The many writing systems of Mansi: challenges in transcription and transliteration

Jeremy Bradley and Elena Skribnik

[1] University of Vienna
[2] Ludwig Maximilian University of Munich
jeremy.moss.bradley@univie.ac.at

Abstract. The paper at hand presents the recently published COPIUS[1] Orthographic Toolset's Mansi module. This open-source software, part of the COPIUS drive to create necessary international infrastructures for teaching/learning and researching Uralic languages, allows for rule-based transcription between four basic writing systems historically used for Mansi: the Cyrillic alphabet, the Latin-based Unified Northern Alphabet (UNA), Finno-Ugric Transcription (FUT), and the International Phonetic Alphabet (IPA). The software aims to take variation in the usage of these respective writing systems into consideration as best possible in a purely rule-based approach currently lacking lexical support. Section 1 will give a short summary of the history of Mansi literacy and aims to elucidate how changing trends, both local and Russia-wide, influenced the manner in which Mansi was captured in writing by scientists and speakers throughout history. Section 2 will give an overview of (Northern) Mansi phonology and discuss how difficult aspects of it are handled in the writing systems under consideration. Finally, Section 3 will illustrate the transcription software, in its current version, in action, with a sample text transcribed from each of the four writing systems under consideration into the three other ones.

Keywords: Mansi, transcription, writing system

1 History of Mansi literacy and documentation

The Mansi language of Western Siberia are rapidly approaching moribund status. In the 2010 All-Russia Population Census, 938 individuals self-identified as speakers of Mansi[2]. The critical mass of remaining speakers is elderly; transmission to younger generations can no longer be assumed. One exception is a small group of speakers of the Upper Lozva subdialect of the Northern Mansi: they have compact isolated settlements (the Ushma and Treskol'ye villages in the Ivdel District of the Sverdlovsk Oblast), use their native language in the everyday life; even children are competent

[1] COPIUS: Erasmus+ strategic partnership "Community of Practice in Uralic Studies", www.copius.eu; Mansi module of Orthographic Toolset at www.copius.eu/trtr.php?lang=mns
[2] www.perepis-2010.ru/results_of_the_census/results-inform.php

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 12–24.

speakers. Ivdel Mansi, about 100 speakers, make ca. 10% of all Mansi speakers today (Zhornik 2019: 101).

Though Mansi literacy has a weak standing today, it traces its roots to the 18th century. The first records of Mansi were created for administrative and missionary purposes. One early written record of Mansi, among many other languages, is the collection of word lists compiled in different provinces of the Russian Empire at the behest of Catherine II, published in the comparative dictionary of Peter Simon Pallas (Pallas 1787–1789). Some unpublished Mansi word lists are preserved in the St. Petersburg Branch Archive of the Russian Academy of Sciences (Normanskaja, Kosheliuk 2020). All of these resources use the Cyrillic script and follow the basic principles of pre-revolutionary Russian orthography. As they were compiled by non-linguists with no understanding of Mansi phonology, they reflect what the compilers could acoustically perceive and how they thought it most fitting to render what they heard. The Cyrillic script is also used in small dictionaries and texts compiled by orthodox priests in the late 19th and early 20th century (e.g. Azbuka 1903).

Mansi texts collected in the 19th and early 20th century give a more advanced view on the language and its varieties. They were collected by professional philologists, mostly from Hungary and Finland where scientific interest towards related peoples as well as their languages and folklore was bolstered by national romanticism. Most prominently, Bernát Munkácsi (1860–1937) and Artturi Kannisto (1874–1943) documented a vast dialectal landscape of vibrant speaker communities. The numerous materials they elicited have been published; many of them can be found as interlinearized texts in the Ob-Ugric Database[3]. Around the turn of the 19th to 20th century, transcription systems used in Uralic Studies were standardized as the Uralic Phonetic Alphabet (UPA), or Finno-Ugric Transcription (FUT) (Setälä 1901).

The next important text collection is that of Valeriy Chernetsov (expeditions in 1933–1938); his field journals are preserved in the Museum of Archeology and Ethnography of Tomsk State University[4]. This collection uses the Unified Northern Alphabet (Единый Северный Алфавит), which is based on the Latin script and was created from 1926 to 1932 for 16 languages, including Mansi, Khanty, Nenets, Selkup and Saami, by the specialists of the Northern Faculty (since 1930 the Institute of the Peoples of the North) at the Leningrad Oriental Institute (cf. Partanen & Rießler 2019).

The creation of the Unified Northern Alphabet (UNA) is connected with the incipient literacy of the speaker community itself. After the revolutions of 1917 a short period of support for languages and cultures of indigenous peoples followed: throughout Russia, literary standards were created, indigenous languages were introduced into education, teaching materials and readers for national schools were translated into indigenous languages. The Latin base for the UNA was selected with the goal of internalization; publications from this time period (e.g. Ƶuḷov 1933, cf. Partanen & Rießler 2019) can be found online in the National Library of Finland's Fenno-Ugrica Collection[5]; an

[3] OUDB, www.babel.gwi.uni-muenchen.de/.

[4] The digitalized version can be found at www.museum.tsu.ru/collection

[5] fennougrica.kansalliskirjasto.fi/

ABC-book (1932), several small readers and a dictionary for Mansi primary schools were compiled in UNA by Valeriy Chernetsov.

The UNA was abandoned in 1937 when it was mandated that Cyrillic alphabets must be used for the languages of Russia. Rapid Cyrillization followed, at first with no additional characters for specific Mansi phonemes and adhering to the syllable principle of the Russian orthography, where the palatalization/palatal expression of consonants is marked by the following special (iotated) vowel letter (so called йотированные гласные буквы), e.g. IPA /nas/ 'simply' > Cyrillic <нас>, IPA /nʲas/ 'hook' > Cyrillic <няс>. This first phase of Mansi Cyrillic orthography is exemplified by and documented in (Balandin 1958). The letter <ӈ> for the velar nasal /ŋ/ was introduced in the end of 1950s (previously the digraph <нг> had been used); the reform of 1979 introduced the marking of long vowels via macron (Rombandeeva 1982, 1985), e.g. IPA /sam/ 'eye' > Cyrillic <сам>, IPA /saːm/ 'corner' > Cyrillic <са̄м>. In the case of <ё> (the iotated counterpart of <о>), this convention resulted in a letter with multiple diacritics: <ё̄>. For quite a long time the realization of the alveo-palatal fricative /sʲ/ varied between <щ> and <сь> (word-finally; before the vowels it was <с> + iotated vowel letters) until the latter was established as the norm in 1979. A special problem is presented by the use of letters <и> and <ы>: both are used for short /i/ and for schwa; the choice of character differentiates between dentopalatal and alveo-palatal consonants like IPA /t/ and /tʲ/ (see the table 1 below), e.g. IPA /ti/ 'this' > Cyrillic <ты>, IPA /aːtʲi/ 'no' > Cyrillic <а̄ти>. However, as there are only four such pairs, after all other consonants the two letters are often used indiscriminately.

Between the 1930s and 1960s about 50 books in Mansi were published, both original works and translations, many of them adapted for children. In the 1970s and 1980s, such publications were rare. Some Mansi materials in a simple variant of the orthography were published in the Khanty language newspaper "Lenin pant xuvat" ('Along Lenin's way'). A factor in this shift was that in the late 1950s, several large Russian-language boarding schools replaced the network of small national schools, while a government policy of settling nomadic communities in larger villages was implemented, leading to increased Russification (Skribnik, Koshkaryova 1996).

After Perestroyka, a period of language revitalization started: the newspaper "Lūjimā sēripos" ("The Northern Sunrise") has appeared roughly twice a month since 1988[6]. In 1992, the publishing house "Северный дом" ('The Northers House') was established in Surgut (Khanty-Mansi Autonomous Okrug); it aimed to foster publications in national languages. Inspired by the work of the Hungarian folklorist Éva Schmidt (1948–2002) in Beloyarsk, local researchers have published several new folklore collections; the materials of Munkácsi and Kannisto have been reprinted in Cyrillic orthography with Russian translations[7]. A new generation of school textbooks and readers have been appearing in recent years. It remains to be seen if these efforts will help turn the tide on the language's decline.

[6] Available online at khanty-yasang.ru/luima-seripos

[7] Available online at ouipiir.ru/collections?field_sel_ethnos_value=mansi&field_authors_tid=All

For more detailed information on Mansi literacy see (Parfenova 2003; Riese & Bradley; Burykin 2000); for general trends pertaining to Uralic languages of Russia, see e.g. (Siegl & Rießler 2015; Rueter & Ponomareva 2019).

In summation, any linguist working with Mansi today has to deal with a great variety of transcriptions and orthographies:

- non-standardized variants of the old Cyrillic alphabet in early documentation of Mansi;
- several idiosyncratic variants of Latin-based transcriptions, e.g. by Munkácsi and Kannisto, before Finno-Ugric Transcription (FUT) was established;
- FUT in the later scientific publications (e.g. Kálmán 1976a & 1976b, Riese & Bradley)
- Latin-based Unified Northern Alphabet (e.g. Chernetsov; Ƶuļov 1933);
- IPA in the recent publications on Mansi (e.g. Ob-Ugric Database; Bakró-Nagy, Sipőcz & Skribnik: to appear);
- Several variants of Cyrillic-based orthography, both simplified (1937-1979, e.g. Balandin 1960) and reformed (since 1979, e.g. Skribnik & Afanasyeva 2007).

The COPIUS Orthographic Toolset's Mansi module aims to alleviate difficulties that arise due to this complex situation by allowing automatic rule-based transcriptions between these writing systems, as best allowed by their respective limitations. The following sections will document and justify the design choices made in its creation, prompted by the circumstances detailed above. A systematic overview of the variation and discrepancies between the myriad Mansi writing systems is essential for future enterprises hoping to subsume diverse Mansi materials in one infrastructure.

2 (Northern) Mansi phonology

Contemporary accounts (e.g. Rombandeeva 1973, Kálmán 1976a & 1976b, Riese & Bradley, Skribnik & Afanasyeva 2007, Bakró-Nagy, Sipőcz & Skribnik: to appear) distinguish the following phonemes in Northern Mansi (disregarding sounds only found in newer Russian loan words), provided here with their IPA values and disregarding differences between the first syllable and non-first syllable vocalism:

Table 1. Consonants

		Bilabial	Dento-alveolar	Alveolo-palatal	Palatal	Velar	Labio-velar
Stop	Voiceless	p	t	tʲ		k	kʷ
	Voiced						
Fricative	Voiceless		s	sʲ		x	xʷ
	Voiced	β[8]			j	ɣ	
Affricate							
Nasal			m	n	nʲ		ŋ
Lateral				l	lʲ		
Trill							
Approximant							

Table 2. Vowels

	Front	Central	Back
Close	iː, i		uː, u
Mid	eː, e	ə	oː, o
Open		aː, a	

The writing systems we are considering differ in one important aspect already mentioned above: while all Latin-based alphabets have separate characters for four alveolo-palatal consonants, the Cyrillic versions use the Russian “syllable principle” and mark alveolo-palatals by using a iotated letter for the following vowel.

Table 3 illustrates how Mansi phonemes are prototypically rendered in the writing systems under consideration. Variants given in parentheses are used in older versions of the script or occur sporadically. Subsequently, some finer points of the individual writing systems will be discussed.

Table 3. Prototypical rendering of vowels in the writing systems

IPA	FUT	UNA	Cyrillic
i	i	i	ы, и
iː	ī	i	ӣ (и), ы̄ (ы)
e	e	e	э, е
eː	ē	e	э̄ (э), е̄ (е)
u	u	u	у, ю
uː	ū	u	ӯ (у), ю̄ (ю)
o	o	o	о, ё
oː	ō	o	о̄ (о), ё̄ (ё)
a	a	a	а, я
aː	ā	a	а̄ (а), я̄ (я)
ə	ə	ь (i)	ы, и, а, у, я, ю

[8] Often also transcribed as /w/

Table 4. Prototypical rendering of vowels in the writing systems

IPA	FUT	UNA	Cyrillic
p	p	p	п
t	t	t	т
t^j	t’	ţ	ть, т + iotated vowel
k	k	k	к
k^w	k^w	kv	кв
β	w[9]	v (u)	в
j	j	j	й, iotated vowel
x	χ	h	х
ɣ	γ	ɧ	г
x^w	χ^w	hv	хв
s	s	s	с
s^j	ś	ꞩ	сь, с + iotated vowel (щ)
l	l	l	л
l^j	l’	ļ	ль, л + iotated vowel
m	m	m	м
n	n	n	н
n^j	ń	ņ	нь, н + iotated vowel
ŋ	η	ŋ	ӈ (нг)

2.1 Vowel length

As can be seen in this overview, vowel length is not indicated in UNA or in older Cyrillic texts (see the forms in parentheses in Table 3). Thus, a rule-based transcription cannot produce reliable results when transcribing from these writing systems into IPA or FUT as this phonologically relevant information is simply missing.

2.2 Realization of alveolo-palatals

Alveolo-palatal consonants are problematic both in FUT and in IPA. IPA differentiates between palatalized and palatal consonants: /s^j/, /t^j/, /n^j/, /l^j/ vs. /ɕ/, /c/, /ɲ/, /ʎ/. There are however no distinct forms for alveolo-palatal consonants that fall into neither category. Thus, the transcription usually used for palatalized sounds must be used. FUT does not distinguish between palatal, alveolo-palatal, and palatalized consonants at all; all of these forms are indicated with an accent: /ś/, /t’/, /ń/, /l’/. As Mansi only has alveolo-palatal consonants, the defects of the transcription systems do not cause any problems language-internally. In UNA they are indicated with distinct letters <ꞩ>, <ţ>, <ņ>, <ļ>. Cyrillic orthography, as discussed above, works with iotated vowel symbols (and if necessary the soft sign <ь>), although older versions sometimes used <щ> for alveopalatal /s^j/.

[9] Sometimes also encountered as /β/

2.3 Realization of /ə/

The reduced vowel /ə/ does not occur in the first syllable; in the non-first syllable its realization is altered by adjacent sounds. Before labial consonants (/m/, /p/, /w/) it is realized as (full or reduced) [u]; in the environment of alveolo-palatal consonants it is pronounced as [i]. In IPA, these vowels are generally rendered in their “shifted” non-phonemic values /u/, /i/, while in FUT transcription both the shifted values and the original value /ə/ can be encountered. In Cyrllic, <у> is used before labial consonants; the choice of <и> or <ы> in the second case is determined by the preceding consonant. In UNA, it seems like <u> and <i> are used (the latter both before and after alveolo-palatal consonants), but we do not have sufficient data to determine if the handling of these cases is systematic.

Table 5. Rendering of /ə/

IPA	FUT	UNA	Cyrillic	Translation
maːxum	māχəm	mahum	мāхум	‘people’
aːmisʲ	āməś	amiꞩ	āмысь	‘riddle’
piːɣrisʲit	pīγriśət	pьriꞩit[10]	пы̄грисит	‘boys’

Note that in rapid speech the sound /ə/ is frequently omitted. Modern Cyrillic texts generally do not show these omissions as they are heavily edited and corrected. In fieldwork materials on the other hand as well as texts in the less standardized UNA oftentimes variants without /ə/ can be found, e.g. мā̄гыс ‘for’ ~ UNA <maђьs/ (Ƶuļov 1933: 1), but минэ̄гыт ~ UNA <mineђt/ (ibid: 3).

2.4 The sound combinations /iɣ/, /əɣ/

Allophony affects the vowels /i/, /ə/ when they precede the voiced velar fricative /ɣ/: in this environment, they are realized further back (Kálmán 1976a: 19–20), ~IPA [ɨ] (Riese & Bradley 2020: 20; transcribed as [i̯] in Rombandeeva 1973: 20). This non-phonemic process is not generally indicated in IPA and only sporadically in FUT (e.g. as [i̯] in Kálmán 1976b). In UNA, both /iɣ/ and /əɣ/ can be found rendered as <ьђ>.

In Cyrillic, this combination can be reflected as either <ыг> or <иг>. As discussed above, the primary function of the choice between <и> and <ы> is to differentiate between alveolo-palatal (/s/, /t/, /n/, /l/) and dento-alveolar consonants (/sʲ/, /tʲ/, /nʲ/, /lʲ/). The question arises how /i/, /iː/, /ə/ are rendered after all other consonants with and without a following /ɣ/ . It seems there are no strict rules here; choices seem to be governed by preferences of individual authors. The general tendencies are as follows:

[10] Variant lacking the fricative /ɣ/

• alveolo-palatal /sʲ/, /tʲ/, /nʲ/, /lʲ/	+ /i/, /ə/		>	<и>
• non-alveolo-palatal /s/, /t/, /n/, /l/	+ /i/, /ə/		>	<ы>
• other non-alveolo-palatals	+ /i/, /ə/		>	<и>, <ы>
• other non-alveolo-palatals	+ /i/, /ə/	+ /ɣ/	>	<ы>
• alveolo-palatal /sʲ/, /tʲ/, /nʲ/, /lʲ/	+ /i/, /ə/	+ /ɣ/	>	<и>

For example, /wit/ <вит> ‘water’, but /wiγər/ выгыр ‘red’; /pil/ <пил> ‘berry’, but /piːɣ/ <пы̄г> ‘boy’; /miŋkʷe/ <миӈкве> ‘to go’ but /am miɣum/ <ам мыгум> ‘I will go’; /kit/ <кит> ‘two’, but /kiɣsʲi/ <кыгси> ‘elder brother’. However, /i/ or /ə/ between the alveolo-palatal and /ɣ/ is rendered as <и> – that is to say, the need for <и> as a marker of alveolo-palatal articulation is stronger than the tendency towards <ы> before /γ/: /anʲiγlaŋkʷe/ <аниглаӈкве> ‘kiss’.

2.5 Minutiae of the UNA

IPA /β/ is rendered as <v> or as <u> in UNA: IPA /βit/ ~ UNA <vit> ‘water’, IPA /oːβəl/ ~ UNA <oul> ‘beginning/end point’, IPA /taβ/ ~ UNA <tau> personal pronoun 3sg (he/she). Generally, the symbol <u> seems to be used after vowels other than /u/ while <v> is used word-initially and after /u/. The sound combination /uw/ is in some (but not all) cases rendered simply as /u/, leading to ambiguity: IPA /xanʲsʲuβlasət/ ~ UNA <haņꞩulasьt> ‘they learned’ (Ƶuļov 1933: 6), but IPA /pusmaltaŋkʷe/ ~ UNA <pusmaltankve> ‘to heal’.

3 Automatic transcription and sample texts

3.1 Mansi and Unicode

Unicode support is essential for any minority language hoping to survive in digital spheres (cf. Rueter & Ponomareva 2019). As of 2021, all characters used in all writing systems under consideration have their own Unicode code points, with the exception of most characters used for long vowels in Cyrillic. Only <ӣ> and <ӯ> have their own Unicode code points thanks to their usage in Tajik. There are currently no Unicode code points for <а̄> <е̄>, <о̄> <ы̄>, <э̄>, <ю̄> <ё̄>, <я̄>; here, a combining character must be used. (Numerous publications will also use visually identical Latin characters where they exist.)

3.2 The COPIUS Orthographic toolset

The Mansi module of the COPIUS Orthographic toolset, found at www.copius.eu/trtr.php?lang=mns (source code at www.copius.eu/ortho.php > www.copius.eu/files/copius_source.zip & https://doi.org/10.5281/zenodo.4596454), was published in March 2021. It is the first attempt at automatic transcription between the different writing systems detailed here and attempts to compensate for the inconsistencies discussed as best possible.

The software is realized as a PHP application which implements rule-based transcriptions by means of simple search-and-replace patterns. An excerpt of the function used to transcribe FUT text into UNA is shown below. Line 578 shows how FUT /^{w}/ (used in /k^{w}/ and /χ^{w}/) is replaced with UNA <v>. Where it is necessary to take the wider context of a sound into consideration for accurate transcription, the software makes use of character sets, as shown in lines 632– 642: `$vlat` includes all vowel symbols in Latin writing systems. By replacing <v> with <u> after any vowel, it is ensured that post-vocalic /w/ is realized as <u>; see above in Section 2.5. Lines 640–642 reverse this process for /u/ as the sound combination /uw/ should be realized as <uv> rather than <uu>.

functions.php (Version 1.1 from 5 March 2021)

```
574 function lat_to_una($text, $lang) // UPA/FUT > UNA
575 {
576     if ($lang == "mns")
577     {
578         $text = str_replace("ʷ","v", $text);
[…]
598         $text = str_replace("χ","h", $text);
599         $text = str_replace("X","H", $text);
[…]
632         global $gr_vlat;
633
634         for ($i = 0; $i < count($gr_vlat); $i++)
635         {
636             $text = str_replace($gr_vlat[$i]."v",$gr_vlat[$i]."u", $text);
637             $text = str_replace($gr_vlat[$i]."V",$gr_vlat[$i]."U", $text);
638         }
640         $text = str_replace("uu","uv", $text);
641         $text = str_replace("Uu","Uv", $text);
642         $text = str_replace("UU","UV", $text);
643
644     }
645     return ($text);
646 }
```

For Mansi, the software has the following transcription mechanisms:

Cyrillic	↔	FUT
FUT	↔	IPA
UNA	↔	FUT

Other transcriptions happen over an intermediary, e.g., if Cyrillic is transcribed into UNA, it is first transcribed into FUT and then from FUT into UNA.

As the software has no lexical support and does not utilize neural networks or similar technologies, it cannot compensate for unpredictable variation and orthographically relevant information omitted in UNA and older Cyrillic texts, most notably vowel length.

3.3 Sample texts

This section aims to show the software in practice and illustrate the possibilities and restrictions of the rule-based approach used here. In the following examples, the leftmost column shows the original text sample while the other columns show the output of the transcription software in its current version (March 2021).

Transcribing from Cyrillic

Table 6. Transcription from Cyrillic (text from Skribnik & Afanasyeva 2007:II: 45)

Cyrillic	> FUT	> IPA	> UNA
Та тӯйтыгпахтас, порыг турн щалтапас, тот та ӯнлы. Хāр-ōйка такос кисхаты, такос кисхаты, мāтāприщ ат хōнтытэ. Хоса кисхатас, вāти кисхатас, аквматэ̄ртн порыг тэ̄ӈкве та лю̄лыс. Порыг та тэ̄г, аквматэ̄ртн хӯнтамластэ – пуки киврэ̄т матыр та рōӈхи: – Аким-ōйка у-у-ӯв, наӈ āнум юв-тāяпаслын!	Ta tūjtiγpaχtas, poriγ turn śaltapas, tot ta ūnli. Xār-ōjka takos kisχati, takos kisχati, mātāpriś at χōntite. Xosa kisχatas, wāt'i kisχatas, akʷmatērtn poriγ tēηkʷe ta l'ūləs. Poriγ ta tēγ, akʷmatērtn χūntamlaste – puki kiwrēt matər ta rōηχi: – Akim-ōjka u-u-ūw, naη ānəm juw-tājapaslən!	ta tuːjtiɣpaxtas, poriɣ turn sʲaltapas, tot ta uːnli. xaːr-oːjka takos kisxati, takos kisxati, maːtaːprisʲ at xoːntite. xosa kisxatas, βaːtʲi kisxatas, akʷmateːrtn poriɣ teːŋkʷe ta lʲuːləs. poriɣ ta teːɣ, akʷmateːrtn xuːntamlaste – puki kiβreːt matər ta roːŋxi: – akim-oːjka u-u-uːβ, naŋ aːnum juβ-taːjapaslən!	Ta tujtiꜧpahtas, poriꜧ turn ꞩaltapas, tot ta unli. Har-ojka takos kishati, takos kishati, matapriꞩ at hontite. Hosa kishatas, vaţi kishatas, akvmatertn poriꜧ teŋkve ta ļulьs. Poriꜧ ta teꜧ, akvmatertn huntamlaste – puki kiuret matьr ta roŋhi: – Akim-ojka u-u-uv, naŋ anum juv-tajapaslьn!

We are not aware of any shortcomings in transcriptions from Cyrillic and cannot find any errors in the output here. In older Cyrillic texts however, vowel length would have been an issue.

Transcribing from UNA

Table 7. Transcription from UNA (text from Ƶuļov 1933: 6)

UNA	> FUT	> IPA	> Cyrillic
Ņavramьt skolat haņꞩulasьt. Juv-johtesьt. Artaļt man savit hul alim oli, nepaken tuv-hansuŋkve eri. Vaꞩiļ tuv-hansuŋkve vermi. Juvan aꜧmьŋ kol ujt pusmaltankve haņsulas. Tau artaļ sali aņa palt aꜧmьŋ salit pusmaļtijane.	Ńawramət skolat χańśulasət. Juw-joχtesət. Artal't man sawit χul alim oli, nepaken tuw-χansuηkʷe eri. Waśil' tuw-χansuηkʷe wermi. Juwan aγməη kol ujt pusmaltankʷe χańsulas. Taw artal' sali ańa palt aγməη salit pusmal'tijane.	nʲaβramət skolat xanʲsʲulasət. juβ-joxtesət. artalʲt man saβit xul alim oli, nepaken tuβ-xansuŋkʷe eri. βasʲilʲ tuβ-xansuŋkʷe βermi. juβan aɣməŋ kol ujt pusmaltankʷe xanʲsulas. taβ artalʲ sali anʲa palt aɣməŋ salit pusmalʲtijane.	Няврамыт сколат ханьсюласыт. Юв-ёхтэсыт. Артальт ман савит хул алым олы, нэпакен тув-хансуӈкве эри. Василь тув-хансуӈкве верми. Юван агмыӈ кол уйт пусмалтанкве ханьсулас. Тав арталь салы аня палт агмыӈ салыт пусмальтыянэ.

Here two issues are visible that cannot be solved with our current approach:

- As vowel length is not indicated in the source writing system, all vowels are transcribed as short vowels into all other writing systems.
- UNA <haņꞩulasьt> ‘they learned’ should be transcribed into FUT as / χańśuwlasət/ rather than /χańśulasət/: the sound combination /uw/ is rendered as <u> in UNA. However, in other cases the sound /u/ alone is rendered as <u>, e.g., in UNA <pusmaļtijane> ‘he/she treats them’, correctly transcribed as /pusmal'tijane/. The ambiguity of UNA does not allow for disambiguation in a purely rule-based application.

Transcribing from FUT

Table 8. Transcription from FUT (text from Riese & Bradley 2020: 85)

FUT	> Cyrillic	> IPA	> UNA
Am χūrəm oχsar alasəm. Mān ńila χāpəl jalsuw. Pīγriśət at χūlpəl χūl χūlpajasət. Akʷ χūlpən low sorəχ l’ikməs.	Ам хӯрум охсар аласум. Мāн нила хāпыл ялсув. Пы̄грисит ат хӯлпыл хӯл хӯлпаясыт. Акв хӯлпын лов сорых ликмыс.	am xuːrum oxsar alasum. maːn nʲila xaːpəl jalsuβ. piːɣrisʲit at xuːlpəl xuːl xuːlpajasət. akʷ xuːlpən loβ sorəx lʲikməs.	Am hurum ohsar alasum. Man ņila hapьl jalsuv. Piђrisьt at hulpьl hul hulpajasьt. Akv hulpьn lou sorьh ļikmьs.

We are not aware of any shortcomings in transcriptions from FUT.

Transcribing from IPA

Table 9. Transcription from IPA[11]

IPA	> Cyrillic	> FUT	> UNA
paːkʷpoːsi wojkan oːtər kaŋke jot kit xum oːleɣ. sʲaːliɣ aːxʷtas, mor aːxʷtas pitraŋ uːsəl uːnleɣ. nʲololuw soːtəre taɣlup mirəŋ uːs onʲsʲeɣ. paːkʷpoːsi wojkan oːtər teːpaŋ aːs xuːlaŋ aːs witnəl neː wis.	пāквпōсы войкан ōтыр каӈке ёт кит хум ōлэг. ся̄лыг āхвтас, мор āхвтас питраӈ ӯсыл ӯнлэг. нёлолув сōтыре таглуп мирыӈ ӯс оньсег. пāквпōсы войкан ōтыр тэ̄паӈ āс хӯлаӈ āс витныл нэ̄ вис.	pākʷpōsi wojkan ōtər kaŋke jot kit χum ōleγ. sʲāliγ āχʷtas, mor āχʷtas pitraŋ ūsəl ūnleγ. nʲololuw sōtəre taγlup mirəŋ ūs onʲsʲeγ. pākʷpōsi wojkan ōtər tēpaŋ ās χūlaŋ ās witnəl nē wis.	pakvposi vojkan otьr kaŋke jot kit hum oleђ. ꞩaliђ ahvtas, mor ahvtas pitraŋ usьl unleђ. ņololuv sotьre taђlup mirьŋ us oņꞩeђ. pakvposi vojkan otьr tepaŋ as hulaŋ as vitnьl ne vis.

We are not aware of any shortcomings in transcriptions from IPA.

[11] http://www.babel.gwi.uni-muenchen.de/index.php?abfrage=view_corpus_file_new&id_text=1138

4 Conclusions and discussion

The purely rule-based approach followed by the COPIUS Orthographic toolset seems highly effective when transcribing from FUT, IPA, or modern Cyrillic into any other writing systems. We are not currently aware of any failings (as the software was under development when this paper was written, any shortcomings found in testing the application were immediately remedied). Cyrillic output in particular might not in all cases mirror forms found in reality, but due to the weak standardization of the Mansi Cyrillic orthography, we believe the created forms can be considered legitimate. It is quite possible that there are fringe cases that we have currently not considered or encountered. If users encounter these, they can report them through a feedback system directly integrated into the software's website; feedback submitted through it is immediately e-mailed to the developers.

When the source text is in UNA or in older Cyrillic, however, a purely rule-based approach is not fully satisfactory: with these writing systems oftentimes not conveying phonologically relevant information (most notably vowel length), lexical support and/or statistical methods would be needed to ensure more reliable and adequate output.

References

1. Azbuka 1903 = Азбука для вогулъ приуральскихъ, составленная Епископомъ Никаноромъ. Москва. (online: https://vivaldi.nlr.ru/bx000000775/view/?#page=1)
2. Bakró-Nagy, Marianne, Katalin Sipőcz & Elena Skribnik (to appear). 'North Mansi', in The Oxford Guide to the Uralic Languages. Oxford: Oxford University Press (in preparation).
3. Balandin 1958 = Баландин А.Н. Основные правила произношения и правописания мансийского языка. Л., 1958.
4. Burykin 2000 = Бурыкин А. А. Изучение фонетики языков малочисленных народов Севера России и проблемы развития их письменности (обзор) // Язык и речевая деятельность. Т. 3 ч. 1. СПб., 2000. 150—180.
5. Kálmán, Béla (1976a). Wogulische Texte mit einem Glossar. Budapest: Akadémiai Kiadó.
6. Kálmán, Béla (1976b). Chrestomatia Vogulica. Budapest: Tánkönyv Kiadó.
7. Kannisto, A. / Liimola, M. Wogulische Volksdichtung. Vol.I (MSFOu 101, 1951), II (MFSOu 109, 1955), III (MFSOu 111, 1956), IV (MFSOu 114, 1958), V (MFSOu 116, 1959), VI (MFSOu 134, 1963).
8. Munkácsi, Bernát (1892–1963). Vogul népköltési gyüjtemény. 8 Bde., Bd. 7–8, hrsg. von B. Kálmán.
9. Munkácsi, B., Kálmán, B. 1986. Wogulisches Wörterbuch. Budapest.
10. Normanskaja, Kosheliuk 2020 = Норманская, Ю.В., Кошелюк, Н.А. Неопубликованный мансийский словарь П. С. Палласа — ранее неизвестный мансийский диалект? Урало-алтайские исследования. 2020. No 1 (36). С. 92—100.
11. Pallas 1787–1789 = Паллас, П.С. Сравнительные словари всѣхъ языковъ и нарѣчій, собранные десницею всевысочайшей особы (В двух томах). Санкт-Петербург.
12. Parfenova 2003 = Парфенова, О. С. Мансийский язык. In: Письменные языки мира: Языки Российской Федерации. — М.: Academia, 2003. — Т. 2. 287–306

13. Riese, Timothy & Jeremy Bradley (eds) 2020. A. N. Balandins Einführung in das Mansische. Vienna: University of Vienna/COPIUS [published online at www.copius.eu]
14. Rombandeeva 1982 = Ромбандеева Е.И. К вопросу об усовершенствовании алфавита, графики и орфографии. Опыт совершенствования алфавитов и орфографий языков народов СССР. М.: Наука, 1982. С. 176-181.
15. Rombandeeva 1985 = Ромбандеева Е.И. Вопросы совершенствования алфавита, графики и орфографии мансийского языка. Просвещение на Крайнем Севере. Вып. 22. Л., 1985. С. 3-8.
16. Rombandeeva 1996 = Ромбандеева Е.И. Фонетические и грамматические процессы в современном мансийском языке. Москва: Икар.
17. Rombandeeva 1973 = Ромбандеева Е.И. Мансийский (вогульский) язык. Москва: Наука.
18. Rueter, J., Ponomareva, L.: Komi latin letters, degrees of unicode facilitation. In: Proceedings of the Language Technologies for All (LT4All) (2019)
19. Setälä, E. N. (1901). Über transskription der finnisch-ugrischen sprachen. Finnisch-ugrische Forschungen 1. Helsingfors, Leipzig. 15–52.
20. Siegl, Florian, Rießler, Michael. Uneven steps to literacy. In: Cultural and linguistic minorities in the Russian Federation and the European Union. Springer, 2015. 189–230.
21. Skribnik & Afanasyeva 2007 = Скрибник, Е.К., Афанасьева, К.В. Практический курс мансийского языка. Ч. 1–2. Ханты-Мансийск: Полиграфист.
22. Skribnik, Elena and Natalya Koshkareva (1996). ‘Khanty and Mansi: the contemporary linguistic situation’, in Juha Pentikäinen (ed.), Shamanism and Northern Ecology. (Religion and Society 36.) Mouton De Gruyter: Berlin - New York. 207–217.
23. Zhornik 2019 = Жорник Д.О. ‘Мансийский рассказ о медвежьем празднике’, Родной язык, 2019/2, 100–127.
24. Ƶuḷov, P. N. 1933. Lovintan maɧьs lovintanut: oul lomt: oul hanisⱦan tal maɧьs. Fenno-Ugrica collection. The National Library of Finland http://urn.fi/URN:NBN:fi-fe2014060426213. https://fennougrica.kansalliskirjasto.fi/handle/10024/67112