The Unicode Standard and the Unicode Approval Process

DEBORAH ANDERSON, SCRIPT ENCODING INITIATIVE, UC BERKELEY
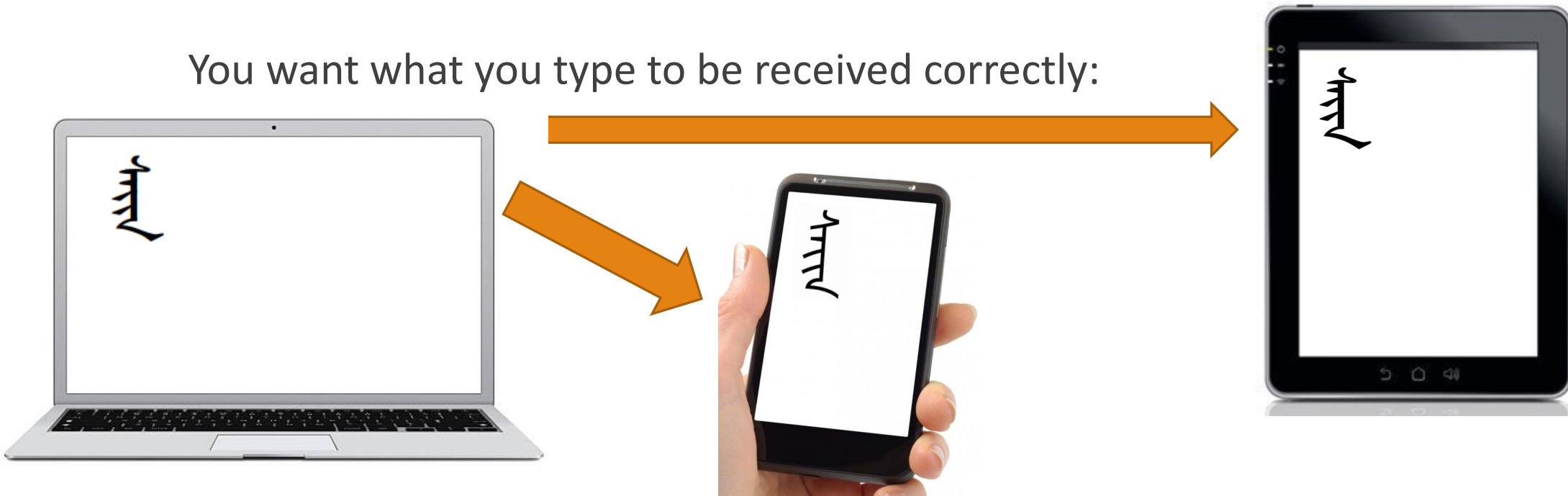
ULAANBAATAR, MONGOLIA
APRIL 3-5, 2019

Background on the Unicode Standard

Makes it possible so
◦ text and documents are interchangeable
◦ platforms supporting Mongolian are interoperable

You want what you type to be received correctly:

Background on the Unicode Standard

What users don't want…

You type:

Users get instead:

Unicode Basics

Unicode assigns a number ("code point") to the letters and symbols in a script. It also includes a representative glyph and name.

Glyph ←

1820 ← Code point
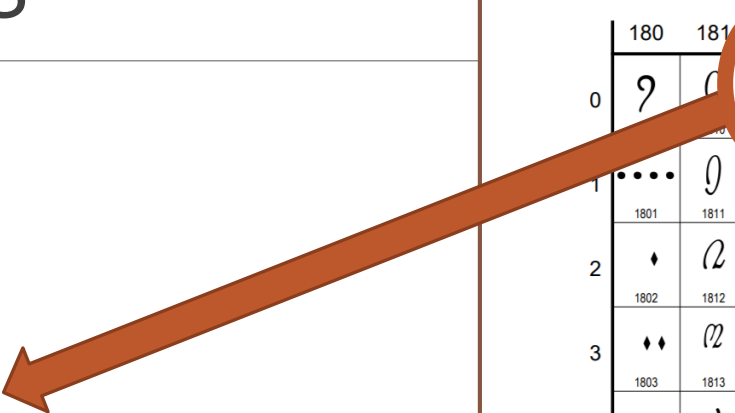
MONGOLIAN LETTER A ← Name

Unicode Basics

Code chart

Unicode Basics

Code chart

Unicode Basics

Names list

Punctuation

Additional birgas are encoded in the Mongolian Supplement block at 11660-1167F.

1800   ᠀   MONGOLIAN BIRGA
- → 0F04 ༄ tibetan mark initial yig mgo mdun ma

1801   ⁘   MONGOLIAN ELLIPSIS
1802   ᠂   MONGOLIAN COMMA
1803   ᠃   MONGOLIAN FULL STOP
1804   ᠄   MONGOLIAN COLON
1805   ᠅   MONGOLIAN FOUR DOTS
- • marks the end of a chapter

1806   ᠆   MONGOLIAN TODO SOFT HYPHEN
- • not a format control character, but simply a hyphen for Todo
- • also used in the Hudum (modern Mongolian) writing system
- → 00AD ⬚ soft hyphen

1807   ᠇   MONGOLIAN SIBE SYLLABLE BOUNDARY MARKER
1808   ᠈   MONGOLIAN MANCHU COMMA
1809   ᠉   MONGOLIAN MANCHU FULL STOP
180A   ᠊   MONGOLIAN NIRUGU

Format controls

180B   ⬚   MONGOLIAN FREE VARIATION SELECTOR ONE
- • abbreviated FVS1

180C   ⬚   MONGOLIAN FREE VARIATION SELECTOR TWO
- • abbreviated FVS2

180D   ⬚   MONGOLIAN FREE VARIATION SELECTOR THREE
- • abbreviated FVS3

180E   ⬚   MONGOLIAN VOWEL SEPARATOR
- • abbreviated MVS

Digits

1810   ᠐   MONGOLIAN DIGIT ZERO
1811   ᠑   MONGOLIAN DIGIT ONE
1812   ᠒   MONGOLIAN DIGIT TWO
1813   ᠓   MONGOLIAN DIGIT THREE
1814   ᠔   MONGOLIAN DIGIT FOUR
1815   ᠕   MONGOLIAN DIGIT FIVE
1816   ᠖   MONGOLIAN DIGIT SIX
1817   ᠗   MONGOLIAN DIGIT SEVEN
1818   ᠘   MONGOLIAN DIGIT EIGHT
1819   ᠙   MONGOLIAN DIGIT NINE

Basic letters

*Cross-references show commonly used Cyrillic transliterations.*

Letters unspecified for distinct positional forms are shown without them. Letters with unspecified positional forms use a squared question mark to indicate the gap. All vowels are expected to have explicit isolate forms. For consonants the isolate first form is the initial first form unless specified otherwise.

1820   ᠠ   MONGOLIAN LETTER A
- → 0430 а cyrillic small letter a
- ~ 1820 ᠠ first form (isolate)
- ~ 1820 ᠊ first form (initial)
- ~ 1820 ᠊ first form (medial)
- ~ 1820 ᠊ first form (final)
- ~ 1820 180B ᠊ second form (isolate)
- ~ 1820 180B ᠊ second form (medial)
- ~ 1820 180B ᠊ second form (final)
- ~ 1820 180C ᠊ third form (medial)

1821   ᠡ   MONGOLIAN LETTER E
- → 044D э cyrillic small letter e
- ~ 1821 ᠡ first form (isolate)
- ~ 1821 ᠊ first form (initial)
- ~ 1821 ᠊ first form (medial)
- ~ 1821 ᠊ first form (final)
- ~ 1821 180B ᠊ second form (initial)
- ~ 1821 180B ᠊ second form (final)

1822   ᠢ   MONGOLIAN LETTER I
- → 0438 и cyrillic small letter i
- ~ 1822 ᠢ first form (isolate)
- ~ 1822 ᠊ first form (initial)
- ~ 1822 ᠊ first form (medial)
- ~ 1822 ᠊ first form (final)
- ~ 1822 180B ᠊ second form (medial)

1823   ᠣ   MONGOLIAN LETTER O
- → 043E о cyrillic small letter o
- ~ 1823 ᠊ first form (isolate)
- ~ 1823 ᠊ first form (initial)
- ~ 1823 ᠊ first form (medial)
- ~ 1823 ᠊ first form (final)
- ~ 1823 180B ᠊ second form (medial)
- ~ 1823 180B ᠊ second form (final)

1824   ᠤ   MONGOLIAN LETTER U
- → 0443 у cyrillic small letter u
- ~ 1824 ᠊ first form (isolate)
- ~ 1824 ᠊ first form (initial)
- ~ 1824 ᠊ first form (medial)
- ~ 1824 ᠊ first form (final)
- ~ 1824 180B ᠊ second form (medial)

1825   ᠥ   MONGOLIAN LETTER OE
- → 04E9 ө cyrillic small letter barred o
- ~ 1825 ᠊ first form (isolate)
- ~ 1825 ᠊ first form (initial)
- ~ 1825 ᠊ first form (medial)
- ~ 1825 ᠊ first form (final)
- ~ 1825 180B ᠊ second form (medial)
- ~ 1825 180B ᠊ second form (final)
- ~ 1825 180C ᠊ third form (medial)

Unicode Basics

Names list

1801 ···· MONGOLIAN ELLIPSIS
1802 · MONGOLIAN COMMA
1803 ·· MONGOLIAN FULL STOP
1804 ⁚ MONGOLIAN COLON
1805 ❖ MONGOLIAN FOUR DOTS
• marks the end of a chapter
1806 – MONGOLIAN TODO SOFT HYPHEN
• not a format control character, but simply a hyphen for Todo
• also used in the Hudum (modern Mongolian) writing system
→ 00AD SHY soft hyphen
1807 ⊤ MONGOLIAN SIBE SYLLABLE BOUNDARY MARKER
1808 › MONGOLIAN MANCHU COMMA
1809 ›› MONGOLIAN MANCHU FULL STOP
180A - MONGOLIAN NIRUGU

1800   **Mongolian**   **1825**

Punctuation
Additional birgas are encoded in the Mongolian Supplement block at 11660–1167F.
1800 ? MONGOLIAN BIRGA
→ 0F04 ꣟ tibetan mark initial yig mgo mdun
1801 ···· MONGOLIAN ELLIPSIS
1802 · MONGOLIAN COMMA
1803 ·· MONGOLIAN FULL STOP
1804 ⁚ MONGOLIAN COLON
1805 ❖ MONGOLIAN FOUR DOTS
• marks the end of a chapter
1806 – MONGOLIAN TODO SOFT HYPHEN
• not a format control character, but simply a hyphen for Todo
• also used in the Hudum (modern Mongolian) writing system
→ 00AD SHY soft hyphen
1807 ⊤ MONGOLIAN SIBE SYLLABLE BOUNDARY MARKER
1808 › MONGOLIAN MANCHU COMMA
1809 ›› MONGOLIAN MANCHU FULL STOP
180A - MONGOLIAN NIRUGU

Format controls
180B [FV1] MONGOLIAN FREE VARIATION SELECTOR ONE
• abbreviated FVS1
180C [FV2] MONGOLIAN FREE VARIATION SELECTOR TWO
• abbreviated FVS2
180D [FV3] MONGOLIAN FREE VARIATION SELECTOR THREE
• abbreviated FVS3
180E [MV] MONGOLIAN VOWEL SEPARATOR
• abbreviated MVS

Digits
1810 0 MONGOLIAN DIGIT ZERO
1811 9 MONGOLIAN DIGIT ONE
1812 2 MONGOLIAN DIGIT TWO
1813 3 MONGOLIAN DIGIT THREE
1814 4 MONGOLIAN DIGIT FOUR
1815 5 MONGOLIAN DIGIT FIVE
1816 6 MONGOLIAN DIGIT SIX
1817 7 MONGOLIAN DIGIT SEVEN
1818 8 MONGOLIAN DIGIT EIGHT
1819 9 MONGOLIAN DIGIT NINE

Basic letters
Cross-references show commonly used Cyrillic transliterations.
Letters unspecified for distinct positional forms are shown without them. Letters with unspecified positional forms use a squared question mark to indicate the gap. All vowels are expected to have explicit isolate forms. For consonants the isolate first form is the initial first form unless specified otherwise.
1820 ᠠ MONGOLIAN LETTER A
→ 0430 а cyrillic small letter a
~ 1820 first form (isolate)
~ 1820 first form (initial)
~ 1820 first form (medial)
~ 1820 first form (final)
~ 1820 180B second form (isolate)
~ 1820 180B second form (medial)
~ 1820 180B second form (final)
~ 1820 180C third form (medial)
1821 ᠡ MONGOLIAN LETTER E
→ 044D э cyrillic small letter e
~ 1821 first form (isolate)
~ 1821 first form (initial)
~ 1821 first form (medial)
~ 1821 first form (final)
~ 1821 180B second form (initial)
~ 1821 180B second form (final)
1822 ᠢ MONGOLIAN LETTER I
→ 0438 и cyrillic small letter i
~ 1822 first form (isolate)
~ 1822 first form (initial)
~ 1822 first form (medial)
~ 1822 first form (final)
~ 1822 180B second form (medial)
1823 ᠣ MONGOLIAN LETTER O
→ 043E о cyrillic small letter o
~ 1823 first form (isolate)
~ 1823 first form (initial)
~ 1823 first form (medial)
~ 1823 first form (final)
~ 1823 180B second form (medial)
~ 1823 180B second form (final)
1824 ᠤ MONGOLIAN LETTER U
→ 0443 у cyrillic small letter u
~ 1824 first form (isolate)
~ 1824 first form (initial)
~ 1824 first form (medial)
~ 1824 first form (final)
~ 1824 180B second form (medial)
1825 ᠥ MONGOLIAN LETTER OE
→ 04E9 ө cyrillic small letter barred o
~ 1825 first form (isolate)
~ 1825 first form (initial)
~ 1825 first form (medial)
~ 1825 first form (final)
~ 1825 180B second form (medial)
~ 1825 180B second form (final)
~ 1825 180C third form (medial)

Unicode Basics

Names list


```
1820    ᠠ    MONGOLIAN LETTER A
        → 0430 a  cyrillic small letter a
        ~ 1820 ᠠ  first form (isolate)
        ~ 1820 ᠠ  first form (initial)
        ~ 1820 ᠠ  first form (medial)
        ~ 1820 ᠠ  first form (final)
        ~ 1820 180B ᠠ  second form (isolate)
        ~ 1820 180B ᠠ  second form (medial)
        ~ 1820 180B ᠠ  second form (final)
        ~ 1820 180C ᠠ  third form (medial)
```

Including variant information in names list poses problems:

- Not the usual approach

- Can't capture all the details of writing system

- Hard to maintain

- Recommend a *different* approach

Unicode Basics

Names list

1820   **MONGOLIAN LETTER A**
    → 0430 **a**   cyrillic small letter a
    ~ 1820   first form (isolate)
    ~ 1820   first form (medial)
    ~ 1820   (medial)
    ~ 1820   (final)
    ~ 1820   d form (isolate)
    ~ 1820   orm (medial)
    ~ 1820 180B   second form (final)
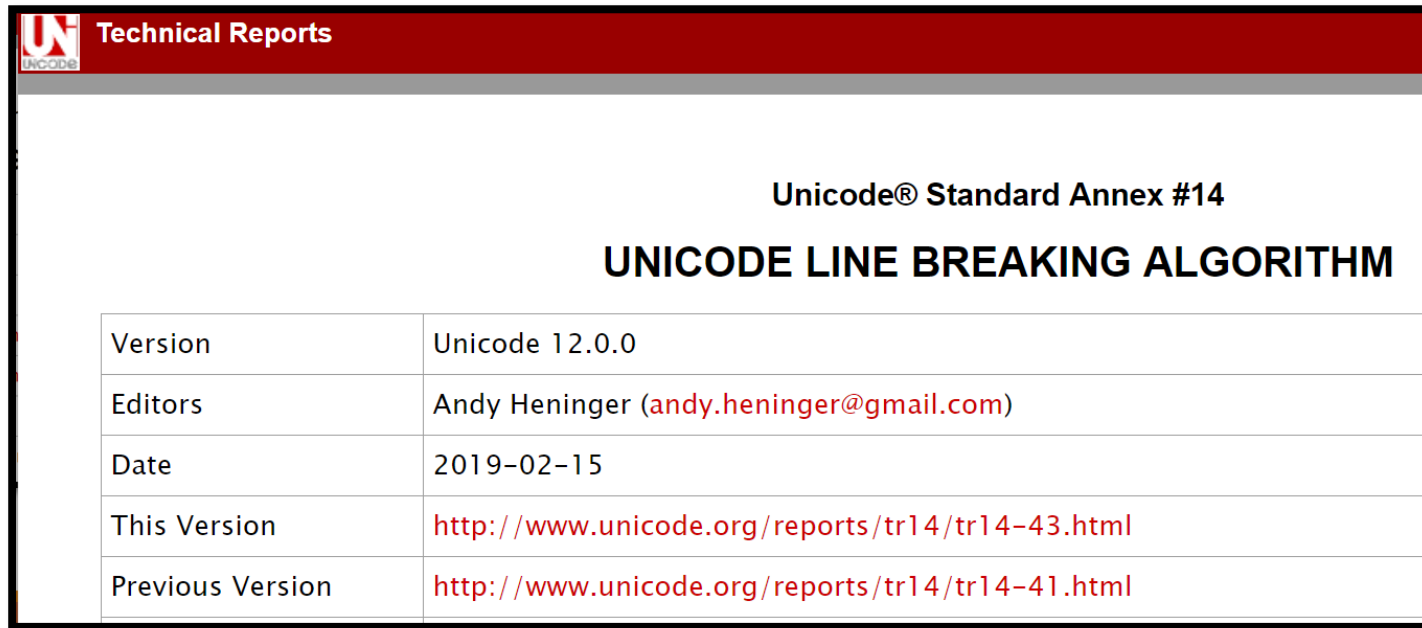   ~ 1820 180C   third form (medial)

Unicode® Technical Notes

- Separate from Unicode publication and its publishing cycle

- Easy to edit

- Contents could become a Unicode Technical Standard, after community agreement

https://unicode.org/notes/

Unicode Basics

The Unicode Standard *also* includes information on characters' behavioral properties (such as line breaking, text segmentation, etc.)

Technical Reports

Unicode® Standard Annex #14

UNICODE LINE BREAKING ALGORITHM

Version	Unicode 12.0.0
Editors	Andy Heninger (andy.heninger@gmail.com)
Date	2019-02-15
This Version	http://www.unicode.org/reports/tr14/tr14-43.html
Previous Version	http://www.unicode.org/reports/tr14/tr14-41.html

https://www.unicode.org/reports/tr14/

Character Approval and Support Process

Character Approval Process

Preparation of proposal

Example from a 2017 proposal to add one character:

MONGOLIAN LETER CHA WITH TWO DOTS

Note that proposals should include evidence:

Fig. 2: Tsydendambaev 1972 p. 304

диакритический знак. Правда, стремление как-то разграничить буквы g и k, омонимичные по своему изображению в классическом монгольском языке (ᠴ · ᠴ · ᠭ), имеет известное основание: в живой речи они приобрели фонематические различия. В "Истории перекочевки в Баргузин..." путем постановки точек справа от знака sādhē (ᠴ) показано, что монгольская буква č в бурятском языке произносится как звук š.

Character Approval Process

Preparation of proposal

Review by Unicode Script Ad Hoc

- Reviews proposals and makes recommendations to standards committees

Character Approval Process

Preparation of proposal

Review by Unicode Script Ad Hoc

Unicode Technical Committee
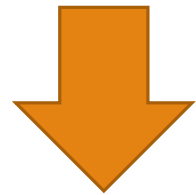- Reviews proposals and makes decisions on proposals
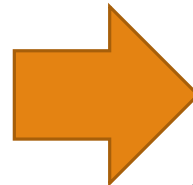
Character Approval Process

Preparation of proposal

Review by Script Ad Hoc

Unicode Technical Committee ISO SC2/WG2

- Reviews proposals and decides whether to put on ballot

Character Approval Process

Preparation of proposal

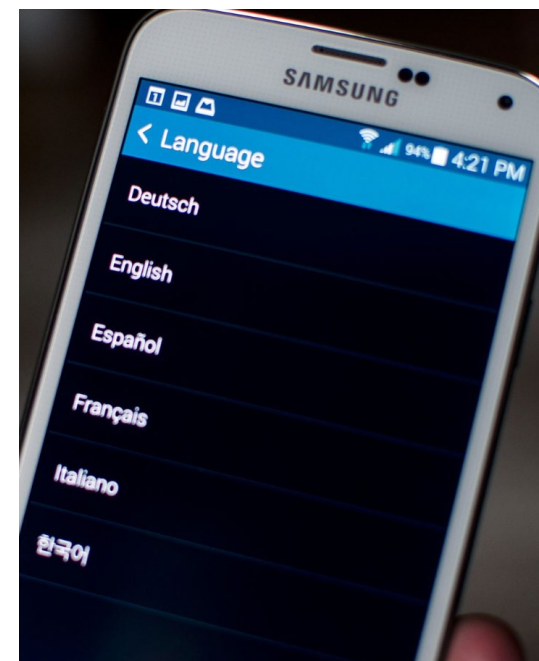Review by Script Ad Hoc

Unicode Technical Committee ISO SC2/WG2

PUBLISHED

Character Support Process

- After publication need:

  Fonts and keyboard definition

  Locale data

Character Approval and Support Process - Summary

- Many steps are involved in getting characters approved
- Can take several years for characters to be approved and implemented on devices

- Working together helps the process go smoothly

Character Approval and Support Process - Summary

- Many steps involved in getting characters approved
- Can take several years for characters to be approved and implemented on devices
- Working together helps the process go smoothly

Баярлалаа!