Proceedings of 1st International Workshop in Computational Linguistics for Uralic Languages (IWCLUL 2015); ‹http://dx.doi.org/10.7557/scs.2015.2›

Corpus.mari-language.com: A Rudimentary Corpus Searchable by Syntactic and Morphological Patterns

Jeremy Bradley
Ludwig Maximilian University of Munich
Institute for Finno-Ugric and Uralic Studies
&
Koneen Säätiö
J.Bradley@lmu.de

December 16, 2014

Abstract

This paper introduces a rudimentary infrastructure for a searchable corpus of Mari, a highly agglutinative Uralic language spoken in the Volga and Ural regions of the Russian Federation. This infrastructure allows users to search the corpus by syntactic and morphological patterns. It makes use of the University of Vienna's digital Mari-English dictionary, published under a Creative Commons License in 2014, and a morphological analyser following a simple item-and-arrangement approach. Texts fed into the corpus are subjected to a morphological analysis, the results of which are saved into the application's database with the corpus materials and are accessed by the search algorithm. A demonstration of this open-source tool, covering 994,097 tokens taken from works not subject to copyright, can be found at `corpus.mari-language.com`, the source code at `source.mari-language.com`. While a non-representative text collection of this scope can only serve demonstrative purposes, the infrastructure could enable quantitative diachronic or sociolinguistic comparisons, if fed with a sufficiently wide text collection annotated with adequate metadata.

1 Introduction: Structure of the Paper

Section 2 introduces the language (or languages) dealt with in the paper. Section 3 gives a brief overview of the language data available and obtainable to those interested in a corpus-linguistic approach towards Mari. Section 4 introduces the manner in which data is stored and manipulated in the corpus (a demonstration of which can be found at `corpus.mari-language.com`), Section 5 briefly explains how data of this kind can be searched in non-trivial manners. Finally, Section 6 outlines the technical framework upon which this tool is based. Given practical constraints, only a rough overview of these tools can be given. However, the source code can be accessed in its entirety by anyone who is interested (`source.mari-language.com`) and extensive documentation is in preparation.

2 What is Mari, who are the Maris?

The Mari language, referred to by the exonym Cheremis in older sources, is a highly agglutinative () Uralic language native to the Volga and Ural Regions of the Russian Federation. It shares official status with Russian in the Mari El Republic, a subject of the Russian Federation slightly smaller in area than Macedonia that is located some 500–800 kilometres east of Moscow, near the confluence of the Volga and Kama rivers. It is a pluricentric language with two distinct literary norms, the dominant Meadow Mari and the critically endangered Hill Mari. In the 2010 All-Russia population census [1], 365,127 people claimed to be Mari speakers, and 23,062 of them identified themselves as speakers of Hill Mari. Linguists generally divide Mari into four dialect groups: Meadow Mari, Hill Mari, Northwestern Mari, Eastern Mari [2, p. 15]. The two aforementioned literary norms are based on the dialect groups of the same name. Speakers of dialects belonging to the other two groups use the Hill Mari and Meadow Mari literary norms in writing. The UNESCO Atlas of the World's Languages in Danger [3] classifies Meadow Mari as "definitely endangered" and Hill Mari as "severely endangered".

Mari uses a variant of the Cyrillic alphabet slightly different from the Russian alphabet, and Mari data is stored using the Cyrillic alphabet in the corpus infrastructure at hand. Using the Vienna project's transcription and transliteration toolkit (found at `transcribe.mari-language.com` [4]), however, on-the-fly transcriptions into UPA and IPA are possible. All examples used in this paper are given in IPA transcription.

3 How much data do/can we have?

In spite of the socio-political hardships facing the Mari language and indigenous languages in Russia in general, sufficient amounts of language data are obtainable to make Mari attractive for serious corpus-linguistic research. Both literary norms are still comparatively widely used: Novels, daily newspapers, magazines, textbooks and scientific theses are still published today in both language norms, in Russia and abroad (thanks in large part to funding from the Finnish M.A. Castrén Society ["M. A. Castrénin seura"] and the Estonian Kindred Peoples' Programme ["Hõimurahvaste Programm"]). A corpus containing millions of tokens of modern-language texts written by native speakers of different speech variants would be viable in principle, were it not for practical and legal constraints. It would not be possible to adhere to guidelines followed when creating corpora on large European languages, because, for example, texts on medicine (as the Oxford Guide to Practical Lexicography [5, p. 222] suggests as a building block of a corpus) simply do not exist in Mari. But a representative corpus covering actual usage domains of literary Mari appears to be a valid goal.

Historical texts are available as well. Mari literacy traces its roots back to the first grammar of Mari, published in Saint Petersburg in 1775 and widely accessible today thanks to the publication of an extensively commented facsimile edition in 1956 [6]. It did not, however, take off in a serious fashion until the 20th century. The Mari elementary school teacher Timofey Yevseyev [Тимофей Евсеев] (1887-1937) provided the Helsinki-based Finno-Ugrian Society ["Suomalais-Ugrilainen Seura"] with a wide range of Mari-language texts between 1908 and 1929; these have since been published with German translations [7]. The Hungarian linguist Ödön Beke was able to collect a large body of texts working with Mari-speaking prisoners of war during the First World War [8]. More recently, a substantial body of Mari-language newspapers and textbooks from the early twentieth century, covering a wide geographic range, has been digitized, and made available on the National Library of Finland's website [9]. If these historical materials were integrated into a joint infrastructure with texts in modern Mari, a wide range of analyses would become possible: diachronic (Hill Mari today vs. Hill Mari around 1920), dialectological (Mari in Mari El around 1920 vs. Mari in Bashkortostan around 1920), genre-based (newspapers vs. schoolbooks), sociolinguistic (articles written by men vs. articles written by women), etc. As this is currently beyond the scope of my capacities, I will restrict myself here to presenting the infrastructure that would make such an analysis possible, if it was fed with the correct texts. For the time being, a non-representative body of texts was fed into the infrastructure. This is discussed in Section 6.3.

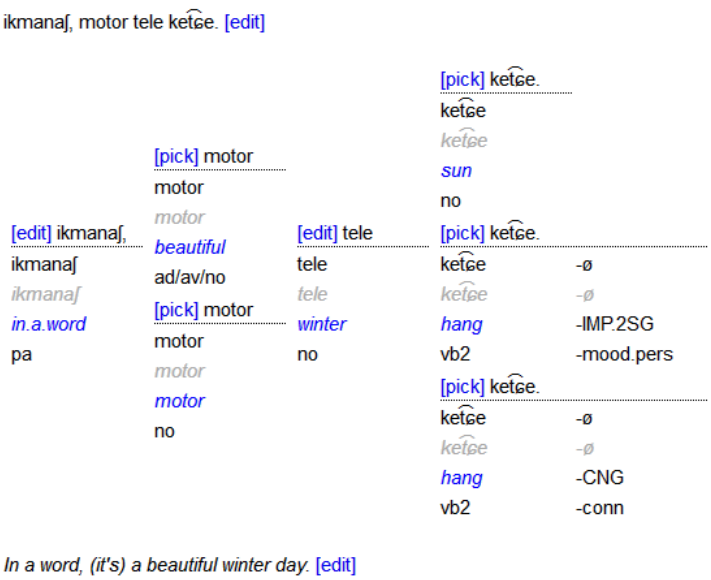

Figure 1: A glossed sentence, not disambiguated

4 Semi-automatic annotation of texts

For a corpus to be maximally useful, it has to be searchable in non-trivial manners. The amount of annotation needed to make this possible differs greatly from language to language. Corpora of morphology-poor English, for example, rely heavily on part-of-speech tagging, where individual words of English strings are classified by their word class: "The house is on fire." could be tagged as "The[article] house[noun] is[verb] on[preposition] fire[noun]." For English, this annotation already suffices to allow users to search for a wide range of grammatical structures. For example, a linguist researching the proliferation of the split infinitive ("to boldly go") could simply search for the lexeme "to", followed by an adverb, followed by a verb, to uncover examples of the structure of interest. In languages where words have more internal structure, however, a morphological analysis is indispensable.

Using the mechanism detailed below, the resources in this demonstration infrastructure were run through an automated morphological analyser when imported into the infrastructure. The result of the analysis of a simple string, a rudimentary interlinearization following the Leipzig Glossing Rules [10] as best possible, can be seen in Figure 1.

Figure 2: The same sentence as above, disambiguated

The morphological analysis is not deterministic, i.e. is no morphological disambiguation is performed. When morphological or lexical ambiguity is encountered and several interpretations of a word are found, the analyser yields and saves all possible interpretations. No attempts to cut down on ambiguity using, for example, collocation data or syntactic models have been made so far. Thus, when the analyser encounters the Mari word form *ketɕe* it can either be an uninflected noun meaning "sun" or "day", or the imperative (second person singular) or a connegative form of a verb meaning "to hang". Authorized users (the screenshots in this paper show the interface as seen by an authorized user) can pick the correct glossing by pushing the button titled "[pick]" beside the correct gloss. Their choice is then stored in the database; any subsequent users who encounter this string will see the disambiguated glossing seen in Figure 2. Authorized users can reset the glossing by pressing the button titled "[(re)analyse sentence]".

Note that the disambiguation of morphological or lexical ambiguity does not disambiguate polysemy. The noun *ketɕe* has several aspects of meaning, the most important of which are "sun" and "day". The analyser puts the very first translation given in the lexical base as a glossing for the stem by default. If users move their mouse over the English gloss, all translations contained in the lexcial base show up as a tooltip, as seen in Figure 3. Users can alter the gloss of individual words by clicking the button titled "[edit]" beside a gloss.

Resources fed into the corpus can be sanitized manually as shown here, or they can be left in the corpus in a raw form. Obviously, sanitized data is preferable, but sanitizing data in this manner is time-consuming. Especially if an infrastructure of this sort was realized as a monitor corpus automatically updated on a regular basis, it

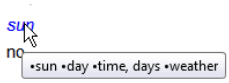

Figure 3: A tooltip

might not be realistic for a human user to disambiguate all the glosses in this manner. However, this data sanitization is not a requirement for the strings to be searchable by syntactic or morphological patterns. The presence of multiple interpretations of a string simply means that the false positive rate of search algorithms will be somewhat higher: grammatical structures will be found in places where they are not present. The false negative rate is not affected: where grammatical structures do occur, the search algorithms do not miss them.

5 Searching the corpus

Users can browse the text collection, or search for specific grammatical constructions within it. The "[Search]"-button on the main page of the corpus infrastructure (`corpus.mari-language.com`) and on pages of indidividual resources[1] and chapters[2]. Depending on where the button is pressed, the entire inventory or one specific resource is searched.

The interlinearization created by the morphological analyser has several layers or tiers. Table 1 gives another example of an interlinearization, with the layers marked. Users can search for full or partial matches on all layers. For example, they can search for all occurences of the base (lemma) form *ɤʃte* "to do", and still find these string even if the allomorph *ɤʃt* occurs in this particular example.

To search for more complicated structures, users can specify additional features that must - or may not - co-occur with the first search specified. It must be specified where addition features must - or may not - occur: in the same word, in the next word, in the previous word, later in the sentence, earlier in the sentence, anywhere in the sentence.

Some examples of possible queries, and the structures they would return:

- "base form" "equals" "ida" - "next word" - "gloss" "equals" "-CNG" "negated' (see Figure 4)': This input would search for occurences of the word "ida", followed by anything but the so-called connegative form. Mari uses a negation

[1] e.g. `www.univie.ac.at/maridict/site-2014/corp_chapters.php?book=1`

[2] e.g. `www.univie.ac.at/maridict/site-2014/corp_content.php?book=1&chapter=16`

(string)	tudo mom ɤʃta?				
unglossed	tudo	mom		ɤʃta?	
morpheme	tudo	mo	-m	ɤʃt	-a
base form	tudo	mo	-m	ɤʃte	-a
gloss	(s)he	what	-ACC	do	-3SG
part of speech	pr	pr	-case	vb2	-pers
(free translation)	What does (s)he do?				

Table 1: The layers of interlinearization, with an example

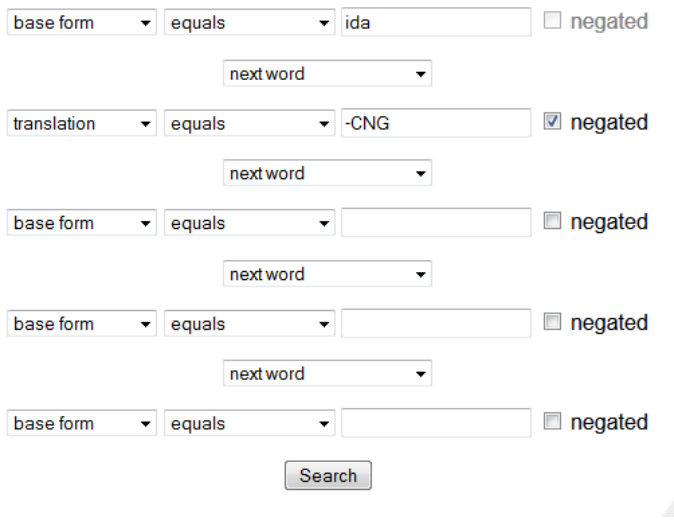

Figure 4: The search interface with a sample query

verb [11, p. 115] typically followed by the connegative form, in the same manner that Finnish does. The form “ida” is the second person plural imperative of the connegative verb, and this query would find all occurrences of it where it is atypically not followed by the negation verb, but rather .

- “gloss” “equals” “-PTCP.FUT” - “next word” - “part of speech” “equals” “po”: This query would return postpositional constructions using the future future participle, which are quite rare compared to postpositional constructions using the passive participle.

Users can use the morphological analyser at `morph.mari-language.com` to determine how exactly the structure they are interested in is glossed by the software. A complete overview of the suffixes processed by the analyser will be included in the documentation.

6 The architecture behind the infrastructure

Three fundamental building blocks were necessary for the creation of this demonstration: a lexical base, a morphological analyser, and a text collection.

6.1 The lexicon

The Mari-English Dictionary [12] created by my project team at the University of Vienna is one of the ingredients needed for this resource. It currently includes 42,560 lexemes. The entries covering these are saved in a systematic format (XML) and are annotated as needed by the analyser - the word class of all lemmas is defined, etc.

6.2 The morphological analyser

The morphological analyser is based on a morphological analyser of Mari I wrote using Java several years ago [13]. Due to repeated difficulties with Java related to security updates, and more general problems resulting from having such a program operate on the client side (i.e., the user’s computer), I recently reimplemented the same infrastructure using server-side PHP. The source code can be found in its entirety at `source.mari-language.com`. Individual strings can be interlinearized at `analyser.mari-language.com`.

A detailed overview of the workings of the analyser will be included in its documentation, which is currently still work in progress. Roughly speaking, the analyser

Suffix	Gloss	PoS	Type	Class	...
∫	X	«case-g2»	LAT	case	...
...	...	...	...	...	...
na	N	«poss»	1PL	poss	...
...	...	...	...	...	...

Table 2: Excerpt of the analyser's inflectional morpheme inventory

follows a naïve item-and-arrangement architecture. The analyser has access to an inventory of inflectional suffixes. Three illustrative entries on this list - which contains over a hundred entries in its entirety - can be seen in table 2.

The field **Suffix** indicates the suffix itself, the fields **Gloss** and **PoS** contain the glosses used for the morphemes in question in the *gloss* and *part of speech* layers respectively (see Table 1).

The field **Type** indicates how a suffix is connected to a stem. The value *E* indicates that this suffix is in some cases preceded by an epenthetic *e*, the value *N* indicates that the suffix is not preceded by an epenthetic vowel. Every suffix is assigned to one type and all suffixes of a type behave in the same way morphologically. The analyser has a separate extraction mechanism for every suffix type that it uses to derive possible base forms when extracting a potential suffix.

The field **Class** assigns every suffix to a grouping. There are complex constraints governing suffix arrangement in Mari [11, p. 75]. For example, possessive suffixes (the class *«poss»*) follow locative, illative, lative, and inessive case suffixes (the class *«case-g2»*), but precede the genitive, accusative, and comitative case suffixes (the class *«case-g1»*), whereas both arrangements are theoretically possible with the dative and comparative case suffixes (the class *«case-g3»*). The frequency of different suffix arrangements has been studied extensively [14], but as the morphological analyser is intended to be possibilistic, not probabilistic, all hypothetically possible arrangements were allowed.

The analyser was equipped with a list of possible suffix arrangements, an excerpt of which can be seen in Figure 5. Every arrangement shows which suffix classes can be connected to which stems (*n* for nominal stems, *v* for verbal stems, etc.) in which order. Suffix classes given in «guillemets» can occur optionally; suffix classes in {braces} must occur exactly once. (The class *{tmp}* represents tense/mode suffixes; one suffix of this type must occur in a finite verbal form.) The analyser only accepts interpretations of words that are compatible with one or more of the valid arrangements known to the computer.

The morphological analyser is, moreover, capable of extracting productive deriva-

```
n + «comp»«gen»«poss»«plur»«case-g1»«p3»«enc»
n + «comp»«gen»«plur»«case-g2»«poss»«p3»«enc»
...
v + {tmp}«comp»«p3»«enc»
...
```

Figure 5: Valid suffix arrangements

Name	Type	Date	Tokens	Eng. Trans.?
Oŋaj marij jʏlme [16]	textbook	2010	2,508	yes
Elnet [17]	novel	1937	63,918	no
The New Testament [18]	religious text	2007	127,717	yes
Marʲijskʲij-russkʲij slovarʲ [19]	dictionary	1990-2005	585,431	no
Mari-English dictionary [12]	dictionary	2014	214,523	yes
Sum	-	-	**994,097**	-

Table 3: Contents of illustrative corpus

tional suffixes from words. Mari morphology generally adheres to the universal principle that derivational suffixes are closer to the base than inflectional suffixes [15, p. 95], though it is possible for plural suffixes to precede derivational suffixes: *verlase* "local" < *ver* "place" + *-la* "-PL" + *-se* "-ADJ". When looking up prospective stems in the lexicon after the inflectional morphology has been extracted, the analyser also attempts to extract any of a number of productive derivational suffixes from the stem that produce words of a valid part of speech (e.g. nominal derivational suffixes when looking up a word that according to the arrangement patterns must be a nominal). Note that for practical reasons, participles are treated as deverbal nominal derivational suffixes by the analyser.

6.3 The texts

Table 3 shows the range of texts included in the demonstration, with some basic data. The New Testament suggested itself as an open-source English counterpart to freely available Mari strings and was included here as well. The content of my work group's textbook *Oŋaj marij jʏlme* [16] was sanitized, but the contents of the other resources were not.

Acknowledgments

I would like to thank Paul Trilsbeek of the Max Planck Institute for Psycholinguistics for providing me with a digital version of the New Testament for non-commercial purposes. Thanks are due to the Kone Foundation for funding the research project that is giving me the opportunity to work on these resoureces ("The Mari Web Project: Phase 2."). Furthermore, I owe gratitude to my former employer, the University of Vienna, for continuing to host my work even after I left its active duty roster. Finally, I am thankful to Timothy Riese and Elaine Bradley for proofreading this paper, and to Nele Lond for help related to hardware matters.

References

[1] Федеральная служба государственной статистики. *Всероссийская перепись населения 2010 года.* [www.gks.ru/free_doc/new_site/perepis2010/croc/perepis_itogi1612.htm, accessed 27 October 2014], 2011.

[2] Коведяева Е. И. Марийский язык. *Основы финно-угорского языкознания*, pages 3–96, 1976.

[3] Unesco. *UNESCO Atlas of the World's Languages in Danger.* Unesco [www.unesco.org/culture/languages-atlas/en/atlasmap.html, accessed 27 October 2014], 1996-2014.

[4] Jeremy Bradley. *The Mari Web Project's Orthography Helper(s).* University of Vienna [transcribe.mari-language.com], 2014.

[5] B. T. Sue Atkins and Michael Rundell. *The Oxford Guide to Practical Lexicography.* Oxford University Press, 2008.

[6] Thomas A. Sebeok and Alo Raun. *The First Cheremis Grammar (1775).* The Newberry Library, 1956.

[7] Alho Alhoniemi and Sirkka Saarinen. *Timofej Jevsevjevs Folklore-Sammlungen aus dem Tscheremisschen I-IV.* Suomalais-Ugrilainen Seura, 1983-1994.

[8] Ödön Beke. *Mari Szövegek I-IV.* Akadémiai Kiadó, 1957-1995.

[9] National Library of Finland. *Kansalliskirjasto Uralica.* National Library of Finland [uralica.kansalliskirjasto.fi, accessed 26 October 2014], 2013.

[10] *Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses.* Max Planck Institute for Evolutionary Anthropology, Department of Linguistics [www.eva.mpg.de/lingua/resources/glossing-rules.php, accessed 27 October 2014], 2008.

[11] Alho Alhoniemi. *Marin kielioppi.* Suomalais-Ugrilainen Seura, 1985.

[12] Timothy Riese, Jeremy Bradley, and Elina Guseva. *Mari-English Dictionary.* University of Vienna [dict.mari-language.com], 2014.

[13] Jeremy Bradley. «mari web project» и его марийский морфоанализатор. *Языки меньшинств в компьютерных технологиях: опыт, задачи и перспективы*, pages 82–89, 2011.

[14] Jorma Luutonen. *The variation of morpheme order in Mari declension.* Suomalais-ugrilainen seura, 1997.

[15] Martin Haspelmath and Andrea Sims. *Understanding Morphology: Second Edition.* Routledge, 2010.

[16] Timothy Riese, Jeremy Bradley, Emma Yakimova, and Galina Krylova. *Оҥай марий йылме: A Comprehensive Introduction to the Mari Language (Release 2.1).* University of Vienna [omj.mari-language.com], 2012.

[17] С. Г. Чавайн. *Элнет.* Книгам лукшо марий издательство, 1967.

[18] *У Сугынь.* Библийым кусарыме институт, 2007.

[19] И. С. Галкин et al. *Словарь марийского языка I-X.* Марийское книжное издательство [dict.komikyv.ru/index.php/list/8/index.xhtml, accessed 27 October 2014], 1990-2005.