Acta Linguistica Academica Vol. **64** (2017) 3, 369–382
DOI: 10.1556/2062.2017.64.3.3

Transcribe.mari-language.com

Automatic transcriptions and transliterations for ten languages of Russia

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

Abstract: The aim of this paper is to introduce my efforts to create server-sided (i.e., platform independent web-based, from a user's perspective) automatic transcription and transliteration software for Uralic and non-Uralic languages of Russia. For ten literary standards – Meadow Mari, Hill Mari, Komi, Udmurt, Erzya, Moksha, Russian, Tatar, Bashkir, Chuvash – an operational interface can be found at transcribe.mari-language.com and the source code at source.mari-language.com, published under a Creative Commons license. This paper details many of the fine aspects of writing systems used for (Meadow) Mari that I had to take into consideration when creating transcription mechanisms for that language.

Keywords: transcription; transliteration; IPA; UPA; ISO-9

1. Structure of this paper

Section 2 describes the circumstances that motivated the creation of the transcription/transliteration infrastructure presented in this paper. Section 3 describes how the software 'normalizes' inputs: ads diacritic symbols users might not have on their keyboard, etc. Sections 4–6 introduce the transcriptions between the relevant writing systems for (Meadow) Mari that my software is capable of handling: Cyrillic (regardless of language) $\longleftrightarrow$ ISO 9:1995, Meadow Mari Cyrillic (including historical, defunct orthographies) $\longleftrightarrow$ UPA, UPA $\longleftrightarrow$ IPA, respectively. (All other transcriptions/transliterations can be achieved by stringing these mechanisms together: for example, ISO 9:1995 can be converted into IPA by a transliteration into Cyrillic, a transcription from Cyrillic into UPA, and a transcription from UPA into IPA). I cannot give a comprehensive overview of all the transformation mechanisms within the limited scope of this paper, but can only provide a brief illustration of some of the more difficult aspects of creating software of this kind. Extensive documentation

can be found on the site where this software is found, transcribe.mari-language.com.

Section 7 illustrates the manner in which I have evaluated the accuracy of the transcription mechanisms for Meadow Mari, and suggests how such testing should happen for the other languages in future. Finally, section 8 briefly discusses equivalent tools to those described in the previous sections for other languages.

2. Why do we need web-based transcription software?

When dealing with languages of the Russian Federation, the choice of a writing system or transcription can be daunting or even politically charged. (Turkic) Tatar can serve as an anecdotal example of a language with a complex past (and present): literary Tatar used the Arabic script until 1927, then Latin-based orthographies until 1939, and since that time the Cyrillic alphabet (Berta 1998, 285). Post-Soviet attempts to reintroduce a Latin-based orthography were rendered moot by a 2002 decision of the Russian constitutional court declaring that all state languages of the Russian Federation must be written in the Cyrillic alphabet (Spolsky 2004, 2).

The situation with regard to (Uralic) Mari is a bit less complex, but not greatly so. Mari literacy traces its roots back to the first Mari grammar, published in 1775 (an extensively annotated facsimile edition of which was published in 1956, Sebeok & Raun 1956); from then until the present day Mari orthographies have predominantly used the Cyrillic alphabet. There are two literary norms of Mari that continue to be actively used, Meadow Mari and Hill Mari. Recent orthographic dictionaries demarcating the rules of the literary standard are available for both Meadow Mari (Иванов et al. 2011) and Hill Mari (Васикова 1994). However, great differences exist between the contemporary literary norms and historical orthographies used in numerous resources. Uralic sources traditionally use the so-called Finno-Ugric Transcription (or UPA – Uralic Phonetic Alphabet) presented in 1901 by Eemil Nestor Setälä (Setälä 1901), with a number of relatively recent high-impact publications (e.g., Alhoniemi 1985; Beke 1997–2001; Alhoniemi & Saarinen 1983–1994) establishing what one might consider an unofficial standard for Latin transcription of Mari. However, competent transcription from Cyrillic into UPA (and vice versa) requires good knowledge of the idiosyncratic aspects of the Mari Cyrillic orthography.

Non-Uralic publications might ask contributors to use the ISO 9:1995 transliteration standard,[1] or the International Phonetic Alphabet (IPA).[2] Whereas an ISO 9 transliteration of literary (Cyrillic) Mari is straightforward and trivial for computer-literate scholars (albeit potentially time-consuming, as online applications for the ISO-9 transliteration of Cyrillic texts (e.g., translit.cc) cannot handle the additional characters found in Mari orthographies that are not part of the Russian alphabet: *ӓ*, *ҥ*,*ӧ*, *ӱ*, *ӫ*), using IPA for Mari is not. I am not aware of any publications other than my own (Bradley 2015; Riese et al. 2014) that use IPA for Mari, and deriving IPA from UPA can be challenging due both to the fact that UPA is not as stringently standardized as IPA is and to a lack of information on the exact pronunciation of sounds in relevant sources. For example, Alho Alhoniemi's Finnish-language grammar of Mari (Alhoniemi 1985), which thanks to its German translation (Alhoniemi 1993) is still the most extensive and modern resource on Mari grammar at least marginally accessible to the international linguistic community, introduces the system of Meadow Mari vowels as seen in Figure 1.

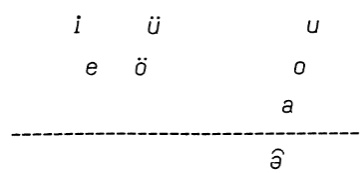

Figure 1: Meadow Mari Vowels (Alhoniemi 1985)

The sound /ə̂/ is especially challenging here. Alhoniemi's graphic representation, which resembles the vowel trapezium, does not give detailed information concerning either the exact position of the vowel or rounding. According to Pekka Sammallahti, UPA /ə̂/ is a reduced mid central unrounded vowel (Sammallahti 1998, 174) (/ə/ in IPA), but even an inspection by ear casts that classification into doubt. My work group rather identified the sound as a mid back unrounded vowel (/ɤ̞/ in IPA), and we marked it as such in our materials (Riese et al. 2014; 2017).

In summary, there are numerous writing systems that scholars dealing with Mari might encounter and in which they might be expected to be able

[1] www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=3589

[2] https://www.internationalphoneticassociation.org/

to produce texts. Transcriptions from some systems into others might be more or less straightforward from a technical standpoint (e.g., UPA ⟷ IPA), but require a very good understanding of the language (which general Uralicists or typologists dealing with Mari superficially might not have). Transliteration between Cyrillic and ISO 9:1995 is, technically speaking, absolutely trivial, but time-consuming for scholars not capable of writing their own transliteration scripts. Motivated by these circumstances, I have created a web-based interface that for ten languages (or language standards) with which I am either familiar, or for which I could consult competent scholars – Meadow Mari, Hill Mari, Komi, Udmurt, Erzya, Moksha, Russian, Tatar, Bashkir, Chuvash – that allows the transcription or transliteration of text from all relevant writing systems that I know of into (almost) all other writing systems, as accurately as the orthographies and transcription systems in question allow. It should be noted that while I have extensively tested the mechanisms for Mari, Tatar, and Russian, those for other languages are in an earlier stage of development and might still be comparatively error-prone.

An operational interface can be found at transcribe.mari-language.com and the PHP source code at source.mari-language.com. The source code is published under a Creative Commons license and can be repurposed for non-profit purposes under the condition of attribution. (If the license we chose is inconvenient for users, we would request they contact us directly.) The relevant procedures can be found in the file *functions.php*; the user interfaces can be found in the files *transcription-general.php*, *transcription-specific.php*, and *transcription-universal.php*. Where relevant, sections of this paper include a footnote containing the name(s) of the function(s) in *functions.php* carrying out the operations detailed in it.

By integrating Mari transcription mechanisms into our work group's electronic Mari-English Dictionary (Riese et al. 2014), which was compiled using contemporary Cyrillic orthography (with additional annotation compensating for defects in the orthography), it became usable in UPA and IPA, depending on a scholar's needs. Moreover, the dictionary's interface allows entries to be displayed using reverse sorting, i.e., sorted right-to-left, starting with the last letter of the word, then the penultimate letter, etc. This is especially useful due to the fact that the same vowel sound is indicated by different Cyrillic characters depending on its environment (UPA/IPA /a/ ⟷ Cyrillic ⟨*а*⟩, ⟨*я*⟩; UPA/IPA /e/ ⟷ Cyrillic ⟨*э*⟩, ⟨*е*⟩; UPA/IPA /u/ ⟷ Cyrillic ⟨*у*⟩, ⟨*ю*⟩) – a reverse-sorted list of lexemes is more useful in UPA or IPA than it is in the Cyrillic orthography we used when creating our dictionary.

3. Orthographic normalization

All software tools found on our website should be usable using an arbitrary Cyrillic (e.g., Russian) or Latin (e.g., English) keyboard layout – i.e., keyboard layouts that only contain the 26 letters of the basic Latin alphabet (and punctuation marks, numbers, etc.), and keyboard layouts that cover all letters used by the Russian Cyrillic alphabet, but not the additional Mari characters *ӓ, ҥ, ӧ, ӱ*, and *ӵ*. To facilitate this, the software includes a number of mechanisms allowing orthographic normalization, for both Cyrillic and Latin inputs. Users can access these by setting the same writing system as the input and the output in the user interface – "Cyrillic to Cyrillic", etc. These same normalization procedures are also carried out on inputs if other options are chosen – e.g., if users ask the software to transcribe Cyrillic to IPA, the input is subjected to the orthographic normalization procedures illustrated here.

Unfortunately, I have not yet been able to implement fully automatic orthographic normalization with dictionary support – i.e., procedures that would restore lacking diacritical markings that are not indicated by the user in any way, but that could be assumed to be necessary in a given place with knowledge of the language.

3.1. Cyrillic

The strategies used by the software to normalize Cyrillic input are based on strategies used by Mari native speakers in colloquial contexts (e.g., in e-mails, on social network sites). To indicate a special Mari character, users can either place a colon *:* after the letter from which it is derived (i.e., *а:* ⟶ *ӓ*, *н:* ⟶ *ҥ*, *о:* ⟶ *ӧ*, *у:* ⟶ *ӱ*, *ы:* ⟶ *ӹ*), or capitalize the letter from which the special character is derived inside a word (e.g., *шУм* ⟶ *шӱм* /šüm/ 'heart') (function *cyrprep*).

3.2. UPA

In Latin-based UPA inputs, users can place a colon *:* after a letter to create UPA-characters that are not part of the basic Latin alphabet, or can use a number of digraphs. In some cases, simple letters can be used to produce UPA symbols, as these simple letters (*y, q, h*) have no UPA value of their own. Table 1 gives an overview of normalization procedures.

If users wish to prevent two letters from being read as a digraph, they can place a vertical bar / between the two words: The input *sheme*

Table 1: Orthographic normalization of UPA inputs

Input(s)	Output
a:, æ	ä
o:, ø	ö
u:	ü
y, õ	ə̂
y:, q	ə
z, zh	ž
s, sh	š
c, ch	č
n, ng	ŋ
h, x	χ

produces the output *šeme* 'black', while the input *s/heme* produces the output *sχeme* 'diagram' (a Russian loan word – the sound *χ* is not found in indigenous Mari vocabulary) (function *latprep*).

4. Cyrillic ⟷ ISO 9:1995

ISO 9:1995[3] is a transliteration system: there is a deterministic 1:1 relationship between Cyrillic characters and Latin characters (e.g., Cyrillic *э* ⟷ ISO 9 *è*); the transliteration occurs completely independent of the pronunciation rules of the language(s) in question. As such, an ISO 9:1995 transliteration can be realized by simply replacing Cyrillic characters with the corresponding Latin characters. Due to the simplicity and language-independence of the task, I have expanded the function responsible for transliteration between Cyrillic and ISO 9:1995 to cover all contemporary (and some non-contemporary) written languages using the Cyrillic alphabet that I am aware of.

5. Meadow Mari Cyrillic Orthographies ⟷ UPA

Alho Alhoniemi's grammar of Mari gives a good overview of the relationship between the modern Meadow Mari Cyrillic Alphabet and UPA (Alhoniemi 1985, 28–29); I. G. Ivanov's handbook on the phonetics of contem-

[3] www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=3589

porary Mari (Иванов 2000) provides detailed accounts of the exact pronunciation of individual sounds. While the orthography is mostly straightforward and there is a 1:1 relationship between many consonant symbols and consonant sounds (e.g., Cyrillic ⟨ш⟩ ⟷ UPA /š/), there are a number of difficult aspects (where programming a transcription script is not trivial), and also some actual defects (where programming a fully automatic and accurate transcription script is not possible unless the user gives disambiguating information) (function *cyr_to_lat*).

5.1. Vowel signs, palatalness, /j/

The most critical aspect is the usage of different vowel signs to indicate palatalness and the phoneme /j/. The realization of vowel sounds in Mari orthography differs depending on their position in a word; the marking of palatalness (there is a phonological distinction /n/ ∼ /ń/ and /l/ ∼ /l'/ in Mari, and there are numerous other distinctions in Russian loan words) and the phoneme /j/ depends on the vowel sound, if any, following the consonant. Table 2 gives an overview of how the eight vowel sounds of Mari are realized in orthography, depending on whether they follow a non-palatal consonant /C/, a palatal consonant /C'/ (either /ń/ or /l'/), the sound /j/, if they are in the initial position, etc. Note that some of these combinations are rarely encountered or only occur in compounds and/or Russian loan words.

Table 2: Vowel signs, palatalness, and the phoneme /j/

	/**C**_/	/**C'**_/	/**j**_/	/#_/	/**V**_/	/**Cj**_/	/**C'j**_/
/**a**/	⟨Са⟩	⟨Ся⟩	⟨я⟩	⟨а⟩	⟨Vа⟩	⟨Съя⟩	⟨Сья⟩
/**u**/	⟨Су⟩	⟨Сю⟩	⟨ю⟩	⟨у⟩	⟨Vу⟩	⟨Съю⟩	⟨Сью⟩
/**e**/	⟨Се⟩	⟨Се⟩	⟨е⟩	⟨э⟩	⟨Vэ⟩	⟨Съе⟩	⟨Сье⟩
/**i**/	⟨Си⟩	⟨Си⟩	⟨и⟩	⟨йи⟩	⟨Vи⟩	——	——
/**o**/	⟨Со⟩	⟨Сьо⟩	⟨йо⟩	⟨о⟩	⟨Vо⟩	⟨Сйо⟩	⟨Сьйо⟩
/**ö**/	⟨Сӧ⟩	⟨Сьӧ⟩	⟨йӧ⟩	⟨ӧ⟩	⟨Vӧ⟩	⟨Сйӧ⟩	——
/**ə**/	⟨Сы⟩	⟨Сьы⟩	⟨йы⟩	⟨ы⟩	⟨Vы⟩	⟨Сйы⟩	——
/**ü**/	⟨Сӱ⟩	——	⟨йӱ⟩	⟨ӱ⟩	⟨Vӱ⟩	⟨Сйӱ⟩	——

The grey cells in Table 2 are especially problematic: before the vowels /e/ and /i/, there is no orthographic distinction between palatal consonants and their non-palatal counterparts. Modern Mari orthography does not distinguish between /le/ and /l'e/, for example, and homographs that are not homophones (i.e., words that are spelled, but not pronounced, in the same way) can be found. For example ⟨*неле*⟩: /nele/ 'difficult' ∼ /nel'e/ '(s)he swallowed'. In these cases, users must manually indicate an orthographically unmarked palatalness with an apostrophe (i.e., ⟨*л'е*⟩ ⟶ /l'e/) to get a correct transcription. We indicated palatalness by these means in our Mari-English dictionary (Riese et al. 2014).

5.2. Orthographically unmarked features

Whereas palatalness, discussed above, is sometimes marked in orthography and sometimes not, there are a number of processes and features that are systematically not marked in the contemporary orthography (presented here without full historical explanations for the phenomena involved):

- The letter ⟨*ð*⟩, while historically generally pronounced as UPA /δ/ / IPA /ð/ and today generally pronounced as /d/ (e.g., ⟨*кид̲ем*⟩ /kid̲em/ 'my hand, my arm'), is pronounced as /t/ in syllable-final position (e.g., ⟨*кид̲*⟩ /kit̲/ 'hand, arm' (Alhoniemi 1985, 33–34).

- The letters ⟨*ð*⟩ and ⟨*г*⟩, which have the prototypical values /d/ and /g/ (historically UPA /δ/ / IPA /ð/ and UPA /γ/ / IPA /ɣ/), are pronounced as /t/ and /k/ respectively after voiceless obstruents. For example, the negative gerund in /-de/ ∼ /-te/ (Alhoniemi 1985, 144–146) (orthographically always ⟨*-ðe*⟩): ⟨*мол-*⟩ /tol-/ 'to come' ⟶ ⟨*молðe*⟩ /tol̲de/ 'without coming', but ⟨*поч-*⟩ /poć/ 'to open' ⟶ ⟨*почð̲e*⟩ /poć̲te/ 'without opening' (Alhoniemi 1985, 33–34). This process occurs across orthographic word boundaries, e.g., the postposition /gə̂ć/ ∼ /kə̂ć/ 'from' (orthographically always ⟨*гыч*⟩): ⟨*ола г̲ыч*⟩ /ola g̲ə̂ć/ 'from town', but ⟨*мут г̲ыч*⟩ /mut k̲ə̂ć/ 'from a word' (Иванов 2000, 90).

- A number of consonant clusters are pronounced in manners that diverge from their orthographic realization, thanks to assimilation (Иванов 2000, 99–105): ⟨*жс̲т*⟩ /š̲t/, ⟨*з̲т*⟩ /s̲t/, ⟨*жс̲ш*⟩ /š̲š/, ⟨*з̲ш*⟩ /s̲š/, ⟨*в̲к*⟩ /p̲k/, ⟨*г̲к*⟩ /k̲k/, ⟨*нк̲*⟩ /ŋ̲g/, ⟨*н̲г*⟩ /ŋ̲g/, ⟨*н̲ч*⟩ /ń̲č/.

- Orthographically unmarked word stress tends to fall on the last full vowel of Mari words (Alhoniemi 1985, 17), where a full vowel is anything but the reduced vowel /ə̂/, and final unstressed /e/, /o/, and /ö/ (Alhoniemi 1985, cf. 20–21; 39–40). However, /e/, /o/, and /ö/ can occur as stressed full vowels in the final position, and there are examples where words are spelled the same, but are pronounced differently: ⟨*шерге*⟩: /še•rge/ 'expensive' ∼ /šerge•/ 'comb'. That is to say, stress is a phonologically relevant feature that is not orthographically marked. It is usually, but not always, predictable; it is in cases where it is unpredictable that it might be phonologically relevant.

- Final unstressed /e/, /o/, and /ö/ are slightly reduced (Иванов 2000, 58–59), e.g., ⟨*йыламе*⟩ /jə•lmeə̂/ 'tongue; language', ⟨*тумо*⟩ /tu•moə̂/ 'oak tree' ⟨*шӱдӧ*⟩ /šü•dөə̂/ 'hundred'.

- More recent Russian loan words, and Russian names in particular, might be pronounced in accordance with Russian, rather than Mari, pronunciation rules.

With many of these features, it is questionable whether or not automatic transcription software should take them into consideration, even if it would be possible for such a system to handle them. They would make back-transformation more difficult, and the orthography can in some cases have a disambiguating function. There are words that are pronounced the same due to the rules detailed above, but are not spelled the same, e.g., ⟨*кид*⟩ /kit/ 'hand, arm' ∼ ⟨*кит*⟩ /kit/ 'whale' (a Russian loan word). Thus an accurate transcription with respect to pronunciation rules is not loss-less and might ultimately be considered unnecessary for many purposes: scholars acquainted with the rules of Mari pronunciation can derive the correct pronunciation from a transcription that retains some aspects of the orthography. It is left up to the user to decide whether or not the features described above are taken into consideration:

- If users activate the checkbox labelled "Orthographically unmarked features (assimilation, etc.)", the system will take the features discussed into consideration to the best degree possible.

- With respect to word stress, the system will assume that the stress falls on the last full vowel (see above) unless specified otherwise. Users can manually define the stress for a particular word by placing an asterisk * after the unpredictably stressed vowel.

- Square brackets [ ] can be used to indicate Russian words, names, and text segments as such. Any text enclosed in square brackets will be transcribed in accordance with the rules of Russian, rather than Mari, orthography (these rules are detailed in the documentation). For example, if the name ⟨*Домодедово*⟩ is placed in square brackets, it is transcribed as /domod̲'ed̲o̲v̲o/ rather than /domo̲d̲e̲do̲β̲o/ – with a Russian palatalized consonant /d'/ and the letter ⟨*в*⟩ having its Russian value /v/ rather than Mari /β/. It would be desirable for the transcription mechanisms to recognize Russian words and text segments automatically, but I have not yet implemented mechanisms capable of doing so – for the time being, users must manually indicate Russian text for it to be processed correctly.

5.3. Early 20th century orthographies

Mari was subjected to an extensive orthographic reform in 1938 (Иванов 2003, 291). Numerous Mari-language newspaper texts from the 1920s and 1930s made available on the National Library of Finland's website (uralica.kansalliskirjasto.fi) use the orthography that become obsolete with this reform, which differs significantly, but systematically, from the contemporary orthography. Before 1938 the letter ⟨*е*⟩ was only used after palatal consonants and the sound /e/ was otherwise consistently marked by the letter ⟨*э*⟩. Moreover, the earlier orthography consistently marked the phoneme /j/ with the letter ⟨*й*⟩. Table 3 shows the various manners in which different sound combinations are indicated in the old and contemporary orthographies respectively, and illustrates that defects regarding the marking of palatalness in modern orthography were not found in pre-1938 writing systems (function *thirtiesprep*).

Table 3: Mari orthographies: pre-1938 and today

UPA	1930s	Contemporary	UPA	1930s	Contemporary
/**ja**/	⟨йа⟩	⟨я⟩	/J̲aponij/	⟨Йапоний⟩	⟨Японий⟩ 'Japan'
/**C'a**/	⟨Сьа⟩	⟨Ся⟩	/okt'a̲br'/	⟨октьа̲брь⟩	⟨октя̲брь⟩ 'October'
/**je**/	⟨йэ⟩	⟨е⟩	/mij̲e̲n/	⟨мийэ̲н⟩	⟨мие̲н⟩ '(s)he went'
/**C'e**/	⟨Сьэ⟩	⟨Се⟩	/ə̂l'e̲/	⟨ыльэ⟩	⟨ыле̲⟩ '(s)he was'
/**ju**/	⟨йу⟩	⟨ю⟩	/j̲umo/	⟨йумо⟩	⟨ю̲мо⟩ 'god'
/**C'u**/	⟨Сьу⟩	⟨Сю⟩	/pol'u̲s/	⟨польус⟩	⟨полю̲с⟩ 'pole'
/**Ce**/	⟨Се⟩	⟨Се⟩	/d̲e̲n/	⟨дэ̲н⟩	⟨де̲н⟩ 'and'

The software is capable of transcribing texts from the old orthography into the contemporary one. As the old orthography is less ambiguous, this is not difficult from a technical point of view. Because the old orthography is now defunct, the software does not offer transcriptions into it, despite its better handling of palatalness.

6. UPA ⟷ IPA

Once correspondences were established between UPA and IPA values, a transcription from UPA into IPA (or from Cyrillic into IPA via UPA) was more or less straightforward. One problem that arose here, however, is that UPA does not distinguish between palatal and palatalized consonants: UPA /ń/ corresponds to both IPA /ɲ/ and IPA /nʲ/. As Mari has palatal rather than palatalized consonants, I configured the software, by default, to transcribe UPA /ń/ as IPA /ɲ/ and to transcribe /ń/ as /nʲ/ only within words or phrases marked as Russian by users by means of square brackets (see above). Thus ⟨*сугынь*⟩ 'blessing' is transcribed into UPA as /sugə̑ń/ and then into IPA as /sugɤɲ/, but Russian ⟨*июнь*⟩ 'June', if placed within brackets, is transcribed into UPA as /ijuń/ and then into IPA as /ijunʲ/ (function *upa_ to_ ipa,* function *ipa_ to_ upa*).

7. Evaluating the mechanisms

When developing the mechanisms presented in this paper, I evaluated and reworked them incrementally by feeding large amounts of texts into the algorithm, seeing if the output was satisfactory, and refining the mechanisms to compensate for any shortcomings uncovered in the process. I am confident that, within the framework at hand, the mechanisms for Meadow Mari are optimal at this point. Table 4 – a paragraph from a 1931 agricultural report[4] which has been transferred into modern orthography, UPA, and IPA – illustrates, however, a number of shortcomings that remain that cannot be fixed within the framework at hand, and which will require expansions. Emphasis was added to show some problematic words.

[4] fennougrica.kansalliskirjasto.fi/handle/10024/67712

Table 4: Evaluating the mechanisms: sample paragraphs

1930s	Contemporary	UPA	IPA
Колхоз озанлыкын кажнэ участкыштыжэ палэмдымэ паша планым да **мачэрйалым** йодын **зайавкым** ыштыман. Мут гыч, **имньэ** ончышо-влак ...	Колхоз озанлыкын кажне участкыштыже палемдыме паша планым да **мачеръялым** йодын **заявкым** ыштыман. Мут гыч, **имне** ончышо-влак ...	Kolχoz ozanlə̑kə̑n kažne uć̆astkə̑štə̑že palemdə̑me paša planə̑m da **maćerjalə̑m** jodə̑n **zajaβkə̑m** ə̑štə̑man. Mut gə̑č̆, **imńe** onč̆ə̑šo-βlak ...	kolxoz ozanlɤ̈kɤ̈n kaʒne ut͡ɕastkɤ̈ʃtɤ̈ʒe palemdɤ̈me paʃa planɤ̈m da **mat͡ɕerjalɤ̈m** jodɤ̈n **zajaβkɤ̈m** ɤ̈ʃtɤ̈man. mut gɤ̈t͡ɕ, **imɲe** ont͡ɕɤ̈ʃo-βlak ...

The problems arising in this sample are:

- ***мачэрйалым*–*мачеръялым*–/maćerjalə̑m/–/mat͡ɕerjalɤ̈m/**: This Russian loan word was adapted in an unpredictable manner in the pre-1938 orthography, but is not in the modern orthography: today, it would be spelled ***материалым***, and be transcribed into UPA and IPA differently. This incorrect transfer into modern orthography cannot be prevented without lexical data – correspondences between older orthographies and the current orthography.

- ***зайавкым*–*заявкым*–/zajaβkə̑m/–/zajaβkɤ̈m/**: While the letter ⟨*в*⟩ in Mari traditionally represents a voiced bilabial fricative /β/, it in Russian represents a voiced labiodental fricative /v/. In this Russian loan word, it would presumably be more appropriate for ⟨*в*⟩ to be transcribed as /v/ in both UPA and IPA, but the software is currently not able to identify Russian loan words independently–users must indicate Russian loan words as such (see above).

- ***имньэ*–*имне*–/imńe/–/imɲe/**: No problems occur here as the source text from 1931 contains clear orthographic marking of palatal pronunciation, which is lacking in modern orthography – see above. If the source text had been in mordern orthography, the software would not have been able to identify the palatal pronunciation of ***н***. Here again, lexical support would be necessary to make the software reliable.

To sum up, the software is currently as reliable as it can be without lexical support – for results to be satisfactory in all cases, users must manually add features that are relevant to pronunciation, but not orthographically marked.

Transcription mechanisms for Russian and Tatar have been tested in a similar fashion, though not to the same extent. Transcription and transliteration mechanisms for all other languages have not yet been evaluated in the same manner, and should optimally be evaluated and refined in cooperation with experts on these languages.

8. Conclusions and prospects

For the time being, I have created language-specific diacritic helpers for a total of 102 languages of Eurasia, roughly half of which use the Cyrillic alphabet. Like the Mari-related mechanisms detailed in this paper, these can be found at transcribe.mari-language.com. In my own evaluation, the transcription mechanisms I have implemented are reliable for Mari, Tatar, and Russian: if a text consistently follows the rules of the input writing system, the output is, as a rule, correct. Transcription mechanisms implemented for other languages will require more testing before similar claims can be made in respect to them.

The diacritic helpers allow users to access the specific special characters used in a language's alphabet using shortcuts. For example, the diacritic helper for Kalmyk – which uses six characters not found in the Russian alphabet, *ә, ө, ү, ң, җ*, and *h*) – carries out the following transformations (on both lower-case and upper-case characters): *a:* ⟶ *ә, o:* ⟶ *ө, y:* ⟶ *ү, н:* ⟶ *ң, жс:* ⟶ *җ, x:* ⟶ *h*.[5] Given time and assistance from scholars of other languages, I hope to widen the scope of languages handled – and improve the quality of mechanisms for those languages already offered – in the future, and market my tools to a wider audience.

[5] These mechanisms are available for Mari as well if one asks the infrastructure to transcribe from Cyrillic into Cyrillic.

References

Alhoniemi, Alho. 1985. Marin kielioppi [Mari grammar]. Helsinki: Suomalais-Ugrilainen Seura.

Alhoniemi, Alho. 1993. Grammatik des Tscheremissischen (Mari). Mit Texten und Glossar. Hamburg: Helmut Buske.

Alhoniemi, Alho and Sirkka Saarinen. 1983–1994. Timofej Jevsevjevs Folklore-Sammlungen aus dem Tscheremisschen I–IV. Helsinki: Suomalais-Ugrilainen Seura.

Beke, Ödön. 1997–2001. Mari nyelvjárási szótár I–IX (Bibliotheca Ceremissica 4) [Mari dialectal dictionary 1–9]. Szombathely: Savariae.

Berta, Árpád. 1998. Tatar and Bashkir. In Éva Ágnes Csató and L. Johanson (eds.) The Turkic languages. London: Routledge. 283–300.

Bradley, Jeremy. 2015. Mari converb constructions – Interpretation and translation. In M. Hilpert, J.-O. Östman, C. Mertzlufft, M. Rießler and J. Duke (eds.) New trends in Nordic and general linguistics. Berlin: Walter de Gruyter. 141–161.

Иванов, Иван Григорьевич 2000. Кызытсе марий йылме – фонетика [Contemporary Mari: Phonetics]. Йошкар-Ола: Марий книга савыктыш.

Иванов, Иван Григорьевич 2003. Марий литературйылме историй [History of Mari literacy]. Йошкар-Ола: Марий кугыжаныш университет.

Иванов, Иван Григорьевич 2011. Марий орфографий мутер [Mari orthographic dictionary]. komikyv.ru/pdf/orfografi_muter.pdf

Riese, Timothy, Jeremy Bradley and Elina Guseva. 2014. Mari–English dictionary. Vienna: University of Vienna. dict.mari-language.com

Riese, Timothy, Jeremy Bradley, Emma Yakimova and Galina Krylova. 2017. Оҥай марий йылме: A comprehensive introduction to the Mari language (release 3.2). Vienna: University of Vienna. omj.mari-language.com

Sammallahti, Pekka. 1998. The Saami languages. An introduction. Karasjok: Davvi Girji OS.

Sebeok, Thomas A. and Alo Raun. 1956. The first Cheremis grammar (1775). Chicago: The Newberry Library.

Setälä, Eemil Nestor. 1901. Über transskription der finnisch-ugrischen sprachen [sic]. Finnisch-ugrische Forschungen 1. 15–52.

Spolsky, Bernard. 2004. Language policy: Key topics in sociolinguistics. Cambridge: Cambridge University Press.

Васикова, Лидия Петровна 1994. Кырык марла орфографи лӹмдер [Hill Mari orthographic dictionary]. Йошкар-Ола: Мары Элӹн периодика.