BRILL

JOURNAL OF LANGUAGE CONTACT 14 (2021) 684–719

JOURNAL OF LANGUAGE CONTACT
brill.com/jlc

Vocabulary-Based Classification and Contact-Induced Formation of Neologisms in Two Standard Varieties of Karelian

Susanna Tavi
Early Stage Researcher, Finnish language, School of Humanities, University of Eastern Finland, Joensuu, Finland
susanna.tavi@uef.fi

Lauri Tavi
Postdoctoral Researcher, General linguistics, School of Humanities, University of Eastern Finland, Joensuu, Finland
lauri.tavi@uef.fi

Abstract

This paper investigates the lexical similarities and formation of neologisms of two written standard varieties of Karelian, North and Livvi Karelian, spoken in the Republic of Karelia, Russia. Firstly, a naïve Bayes statistical model was generated to classify North and Livvi Karelian newspaper texts automatically. Secondly, the word formation strategies of neologisms from the classified newspaper texts were studied. The strategies between the two varieties were compared in terms of the code-copying framework. The results from the automatic classification and the investigation of neologisms show that the standards differ in lexicon and phonology, but the strategies of forming neologisms are similar: the most common strategy is to form words by language-internal means, and the other strategies are selective and global copying from Finnish, Russian, and English. The similar strategies in both standards suggest similar language planning.

Keywords

neologisms – standardization – Karelian – naïve Bayes classification – code-copying framework

1 Introduction

The Karelian language is an endangered Finnic minority language that, along with Hungarian, Finnish and Estonian, belongs to the Finno-Ugric language group of Uralic languages. Karelian is spoken in the Republic of Karelia (North-West Russia), Tver Oblast (Russia), and Finland. Since the 1990s, the Karelian language has been in a process of revitalization, which, by definition, is the process of promoting a language to become used in all domains (Huss, 2001: 278). Revitalization often begins with creating a written standard variety, which has a symbolic value (Tánzcos, 2015: 94) and the capacity to open new domains for language use (Sulkala, 2010). Creating a written standard variety requires standardization of existing grammar and the creation of a new vocabulary, since endangered languages, such as Karelian, often lack words for novel concepts of society and inventions. However, the Karelian language has several differing dialects, and the speakers of the Karelian dialects have found the differences too extensive for the creation of one written standard variety (e.g., Kunnas and Arola, 2010).

This study focuses on Karelian in the Republic of Karelia where two written standard varieties, North and Livvi Karelian, have developed and are currently in use. North and Livvi Karelian dialects have the most developed standards, and a weekly newspaper, *Oma Mua*, is published in both varieties. *Oma Mua* plays an important role in the revitalization of the Karelian language in serving as an example of language use and as a source of neologisms. Neologisms are essential to revitalizing endangered languages by providing speakers with a lexicon to describe modern-day phenomena. Karelian neologisms were studied as early as in the 1990s when the revitalization and conscious development of vocabulary began after decades of the assimilation policies of the Soviet Union. The first study was published in 1997 (Õispuu, 1997), and an extensive glossary was published in 2003 (Õispuu, 2003a, 2003b). Ever since, however, research on the linguistic strategies used in neologism formation has only been a minor part of the studies (e.g., Karjalainen et al., 2013; Tánzcos, 2015).

The aim of this study was to: 1) examine the similarities and differences in the vocabularies of the two standard varieties of Karelian and 2) compare the word formation strategies of the neologisms in the two varieties in order to examine whether they are similar or divergent and evaluate what ideologies may have affected the word formation strategies. Neologisms were focussed on in order to assess whether the conscious development of a new lexicon involves new strategies or whether the development is similar to the established lexical features of the two varieties.

The methodology of the study was thus two-fold. Firstly, text corpora from two Karelian newspapers (*Oma Mua* and *Vienan Karjala*) were used to generate a simple naïve Bayes classification model for North and Livvi Karelian standard varieties. Then, the model was used to automatically classify the editions of *Oma Mua* issued over one year, including texts in both varieties. The classification was done automatically in order to process the large linguistic data sets from an objective point of view; another purpose was to demonstrate the use of an automatic classifier, which has rarely been utilized in previous studies of language contacts or the Karelian language. Secondly, the result of the classification was divided into two corpora, North and Livvi Karelian. The two-fold approach to the lexicon and lexical renewal of Karelian was chosen, firstly, to determine the lexical similarities and differences of the varieties before analysing the bidialectal newspaper and, secondly, to compare the two and to identify what strategies for creating new words are at work in each variety. Are they coming closer to each other in terms of standardizing the written form, or are the development patterns different? It was also considered whether ideologies may regulate the process. The investigation also aimed to determine whether the neologisms were formed by language-internal means or by copying from Finnish, Russian or English models. The distribution of these categories was investigated within the code-copying framework.

The research questions are:

1) What variety-related lexical characteristics are revealed in the automatic classification of North and Livvi Karelian varieties? Which of the two varieties is dominant in the Karelian newspaper corpus based on automatic classification?
2) What are the similarities or differences in neologisms between the automatically classified North and Livvi Karelian varieties? How are the neologisms formed?
3) What do the neologisms and the other lexical characteristics reveal about the development of the two different written standard varieties of Karelian?

The article is structured as follows. Section 2 presents the Karelian language and its history and contact settings. Additionally, standardization and revitalization are explained with a focus on North and Livvi Karelian varieties. Section 3 discusses the creation of Karelian neologisms; the concepts of conscious vocabulary development, ideologies, word formation strategies, and the code-copying framework are described. Section 4 presents the newspaper corpora, their aggregation and metrics and research methods, including the naïve Bayes model and comparison of neologisms. The classification results

are presented in Section 5, and findings on the neologisms are presented in Section 6. The discussion in Section 7 concludes the article.

2 The Karelian Language

The Karelian language and its dialect taxonomy have several definitions depending on whether the research is conducted within a Russian or Finnish tradition (Karjalainen et al., 2013: 47). Section 2.1 briefly introduces the development of Karelian with a focus on the North and Livvi varieties. Language revitalization and standardization are discussed in Section 2.2. Language contacts have played an important role in the development of Karelian; hence, this section is presented from the point of view of language contacts and contact-induced variation and change.

2.1 *The Development of North and Livvi Karelian Varieties*

According to the traditional Finnish taxonomy, Karelian has two main dialects, Karelian Proper and Livvi Karelian. The former has two subgroups, North and South Karelian. All varieties are spoken in the Republic of Karelia, Russia. In addition, South Karelian is spoken in Tver Oblast, Russia, as a consequence of migration caused by wars in the 17th century (Karjalainen et al., 2013: 47; Laakso et al., 2016: 96–97; See Fig.1). According to the 2010 census of Russia, 12,369 people defined Karelian as their mother tongue (Laakso et al., 2016: 97). In Finland, Karelian is a non-regional minority language; it is mainly an autochthonous and partly immigrant language. Speakers of all varieties live dispersed across the country. Approximately 5,000 people speak Karelian fluently, and 20,000 understand Karelian in Finland (Laakso et al., 2016: 105, 108).

The development of the Karelian language into its current forms (see Fig. 1) begins with the first contacts between Proto Finnic and East Slavic peoples around Lake Ladoga in current North-Western Russia between 300–700 AD (Kallio, 2006: 157). Throughout its existence, the Karelian language has been located in the borderland of (Novgorodian) Russia and Sweden (to which Finland belonged until 1809), and Karelians have never formed a state of their own (Laakso et al., 2016: 96). Several wars have shifted the location of this border, influencing the development of the Karelian language. However, the easternmost parts of Karelia have always been under Russian rule (Palander, Opas-Hänninen, and Tweedie, 2003: 359–360), resulting in a heavy linguistic influence of Russian on Karelian. Despite this strong influence, the Karelian language and culture have endured.

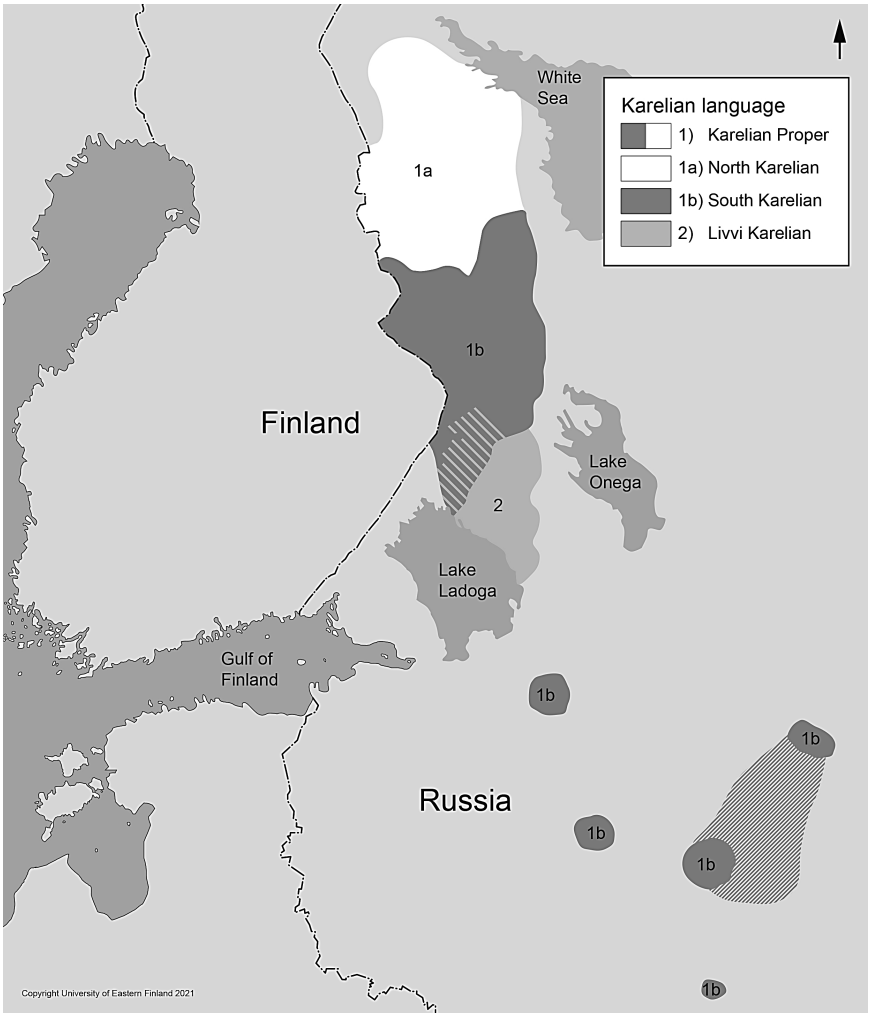

FIGURE 1 The geographical distribution of the dialects of the Karelian language.

At the beginning of the 20th century, only 10% of Karelians residing in the Soviet Union knew Russian. This subsequently changed, however, as the Soviet Union's language policies prohibited the use of Karelian in official contexts. The number of speakers decreased rapidly, and the speakers became, to a greater extent, bilingual with Russian (Karjalainen et al., 2013: 11). As a result, no language planning for Karelian existed in the 1970s and 80s, which strengthened russification. Finally, during the free atmosphere of perestroika and glasnost, the language minorities of Russia started receiving official support, which led to Karelian being recognized as one of the administrative languages

(Sarhimaa, 1996: 77–79). However, currently, less than half of ethnic Karelians speak Karelian (Karjalainen et al., 2013: 11).

North and Livvi Karelian are the main varieties used in the Republic of Karelia. They both have a written standard variety that is used, for instance, in the minority media in the area (Tánzcos, 2015). These written standard varieties were developed during the 1990s revitalization of Karelian in the Republic of Karelia. North Karelian is spoken in the region from the White Sea to the present-day Finnish border (see Fig. 1). The speakers of North Karelian are bilingual with Russian. In addition, contact with eastern dialects of Finnish via intensive trading and mobility has taken place from the 16th century onwards. For North Karelians, standard Finnish was a written standard variety during the Soviet period as well as the language of literature and, to some extent, radio and television. This use of Finnish was due to Soviet language policies rather than linguistic reasons (Kunnas, 2007: 58–61; Pasanen, 2006: 122–123). After the Second World War, the status of Finnish weakened in Karelia (Sarhimaa, 1996: 77–79). Due to linguistic similarities between North Karelian and Finnish (Anttikoski, 2003: 29), the languages are mutually intelligible to some extent. The eastern dialects of Finnish, in particular, are closely related to North Karelian.

Livvi Karelian, on the other hand, is spoken in the Olonets Isthmus on the eastern and north-eastern shores of Lake Ladoga (Karjalainen et al., 2013: 47; see Fig. 1). The speakers of Livvi Karelian are also bilingual with Russian. Whereas Finnish was the standard used for North Karelian during the Soviet period, Russian was used as the standard for speakers of Livvi Karelian (Anttikoski, 2003: 30). Livvi Karelian has had long and intensive contact with Russian. For instance, words and other linguistic features, such as the development of voiced plosives and sibilants, have been copied from and developed under the influence of Russian (Sarhimaa, 1995: 212). Moreover, the influence of Russian on the Finnic languages seems to increase gradually variety by variety eastwards (Koivisto, 1990: 20); the influence of Russian is thus greater in Livvi Karelian than in North Karelian. Other differences from North Karelian are also evident.The following excerpts written in North and Livvi Karelian demonstrate the differences and similarities between the written standards of the varieties.

> North Karelian: *Tule opaštumah kiäntämäh karjalakši!*
> *Mimmoni on kiäntäjän ruato? Ken, kuin, mitä ta konša on kiäntän karjalakši? Ta mitä pitäis karjalakši kiäntyä? Mimmoista kirjallisutta karjalakši kiännetäh? Kuin kiännöštieto opaštau kaunokirjallisuon kiäntämiseh?*

Mistä kiäntäjä šuau tietuo, konša vaštah tulou tuntematoin šana? Tahtositko kiäntyä Wikipedijua karjalakši?
Livvi Karelian: *Tule opastumah kiändämäh karjalakse!*
Mittuine on kiändäjän ruado? Ken, kui, midä da konzu on kiändänyh karjalakse? Da midä pidäs karjalakse kiändiä? Mittustu kirjalližuttu karjalakse kiännetäh? Kui kiännöstiedo opastau kaunehkirjalližuon kiändämizeh? Kuspäi kiändäi suau tieduo, konzu vastah tulou tundematoi sana? Tahtozitgo kiändiä Wikipediedu karjalakse?
Translation in English: Come and learn how to translate into Karelian! What is the translator's work like? Who has translated what into Karelian and how and when? And what should be translated into Karelian? What kind of literature is translated into Karelian? How do translation studies guide in translating prose? Where can a translator get information when encountering a foreign word? Would you like to translate Wikipedia into Karelian*?* (excerpts sourced from http://kianna-hanke.blogspot.com/search?updated-max=2018-10-06T16:28:00%2B03:00andmax-results=7, the English translation is the authors' own.)

The differences between the varieties occur especially in phonetics and phonology (Anttikoski, 2003: 33). The same lexemes sound different in each variety, such as the words *ruato* ~ *ruado* 'work' and *konša* ~ *konzu* 'when' in the above excerpts. The differences are often between voiced and unvoiced consonants, for instance /*d*/ ~ /*t*/, /*g*/ ~ /*k*/, /*b*/ ~ /*p*/, and /*z*/ ~ /*s*/ and in the quality of the sibilants, for instance /*s*/ ~ /*š*/ and /*z*/ ~ /*ž*/. In addition, differences in the quality of diphthongs exist in the last syllable, for instance ***kiäntyä*** ~ ***kiändiä*** in the excerpts. The speakers of North and Livvi dialects often perceive these differences as lexical and state that they do not understand all of the words (Kunnas and Arola, 2010: 126–127). Nevertheless, the varieties also have various similarities, for instance, North and Livvi Karelian share a rich suffixal morphology, basic grammar and lexicon (Karjalainen et al., 2013: 50). The written standard varieties reflect the spoken varieties to a great extent even though internal variation in the spoken varieties naturally exist (Section 3.1 briefly discusses the role of Russian in spoken Karelian, especially Livvi Karelian; see also Sarhimaa, 1999; Pyöli, 1996).

2.2 *Revitalization and Standardization*

Language revitalization began several decades ago as minorities and autochthonous peoples that had been under severe linguistic assimilation policy experienced ethnic awakening. Today, language revitalization has spread to all continents. At the societal level, revitalization of a language means speakers

are able to use their language actively in all domains, which includes domains where the use of the language has either disappeared or never existed (Huss, 2001: 278–279). Starting to use a minority language in novel domains also supports language development as it forces language users to develop new vocabulary.

Considering language use, several domains open to minority languages as a written standard variety have developed (Sulkala, 2010: 20). In Karelia, the development of written standard varieties has gone hand-in-hand with the publication of Karelian-language newspapers. For instance, the *Oma Mua* newspaper continues to play an important role in the revitalization process of Karelian by serving as a linguistic example for the language users; some Karelian dictionaries and glossaries are even based on the vocabulary used in *Oma Mua* (e.g., Õispuu 2003a, 2003b; Pyöli 2016). In addition, minority language media are important drivers in maintaining, developing, planning, and reviving languages as well as providing information. They signal competence in reporting on modern-day phenomena (Tánczos, 2015: 91–94). Standardization in the planning of a minority language thus makes the language visible (Mantila, 2010: 68) and, for this reason, standardization is often the first step in the revitalization process. The acts of standardization (i.e., development of written language) include creating new terms, renewing, establishing orthography, and choosing between alternatives (Sulkala, 2010: 17).

The greatest challenge to the standardization of Karelian seems to be the geographical variation between its dialects (Anttikoski, 2003: 34–35). The standardization of a language with substantial geographical variation has three different possible options. The first is to choose one dialect as the basis of the written standard variety. The second option is to find compromises between the different dialects. The third option is to accept the use of all the dialects in written form and use several written standard varieties (Mantila, 2010: 59–61). In Karelian, a combination of the first two of these options was briefly experimented with in the 1930s by language-internal means of creating new vocabulary and by copying from Russian using the Cyrillic alphabet. The resulting new standard Karelian was discarded soon after its creation due to political reasons. Linguistic problems also existed; the original morphology and lexicon of Karelian were reduced and modified to more closely resemble Russian. The standard was also not clearly based on any of the Karelian dialects and, thus, it was difficult to learn. (Sarhimaa, 1996: 74–76.)

Language activists' desire for a single, common written standard variety of Karelian has long existed (Kunnas and Arola, 2010: 126); however, speakers' experiences of substantial dialectal differences and the difficulty in finding compromises based either on the standard or on the selected forms have

prevented the creation of such a standard. However, the differences between the dialects are not as substantial as in Finnish, which has a common written standard variety based on both main dialects. Rigorous language planning can overcome dialectal differences in creating a common written standard variety. Nevertheless, the third option is the current state of the standardization process, with four different standard varieties: Livvi Karelian, North Karelian, Tver Karelian, and South Karelian. The first three have been developed in Russia since the beginning of the 1990s, and the fourth has been developed in Finland in recent years. The first two varieties are the most established, whereas the latter two are in the process of standardization.

3 Karelian Neologisms

In this study, a neologism is defined as a new word emerging in the Karelian language (for definition of new word, see Section 4.4). Thus, the term encompasses only new words and not changes in meaning of an old, established word. The term also encompasses occasionalisms, meaning words that have not yet been conventionalized but are in the propagation process. In addition, the term encompasses new global, selective and mixed lexical copies, that is, contact-induced word formation. This definition is used in order to reveal the developments of new concepts central to the revitalization and standardization process (see e.g., Belentschikow, 2015 for definitions of neologism). This section introduces conscious development and the ideologies behind neologisms (3.1) and Karelian neologism formation strategies, especially copying (code-copying framework; 3.2).

3.1 *Conscious Development of Vocabularies in Karelian Standard Varieties*

At the beginning of the revitalization process of the Karelian language in the 1990s, a special commission for creating neologisms was established. The commission's task was to invent new words as Karelian lacked many layers of lexicon because the language had never been used in areas such as science, technology, urban life, or politics (Markianova, 2005: 59–60). The commission published new words in special bulletins that were distributed to schools, educational institutions, mass media, and other organizations. Overall, 8,000 socio-political terms, 1,500 terms for the natural world and species, and 700 linguistic terms were introduced in the bulletins by the year 2005. The work of the commission lasted until 2011 (Karjalainen et al., 2013: 51).

The source of neologisms can be original lexemes of different dialects that might have been forgotten or other languages' lexemes. International words

common to most languages that are currently copied from the English model tend to be introduced through Russian. Other languages, such as the close cognate language Finnish, are sources of other new words (for more on the model codes for Karelian, see Markianova, 2005: 59). However, global copying is usually avoided; instead, derivation from the Finnish model is a common strategy. Neologisms independent of foreign models are formed by language-internal means, i.e., derivation and composition (Karjalainen et al., 2013: 51).

Language ideologies may direct the revitalization process of a language, i.e., whether the word formation is done by language-internal means or copied from a foreign model. In the case of Karelian, the ideology of linguistic purism has affected the conscious development of the standard varieties. Purism is a language ideology according to which foreign elements should be avoided at all levels of the language and its development. The ideology is based on the notion that all languages were pure at some point and that this status should be pursued in order to protect languages. Purism is a common ideology in the revitalization of languages as well as in all language development. Moreover, purism can be considered an important tool for maintaining languages, especially in the revitalization of minority languages (Spolsky, 2004: 22; Puura, 2019: 38–39). Language ideologies thus play an important role in the revitalization and standardization of a language.

Finnish as a source language was favoured over Russian at the beginning of the revitalization process. For instance, Õispuu (1997: 92) examined the creation of neologisms during the 1990s and observed that the written standard varieties used Finnish as a source language. According to him, the fact that Finnish is the closest cognate language, and that it has a stable written standard variety and status can be considered as a purist choice as a source of new words. In addition, Õispuu observed that the attitude towards Russian global copies was negative, which resulted in replacing earlier Russian copies with neologisms and Finnish copies in early issues of *Oma Mua*. For example, in early issues of *Oma Mua* the names of months were globally copied from Finnish (Õispuu, 1997: 92). However, at present, Karelian original names are in use.

Russian global copies are common and established in all varieties of Karelian, for instance, many of the Karelian particles, such as *ta* ~ *da* 'and' and *hot'* 'although', are global copies from Russian and also established in written varieties (Karjalainen et al., 2013: 56; Tavi and Tavi, 2019). In addition, the current speech practices of Russian Karelians contain code alternation, i.e., 'a switching between two basic codes yielding "mixed discourse"' (Johanson, 1999: 55), and global copies that are used especially as occasionalisms, for

example when discussing work, politics and professions. Karelian can be defined as a high-copying code, i.e., a code that contains a very high amount of copies, usually due to bilingualism (see Johanson, 2002b), as illustrated in Example 1 below in which global copies are shown in bold (Livvi Karelian, taken from Pyöli, 1994: 250).

(1)	*Sie*	*ol'i*	*suuri*	***konkursu***,	*minä*
	there	was	great	competition	I
	pyta-la-s'	***postupit'***,	*en*	***postupinnuh***,	***po konkurs-u***.
	try-f.pst-refl [russian inflection]	apply [russian inflection]	no-1.sg	get in-ptcp. perf [karelian inflection]	because of competition-m.sg.acc

'There was great competition, I tried to apply, I did not get in, because of competition'.

The negative attitude towards Russian copies is likely focused on new copies due to an attempt to avoid replicating the common code alternation and global copying of the spoken discourse, in keeping with the purist ideology of standardisation. In standardization, for instance Finnish is a preferred model as it helps to maintain the original elements in the vocabulary. For example, in our data the term *yhteistyösopimuš* 'co-operation agreement' is globally copied and phonologically adjusted to the North Karelian standard variety. The parts of the compound are *yhteis* 'common', *työ* 'work' and *šopimuš* 'agreement'. The first two parts are known in some form in all Karelian varieties, but the third part is known only in North Karelian. In other lexical contexts in Karelian, a global copy from Russian, *roato* 'work', would be used to refer to work instead of the word *työ*.

3.2 *Word Formation and Code-copying in Karelian*

Forming neologisms in Karelian standards is a complex, contact-induced process that combines language-internal means of word formation, i.e., derivation and composition (for North Karelian strategies, see Zaikov, 2013: 103–112; and for Livvi Karelian strategies, see Pyöli, 2011: 177–186), and use of a foreign model. The *code-copying framework* offers sophisticated terminology for contact-induced changes, i.e., different types of borrowing and code-switching, within one framework. The code-copying framework is applied to describe the contact-induced features of neologisms of Karelian because when creating new words it is impossible to separate a switch from a borrowing. In fact, *copying* means all kinds of contact-induced changes of language, synchronic and diachronic. A copy is never identical to its original because they belong to the

recipient language and are subject to its internal developments. Thus, copies are not 'transferred' or 'borrowed' from another code (Verschik, 2016: 193).

In the code-copying framework, *code* refers to any variety. The source language is the *model code*, in this study Russian, Finnish or English, and the receiving language the *basic code,* in this study Karelian. The term *global copy* corresponds to *loanword* and *insertional code-switching*. Most material, phonetic, semantic, combinational, and frequential features of an item are copied. For instance, the noun *ploatja* (North Karelian form) 'dress' (< Russian *plát´e*) corresponds to a loanword as is a conventionalized global copy (see Dictionary of Karelian (KKS), s.v. *ploatja*) and *pipo* 'knitted hat' (< Finnish *pipo*) corresponds to insertional code-switching as it is not conventionalized in Karelian and is used in the data of this study in quotation marks. Both of these instances are defined as global copies according to the code-copying framework. *Selective copy*, in contrast, is used for a single copied feature of an item, such as meaning. Thus, selective copy replaces, for instance, the terms *calque* and *loan translation* (Verschik, 2016: 193–194, 2008: 54; Johanson, 2002a: 291–292). An example of a selective copy (from Pyöli, 1994: 250) in (Livvi) Karelian is the word order in numerals, as shown in Example 2a, where the numeral is followed by the header. The expression of time corresponds to the Russian expression in Example 2b.

(2) a. *vuot-tu* *kakskymmen*
year-sg.part twenty
'(approximately) twenty years'
b. *let* *dvadcat'*
years twenty
'(approximately) twenty years'

Correspondingly, placing the numeral first in Karelian (Example 3a) corresponds to the Russian meaning of exact time (Example 3b).

(3) a. *kakskymmen* *vuot-tu*
twenty year-sg.part
'(exactly) twenty years'
b. *dvadcat'* *let*
twenty years
'(exactly) twenty years'

Moreover, in some cases both global and selective copying are employed at the same time, constituting a *mixed copy* (Verschik, 2016). For instance, the conventionalized Karelian word *esiniekka* (North Karelian form, see KKS) 'apron' consists of two morphemes *esi* 'front' + *niekka* (person-related suffix).

The model word is the Russian for 'apron', *perednik*, which also has two parts: *pered* 'front' + *nik* (person-related suffix). In Karelian, the suffix is a conventionalized global copy and the word for front is original, which makes the Karelian word *esiniekka* a mixed copy. Within endangered Finno-Ugric minority languages, selective and mixed copying of compounds are particularly common strategies for forming neologisms (Olthuis, 2003: 531–536).

The code-copying framework is often used to describe spoken language (Verschik, 2008), but it is also applicable to written languages (Verschik, 2016). Anna Verschik (2016) has studied Estonian-Russian language contacts in literary contexts, namely blogs. She argues that,

> […] written texts are in their nature more deliberate than oral texts and the appearance of changes there hints at a greater degree of conventionalization (that is, systematic use of other language content words and patterns is more visible and cannot be dismissed just as a slip of the tongue or anecdotal occurrence […]
> (2016: 187)

Thus, the code-copying framework is applicable to newspaper data where copied neologisms may not at first be conventionalized in a language, although their appearance often indicates a lexical gap that requires a new word. The examples presented here included old, conventionalized copies, but also neologisms. Hence, the framework is explanatory of both old and new word formation in Karelian. In this study, the term neologism refers to any new word formed by either language-internal means or code-copying, and all analysed code-copies are considered neologisms.

4 Data and Methods

This section introduces the data and methods of the study. Sections 4.1 and 4.2 present the Karelian newspapers as data and the aggregation of the newspapers as a corpus. Section 4.3 presents the generation of naïve Bayes classifier and Section 4.4 the analysis methods of neologisms.

4.1 *The Karelian Newspapers*

The newspaper *Oma Mua* 'Own Land' was established in Russia in 1990 and was published in the Livvi Karelian and North Karelian written standard varieties until 1999. At the beginning of 2000, the newspaper was divided into two separate papers according to the standard varieties. Consequently, *Oma*

Mua continued publishing articles in Livvi Karelian, and a new newspaper, *Vienan Karjala,* 'North Karelia' published articles in North Karelian (Viinikka-Kallinen, 2010: 194). Both newspapers were published once a week; however, *Vienan Karjala* was published only twice a month from the second half of 2012 until the end of 2013. The printing company Periodika published both papers, and the newspapers had a joint editorial office in Petrozavodsk, the capital of the Republic of Karelia, Russia. The period of two separate newspapers continued until the end of 2013. From the beginning of 2014, *Oma Mua* and *Vienan Karjala* merged, and, once again, *Oma Mua* began to publish in North Karelian in addition to Livvi Karelian (Viinikka-Kallinen, 2010: 194; Tánczos, 2015: 96).

The first issues of both newspapers had only four pages. During the initial years the papers grew to eight-pages and, after merging in 2014, the issues have since consisted of twelve pages. Both newspapers and the merged volumes focus on Karelians and the Karelian language. They contain different genres, including news reports, prose, poetry, and opinion. In some of the issues children have a dedicated page with facts, stories, and assignments in Karelian. All of these genres include translations from Russian and Finnish. The newspapers have received material from readers and freelance writers, especially during their early years of publication (Tánczos, 2015: 96). Occasionally, especially in the paper's current merged form, South Karelian and Karelian´s close cognate languages, Ludic and Veps, have been used in the texts. The use of these varieties is sometimes mentioned in the texts, unlike the use of North and Livvi Karelian varieties. The two main varieties are used concurrently without naming them on each occasion. Different articles are written in different varieties with no mixing of the two varieties. Parallel translations of the articles are rare, and are provided only, for example, for official bulletins. In addition, advertisements, for instance of elections, are occasionally published in Russian.

The distribution of both newspapers prior to merging reached 700 copies in total (Tánczos, 2015: 96) in addition to orders for PDF copies. The actual number of readers of the PDF versions is not known.[1] The PDF copies are free online. Backdated issues of *Oma Mua* are available from 2010 to 2013[2] and of *Vienan Karjala* from 2011 to 2013.[3] The merged *Oma Mua* has also been available online since the beginning of 2014, although the latest two years' editions have been available only to paying subscribers. Availability online was one of the criteria in designing the corpus for the present study.

1 http://omamua.ru/ob_izdanii/.
2 http://omamua.ru/issues/.
3 http://omamua.ru/liv/issues/.

4.2 *Corpus Aggregation and Metrics*

The authors collected several Karelian newspaper volumes in order to compile the Karelian newspaper text corpus (KNP). The KNP consists of several subcorpora, which were used in different stages of the study. One of the main ones is OMVK, which encompasses the volumes of *Oma Mua* (OM) and *Vienan Karjala* (VK) from 2011 to 2013. However, since VK contained fewer words and was released less frequently than OM, every other issue from each volume of OM was excluded in order to maintain a rough balance between the varieties in the OMVK subcorpus. In addition to OMVK, the KNP includes one volume of *Oma Mua* from 2017 (OM17), which is a merged version of the two former newspapers. Finally, based on the automatic classification (see Introduction), OM17 was divided into two subcorpora, $OM17_{LIVVI}$ and $OM17_{NORTH}$. The KNP and its subcorpora are summarized in Table 1.

All of the newspapers were converted from PDF to plain text format. Russian-language TV programme names and numerical weather report information were excluded from the conversion. The total number of tokens of the OM and VK subcorpora were similar: 414,815 for OM, and 412,398 for VK. The OM17 subcorpus contained slightly fewer tokens, i.e., 372,693. The size, word types (i.e., types of word forms that occur in the corpus), and the lexical diversity index in the subcorpora are presented in Table 2.

TABLE 1 The Karelian newspaper corpus and its subcorpora. The first column indicates the volumes of the newspapers, the second which variety was used in the issues, and the third the name of the subcorpora. The name is formed with capital letters on the basis of the newspapers and the variety in the subtext.

Volumes of the newspapers	**Standard varieties used in the volumes**	**Name of the (sub) corpora**
Oma Mua 2011–2013, *Vienan Karjala* 2011–2013, and *Oma Mua* 2017	Livvi Karelian and North Karelian	KNP
Oma Mua 2011–2013	Livvi Karelian	OM
Vienan Karjala 2011–2013	North Karelian	VK
Oma Mua 2011–2013 and *Vienan Karjala* 2011–2013	Livvi Karelian and North Karelian	OMVK
Oma Mua 2017	Livvi Karelian and North Karelian	OM17
part of *Oma Mua* 2017	Classified as Livvi	$OM17_{LIVVI}$
part of *Oma Mua* 2017	Classified as North	$OM17_{NORTH}$

TABLE 2 Corpus metrics. The first column indicates the subcorpus, the second the number of all words, the third the number of different word forms and the fourth the lexical diversity index value in the corpus.

Subcorpus	Tokens	Types	Lexical diversity index
OM	414,815	66,916	0.161
VK	412,398	72,342	0.175
OM17	372,693	71,565	0.192

The lexical diversity index is one of the various metrics for comparing different kinds of text types. It is used in text corpus analysis to define the richness of a lexicon, and it is calculated as a type-token relationship of the words. The index value for OM is 0.161, VK 0.175, and OM17 0.192, showing the text in OM17 to be the richest. It should be noted that these indexes were calculated from unlemmatized word types; in the present study, unlemmatized word types are used because the KNP contains only raw text, and, currently, no complete automatic lemmatization tools exist for the Karelian varieties. Nevertheless, some tools for analysing Livvi Karelian are in progress.[4]

As described above, OMVK was divided into $\text{OM17}_{\text{LIVVI}}$ and $\text{OM17}_{\text{NORTH}}$ automatically using the Naïve Bayesian classifier.

4.3 *Naïve Bayes Classifier*

The naïve Bayes classifier is a conventional probabilistic model, which is commonly utilized in text classification studies, such as sentiment analysis of written texts. Even though the naïve Bayes classifier is considered a simple, or 'naïve', and non-state-of-the-art machine learning algorithm, previous studies have shown that it achieves good text classification results with low computational costs (e.g., Pang et al., 2002). The term 'naïve' indicates that the classifier assumes all features (e.g., words in a text) to be independent of each other. Although this assumption is linguistically incorrect, text classification based on independent word probabilities has been sufficiently powerful with many real data applications (Feng et al., 2015). However, the naïve Bayes classifier has been rarely used for the variety classification of an endangered and underresourced language.

As the name indicates, the naïve Bayes classifier is based on Bayes' rule, or Bayes' theorem, which in this context is defined as in Equation 1.

4 See for example: http://giellatekno.uit.no/cgi/index.olo.eng.html.

$$P(\text{variety}|\text{features}) = \frac{P(\text{variety}) \times P(\text{features}|\text{variety})}{P(\text{features})}$$

Equation 1. Bayes' theorem utilised in language variety classification

Equation 1 shows the probability of language variety (i.e., North Karelian or Livvi Karelian) given features (i.e., word occurrences). In other words, based on the given text, the classifier shows the maximum probability between varieties, which is, in turn, based on the vocabulary of OMVK.

The naïve Bayes model for the automatic classification of North and Livvi Karelian was implemented in Python 3 using the nltk.classify.naivebayes module in Natural Language Toolkit's (Bird, Loper and Klein, 2009) Classify package. The classifier was generated as follows. Firstly, the 2,000 most frequent words in OMVK were collected for feature extraction. Then, the OMVK subcorpus (189 issues) was divided into training (149 issues) and test (40 issues) data. Instead of comparing the whole OMVK subcorpus word-for-word, the feature extractor was used to indicate whether the most common 2,000 unlemmatized word forms occur in the given newspaper and to train the naïve Bayes classifier. Since the extractor searched only for the presence of a word form, the frequency of the word form was excluded from the classification. After training the classifier and calculating the classification accuracy for test data, a list of the most informative words based on the classifier was generated using the *most_informative_features* function. Finally, the classifier was applied to classify OM17 into two subcorpora: $\text{OM17}_{\text{LIVVI}}$ and $\text{OM17}_{\text{VIENA}}$. The most informative words in the OMVK classification were used to compare the lexical features of the two varieties. The classification results of OM17 were used to study new word formation strategies of neologisms (see Section 4.4). The classification results are presented in Sections 5.1 and 5.2.

4.4 *Corpus-based Selection and Categorization of Neologisms*

Neologisms in OM17 were investigated as follows. Based on the division of OM17 into $\text{OM17}_{\text{LIVVI}}$ and $\text{OM17}_{\text{VIENA}}$, two word lists were generated accordingly. Then, the words in the lists were compared to the following dictionaries and glossaries:

1) Pyöli (2016) for Livvi Karelian in Finland
2) Fedotova and Bojko (2009) for North Karelian in Russia
3) Õispuu (2003a, 2003b) for the neologisms of *Oma Mua* in the 1990s
4) Makarov (1990) for Livvi Karelian in Russia
5) KKS for Karelian dialects at the turn of the 20th century.

The first and third dictionaries utilize *Oma Mua* as a source, which facilitated the identification of neologisms. In addition, the lists were compared to the word lists of OMVK. Similar words with the dictionaries, glossaries, the word lists of OMVK and words with five or more occurrences were then deleted from the two lists, resulting in the final word lists of Karelian neologisms. Finally, the origin of the neologisms was categorized using the following four categories:

1) Language-internal means of word formation via derivation and composition
2) Words copied from or via Finnish
3) Words copied from or via Russian
4) Words copied from English.

Each category was analysed from the perspective of the code-copying framework. Section 6 presents the neologisms in OM17 and the results regarding their origin.

5 Newspaper Classification

This section presents the results of the automatic variety classification of the newspaper issues in the KNP. Section 5.1 presents the classification accuracy of the KNP. The classification of OM17 is presented in Section 5.2

5.1 *Classification Accuracy and the Most Informative Words*

The naïve Bayes classifier obtained an overall classification accuracy of 100% on the test set of 40 newspapers from OMVK. Rather than demonstrating the high performance of a simple naïve Bayes algorithm, the perfect classification accuracy in this case reveals major differences between the vocabularies of the North Karelian and the Livvi Karelian newspapers. Table 3 shows 10 words that were the most informative for the naïve Bayes classifier. For instance, the word *päivy* 'day' is approximately 60 times more likely to occur in the Livvi Karelian newspaper. The counterpart in North Karelian would be *päivä*, which is an example of the many lexical differences that are actually phonological in nature (see Section 2.1 for a description). Interestingly, the most informative features in automatic classification were defined as the presence (i.e., the word form received a true value) of Livvi Karelian instead of the presence of North Karelian words. Table 3 also shows 'false' values, which indicate that a word is more likely to be absent in comparison to another variety, for instance *šitä* is approximately 60 times more likely to not occur in Livvi than in North

TABLE 3 The 10 most informative word forms in naïve Bayes classification. The word form is presented in the left column with a true or false value. The centre column indicates how to read the rightmost column, which shows the likelihood of a wordform to appear in one variety in contrast to the other variety.

päivy = True	Livvi: North	60.2: 1.0
kumpani = False	Livvi: North	60.2: 1.0
šitä = False	Livvi: North	60.2: 1.0
mukah = False	Livvi: North	60.2: 1.0
sendäh = True	Livvi: North	59.1: 1.0
rubei = True	Livvi: North	59.1: 1.0
libo = True	Livvi: North	59.1: 1.0
niškoi = True	Livvi: North	59.1: 1.0
ruadoh = True	Livvi: North	59.1: 1.0
oltih = False	Livvi: North	59.1: 1.0

Karelian, i.e., the word is North Karelian. However, the majority of the informative features have 'true' values.

The 2,000 most informative words contain several Russian global copies. The Russian global copies in Table 3 are all conventionalized words, i.e., they were copied at least some decades ago and have become established (e.g., Backus, 2010) in spoken Karelian (e.g., Tavi, 2018) and in several dictionaries (e.g., KKS; Pyöli, 2016). Most of them are more likely to appear in Livvi Karelian (see word forms 7 and 9 in Table 3: *libo* 'or', *ruado-h* 'work-sg.ill.'), and the absence of some words is characterized as a North Karelian feature in the classifier (see word forms 10–13 in Table 4: *ruod-uo* 'work-sg.part.', *libo* 'or', *ruado-h* 'work-sg.ill.', *hos* 'although'). Livvi Karelian is often considered as having more Russian influence than other dialects of Karelian largely for geographical reasons (see Section 2.1) and Table 3 and 4 demonstrate that old, conventionalized Russian global copies are more likely to appear in Livvi Karelian texts than North Karelian texts, which along with phonology create the lexical differences between the varieties (for an example of same text written in North and Livvi Karelian, see Section 2.1). Russian global copies of Livvi Karelian constitute about 4% of the most informative words. Fifteen of the most frequent Russian global copies of the 2,000 most informative words are listed in Table 4.

Next, the naïve Bayes classifier was used to test whether the vocabulary of the issues of the new merged Karelian newspaper is closer to the North or Livvi Karelian variety. Since the classification accuracy achieved 100% on the OMVK test set, the classification of OM17 into $\text{OM17}_{\text{LIVVI}}$ and $\text{OM17}_{\text{NORTH}}$ was expected to be reliable.

TABLE 4 15 of the most frequent word forms copied from Russian of the 2,000 most informative words. The word form is presented in the left column with a true or false value. The centre column indicates how to read the rightmost column, which shows the likelihood of a wordform to appear in one variety in contrast to the other variety.

libo = True	Livvi: North	59.1: 1.0
ruadoh = True	Livvi: North	59.1: 1.0
rodieu = True	Livvi: North	52.1: 1.0
programmu = True	Livvi: North	51.1: 1.0
rubl'ua = True	Livvi: North	51.1: 1.0
školan = True	Livvi: North	50.0: 1.0
ruadajat = True	Livvi: North	49.0:1.0
raudau = True	Livvi: North	48.0: 1.0
poliitiekan = True	Livvi: North	43.0 = 1.0
ruaduo = False	North: Livvi	39.8: 1.0
libo = False	North: Livvi	39.3: 1.0
ruadoh = False	North: Livvi	39.3: 1.0
hos = False	North: Livvi	38.9: 1.0
rodih = False	North: Livvi	38.9: 1.0
onnuako = True	Livvi:North	37.9: 1.0

5.2 *Classification of* OM17

The automatic classification results for OM17 show that the majority of the issues contained more linguistic content written in Livvi Karelian than in North Karelian. Of the total number of issues, 80% were classified as Livvi Karelian, whereas only 20% were classified as North Karelian. It should be noted that even though all issues of volume 2017 include linguistic content from both varieties, the classification result indicates the dominance of the North or Livvi variety in a classified issue, i.e., in 80% of the issues, most articles were written in Livvi variety and in 20% of the issues, most articles were written in North Karelian variety. This was confirmed by checking several classification results manually.

Based on the automatic classification, $\text{OM17}_{\text{LIVVI}}$ and $\text{OM17}_{\text{NORTH}}$ were formed. To reveal whether the new word formation strategies follow the older, conventionalized vocabulary or whether new strategies have emerged, neologisms were collected manually from the $\text{OM17}_{\text{LIVVI}}$ and the $\text{OM17}_{\text{NORTH}}$ corpora.

6 Neologisms of OM17

Two word lists were compiled from the $\text{OM17}_{\text{LIVVI}}$ and $\text{OM17}_{\text{NORTH}}$ corpora. The word lists were reviewed for neologisms by comparing them to the dictionaries

TABLE 5 The quantitative distribution of the types of neologisms. The number indicates all forms of neologisms occurring in the corpora.

The types of neologisms	OM17$_{NORTH}$	%	OM17$_{LIVVI}$	%
Language-internal means and selective copying	1,155	61	4,890	65
Finnish global copies	543	29	1,777	24
Russian global copies	170	9	757	10
English global copies	24	1	57	1
Total	1,892	100	7,481	100

and glossaries mentioned in Section 4.4. The categories of the detected neologisms were based on the originality of the neologism, i.e., the language that it was copied from. The focus was on form and, hence, on global copying. Secondarily, selective copying was included in the language-internal means category as it is difficult to differentiate Finnish model and Karelian language-internal means of word formation because of the similarities of the two languages. Table 5 illustrates the categories of neologisms in both OM17$_{LIVVI}$ and OM17$_{NORTH}$. Each neologism can occur in various word forms, yet the maximum occurrence in one corpus was set to only four occurrences as the focus of the study is on new strategies of neologism creation. The rareness of the neologism may refer to its non-conventionalized status. Practically all the neologisms were nouns, which may be due to the nature of neologisms. Nouns often describe new innovations and concepts.

Table 5 shows that although OM17$_{LIVVI}$ and OM17$_{NORTH}$ differ in their vocabularies due to the dominance of a particular variety, the types of neologisms are similar in both corpora. Language-internal means of word formation is the most frequent strategy for forming new words, and copying from Finnish is the second most frequent strategy. Copying from Russian is clearly a less common strategy for the formation of neologisms, even though the newspaper was published in Russia by Karelian speakers who are bilingual with Russian. The following Sections 6.1–6.4 introduce each category of neologisms.

6.1 *Language-Internal Means or Selective Copying?*

Language-internal means in word formation encompasses derivation and composition (see Section 3.2). Therefore, derivation and composition are summarized as language-internal means in Table 5. Karjalainen et al. (2013: 51) give derivation and lexical calques as examples of original word formation. Lexical calques that are considered selective copies are under the category of original

word formation strategies. Many of the neologisms in this category have a model from another language; about 6% of the compounds and derivatives in OM17$_{\text{NORTH}}$ and about 5% in OM17$_{\text{LIVVI}}$ are selective copies formed on the basis of the Finnish model. These neologisms were possible to define as selective copies despite the similarities in the word formation of Finnish and Karelian; however, the real percentage of the selective copies may be higher. In some of the cases, the line between language-internal means of word formation and the Finnish model is ambiguous as the two languages are very closely related and share a great deal of lexicon, derivative items, and word formation strategies.

Example 4a shows an instance of a selective copy that has a Finnish derivative as the model. The Finnish model word is the derivative *samanlaisuus*. Examples of compound words with the meanings of their components selectively copied from Finnish are the North Karelian determinative compound in Example 4b and the Livvi Karelian determinative compound in 4c. Example 4b is formed with Karelian lexical items *noja* 'a rest' and *stuula* 'a chair'. *Stuula* is an old Russian global copy whereas in Finnish the word is *tuoli.* Similarly, in 4c, *laukku* 'a store' is an old Russian global copy whereas in Finnish the word would be *kauppa* 'a store'. *Keskus* 'a centre' is a Finnish global copy in Karelian. The examples of the OM17$_{\text{LIVVI}}$ and OM17$_{\text{NORTH}}$ corpora indicate that the compounds have a Finnish model in meaning, which indicates that they are selective copies.

(4) a. *samanlažuos* [OM17$_{\text{LIVVI}}$]
'uniformity'
< Finnish: samanlaisuus 'uniformity'

b. *noja-stuula-šša* [OM17$_{\text{NORTH}}$]
rest-chair-sg.iness
'armchair'
< Finnish: nojatuoli 'armchair'

c. *laukku-keskus* [OM17$_{\text{LIVVI}}$]
store-centre
'shopping centre'
< Finnish: kauppakeskus 'shopping centre'

Selective copying from Russian is rarer, which may be due to slightly different word formation strategies. Russian compounds are open, unlike in Finnish and Karelian. For instance, 'Minister for Foreign Affairs' is in Russian *ministr inostránnyh del,* whereas both Finnish and Karelian use the determinative compound *ulkoministeri* ('outer+minister'). In addition, Russian compounds may be appositional with two parts; the second head is semantically dominant and the first is a specifier, or the first head defines the type, and the second

noun defines the specifier. The parts are joined with a hyphen, for example, *shkóla-internát* 'boarding school', where the first head is semantically dominant and the second is the specifier (an example of this kind of selective copying is presented in 5a). This type of use of compounds is fixed in the Russian language (Timberlake, 2004: 151).

Russian derivational suffixes have been copied to Karelian via globally copied lexemes without gender marking in the masculine form. For instance, the Russian *nik* occurs in many conventionalized global copies of Karelian, such as *pruasniekka* (< *prázdnik* 'feast'), and is used to derivate new words, such as *pruasniekkaniekka* 'a guest at a feast'. However, neologisms derived using old suffixes copied from Russian are rare in the data as only one example occurs in $\text{OM17}_{\text{LIVVI}}$ and $\text{OM17}_{\text{NORTH}}$ (see Example 5b).

(5) a. *viruo-huogavuo* [$\text{OM17}_{\text{LIVVI}}$]
lay.inf-rest.inf
'lay-rest'

b. *račču**nieka-t* [$\text{OM17}_{\text{LIVVI}}$]
hackney-pl
'horsemen'

The compounds can be formed by global copying, or one part of the compound can be an original neologism or a global copy. Thus, the resulting neologism can also be a mixed copy. In OM17, one component of the compound and the entire meaning of the compound can be copied in a neologism. The Livvi Karelian neologism *sotsialurahasto* in Example 6 is a mixed copy (see also Example 4c for a mixed copy). The first semantic component, the modifier *sotsiualu,* is a copy from the Russian *sotsialnyj* and has already become a conventionalized word in Karelian. The word has an ending *-alu* formed based on the Finnish model *-aali* used in compounds. The second semantic component, the head *rahasto,* is a global copy from Finnish. The meaning of the endocentric compound has been copied from Finnish.

(6) *sotsiualu-rahasto* [$\text{OM17}_{\text{LIVVI}}$]
'social-fund'

No model can be traced in most of the instances of the neologisms created by derivation and composition. Of these two strategies, composition is more common than derivation. Instances of neologisms created by derivation are given in Examples 7a and 7b and by composition in 7c. Examples 7a and 7b are not new concepts, but they are not included in dictionaries and may not be conventionalized. Thus, they can be interpreted as neologisms.

(7) a. *kyykkäri-t* [OM17$_{\text{NORTH}}$]
'player of Karelian traditional game Kyykkä'-pl
b. *halviste-ttih* [OM17$_{\text{LIVVI}}$]
cheapen-pass.pst
c. *arvostus-joukko-h* [OM17$_{\text{NORTH}}$]
evaluative-team-sg.ill

Of the neologisms formed on the original basis, 3% are derivations and 97% are compounds in OM17$_{\text{NORTH}}$, and 2% are derivations and 98% are compounds in OM17$_{\text{LIVVI}}$. The strategies for word formation are similar to Finnish selective copies, which may be due to a selective copying of the Finnish word formation model and to similarities with word formation typology. Because of the similarities and close relation between Karelian and Finnish, interpreting whether the neologism is formed by language-internal means or selective copying with Finnish model is elusive.

6.2 *Finnish Global Copies*

Finnish has served as an example and a source of new words for both North and Livvi Karelian during the revitalization process. In OM17$_{\text{NORTH}}$, Finnish global copies are the second largest group representing 29% of neologism occurrences. In OM17$_{\text{LIVVI}}$, Finnish global copies also form the second largest group of neologisms with 24% of the total occurrences of neologisms. This result is rather surprising as Karelian neologisms are often reported to be Russian global or selective copies; however, the assumption that Karelian extensively copies from Russian may have its roots in spoken Karelian and its code alternation, as mentioned in Section 3.2. Using the Finnish model is, however, supported by the ideologies of the conscious development of Karelian written standard varieties.

The Finnish global copies resemble words formed by language-internal means; Finnish global copies are Finnish derivatives and compounds. Example 8 gives instances of Finnish derivations that are global copies. The word in Example 8a is derived in Finnish with the suffix *-ja*, which refers to an actor. In Example 8b, the suffix describing collectiveness is *-isO*. This word has already been lexicalized in the Finnish language.

(8) a. *vuatija-t* [OM17$_{LIVVI}$]
demander-pl
< Finnish: vaatijat 'demanders'
b. *yhteisö* [OM17$_{LIVVI}$]
'community'
< Finnish: yhteisö 'community'

Example 9 contains four instances (9a-9d) of globally copied Finnish compounds. All of these are determinative compounds; the first component is the modifier, and the second is the semantic head.

(9) a. *huvi-puisto-n* [OM17$_{NORTH}$]
amusement-park-sg.gen
< Finnish: huvi-puisto 'amusement park'
b. *ilmasto-olo-t* [OM17$_{NORTH}$]
climatic-condition-pl
< Finnish: ilmasto-olot 'climatic conditions'
c. šinivalaš [OM17$_{NORTH}$]
'blue whale'
< Finnish: sinivalas 'blue whale'
d. *luonno-n-olo-i-ssa* [OM17$_{LIVVI}$]
natural-sg.gen-condition-pl-iness
< Finnish: luonnonolot 'natural conditions'

Copying international words from Finnish is also a strategy for creating neologisms in OM17$_{NORTH}$ and OM17$_{LIVVI}$. According to Section 3.1, the source for international words is expected to be Russian, although favouring Finnish over Russian has been a strategy since the 1990s. Examples 10a-c demonstrate international words copied via Finnish. For instance, in Example 10c, the word in Russian is *bum*, which would most likely receive the Karelian nominative ending *-a,* and the resulting word would be *bu*(*u*)*ma*. The nominative ending *-i* in the word in Example 10c is Finnish and is typical for adapting global copies. The ending reflects that the model language is Finnish. The model code is more ambiguous in Example 10a, where only the case-ending similar to Finnish is a clue to the Finnish model.

(10) a. *esse-i-ta* [OM17$_{\text{NORTH}}$]
essay-pl-part
< Finnish: essee 'essay'

b. *revitalisointi-h* [OM17$_{\text{NORTH}}$]
revitalization-sg.ill
< Finnish: revitalisointi 'revitalization'

c. *buumi* [OM17$_{\text{NORTH}}$]
'boom'
< Finnish: buumi 'boom'

Finnish global copies have preserved their phonological structure in Karelian, which can be seen, for instance, in the nominative case-ending *-i* instead of the Karelian ending *-a*. Thus, the type of copying can be defined as global. Contact-induced change is difficult to define in the contacts between close cognate languages (Riionheimo, 2007: 33). The novelty of an item may be the only clue to the change being contact-induced.

6.3 *Russian Global Copies*

Due to the bilingualism of Karelians in Russia, Russian is a natural choice as a source of new words that are difficult to form by language-internal means. However, Russian is not the first choice in copying, as stated in Section 6.2. Russian global copies form only 9% of the neologisms in OM17$_{\text{NORTH}}$ and 10% in OM17$_{\text{LIVVI}}$. Since Russian has influenced Livvi Karelian more than North Karelian (Section 2.1), the small difference in the percentages of the Russian copies is unexpected. As shown in Tables 3 and 4, Livvi Karelian has more Russian global copies than North Karelian.

Several of the Russian global copies are international, as in Example 11. Consequently, Russian is a source of international words along with Finnish.

(11) a. *sessii* [OM17$_{\text{NORTH}}$]
'session'
< Russian: *sessija* 'session'

b. *saiti-lla* [OM17$_{\text{NORTH}}$]
site-sg.ill
< Russian: *sajt* 'site'

c. *gormon-i-en* [OM17$_{\text{NORTH}}$]
hormone-pl-gen
< Russian: *gormon* 'hormone'

d. *treenera-t* [OM17$_{\text{LIVVI}}$]
trainer-pl.nom
< Russian: *trener* 'trainer'

e. *migrant-oin* [OM17$_{LIVVI}$]
migrant-pl.gen
< Russian: *migrant* 'migrant'

f. *revitalizatsie-s* [OM17$_{LIVVI}$]
revitalization-sg.iness
< Russian: *revitalizatsija* 'revitalization'

The phonology of international words reveals the model code of the copy. For instance, in Example 11c, the Finnish word is *hormoni*, and, hence, the /g/ ~ /h/ alternation originates from the Russian model. Additionally, Example 13b has *-i-* in the ending of the word, but in Finnish, the original word *sivusto* is more common than the English model. In Russian, *sajt* is used for site. Examples 11a and 11f come from the Russian nominative inflection in singular and plural forms *sessija*: *sessii*, *revitalizatsija*: *revitalizatsii*. Another possible explanation for Example 11a is frequency-based copying because the form *sessii* here represents the nominative singular form in Karelian. In Russian, also the genitive singular form is *sessii* and this is used frequently also in the accusative case. Similarly, most Russian adjectives globally copied in Karelian have the ending *-oi*, for instance *pervoi* 'the first' (< Russian *pervyj*). The ending *-oj* appears in Russian in the nominative case of only some adjectives of masculine gender but it appears frequently in many other cases in all genders. Thus, frequency of the ending is also a possible explanation for the form of adjectives. The voiced sibilant /z/ in Example 11f originates from the Russian model since the Finnish language has only one sibilant /s/.

In addition to international words, some administrative words have been copied from Russian (Examples 12a and 12c), which is typical considering that these types of words are culture-specific. Russian global copies can also be words that are first copied to the spoken language and then used in the standardization of the language, such as *polučči-u* '(s)he receives' in Example 12b.

(12) a. *mikro-rajona-n* [OM17$_{NORTH}$]
micro-area-gen.sg
< Russian: *mikrorajon* 'micro area'

b. *polučči-u* [OM17$_{LIVVI}$]
receive-pr.3sg
< Russian: *poluchit´* 'receive'

c. *buklettu* [OM17$_{LIVVI}$]
'flyer'
< Russian: *buklet* 'flyer'

Even though most neologisms are nouns, some verbs have been copied, as in Example 12b. The verb in the example may have been copied previously to spoken language but is not listed in dictionaries. This may be because Karelian has language-internal means to convey the same meaning, for example, *ottoa vastah* 'take in' (KKS s.v. *vassassa, vastah*). These types of global copies are, however, rare among neologisms in $\text{OM17}_{\text{NORTH}}$ and $\text{OM17}_{\text{LIVVI}}$.

The ideology of purism in the revitalization process explains the overall results of Russian global copies; they are only the third largest group, and many of the words are of English origin conveyed by way of the Russian model. The words also describe culture-specific referents in Russian. The motivation for using other means than the Russian model to create words may be found in the desire to maintain a language's originality and distinguish it from Russian. This supports the purist ideology of language revitalization.

6.4 *English Global Copies*

Copying international words straight from English under the influence of social media and the internet is possible for all languages of the world. However, this strategy has not been considered in previous literature concerning the Karelian language (e.g., Õispuu, 1997; Markianova, 2005; Karjalainen et al., 2013). As stated in Sections 6.2 and 6.3, English words are often copied using the Finnish or Russian model. Nevertheless, some of the global copies with an English origin do not seem copied from a Finnish or Russian model, which can indicate that they originate from the English model. The newspapers in the KNP indicate that copying from English is possible, yet rare; English global copies constitute 1% of the neologisms in both $\text{OM17}_{\text{NORTH}}$ and $\text{OM17}_{\text{LIVVI}}$.

Four instances (a-d) of English global copies are shown in Example 13, with Russian counterparts. The Finnish counterparts are written according to the English form.

(13)	a.	*flashmobi-ja* flashmob-pl.part comp. Russian *flesh-mob*	[$\text{OM17}_{\text{NORTH}}$]
	b.	*online* comp. Russian *onlajn*	[$\text{OM17}_{\text{NORTH}}$]
	c.	*wifi* comp. Russian *vaj-faj*	[$\text{OM17}_{\text{LIVVI}}$]
	d.	*offline* comp. Russian *oflajn*, *ne v seti* 'not in the web'	[OM17]

Example 13a is morpho-syntactically adapted, whereas the word in Examples 13b-13d is in the basic form. The word in 13c, *wifi*, is common to many languages; for example, the word has exactly the same form in Finnish, indicating that Finnish or English is at least the orthographic model for the words instead of Russian. Example 14 describes the syntactic use of the words *online* and *offline*.

(14)

Virallis-ien	*tieto-jen*	*muka-h*	*2016*	*vuote-na*
official -PL.GEN	information -PL.GEN	according -SG.ILL	2016	year-SG.ESS
oli	*2185*	*kirjuttamis-kentty-ä,*	*145 000*	*ošallistuju-a*
be-PST.3SG	2185	writing-field -SG.PART,	145 000	partici-pant-SG.PART
kirjutt-i	*ši-tä*	*online*	*ta*	*offline.*
write.PST-3SG	it-SG.PART	online	and	offline.

'According to official data, in 2016 there were 2185 writing fields in which 145,000 participants wrote online and offline.' [*Oma Mua* (14) 2017.]

The words *online* and *offline* should have a different case government in the context. The meaning 'to write somewhere' requires the inessive case in Karelian and, therefore, the correct syntax would be *kirjutti šitä onlinessa ta offlinessa*. Thus, the words have not been morpho-syntactically adapted, which indicates that the words are global copies but not yet conventionalized.

Indications that English may be a new model language for Karelian word formation exist in OM17. However, Finnish and Russian remain the models for copying international words. This may be due to the ideologies of revitalization, according to which models should be as close to Karelian as possible. In the future, it is possible that more words will be copied from English as contact between the languages becomes more intense.

7 Conclusion and Discussion

This study investigated two written standard varieties of Karelian that are based on two different dialects used in the Republic of Karelia, Russia. Although attempts have been made to generate a common written standard variety, the variation between the two dialects seems too substantial for speakers of Karelian to find a compromised standard form (Kunnas and Arola, 2010;

Anttikoski, 2003). The study aimed to examine lexical dialectal differences using automatic text classification methods and trends in the development of new words, i.e., neologisms. The neologisms were chosen as the focus of the study in order to evaluate lexicon development in contrast to known strategies of word formation. Consequently, three research questions were raised.

To answer the first question regarding what variety-related lexical characteristics are revealed in the automatic classification of North and Livvi Karelian varieties, a naïve Bayes classification model was generated based on North and Livvi Karelian raw text vocabulary (the OMVK subcorpus). As expected, the classifier achieved high accuracy (100%) in recognizing newspaper texts in North and Livvi Karelian due to major lexical and phonological differences between the varieties. The classifier was then applied to the 2017 volume of *Oma Mua* (OM17) to determine which of the two varieties is dominant in the merged Karelian newspaper. As a result, 20% of OM17 was classified as North Karelian and 80% as Livvi Karelian, indicating the dominance of Livvi Karelian vocabulary in the merged newspaper. In the merged newspaper, North and Livvi Karelian varieties are used in separate articles; the classification thus showed in which variety most of the newspaper articles were written.

Based on the aforementioned classification results, OM17 was divided into two corpora, one dominated by North Karelian elements (i.e., $\text{OM17}_{\text{NORTH}}$) and one by Livvi Karelian elements (i.e., $\text{OM17}_{\text{LIVVI}}$). Two word lists were created from the corpora and Karelian neologisms collected from the word lists to answer the second research question regarding the similarities or differences in neologisms between the automatically classified North and Livvi Karelian varieties and how neologisms are formed. Older dictionaries and glossaries were used as a parallel corpus in the collection of neologisms. The total numbers of collected occurrences of Karelian neologisms were 1,892 in $\text{OM17}_{\text{NORTH}}$ and 7,481 in $\text{OM17}_{\text{LIVVI}}$.

The neologisms were categorized by the origin of the words and analysed within the code-copying framework, as the formation of neologisms in Karelian is a contact-induced process. Also, the effect of purist ideology typical of language standardization processes was considered as a factor behind the conscious development of neologisms. In both corpora, more than half of the neologisms were created by language-internal means, i.e., derivation and composition. Language-internal means as the most common strategy shows that the purist ideology of using languages' own means and thus maintaining the originality of the language is important in the standardization of both North and Livvi Karelian. However, since the Finnish model, in particular, may have influenced word formation, several neologisms in this category may have actually been selective copies. Selective copies are difficult to differentiate from

original word formation due to the similarities of the two closely related languages. Because of the closeness of the languages, the Finnish model instead of Russian or English models supports the purist strategies. Also, selective copying from Russian is rarer, which may be due to slightly different word formation strategies of Karelian and Russian.

The second largest group was Finnish global copies, which was unexpected because Karelian-Russian bilingualism has had a major impact on Karelian and would thus be a natural choice as a source for new words that are difficult to form by language-internal means. Similarly, as in the use of language-internal means and selective copying from Finnish, one prominent explanation for the number of Finnish copies is the close relation of the two languages together with the fact that Finnish has been used as an example in the standardization process from the beginning in the 1990s in keeping with a purist approach. Russian, on the other hand, is often the source of international words in Karelian, and Russian global copies comprise the third largest group of neologisms. In addition to Russian, Finnish and English are sources of international words, although English copies are rare in OM17. The Russian model seems to be avoided in word formation, based on the distribution of Karelian neologisms; for instance, old, conventionalized globally copied derivational suffixes are not used in the neologisms in OM17.

The third research question concerned the tendencies in the development of two different written standard varieties of Karelian from the point of view of neologisms and other lexical characteristics. Investigation of the Karelian neologisms showed that the difference in the number of Russian copies between the varieties is smaller with respect to neologisms than in the majority of the informative words. Using English as a model code was found to be a new and rare strategy. The main strategy in the generation of neologisms was found to be composition. Finnish global copies and using the Finnish model in original word formation were popular strategies in both written standard varieties. These strategies lead to mixed copying in the creation of neologisms; hence, the resulting neologism can also be a mixed copy. In conclusion, the code-copying framework provides ample explanations for creating neologisms in written language standardization as it is often done by copying. According to the results, it seems that Finnish is a typical model code when copying selectively, and Russian copies are more often global, which is probably due to the word formation strategies of the three languages. These results answer the third research question by indicating cooperation in language planning and similar development patterns of neologisms of the two written standard varieties of Karelian.

Similar lexical development patterns may demonstrate common objectives in language planning and standardization. If a common written standard variety were to be created, similar word formation strategies would benefit the process. As Söderholm (2010: 34) considers, a common lexicon is the most important part of mutual intelligibility, and often phonological differences are intelligible if they are not overly distinctive. Thus, also the different standards may be mutually intelligible. Therefore, further research on the development of Karelian lexical (or phonological) features during standardization would benefit the revitalization process of the Karelian language.

References

Anttikoski, Esa. 2003. The Problem of Dialectal Differences in the Creation of a Unified Karelian Literary Language: The Experience of the 1930s. In Esa Anttikoski (ed.), *Developing Written Karelian. Papers from the Karelian session of the 11th International Conference on Methods of Dialectology*, 29–37. Studies in Languages 38. Joensuu: Joensuun yliopisto.

Backus, Ad. 2010. The role of codeswitching, loan translation and interference in the emergence of an immigrant variety of Turkish. *Working papers in corpusbased linguistics and language education* 5: 225–241.

Belentschikow, Renate. 2015. 131. Dictionaries. In P. O. Müller (ed.), *Word-Formation: An International Handbook of the Languages of Europe*, 626–642. Berlin: De Gruyter Mouton.

Bird, Steven, Edward Loper and Ewan Klein. 2009. *Natural Language Processing with Python.* O'Reilly Media Inc.

Fedotova, Vieno Petrovna and Tat´jana Petrovna Bojko. 2009. *Slovar' sobstvenno-karel'skih govorov Karelii* [*'Dictionary of Karelian Proper'*]. Petrozavodsk: Uchrezhdenie RAN Instituta jazyka, literatury I istorii Karel'skogo hauchnogo centra RAN.

Feng, Guozhong, Jianhua Guo, Bing-Yi Jing and Tieli Sun. 2015. Feature subset selection using naive Bayes for text classification. *Pattern Recogn Lett* 65: 109–115.

Huss, Leena. 2001. Kielen elvyttäminen eli revitalisaatio ['Revitalization of a language']. In H. Sulkala and L. Nissilä (eds.), XXVII Finnish Conference of Linguistics in Oulu 19–20 May 2000, 278–284. Acta Universitatis Ouluensis B Humaniora 41. Oulu.

Johanson, Lars. 2002a. Contactinduced change in a codecopying framework. In M. C. Jones and E. Esch (eds.), *Language change. The interplay of internal, external and extralinguistic factors*, 286–313. Berlin: Mouton de Gruyter.

Johanson, Lars. 2002b. Do languages die of 'structuritis'? On the role of codecopying in language endangerment. *Rivista di Linguistica* 14: 249–270.

Johanson, Lars. 1999. Dynamics of code-copying in language encounters. In B. Brendemoen, E. Lanza, and E. Ryen (eds.), *Language encounters across time and* space, 37–62. Oslo: Novus Press.

Kallio, Petri. 2006. On the earliest Slavic loanwords in Finnic. In J. Nuorluoto (ed), *Slavicization of the Russian north. Mechanisms and chronology*, 154–166. Slavica Helsingensia 27. Helsinki: University of Helsinki.

Karjalainen, Heini, Ulriikka Puura, Riho Grünthal, and Svetlana Kovaleva. 2013. *Karelian in Russia. ELDIA Case-Specific Report*. Mainz: ELDIA.

KKS = *Karjalan kielen sanakirja* [*'Dictionary of Karelian'*]. Kotimaisten kielten tutkimuskeskuksen verkkojulkaisuja 18. Helsinki: Kotimaisten kielten tutkimuskeskus 2009. http://kaino.kotus.fi/cgibin/kks/ kks_etusivu.cgi.

Koivisto, Vesa. 1990. Venäjän vaikutuksesta itäisten itämerensuomalaisten kielten refleksiiviverbeihin. ['About Russian affect on reflexive verbs of Eastern Finnic languages']. In S. Vaula (ed.), *Itämerensuomalaiset kielikontaktit. Itämerensuomalainen symposium 7. kansainvälisessä fenno-ugristikongressissa Debrecenissä 27.8.–1.9.1990*, 20–27. Kotimaisten kielten tutkimuskeskuksen julkaisuja 61. Helsinki: Kotimaisten kielten tutkimuskeskus.

Kunnas, Niina. 2007. *Miten muuttuu runokylien kieli. Reaaliaikatutkimus jälkitavujen a-loppuisten vokaalijonojen variaatiosta vienalaismurteissa.* [*'The change of language in the Viena Karelian villages: A real-time study of phonological variation in Viena dialects'*]. Acta Universitatis Ouluensis. B Humaniora 78. Oulu: Univeristy of Oulu.

Kunnas, Niina and Laura Arola. 2010. Perspectives on the attitudes of minority language speakers in the Swedish Torne Valley and Viena Karelia. In H. Sulkala and H. Mantila (eds.), *Planning a new standard language. Finnic minority languages meet the new millennium,* 119–146. Studia Fennica Linguistica 15. Helsinki: Finnish Literature Society.

Laakso, Johanna., Anneli Sarhimaa, Sia Spiliopoulou Åkermark and Reetta Toivanen. 2016. *Towards openly multilingual policies and practices assessing minority language maintenance across Europe*. Bristol: Multilingual Matters.

Makarov, G. 1990. *Slovar' karel'skogo jazyka* (*Livvikovskij dialekt*). *Okolo 20 tys. slov*. [*'Dictionary of Karelian* (*Livvi dialect*). *About 20,000 words.'*]. Petrozavodsk: Karel'skij nauchnyj centr AN SSSR. Institut jazyka, literatury i istorii.

Mantila, Harri. 2010. The relationship between variation and standardisation in the creation of a new standard language. In H. Sulkala and H. Mantila (eds.), *Planning a new standard language. Finnic minority languages meet the new millennium*, 54–73. Studia Fennica Linguistica 15. Helsinki: Finnish Literature Society.

Markianova, Ljudmila. 2005. Karjalan kieli rajantakaisessa Karjalassa ['The Karelian language in Karelia behind border']. In Paula Kokkonen (ed.), *Sukukansaohjelman arki. Suomalais-ugrilainen perintö ja arkipäivä. Studia Fenno-Ugrica 21.9.–16.11.2004*,

59–66. Castrenianumin toimitteita 64. Helsinki: M. A. Castrénin seura – Suomalais-Ugrilainen Seura – Helsingin yliopiston suomalais-ugrilainen laitos.

Õispuu, Jaan. 2003a. *Oma Mua -lehden uudissanastoa vuosilta 1990–1997 I* [*'Neologisms of Oma Mua from 1990–1997 I'*]. Tallinn: Tallinna Pedagoogikaülikool.

Õispuu, Jaan. 2003b. *Oma Mua -lehden uudissanastoa vuosilta 1990–1997 II* [*'Neologisms of Oma Mua from 1990–1997 II'*]. Tallinn: Tallinna Pedagoogikaülikool.

Õispuu, Jaan. 1997. Puristisia ilmiötä elvytetyssä karjalan kirjakielessä ['Puristic phenomena in the revitalized Karelian literary language']. In L. Nissilä (ed.), *Suomen kielen päivä*, 90–97. Tallinn: Tallinna Pedagoogikaülikool.

Olthuis, Marja-Liisa. 2003. Sanaston aktiivisen kartuttamisen metodiikkaa. Mallina inarinsaameen luodut biologian uudissanat ['A methodology for active enlargement of vocabulary in Inari Sami']. *Virittäjä* 107: 529–544.

Palander, Marjatta, Lisa Lena Opas-Hänninen and Fiona Tweedie. 2003. Neighbours or enemies? Competing variants causing differences in transitional dialects. *Computers and the Humanities* 37: 359–372.

Pang, Bo, Lillian Lee and Shivakumar Vaithyanathan. 2002, July. Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 79–86. Association for Computational Linguistics.

Pasanen, Annika. 2006. Karjalan kielen nykytila ja tulevaisuus ['The present and the future of the Karelian language']. *Suomalais-Ugrilaisen Seuran Aikakauskirja* 91: 115–131. Helsinki: Suomalais-Ugrilainen Seura.

Puura, Ulriikka. 2019. *Sinä iče oled vepsläine, voib sanuda, ka? Vepsäläisyyden rakentuminen ja 2000-luvun vepsän kieli* [*'The construction of Vepsianness and the Veps language in 2000'*]. Helsinki: Helsingin yliopisto. http://urn.fi/URN:ISBN:978-951-51-4878-0.

Pyöli, Raija. 2016. *Sanakirja. Karjala-Suomi* [*'Dictionary. Karelian-Finnish'*]. Lahti: N-Paino Oy. https://www.karjal.fi/avoinkirjasto/wp-content/uploads/2018/03/Sanakirja_karjala-suomi_Py%C3%B6li.pdf.

Pyöli, Raija. 2011. *Livvinkarjalan kielioppi* [*'The grammar of Livvi Karelian'*]. Joensuu: Karjalan Kielen Seura.

Pyöli, Raija. 1996. *Venäläistyvä Aunuksenkarjala. Kielenulkoiset ja -sisäiset indikaattorit kielenvaihtotilanteessa* [*'Olonets Karelian under the pressure of Russian. Extralinguistic and linguistic indicators in a state of language sift'*]. University of Joensuu publications in the humanities. Joensuu: University of Joensuu.

Pyöli, Raija. 1994. Nämil bednoil rajažil. Venäläis-aunuksekarjalaisista kontakteista. ['About Russian-Livvi Karelian contacts']. In T. Hämynen (ed.), *Kahden karjalan välillä. Kahden Riikin riitamaalla.* ['Between the two Karelias. Between the two lands of conflict of the two realms'], 245–252. Studia Carelica Humanistica 5. Joensuu: The University of Joensuu.

Riionheimo, Helka. 2007. *Muutoksen monet juuret. Oman ja vieraan risteytyminen Viron inkerinsuomalaisten imperfektinmuodostuksessa* [*'The many roots of change. The crossing of internal and foreign in past tense formation of the Ingrian Finnish in Estonia'*]. Helsinki: Finnish Literature Society.

Sarhimaa, Anneli. 1999. Syntactic transfer, contact-induced change, and the evolution of bilingual mixed codes. Helsinki: Finnish Literature Society.

Sarhimaa, Anneli. 1996. Language planning and sociolinguistic trends in (Soviet) Karelia 1917–1994. In T. Hickey and J. Williams (eds.), *Language, education and society in a changing world*, 73–85. Clevedon: Multilingual Matters.

Sarhimaa, Anneli. 1995. Karjalan kansat ja kielet kontakteissa. Asutushistoriallista taustaa ja lingvistisiä seurauksia ['Karelian peoples and languages in contact: settlement history and linguistic consequences']. *Virittäjä* 99: 191–223.

Spolsky, Bernard. 2004. *Language policy*. Cambridge: Cambridge University Press.

Sulkala, Helena. 2010. Introduction. Revitalization of the Finnic minority languages. In Helena Sulkala and Harri Mantila (eds.), *Planning a new standard language. Finnic minority languages meet the new millennium*, 8–26. Studia Fennica Linguistica 15. Helsinki: Finnish Literature Society.

Söderholm, Eira. 2010. The planning of the new standard languages. In H. Sulkala and H. Mantila (eds.), *Planning a new standard language. Finnic minority languages meet the new millennium*, 27–53. Studia Fennica Linguistica 15. Helsinki: Finnish Literature Society.

Tánczos, Outi. 2015. Representations of Karelians and the Karelian language in Karelian and Russian local newspapers. *Journal of Estonian and Finno-Ugric Linguistics* 6–1: 91–110.

Tavi, Susanna and Lauri Tavi. 2019. 'Ja'-, 'tai'- ja 'vaikka'-merkityksisten lekseemien kontaktilähtöinen variaatio rajakarjalaismurteissa ['Contact-induced lexical variation of 'and', 'or', and 'though' lexemes in Border Karelian dialects']. *Virittäjä* 123: 401–429.

Tavi, Susanna. 2018. Rajakarjalaismurteiden kielikontaktit venäläiskopioiden taajuuden ja fonologian valossa ['Language contacts of Border Karelian dialects in the light of frequency and phonology of Russian copies']. Lähivõrdlusi. Lähivertailuja 28: 316–356. http://dx.doi.org/10.5128/LV28.10.

Timberlake, Alan. 2004. *A Reference Grammar of Russian*. Cambridge: Cambridge University Press.

Verschik, Anna. 2016. Mixed copying in blogs. Evidence from Estonian-Russian language contacts. *Journal of Language Contact* 9: 186–209.

Verschik, Anna. 2008. *Emerging bilingual speech: From monolingualism to code-copying*. London: Continuum.

Viinikka-Kallinen, Anitta. 2010. Substance through your own language – The minority media conect, strengthen and inform. In H. Sulkala and H. Mantila (eds.), *Planning*

a new standard language. Finnic minority languages meet the new millennium, 178–202. Studia Fennica Liguistica 15. Helsinki: Finnish Literature Society.

Zaikov, Pekka. 2013. *Vienankarjalan kielioppi. Lisänä harjotukšie ta lukemisto* [*'The grammar of North Karelian. Exercises and readings attached'*]. Helsinki: Karjalan Sivistysseura.

