Unicode request for dezh with retroflex hook

Kirk Miller, kirkmiller@gmail.com

Michael Everson, everson@evertype.com

2021 January 03

The original version of L2/20-125r, Unicode request, Miller 2020, expected IPA retroflex letters and similar hooks, had requested a pair of letters, ⟨\dj \ts⟩, that have been used in phonetic descriptions of Polish orthographic cz and dż. Due to an asymmetry in attestation, only ⟨\ts⟩ was recommended by the SAH for approval by the UTC, which approved it in 2020 July. However, the acceptance of one character without its partner makes it essentially useless, as both are needed if either is to be used for any significant quantity of transcription. John Esling and Michael Ashby of the IPA concur. In its 2020 October meeting, the SAH recommended ⟨\dj⟩ based on the current proposal, and left space for it in a list of characters recommended for the Latin Extended-G block.

Background

Phonetic characters for alveolar stops and affricates come in voicing pairs, reflecting the ubiquity of a voicing distinction in European languages. It would thus hinder the utility of Unicode to encode just one member of a pair for reasons of attestation: each is clearly part of a structured set, and anyone using a transcription system with a voiceless letter for a language like Polish or Russian will need the voiced letter as well. If we do not encode both, the one that has been accepted will not serve any such user.

The form of the letter is self-evident. Without the summary above as a guide, could the UTC tell which of the ligatures below has not yet been attested? (The SAH was not able to do so in September, given that several months had passed since they had last seen the proposal.)

\[
\begin{array}{cccccc}
Z & S & Z & S & Z & L \\
\dj & ts & \dj & tc & \dj & ts \\
\dj & ts & \dj & ts & \dj & ts
\end{array}
\]
The component characters, U+1D98 ꝱ and 1D9A ꝲ, were adopted with proposal L2/03-190r by Peter Constable. Just as here, Constable could only attest to the voiceless consonant (his Figure 7), but the UTC accepted both characters as a pair.

John Esling, former president of the IPA, final editor of the IPA Handbook and author of its appendix on the computer coding of IPA symbols, wrote (p.c. 2020 Oct 19),

Laver (1994: 560) gives us (someone’s?) transcriptions of a few Polish words. Only the voiceless fricative and affricate in question appear, because the list is short. They both have voiced counterparts, as do all of the (pre-velar) fricatives and affricates in Polish (Jassem, JIPA, 2003). We could say that Polish is a full set of this genre of sibilant contrast, including post-dental (or whatever it is) as a third series. Although the voiced values may not have been put into print in any particular instance, we can assume that Laver would have had occasion to use them had Polish words containing the voiced fricatives and affricates presented themselves in his list.

Now Jassem calls the /ʧ ʤ/ pair alveolar. He calls the /ʨ ʥ/ pair alveolo-palatal. My view is that trying to nail these sounds down to a place of articulation is extremely limiting; the tongue is doing more than that (not to mention the lips). But as I recall, Hamann (JIPA 2003) found evidence of the first pair being retroflex, and Żygiel echoes that view: (Non)retroflexivity of Slavic affricates and its motivation: evidence from Polish and Czech <č>.

As your comment suggests, not everyone would agree that these sounds are archetypically retroflex (e.g. as in Russian). I like very much that Laver’s transcriptions took a ‘middle road’, between alveolar and retroflex. On this basis, I would support codification of the ‘intermediate’ symbolic representation (of the ‘alveolar’ set with the retroflex hook) for the two fricatives and the two affricates in this series (vl and vd).

That is, I agree with Michael Everson that it only makes sense to include both members of each pair, as it is just as likely (e.g. in Polish) for the vd member to occur as the vl one, and if one’s transcription is so inclined, all four sounds will need symbols to represent them.

Michael Ashby, current president of the IPA, wrote (p.c. 2020 Oct 20),

I take the same view as John and Michael Everson, and I don’t mind putting my name to a brief statement if it will help. It only makes sense to include both the voiceless and the voiced together.

**Requested character**

The referenced codepoint in red is newly approved by the SAH but not yet by the UTC.

ʤ ꝱ LATIN SMALL LETTER DEZH WITH RETROFLEX HOOK. See Figure 1.

Forms a pair with 1DF1C ꝱ LATIN SMALL LETTER TESH WITH RETROFLEX HOOK.

**Properties**

1DF19; LATIN SMALL LETTER DEZH WITH RETROFLEX HOOK; L1; 0; L; ; ; ; N; ; ; ;
Chart

A gap was left for ⟨dʒ⟩ among the Latin Extended-G codepoints (grey) that were recommended by the SAH on 2020 Oct 19. ⟨dʒ⟩ was not included in Everson’s chart of those characters so as not to interfere with their evaluation by the UTC.

|   | .0 | .1 | .2 | .3 | .4 | .5 | .6 | .7 | .8 | .9 | A | B | C | D | E | F |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Latin Extended-G |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| U+1DF0x |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| U+1DF1x |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | dʒ |

References

Peter Constable (2003) ‘Proposal to encode additional phonetic symbols in the UCS’, L2/03-190R.


Figure

This is the self-evident character for people who use ⟨ʃ⟩ and ⟨ʒ⟩ for Polish sz and ść (or for Russian ш and ж) along with the still-common IPA convention of representing affricates by ligatures, as seen for example in Figure 1. Figure 1 shows only the voiceless affricate /ʃ/, but Polish has a corresponding voiced affricate that needs to be transcribed the same way. Note that Polish contrasts affricates such as dż /dʒ/ with plosive-fricative sequences such as drz /dʒ/.  

Figure 1. Laver (1994: 560) uses only one letter of the pair in this Polish example in his textbook on phonetics. Laver was unresponsive last year to queries on the need for both letters ⟨ʃ ʒ⟩ of the pair during the writing of the parent request to this proposal, and died in May 2020.
A. Administrative

1. Title: Additional phonetic click letters

2. Requester's name: Kirk Miller, Michael Everson

3. Requester type (Member body/Liaison/Individual contribution): individual

4. Submission date: 2021 January 03

5. Requester's reference (if applicable): 

6. Choose one of the following:

   This is a complete proposal: yes

   (or) More information will be provided later: 

B. Technical – General

1. Choose one of the following:

   a. This proposal is for a new script (set of characters):

      Proposed name of script: 

   b. The proposal is for addition of character(s) to an existing block:

      Name of the existing block: Latin Extended-G

2. Number of characters in proposal: 1

3. Proposed category (select one from below - see section 2.2 of P&P document):

   A-Contemporary X B.1-Specialized (small collection) 

   C-Major extinct D-Attested extinct 

   E-Minor extinct 

   F-Archaic Hieroglyphic or Ideographic 

   G-Obscure or questionable usage symbols

4. Is a repertoire including character names provided?

   a. If YES, are the names in accordance with the “character naming guidelines” in Annex L of P&P document? yes

   b. Are the character shapes attached in a legible form suitable for review? yes

5. Fonts related:

   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard? Kirk Miller

   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.): SIL (Gentium Release)

6. References:

   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? yes

   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? yes

7. Special encoding issues:

   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? yes

8. Additional Information:

   Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

---

### C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?
   - No

2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?
   - Yes
   - With whom: John Esling, Michael Ashby
   - Available relevant documents: (emails available on request, willing to write formal letters)

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?
   - Reference:

4. The context of use for the proposed characters (type of use; common or rare)
   - Reference: phonetic

5. Are the proposed characters in current use by the user community?
   - Yes
   - Reference: see illustrations

6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?
   - No

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?
   - NA

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?
   - No

9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?
   - Yes
   - Reference: (Unicode disfavors use of combining retroflex hook)

10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character?
    - No
    - If YES, is a rationale for its inclusion provided?
    - If YES, reference:

11. Does the proposal include use of combining characters and/or use of composite sequences?
    - No
    - If YES, is a rationale for such use provided?
    - If YES, reference:

12. Does the proposal contain characters with any special properties such as control function or similar semantics?
    - No
    - If YES, describe in detail (include attachment if necessary)

13. Does the proposal contain any Ideographic compatibility characters?
    - No
    - If YES, are the equivalent corresponding unified ideographic characters identified?
    - If YES, reference: