitletakse metodoloogiat, mida kasutati muutereeglite väljatöötamisel ja rakendamisel päriskarjala verbide jaoks. Materjali moodustasid vepsa ja karjala keele avatud korpusest (VepKar) ning karjala keele sõnaraamatu elektroonilisest versioonist kogutud lemmad ja sõnavormid. Esmakordselt arendati välja reeglite süsteem päriskarjala verbivormide automaatseks genereerimiseks. See on teaduslikult huvipakkuv nii praktilise kui ka teoreetilise poole pealt. Uued reeglid on juba võimaldanud lisada 141 000 päriskarjala sõnavormi VepKar sõnaraamatusse. Uus sõnavormide genereerimise programm on oluliselt vähendanud aega, mis kulub täieliku muuteparadigma lisamisele mingi päriskarjala verbi juurde Vepkar sõnaraamatus. 125 sõnavormi asemel on selleks üksnes vaja täita mallid mõningate parameetritega.

Märksõnad: karjala keel, päriskarjala, korpuslingvistika, verbi muutemorfoloogia, muuteparadigma, sõnavormide genereerimine