LINGUISTICA URALICA LIX 2023 4 https://dx.doi.org/10.3176/lu.2023.4.02

JEREMY BRADLEY (Vienna), *ROGIER BLOKLAND* (Uppsala)

MANSI ET AL. IN PRINT BEFORE AND UNDER UNICODE

Abstract. The Unicode Standard, in its various iterations, aims to provide and guarantee consistent, transparent and reliable encoding for the totality of all codified human writing systems. How successful these aims have been in practice greatly depends on the support individual languages and varieties enjoy in politics and infrastructures. This contribution looks at the realities of the digital and print realization of Uralic minority languages, especially Mansi, in the past and today. Based on interviews and the study of surviving digital files, it aims to make knowledge held by relevant scholars accessible to a broader audience.

Keywords: Mansi, history of literacy, Unicode.

1. Uralic languages and Unicode

The Unicode Consortium,[1] established in 1991, has since its inception aimed to remove erstwhile borders between the digital realization of different writing systems. In 2008, UTF-8 Unicode became the most-widely used character encoding globally; today, it dominates over all other encoding schemes (Davis 2012). From the user's perspective, the principle of Unicode in its various iterations is simple: every character (including diacritic-marked characters) used somewhere in at least one established human writing system should have a code point of its own. This code point, which is assigned a definition and has a four-digit hexadecimal designation (U+XXXX), unambiguously refers to this symbol. For example, the Udmurt language — which has official status in the Republic of Udmurtia, a subject of the Russian Federation — uniquely uses the letter ӵ to denote a voiceless palato-alveolar affricate /tʃ/, and thus this grapheme has since 1993[2] been represented in Unicode, in upper-case and lower-case:

Ӵ CYRILLIC CAPITAL LETTER CHE WITH DIAERESIS U+04F4
ӵ CYRILLIC SMALL LETTER CHE WITH DIAERESIS U+04F5

[1] unicode.org.
[2] Technical specifications and the history of individual Unicode code points can be accessed at www.compart.com/en/unicode.

Received 22 June 2023, accepted 9 October 2023, available online 10 December 2023.

Any problems pertaining to the creation and realization of this symbol in a Unicode-using system pertain to the input side (i.e., users do not have the relevant symbol on their keyboard layout) or the font (e.g., Comic Sans was not designed for Udmurt and cannot render ӵ correctly).

In contrast, no writing system (that we know of) utilizes the letter г̈. It is, however, still possible to create this symbol (albeit it in a manner more susceptible for typographical complications) in Unicode using so-called combining characters through which a specific diacritic mark is added to a character. Specifically, the symbol here was realized as a combination of:

г	CYRILLIC SMALL LETTER GHE	U+0413
◌̈[3]	COMBINING DIARESIS	U+0308

It should not be necessary to use combining characters when rendering established writing systems, but reality can be more complex than this. First of all, it is debatable what an ”established writing system” even is. For many minoritized languages, the codification of the literary norm is weak and contentious; the pool of users of a literary norm is highly restricted. In what follows we will consider the written realization of literary languages as standardized in the COPIUS Transcription & orthography toolset (www.copius.eu/ortho); deviations from the norms set here will be discussed below.

It requires some sort of visibility for a writing system to be covered by Unicode — the Unicode Consortium cannot take writing systems it does not know about into consideration. If this visibility is not a given, lobbying might be required. This is exactly what happened with respect to Kildin Saami, a minoritized language of North-western Russia lacking the same official status that Udmurt has: only through activism by well-established researchers (Trond Trosterud, Michael Everson, Rimma Kuruch) did the following letters used in the Kildin Saami orthography receive Unicode code points in 1999 (cf. Rießler 2013 : 202):

Ҍ	CYRILLIC CAPITAL LETTER SEMISOFT SIGN	U+048C
ҍ	CYRILLIC SMALL LETTER SEMISOFT SIGN	U+048D
Ӭ	CYRILLIC CAPITAL LETTER E WITH DIAERESIS	U+04EC
ӭ	CYRILLIC SMALL LETTER E WITH DIAERESIS	U+04ED

Already Version 1.1 of Unicode,[4] published in 1993, included the necessary special characters to cover both Uralic national languages (Finnish, Hungarian, Estonian) as well as Uralic ”larger small languages” (cf. Riese, Bradley 2011 : 210) that tend to enjoy some kind of official status, irrespective of the base orthography used in a writing system (Latin, Cyrillic), for example ő (Hungarian), ҥ (Meadow Mari), ӹ (Hill Mari), ŧ (North Saami). Later updates would introduce further Uralic characters (for all of these characters, the corresponding capital letter was introduced as well) — characters not indicated here or below were already covered by Version 1.1 in 1993 at the latest:

- 1999: ȯ, ō, ȱ, ț (Livonian), ҍ (Kildin Saami), ӭ (Kildin Saami and Selkup)
- 2002: ӆ (Khanty)
- 2005: ҙ (Selkup)
- 2006: ԑ (Enets)

[3] Combining characters are conventionally displayed on a dotted circle ◌ to show their exact placement.
[4] www.unicode.org/versions/Unicode1.1.0/appI.pdf.

Sometimes, "successes" of language in Unicode can be coincidental: when *ӟ* was introduced to Unicode in 1999 thanks to Kildin Saami-related lobbying, Selkup, which uses this grapheme as well, also "profited". It can be assumed that the 2005 introduction of *ҷ* was more connected to the usage of this grapheme in comparatively structurally strong languages and that the usage of this character in Selkup was not a motivating factor. One character used in Selkup that is not used in other politically more salient languages, *и̇* (suggested in Быконя, Ким, Купер 1994), is still today[5] not covered by Unicode, and must be realized using a combining character. The contentious nature of writing systems for smaller languages naturally complicates the picture: often, orthographic suggestions or even orthographic standards can have short shelf lives and it can be difficult to determine what truly is part of a literary standard. For example, in the case of Võro the grapheme *ô̭* was put forth by Ain Kaalep in 1989 for the raised mid vowel /ɨ/ (cf. Russian *ы* or Polish *y*), but today the grapheme *õ* has become the norm for this sound (Koreinik, Plado 2022 : 323) — thus, the lacking Unicode code point for this grapheme, now looking back, is justifiable. Likewise for Nganasan, Žovnitskaja-Turdagina (Жовницкая-Турдагина 1999) uses *ǯ* (which does not have a Unicode code point) for the voiced dental fricative /ð/ in her primer, but this symbol does not seem to have found usage outside of her work[6] — there is a legitimate debate to be had in such cases when one should advocate for the inclusion of a character in Unicode. In the 2013 Khanty orthographic reform, a ligature consisting of *т* and *ь* was introduced to mark /tʲ/, following the logic of the Serbian Cyrillic alphabet (where the ligatures *љ* ← *л* + *ь* and *њ* ← *н* + *ь* are used — both these graphemes are part of the new Khanty orthography as well). From a technical perspective this solution is doubly problematic: this character is not covered by Unicode, and unlike diacritic-marked characters, there is no Unicode-conform way to create custom ligatures (cf. Skribnik, Laakso 2022 : 97, 100). It is possible to design a font in a manner that specific combinations of letters are always displayed as ligatures (this is the principle used by fonts that imitate handwriting in which individual characters are joined), thus one could design a font in which the letter combinations *ль*, *нь*, and *ть* are always displayed as ligatures. However, this font would do so irrespective the language used and would display these ligatures also in Russian text or text passages, for example rendering the word *цель* 'goal' as *цељ*, making this prospective solution less than optimal. Should the *т* + *ь* ligature remain in usage in Khanty, advocacy for including this font in a future Unicode release would therefore be appropriate.

Following the most recent and reliable orthographic descriptions of Uralic literary standards, the following gaps can be said to persist:

- Beserman (cf. Pischlöger 2022 : 361): *ӳ*
- Mansi (cf. Bradley, Skribnik 2021): *а̄*, *е̄*, *ё̄*, *о̄*, *ы̄*, *э̄*, *ю̄*, *я̄*
- Khanty (cf. Skribnik, Laakso 2022 : 97): *ӗ*, *ӯ*, *ӑ*, *т* + *ь* ligature
- Kildin Saami (cf. Rießler 2013): *а̄*, *е̄*, *ё̄*, *о̄*, *э̄*, *ю̄*, *я̄*
- Selkup (cf. Быконя, Ким, Купер 1994): *и̇*

[5] This manuscript was last edited on 25.10.2023.
[6] Beáta Wagner-Nagy, personal correspondence.

Despite the Kildin Saami-related lobbying discussed above, Kildin Saami is still not adequately covered by Unicode: the indication of (only marginally phonemic, see Rießler 2013 : 203—204) vowel length through a macron can generally only be realized through a combining character. Only for *ӯ* and *ӯ* are there Unicode code points, thanks to the usage of these graphemes in Tajik Cyrillic. According to involved scholars,[7] there was little drive towards the inclusion of these letters from within the (speaker and scholarly) community and little interest from the Unicode Consortium for the inclusion of further Cyrillic variants at the time.

Mansi has, since the 1979 orthographic reform (Bradley, Skribnik 2021 : 3) marked unambiguously phonemic vowel length with macrons as well; as these variants were not included in Unicode at the time, it remains impossible to render the contemporary Mansi orthography without combining characters. This makes the Mansi orthography the well-established Uralic orthography with the most gaps in its Unicode coverage.

A further point in which the Unicode mission of providing unambiguous encoding for all established human writing systems runs contrary to lived experiences within Finno-Ugric/Uralic studies is that much of the time we are not working with established orthographies, but with more or less standardized transcriptions. Finno-Ugric Transcription (FUT) / The Uralic Phonetic Alphabet (UPA), the standard transcription system in Uralic studies, is today adequately covered by Unicode. For example, *ᴞ* was included in Unicode in 2003; different diacritics commonly used in it can be realized using combining characters such as ◌͕ (U+0355 COMBINING RIGHT ARROWHEAD BELOW), introduced in 2003 as well and allowing the creation of combinations such as *ū͕*. There is some inconsistency if one needs to use multiple diacritic marks: there is one combining character for both a right arrowhead and up arrowhead under a character, ◌͖ (U+0356 COMBINING RIGHT ARROWHEAD AND UP ARROWHEAD BELOW), but there is no corresponding combining character for other combinations of arrows. Thus, *u͖* can be produced using only one combining character, but *ü͔̬* can only be created using two distinct combining characters (U+0354 COMBINING LEFT ARROWHEAD BELOW and U+032C COMBINING CARON BELOW) — which many fonts will place on different vertical levels. Thus, for the sake of graphic consistency, it is best to avoid using ◌͖ completely if one also needs to use other combinations of diacritic marks. This curiosity within Unicode presumably came about due to conventional diacritic combinations not being included in the relevant proposal at the time, and the Unicode Consortium afterwards becoming increasingly uneager to introduce new code points for precomposed characters used only in transcription systems.[8]

Furthermore, individual resources only adhere to the abstract standard of FUT to a varying degree; oftentimes, they predate the establishment of this standard (Setälä 1901). This is particularly self-evident in the case of Mansi, where the eminent scholars Artturi Kannisto (1874—1943) and Bernát Munkácsi (1860—1937) used distinct notoriously idiosyncratic and diacritic-laden writing systems in their works (cf. Riese, Bradley 2020 : 14; cf. Stachowski 2011 : 308 for a more general critique of the intense usage of diacritics by Finnish scholars

[7] Michael Rießler, personal correspondence.
[8] Niko Partanen, personal correspondence.

in particular).[9] With an ample amount of patience, combining characters, and some flexibility as regards the optical realization of different graphemes, a Unicode-compatible digitization of these sources is possible.

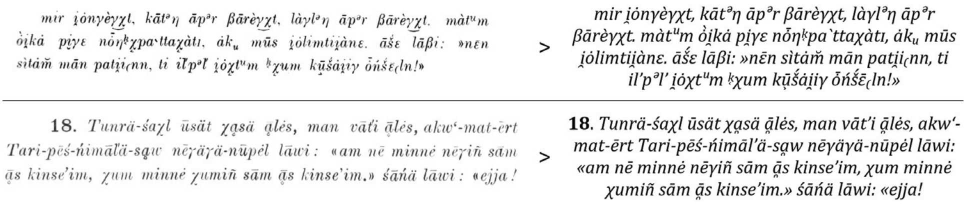

Figure 1. **Text passages by Kannisto and Liimola (Wogulische Volksdichtung 1956 : 15) and Munkácsi (1892—1902 : 21), scanned (left) and rendered in Unicode (right).**

The text shown in Figure 1 was transcribed by hand (or rather, largely using the COPIUS Transcription & orthography toolset at www.copius.eu/ortho), but modern technology makes the (semi-)automatic digitization of such sources viable: artificial intelligence-supported text recognition software such as Transkribus[10] allow for the relatively quick and efficient digitization of texts irrespective of the writing system used (see Partanen, Rießler 2019), as illustrated in Figure 2, where the upper half of the image shows the scanned text from the source (with computer-recognized polygons identifying the text to be digitized) and the lower half showing the recognized texts (with numbers on the left identifying individual lines).

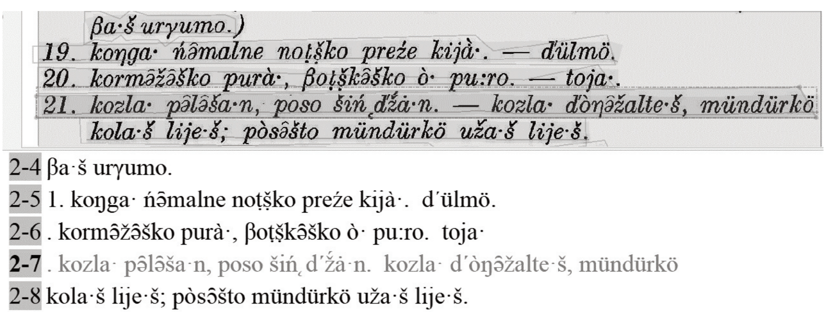

Figure 2. **Digitization of Finno-Ugric transcription (Beke 1961 : 12) with Transkribus on the example of Mari.[11]**

[9] In Finland exaggerated phonetic transcription was occasionally poked fun at as a *fransskriptio*, after the phonetician Frans Äimä, who himself would jokingly say "En minä osaa ääntää, mutta kyllä minä osaan merkitä!" ('Maybe I can't pronounce it, but I can write it down!'), but it was also mocked in 1930s Hungarian linguistic circles as *a finn betegség* 'the Finnish disease' (Kettunen 1939 : 272). Indeed, already at the turn of the previous century Munkácsi (1901 : 227) voiced his displeasure when discussing Setälä's transcription proposal: "[–––] thatsächlich tauchen seit fast einem Jahrzehnt in den finnländischen Mittheilungen die neufabrizirten Buchstaben für die schon längst bekannten und durch geeignete Schriftzeichen unterschiedenen Sprachlaute in solcher Menge auf, dass selbst die Fachleute sich nur schon mit grosser Mühe darin orientiren können." Hungarian dissatisfaction with the transcriptions used by the Finns, however, can be dated back at least 127 years: Móricz Szilasi (1896 : 494), in his review of Wichmann's "Wotjakische Sprachproben", plaintively notes the following: "Ugyanis a finnek most újabban a mienkétől sokszor egészen külömböző betűket használnak, a miből csak zűrzavar támadhat." ('The Finns have recently started using letters that are often quite different from ours, which can only lead to confusion.').

[10] readcoop.eu/transkribus.

[11] Graphic provided by Alexandre Arkhipov, based on work carried out at the REMODUS Winter School 2023 (remodus.univie.ac.at/teaching-events/winter-school-2023).

2. Before Unicode: Garage fonts

Prior to the advent of Unicode — or rather, before Unicode became broadly adopted and a ubiquitous standard — digital infrastructures made use of character encodings designed to cover specific orthographies (e.g., Windows-1253 for Greek), or clusters of orthographies (e.g., Windows-1251 for Slavic Cyrillic). Often it would be a matter of luck if a language could be adequately covered within a character encoding: for example, it was possible to mostly render Estonian appropriately (excluding *š* and *ž*) within the Windows-1252 character encoding designed for Western European languages, as the four diacritics-marked graphemes commonly used in the Estonian alphabet, *ä*, *ö*, *ü*, and *õ*, are all used in major Western European languages: the first three among others in German, the last in Portuguese.[12]

When an orthography was not covered by an established character encoding, creativity was mandated. Speakers and scholars working with Uralic languages, be it in transcription or in orthography, broadly used home-made solutions to have their languages look adequate in print. The basic principle used is the same one historically used by so-called dingbat fonts such as Wingdings: in that font, *a* would be graphically realized as ♋, *N* would be graphically realized as ☠, etc., allowing users to make use of functional equivalents of emojis long before these were codified as a principle (or codified in Unicode). Likewise, Mari speakers wishing to realize Mari, with its graphemes not used by any Slavic language (In Meadow Mari *ӧ*, *ӱ*, *ҥ*, in Hill Mari *ӓ*, *ӧ*, *ӱ*, *ӹ*), made wide use of a font named Mari-Pragmatica:[13] in this font, graphemes used in the Belarusian and/or Ukrainian alphabets (but not Russian) as well as comparatively rarely used signs were graphically realized as the Mari graphemes not found in the Russian alphabet: *є* → *ӧ*, *ў* → *ӱ*, *‰* → *ҥ*, *і* → *ӓ*, *ї* → *ӹ*, *Є* → *Ӧ*, ™ → *Ӱ*, ^ → *Ҥ*, *І* → *Ӓ*, *Ї* → *Ӹ*.[14] This allowed users to, at least graphically (i.e. in print and on computers on which this font was installed, see Figure 3), adequately represent Mari using a character encoding not designed for Mari, notably Windows-1251, the character encoding designed for Slavic Cyrillic alphabets.

Пытартыш коло ий жапыште лийше событий-влак тӱрлӧ калык-влакын самосознанийыштым нӧлталыныт да ятыр национальный общественный движений-влаклан шочаш негызым ыштеныт. Тӱрлӧ кундемлаште ильше марий-влак коклаштат ончыл шонымашан ен-влак «Марий ушем» организаций-влакым ыштеныт. Варажым тӱрлӧ кундемласе марий-влакым иктыш ушышашлан Йошкар-Олаште «Мер канаш» Межрегиональный общественный движений ышталтын. Кызыт марий калыкым тургыжландарыше тӱн йодыш-влак марий «Мер канашыште» ончалтыт. «Мер канашым» руш калыкын Тӱнямбал Соборжо дене, але пошкырт калыкын Тӱнямбал Курултайже дене танастараш лиеш. Тудын пытартыш погынжо (съезд) 2004 ий 4 мартыште эртен.

Пытартыш коло ий жапыште лийше событий-влак тӱрлє калык-влакын самосознанийыштым нєлталыныт да ятыр национальный общественный движений-влаклан шочаш негызым ыштеныт. Тӱрлє кундемлаште ильше марий-влак коклаштат ончыл шонымашан є‰-влак «Марий ушем» организаций-влакым ыштеныт. Варажым тӱрлє кундемласе марий-влакым иктыш ушышашлан Йошкар-Олаште «Мер ка‰аш» Межрегиональный общественный движений ышталтын. Кызыт марий калыкым тургыжландарыше т‰ йодыш-влак марий «Мер ка‰ашыште» ончалтыт. «Мер ка‰ашым» руш калыкын Тӱнямбал Соборжо дене, але пошкырт калыкын Тӱнямбал Курултайже дене та‰астараш лиеш. Тудын пытартыш погынжо (съезд) 2004 ий 4 мартыште эртен.

Figure 3. **When Mari-Pragmatica is installed (left), when it is not installed (right).**[15]

The same approach was followed by scholarly communities wishing to realize transcription in an era when it was not supported by existing character encodings at all. Fonts that appeared in this era follow two different principles:

[12] In the pre-digital era, Estonian publishers in the diaspora would sometimes use the at the time more accessible *ô* in place of *õ*.
[13] See tech.mari-language.com for a dated (composed in 2010), but mostly still relevant, user-oriented guide to problems encountered in the digital encoding of Mari.
[14] For this particular font, the COPIUS Transcription & orthography toolset for Mari (www.copius.eu/trtr.php?lang=mhr) can make text Unicode-conform, if one chooses "Cyrillic → Cyrillic" as the transcription direction: *тӱ‰жє* → *тӱҥжӧ*.
[15] mari-language.univie.ac.at/tech/trouble_en.html.

some were designed to be used on their own (the same principle used by Mari-Pragmatica detailed above), and some were designed to be used in combination with established fonts (e.g., Times). Fonts from the latter category were only used for individual characters representing symbols not covered by the base font. This approach had the advantage that it allowed scholars to use a much wider range of symbols, but the disadvantage that it greatly complicated the writing and editorial processes.[16]

In Finland in the 1980s, Mikko Korhonen created a dot matrix printer font for home usage (i.e., for the dot matrix printer in his home)[17] while Tapani Salminen pioneered the creation of FUT and Tundra Nenets fonts for Macintosh. The Salminen fonts were compatible with laser printers and produced aesthetically pleasing results in print, as shown in Figure 4.

(3.4) N *am naŋėn aśermėl pōl'ilėm* (VNGy I: 164)
I (s) you (o) cold-INSTR freeze(tr.)-OBJC.SG1SG
'I'll freeze you with a frost'

(3.5) N *taw pōl'uŋkwė ta pats* (VNGy I: 24)
he (s) freeze(intr)-INF so start-SBJC3SG
'he started to feel cold'

(4) N *pēlamlawe-* 'catch fire' (1, E 2 = 3 sentences), e.g.

(4.1) LO *ak$_o$ mūs janitētəl kol lap-pēlamlaws* (WV II: 168)
one alike totally house (s) closed-catch-fire-PS3SG
'suddenly the house was totally in flames'

Figure 4. **Mansi data, laser-printed using Tapani Salminen's fonts (Kulonen 1989 : 136).**

Later, Klaas Ruppel would spearhead the creation (for Macintosh) of the custom FUT font Ajatar for the creation of the Finnish etymological dictionary SSA;[18] this font became the industry standard in Finland for a time (it was widely used by the Finno-Ugrian Society and also by the journal Virittäjä) and was also used in the creation of other resources such as the "Tscheremissisches Wörterbuch" (Moisio, Saarinen 2008) and "Wogulisches Wörterbuch" (2013).[19] In Ajatar, which was used in combination with Times, individual characters are used either to represent symbols (e.g., $q \rightarrow ə$) or diacritics (e.g., *V* represents a haček on the following letter, $Vs \rightarrow š$). Scholars working with the font tended to have macros defined allowing them easy access to the necessary symbols and diacritics.

In 1991 Juhani Lehtiranta, perhaps best known in our circles as the author of the "Yhteissaamelainen sanasto" (Lehtiranta 1989) but also very active as a font developer, created the font Fluralic specifically with the aim of supporting the Uralic Phonetic Alphabet. It was used by the present journal from 1993 to the early 2000s and is used to some extent even today; it was also occasionally used by the Estonian journal Keel ja Kirjandus.[20] In Fluralic, the diacritics

[16] Jack Rueter, personal correspondence.
[17] Ulla-Maija Forsberg (Kulonen), personal correspondence.
[18] Johanna Laakso and Ulla-Maija Forsberg, personal correspondence.
[19] The authors of this paper were provided with source files for both of these dictionaries.
[20] Väino Klaus, personal correspondence.

are entered before the main character; in this way a character with more than one diacritic both under and above it can easily be created. Based on the OpenType font JLOT-Fluralic ordered in 2004—2005 by the University of Tokyo, Lehtiranta further developed Fluralic; it is now known as Uvallanne.[21]

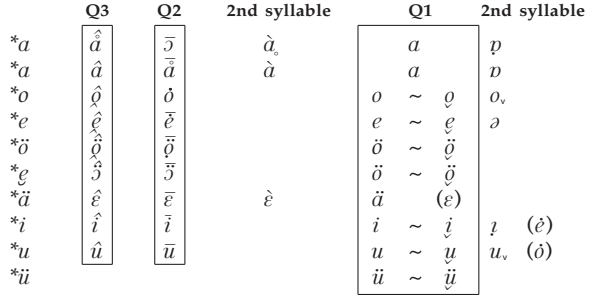

Figure 5. **Vowels of Hiiumaa Estonian dialects using Fluralic (Viitso 2005 : 17).**

Compatibility issues were commonplace at the time: various solutions that were created for Macintosh did not work on PC and vice versa. Sometimes, scholars seemed unaware of such issues, as exemplified by an article by Jorma Koivulehto (2001) where the intended **kečä* appeared as **keVcä* — mirroring the background implementation of hačeks in Ajatar — in print, among other typographical issues. When the Hungarian linguist Márta Csepregi,[22] after spending years as a Hungarian lecturer in generally Macintosh-using Helsinki in the 1980s, returned to generally PC-using Hungary (where László Honti had initiated the creation of a wide range of fonts for Uralic studies) in 1990, she continued using her Macintosh computer to compose her Surgut Khanty chrestomathy. The PC-using Szeged-based publishing house could not use her document and the entire manuscript was retyped on a PC before its eventual publication (Csepregi 1998).

Eventually, True Type fonts (.ttf fonts) — usable on both PCs and Macintosh — became the standard, and it became standard practice for editors and conference organizers to require contributors to send .ttf versions of their fonts along with the Word and PDF versions of their papers and abstracts. The organizers of Congressus IX Internationalis Fenno-Ugristarum, held in Tartu in the summer of 2000, reportedly had to deal with 90 different fonts for the proceedings (CIFU IX).[23] There were also attempts to simplify the phonological transcription of Uralic languages to alleviate character encoding and font issues (e.g., Kortesharju 1999) — analogous to the SAMPA/X-SAMPA writing systems popular at the time for rendering the International Phonetic Alphabet using ASCII characters only. Likewise, there were also idealistic suggestions to introduce low-diacritic Latin-based orthographies for the Uralic languages of Russia that avoid "unpractical" diacritics, for example for the Permic languages (Udmurt, Komi) by introducing the digraph *xh* for the voiced voiced alveolopalatal affricate (FUT /ʒ/, Udmurt *ӟ*, Komi *дз*), with these unusual choices largely motivated by the difficulty of creating established writing systems in computing (Dobó 1996).

A desire to avoid character encoding and font-related headaches, but mostly the wish to be maximally efficient, also motivated Ulla-Maija Kulonen (now Forsberg) to use her own transcription system for Eastern Mansi in her

[21] See www.jltypes.com/Uvallanne/.
[22] Personal correspondence.
[23] Márta Csepregi, personal correspondence.

materials (e.g., Kulonen 2017) which avoids a number of phonologically unnecessary diacritics used by Kannisto (e.g., *o̭* → *o*, *ä̭* → *ä*), utilizes only characters (if necessary digraphs) used in well-established writing systems of Northern Europe (*β* → *w*, *γ* → *g*, *i̯* → *j*, *ə* → *ø*, *ḡ* → *õõ*, *ö̭* → *ü*, *ȯ* → *ö*, *ȧ* → *ä*, *χ* → *x*, *ŋ* → *ng*), indicates vowel length by writing a vowel twice (e.g., *ā* → *aa*), uses the "Hungarian" system of indicating palatalness/palatalization where an *y* is placed after a consonant (*ń* → *ny*, *l'* → *ly*, *ś* ~ *š́* → *sy*, *t'* → *ty*), used the degree sign ° (instead of the more conventional ^{w}) to indicate labialization (thus k^{w} → *k*°, *χ* → *x*°) (Kulonen 2017 : 13—20).

The full range of fonts used during this era is impossible to reconstruct today, given how many were highly specialized and used by a very small set of people, in many cases presumably by one person alone. Even today when handling writing systems that are idiosyncratic beyond what Unicode with combining characters can bear (e.g., Junttila 2022) editors must begrudgingly resort to this approach. The following fonts can, based on direct experience, consultations, and access to source files, be confirmed to have been in heavy use:

- Mansi Font for Cyrillic Mansi (Скрибник, Афанасьева 2007)
- Times New 4Diacritical, Fugora Italic, Fugorb Italic, Fugor3 Italic and presumably others for FUT (used by Nyelvtudományi Közlemények)
- SmolenskSGR for Cyrillic Russian within otherwise Latin documents (Moisio, Saarinen 2008)
- Symbol for Greek characters within otherwise Latin documents (Moisio, Saarinen 2008; Wogulisches Wörterbuch 2013)

Solutions of this type were ingenious within the context of the time and yielded the possibility to make less-used writing systems look good in print with comparatively little effort from the author. More problematic is the fact that these solutions stayed in use, and in some cases stay in use, long after they outlived their usefulness. For example, Mari-Pragmatica-type encoding is still used today,[24] and one author of this paper was asked to use fonts of this type by the editor of a major publication in 2015. It is natural that something as basic as the technical infrastructure and character encoding used will not be changed halfway through a project — thus it is understandable that many non-Unicode sources (e.g., Moisio, Saarinen 2008; Wogulisches Wörterbuch 2013) were published in an era where Unicode had become standard, as it had not been so in when the undertakings were initiated. It should also be remembered that although the Unicode Consortium was established in 1991, Unicode only became dominant globally in 2008. Also, Unicode support for characters did not and does not automatically guarantee font support for a character: while numerous major fonts today adequately support the diverse writing systems used in Uralic studies, this was not historically the case, which undoubtedly also disincentivized scholars from moving away from then-established conventions and methods.

It should be noted here that source files from this bygone era can, with comparatively little effort, be made Unicode-conform and should thus not be disregarded as "useless" by those who still possess them: even if users and editors no longer have access to the fonts used at the time (most of these are not freely available on the Internet and must be acquired from scholars who

[24] E.g., at : https://mari-lab.ru/index.php/%D0%A3_%D1%81%D0%B5%D0%BC_(PDF).

used to use them), the font encoding information is generally still preserved. A modern word processor might not know what *q* in Ajatar (used there to represent *ə*) looks like, but it will know that it is the symbol *q* in the font Ajatar — in contrast to word processors used in the past that generally did not afford the possibility for font-sensitive searches.[25] Even using simple search-and-replace functionalities, modern word processors can replace all *q* in Ajatar with *ə* in a modern font, thus making the realization of this symbol Unicode-conform (this is illustrated in Figure 5); naturally, such procedures can also be streamlined using scripts — set sequences of commands through which programmers can automatize procedures.

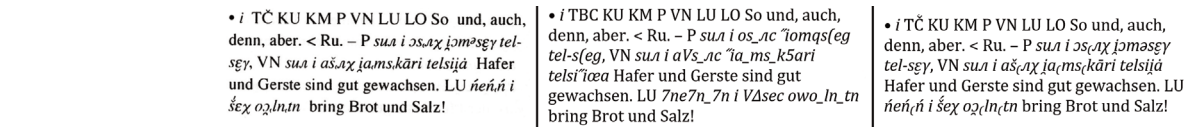

Figure 5. **The Turku-made "Wogulisches Wörterbuch" (2013 : 1): scanned (left), source without font realization (centre), digitally restored and Unicode-conform text (right).**

3. Before Computers: the pre-digital age and early digital typesetting

The advent of digital typesetting — before solutions detailed above became the norm — was extremely difficult for a discipline as reliant on idiosyncratic writing systems as Uralic Studies: paradoxically, early digital solutions made matters more difficult, rather than easier, for scholars of Uralic languages as at the time well-established methodologies were replaced with incipient technologies through the lens of which the realization of comparatively less-used writing systems had not yet been considered. Also, as discussed above, early digital solutions suffered from grave issues in the domain of compatibility, esp. between Mac and PC solutions.

Accounts from the pre-digital era paint a somewhat diffuse picture. At the Institute of Finno-Ugrian Studies at the University of Vienna, the way it was done depended on status.[26] The conventional practice was to write up manuscripts on a standard German-language typewriter and manually insert diacritics with a pen before sending it to the publisher. Letters not found on the typewriter at all (e.g., *ə*, *γ*) were either inserted after the fact or were created on the basis of the most-similar looking letter found on the typewriter (e.g., *χ* was created on the basis of *x*, *η* was created on the basis of *n*) — all using a pen. This is illustrated in Figure 6. When words or passages were to be realized in Cyrillic, the page was taken out of the regular typewriter and placed into a Russian typewriter kept at the institute for such occasions. The workflow, however, was different for Károly Rédei: he would write all his manuscripts by hand (in very legible handwriting) and send them to Budapest for processing, where there were people to take care of such things. For regular scholars in Hungary, however, the standard approach was much the same as in Vienna: people worked with standard typewriters and added by hand what they could not produce.[27]

[25] Johanna Laakso, personal correspondence.
[26] Timothy Riese, personal correspondence.
[27] Marianne Bakró-Nagy, personal correspondence.

Sy. *taml'e mat wārmal' ōnsēɣn, manriɣ at jalēɣn?*
"Ha valami olyasmid van, miért nem mész?" [Kálmán: 64]
So. *l'ul'saŋ ońsəŋkw patilən, tawe ońsənkw at wērmilən.*
"(Wenn) ihr anfangt, sie schlecht zu behandeln, könnt ihr sie nicht behalten." [Kannisto III: 28]
So. *ānəm aluŋkw man kul' joχti man matər joχti, tiɣ ul wos juw!*
"(Wenn) ein Teufel kommt oder (sonst) etwas kommt, um mich zu töten, hierher komme er nicht." [Kannisto II: 115]
LO *toχ mos worateɣn, jalēw.*
"Willst du so sehr, (so) machen wir uns auf." [Kannisto II: 162]

Figure 6. **Mansi data, typed and amended (Riese 1984 : 67).**

In Finland and Scandinavia, we know of a number of scholars who used mechanical typewriters adapted to create FUT characters (see Figure 7 for an example). Later, (customized and expensive) electronic typewriters that could type FUT entered the scene.[28] They had their own keys for characters such as *ə*, and also for diacritic marks such as the haček, allowing for the creation of characters such as *š*. Some scholars would improvise the haček by placing both an acute accent ´ and a grave accent ` over the same letter.[29] Due to the relative simplicity of adding hačeks to letters — specifically on Mikko Korhonen's typewriter at the time — the orthographic solution *ǩ* was introduced to the rendering of Skolt Saami in Latin for the voiceless palatal affricate (not to be confused with the voiceless postalveolar affricate, orthographically realized as *č*, cf. Feist 2015 : 87—88); this character has been part of Unicode since at least 1993, in a beautiful example of the convenient-solution-to-norm pipeline.[30]

Figure 7. **Björn Collinder's typewriter (photographed in Uppsala).**

Later, the introduction of IBM Selectric typewriters would revamp the workflow: these typewriters, with their replaceable typeballs (colloquially: golf balls) brought a number of advantages: users could use special print balls that would have some of the needed special characters (such as *š*, *č*, *ž*, *ə*) and also made it considerably easier to create italic text. But, here too the technology had its restrictions: one could not create one's own combinations of diacritics and letters as had previously been possible, necessitating the manual addition of diacritic

[28] Sirkka Saarinen, personal correspondence.
[29] Johanna Laakso, personal correspondence.
[30] Ulla-Maija Forsberg and Markus Juutinen, personal correspondence.

marks. The text collection "Timofej Jevsevjevs Folklore-Sammlungen aus dem Tscheremisschen" (Alhoniemi, Saarinen 1983—1994) illustrates the rapid pace of progress at the time: The first volume (1983) was composed using an electronic typewriter, the second volume (1989) using an IBM Selectric typewriter, and the last two volumes (1992, 1994) using a computer.

In Helsinki, Finno-Ugric scholars had traditionally relied on a publishing house practising traditional typesetting that went bankrupt in the late 1970s. Finding a replacement just as publishing houses were "going digital" proved to be difficult as early computer infrastructures at first did not have font solutions for the needed writing systems at all. It is presumably due to this circumstance (coupled with a lack of articles as predoc scholars were apprehensive of submitting their work to such a prestigious journal at the time) that the FUF, one of the traditional journals in Finno-Ugric Studies published since 1901, was not published between 1979 and 1982;[31] the publication of several critical resources (notably Wotjakischer Wortschatz 1987) were delayed due to the "typographische Odyssee" (Wotjakischer Wortschatz 1987 : XII) editors had to go through at the time.[32]

The epoch of manual insertions into typed manuscripts was fairly sophisticated in Finland: at times it was ceremonial, at others streamlined. When Tapani Lehtinen completed his dissertation on historic Finnic verbal conjugation (Lehtinen 1979), he organized a *dentaalispiranttitalkoot* 'dental spirant work party' in which he and his friends added all the missing *δ* and diacritics to his manuscript; Figure 8 shows the result.

ims. johtimeen (*-ta-, -tä-* ∿) **-δa-, *-δä-*, jolla myös on voitu johtaa momentaaniverbejä (ESA II s. 96).

Jo Setälä kiinnitti huomiota seuraavanlaisiin *-aise-* ja **-aiδa-* loppuisten verbien suhteisiin:

halkaisen | in *halkaan* | ka *halgoan* | ve *håugaidan*[1]

henkäisen | li *i̯eŋgub, jḛ̄ŋgə̑b* : inf. *jeŋ̄gə̑, ję̰ŋ̄gə̑*

Figure 8. **Results of a *dentaalispiranttitalkoot* (Lehtinen 1979 : 6).**

The Finno-Ugrian Society, on the other hand, had a publication secretary (Anneli Peräniitty, later Honko) who would meticulously add diacritics and special characters with an ink fine liner, producing results hard to distinguish from the results of mechanic typesetting, as shown in Figure 9.

In the early 1980s, *ə* — in frequent need of manual insertion into manuscripts — acquired the nickname *nurmikko* 'lawn' at the Finno-Ugric Department of the University of Helsinki, based on a malapropism by the Hungarian Ob-Ugrist Edit Vértes, who struggled with the term *nurinpäinen e* 'upside-down e'.[33]

[31] Johanna Laakso and Sirkka Saarinen, personal correspondence.

[32] In actual fact the odyssey of the Udmurt dictionary (Wotjakischer Wortschatz 1987) commenced on the 13th of July 1891, when Yrjö Wichmann started on his study tour to the Udmurts. From 1942 to 1947 T. E. Uotila worked on and added to Wichmann's materials, and Mikko Korhonen was the fourth and final person to toil on it, intermittently from 1960 to 1987. The whole chronicle, describing the card catalogues, concordances bound in hefty black binders, Udmurt prisoners-of-war, the intricacies of relief printing and phototypesetting, the aforementioned bankruptcy of the printing house, the plethora of symbols needed, and other details of the dictionary's 95-year gestation period, are set out in remorseless detail on pages XI—XII of the dictionary for those readers perverse enough to want a full account.

[33] Johanna Laakso, personal correspondence.

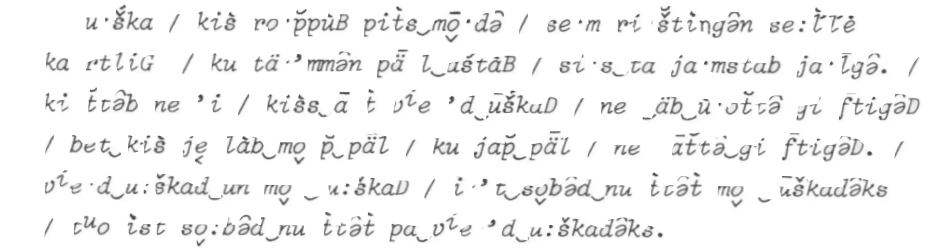

Figure 9. **Ink fine liner diacritics (Muistoja Liivinrannasta 2006 : 59).**

4. Concluding thoughts

While our article aims to be no more than a footnote in the history of Uralic linguistics, we hope to through it have made some lived experiences accessible to current and future generations. While our initial goal had been to write specifically about the technical side of the orthographic realization of Mansi, the veritable treasure trove of interesting anecdotes, titbits, and reminisces we gathered in the process ended up pertinent to the discipline of Uralic studies in general, and too valuable to want to withhold from our readers.

Here we would also like to thank those who consulted us in compiling this paper: Marianne Bakró-Nagy, Márta Csepregi, Ulla-Maija Forsberg, László Honti, Csilla Horváth, Markus Juutinen, Johanna Laakso, Niko Partanen, Timothy Riese, Michael Rießler, Jack Rueter, Sirkka Saarinen, and Beáta Wagner-Nagy. The personal experiences, direct and indirect, they shared with us from the BUE (Before Unicode Era) took our investigation into captivating and unexpected directions. While the introductions to some works from this era give some detail on how they were realized under the difficult circumstances of the time, it is clear that scholars did not usually (and perhaps understandably) consider their hardships and solutions to these worth committing to print, no matter how interesting or even useful they might seem in hindsight. This should be seen as motivation for scholars to put in writing even those aspects of their work processes they consider strange, idiosyncratic, and not worthy of attention.

Acknowledgements. The publication costs of this article were covered by the Estonian Academy of Sciences.

Addresses

Jeremy Bradley
University of Vienna
E-mail: jeremy.moss.bradley@univie.ac.at

Rogier Blokland
Uppsala University
E-mail: rogier.blokland@moderna.uu.se

BIBLIOGRAPHY

Alhoniemi, Alho, Saarinen, Sirkka 1983—1994, Timofej Jevsevjevs Folklore-Sammlungen aus dem Tscheremisschen, Helsinki (MSFOu 184, 199, 211, 219).

Beke Ödön 1961, Mari szövegek IV, Budapest.

Bradley, Jeremy, Skribnik, Elena 2021, The Many Writing Systems of Mansi: Challenges in Transcription and Transliteration. — Multilingual Facilitation, Helsinki, 12—24. doi.org/10.31885/9789515150257.2.

Csepregi Márta 1998, Szurguti osztják chrestomathia, Szeged (Studia Uralo-Altaica Supplementum 6).

D a v i s, Mark 2012, Unicode over 60 Percent of the Web. googleblog.blogspot.com/2012/02/unicode-over-60-percent-of-web.html.
D o b ó, Attila 1996, Towards a Transition from Cyrillic to Latin and Keyboard Scripts in Komi and Udmurt. — Nyelv, nyelvész, társadalom, Pécs, 32—34.
F e i s t, Timothy 2015, A Grammar of Skolt Saami, Helsinki (MSFOu 273).
J u n t t i l a, Santeri 2022, Johdatus kuilinalaiseen kielentutkimukseen. — Tonavan Laakso. Eine Festschrift für Johanna Laakso, Wien, 89—107. homepage.univie.ac.at/jeremy.moss.bradley/jl60/.
K e t t u n e n, Lauri 1939, Kaksi eestiläistä murrenäytekokoelmaa. — Vir. 43, 272—274.
K o i v u l e h t o, Jorma 2001, Merkillistä sananselitystä. — Tieteessä tapahtuu 19 (1), 50—56.
K o r e i n i k, Kadri, P l a d o, Helen 2022, Linguistic and Extra-Linguistic Arguments in Graphization Debates: An Unsettled Standard of Southern Estonian. — Tonavan Laakso. Eine Festschrift für Johanna Laakso, Wien, 309—336. homepage.univie.ac.at/jeremy.moss.bradley/jl60/.
K o r t e s h a r j u, Jouni 1999, Javaslat az uráli nyelvek fonológiai jelöléséhez. — Folia Uralica Debreceniensia 6, 93—100.
K u l o n e n, Ulla-Maija 1989, The Passive in Ob-Ugrian, Helsinki (MSFOu 203).
—— 2017, Itämansin kielioppi ja tekstejä, Helsinki (Apuneuvoja suomalais-ugrilaisten kielten opintoja varten XV).
L e h t i n e n, Tapani 1979, Itämerensuomen verbien historiallista johto-oppia. Suomen *avajaa-*, *karkajaa*-tyyppiset verbit ja niiden vastineet lähisukukielissä, Helsinki (MSFOu 169).
L e h t i r a n t a, Juhani 1989, Yhteissaamelainen sanasto, Helsinki (MSFOu 200).
M o i s i o, Arto, S a a r i n e n, Sirkka 2008, Tscheremissisches Wörterbuch. Aufgezeichinet von Volmari Porkka, Arvid Genetz, Yrjo Wichmann, Martti Räsänen, T. E. Uotila und Erkki Itkonen, Helsinki (LSFU XXXII).
Muistoja Liivinrannasta. Liivin kieltä Ruotsissa. Kerännut Julius Mägiste. Suomentanut ja julkaissut Anneli Honko, Helsinki 2006 (MSFOu 250).
M u n k á c s i Bérnát 1892—1902, Vogul népköltési gyűjtemény I, Budapest.
—— 1901, [Review] Finnisch-ugrische Forschungen. — KSz 2, 223—233.
P a r t a n e n, Niko, R i e ß l e r, Michael 2019, An OCR System for the Unified Northern Alphabet. — Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages, Tartu, 77—89.
P i s c h l ö g e r, Christian 2022, Vom Nutzen und Nachteil der Sprachwissenschaft für das Leben von Minderheitensprachen: Die "Wieso-Sprache" Besermanisch. — Tonavan Laakso: Eine Festschrift für Johanna Laakso, Wien, 348—381. homepage.univie.ac.at/jeremy.moss.bradley/jl60/.
R i e s e, Timothy 1984, The Conditional Sentence in the Ugrian, Permian and Volgaic Languages, Wien (Studia Uralica 3).
R i e s e, Timothy, B r a d l e y, Jeremy 2011, Acquisition of 'Small' Finno-Ugrian Languages and the Mari Web Project. — Языки, литература и культура народов полиэтнического Урало-Поволжья (современное состояние и перспективы развития). Материалы VIII Международного симпозиума «Языковые контакты Поволжья», Йошкар-Ола, 208—216.
—— 2020, A. N. Balandins Einführung in das Mansische, Wien. www.copius.eu/mansi/.
R i e ß l e r, Michael 2013, Towards a Digital Infrastructure for Kildin Saami. — Sustaining Indigenous Knowledge. Learning Tools and Community Initiatives for Preserving Endangered Languages and Local Cultural Heritage, Fürstenberg/Havel, 195—218. dh-north.org/siberian_studies/publications-/sikriessler.pdf.
S e t ä l ä, E. N. 1901, Über die Transskription der finnisch-ugrischen Sprachen. Historik und Vorschläge. — FUF 1, 15—52.
S k r i b n i k, Elena, L a a k s o, Johanna 2022, Graphization and Orthographies of Uralic Minority Languages. — The Oxford Guide to the Uralic Languages, Oxford, 91—100.
S t a c h o w s k i, Kamil 2011, Remarks on the Usefulness of Different Types of Transcription, with a Particular Regard to Turkic Comparative Studies. — JSFOu 93, 303—338.

Szilasi, Móricz 1896, [Review] Yrjö Wichmann Wotjakische Sprachproben I. Lieder, Gebete u. Zaubersprüche. Suom.-Ugr. Seur. Aikakausk. XI. Helsingissä 1893. — NyK XXVI, 493—495.
Viitso, Tiit-Rein 2005, Some Comments about Paul Ariste's Doctoral Dissertation on Phonetics of Hiiumaa Estonian Dialects. — LU XLI, 4—19.
Wogulische Volksdichtung III. Märchen. Gesammelt und übersetzt von Artturi Kannisto. Bearbeitet und herausgegeben von Matti Liimola, Helsinki 1956 (MSFOu 111).
Wogulisches Wörterbuch. Gesammelt und geordnet von Artturi Kannisto. Bearbeitet von Vuokko Eiras. Herausgegeben von Arto Moisio, Helsinki 2013 (LSFU XXXV).
Wotjakischer Wortschatz. Aufgezeichnet von Yrjö Wichmann. Bearbeitet von T. E. Uotila und Mikko Korhonen. Herausgegeben von Mikko Korhonen, Helsinki 1987 (LSFU XXI).
Быконя В. В., Ким А. А., Купер Ш. Ц. 1994, Словарь селькупско-русский и русско-селькупский, Томск.
Жовницкая-Турдагина С. Н. 1999, Ня'' букварь. — Санкт-Петербург.
Скрибник Е. К., Афанасьева К. В. 2007, Практический курс мансийского языка 1—2, Ханты-Мансийск.

ДЖЕРЕМИ БРЭДЛИ (Вена), *РОДЖЕР БЛОКЛАНД* (Уппсала)

МАНСИЙСКИЙ ЯЗЫК И ДРУГИЕ ЯЗЫКИ В ПЕЧАТИ ДО И ПОСЛЕ UNICODE

Разные версии стандарта Unicode создавались с целью обеспечить последовательное, ясное и надежное кодирование для всех терминов систем письменности. Успешное решение этой задачи зависит во многом от объема и уровня государственной деятельности, направленной на технологическое развитие конкретных языков и языковых вариантов. Авторы рассматривают, как в прошлом и сегодня языки уральских национальных меньшинств, прежде всего мансийский, используются в печати и дигитально. Исследование опирается на интервью и анализ сохранившихся дигитальных файлов, его цель — познакомить с полученными научными результатами более широкие слои общественности.

JEREMY BRADLEY (Viin), *ROGIER BLOKLAND* (Uppsala)

MANSI KEEL JA TEISED UURALI KEELED TRÜKIS ENNE JA PÄRAST *UNICODE*'I

Unicode'i standardi eri versioonid on loodud eesmärgiga tagada kõikide kirjasüsteemide järjepidev, selge ja usaldusväärne kodeerimine. Selle ülesande edukus sõltub suuresti konkreetsetele keeltele ja keelevariantidele suunatud riikliku tehnoloogilise arendustegevuse ulatusest ja tasemest. Artiklis vaadeldkse, kuidas on minevikus ja tänapäeval uurali vähemuskeeli, eriti mansi keelt, trükis ja digitaalsel kujul kasutatud. Uurimus põhineb intervjuudel ja säilinud digitaalsete failide analüüsil ning selle eesmärk on tuua asjaomased teadustulemused laiema avalikkuseni.