ISO/IEC JTC1/SC2/WG2 N2656
2003-10-15

Universal Multiple Octet Coded Character Set
International Organization for Standardization
Organisation internationale de normalisation
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Proposal — Phonetic symbols used in dictionaries**
Source: **Asmus Freytag**
Status: **Expert contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Related: **N2655. Responding to N2645 and other proposals**

Document N2645 proposes another character used in phonetic notations in dictionaries. This document intends to extend the research on this type of notation, document additional instances, and propose additional characters.

Phonetic symbols

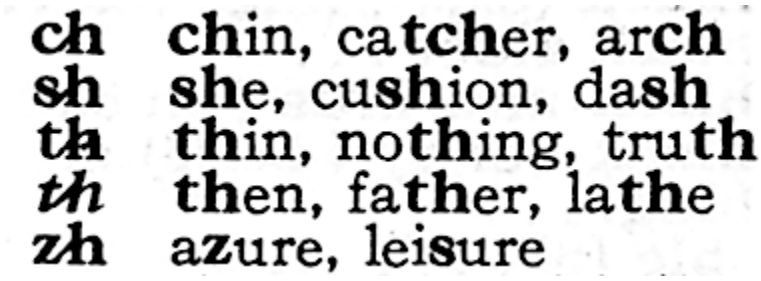
Dictionaries use a number of different methods to indicate the pronunciation of terms. Some are based on IPA, others employ other symbols, in particular barred or ligated di- and trigraphs based on small Latin letters as well as the use of diacritics across two letters. While the systems are different, there is some common ground, and systems for use in monolingual English and monolingual German dictionaries may sometimes use the same symbol for the same sound.

However, the argument made in presenting the character proposed in document N2645 to the Unicode Technical Committee, that only a single character is needed to complete the coverage this type of practice is incorrect. The proposed character may be sufficient to complete the coverage of one particular US American system, but it is not sufficient to cover all systems common use in US dictionaries, let alone cover usages in extremely widely used dictionaries in other languages.

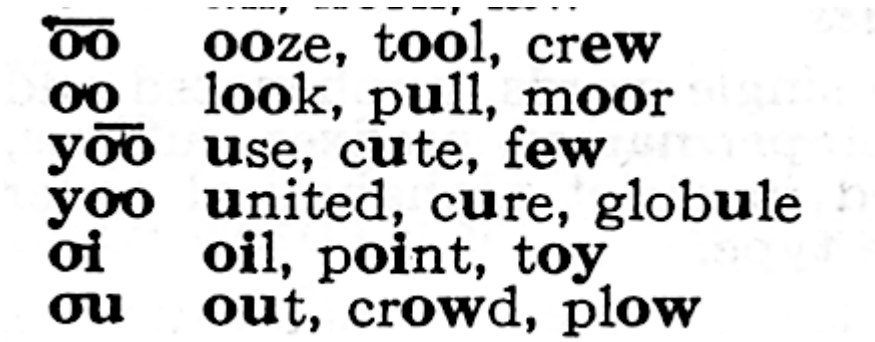
This document researches (and unlike N2645 actually cites) several dictionaries and compares their notational systems to each other and to the available characters in the Unicode standard. Characters that are readily available in Unicode are not separately discussed, as they make up the vast majority of characters in any of the systems investigated.

Widely used US dictionaries

The following two excerpts are from an American dictionary for college use, showing a variation of the phonetic transcription system for which the character th with strike through was requested in document N2645. Instead of strikethrough's, ligatures are used. (Note: there is some dirt on the page at the location of the 'th' ligature, there is *no* bar across that h.)



The full pronunciation listing for that dictionary also shows a kh ligature (not shown here), with the glyph constructed on the same principles. It is used for the ch sound in German 'ach'. In addition, it shows a number of ligatures, some with overbar:



Note that this example shows an oi and an ou ligature, as well as an oo ligature. There is again some dirt on the paper, making it look like the bar across the oo ligature for ooze is wider than the one further below, but that is *not* the case.

Not all dictionaries use either the TH with strike through or a even a ligated th. The following sample is from a dictionary that uses an unligated digraph, but with italics to indicate voiced pronunciation.

tight, stopped	t	t
thin	th	θ
this	*th*	ð
cut	ŭ	ʌ
urge, term, firm, word, heard	ûr	ɜˠ, ɜr
valve	v	v
with	w	w
yes	y	j
zebra, xylem	z	z
vision, pleasure, garage	zh	ʒ
about, item, edible, gallop circus	ə	ə
butter	ər	ɚ

FOREIGN	**AHD**	**IPA**
French **feu** *German* **schön**	œ	œ
French **tu** *German* **über**	ü	y
German **ich** *Scottish* **loch**	KH	ç, x
French **bon***	N	õ, æ̃, ã, œ̃

are shown include **more, glory,** and **borne.** A similar variant occurs in words such as **coral, forest,** and **horrid,** where the pronunciation of *o* before *r* varies between (ô) and (ŏ). In these words the (ôr) pronunciation is given first: **forest** (fôr′ist, fŏr′-).

***The IPA symbols show nasality with a diacritic mark over the vowel, whereas the Dictionary uses N to reflect that the preceding vowel is nasalized.

Note the use of small caps K, H and N. Of these, only U+1D0B LATIN LETTER SMALL CAPITAL K is currently encoded.

Glyph representation in online reference works

Microsoft Office 2000 ships with a font that is used for the on-line reference works included with various versions of Microsoft Office. There are many characters that are provided for phonetic representations and readily correspond to the phonetic notation found in the printed sources, such as:

o͞oo͝ot̶ho͜yɔ͜y a͜u

The ligated and accented digraphs o͞o and o͝o are equivalent to the oo ligature with and without a bar, note the use of both ligation and double wide diacritic, matching the sample above (where the ligation is a bit difficult to spot). The symbol t̶h is equivalent to the th ligature or the TH WITH STRIKE THROUGH from document N2645, but this time realized as an incomplete horizontal strikethrough. The two forms o͜y and ɔ͜y are equivalent to some forms of oi, depending on the precise phonetic value, while a͜u represents the same sound as the ou ligature. The font contains additional ligated digraphs, constructed by the same principle, some of them for non-English sounds:

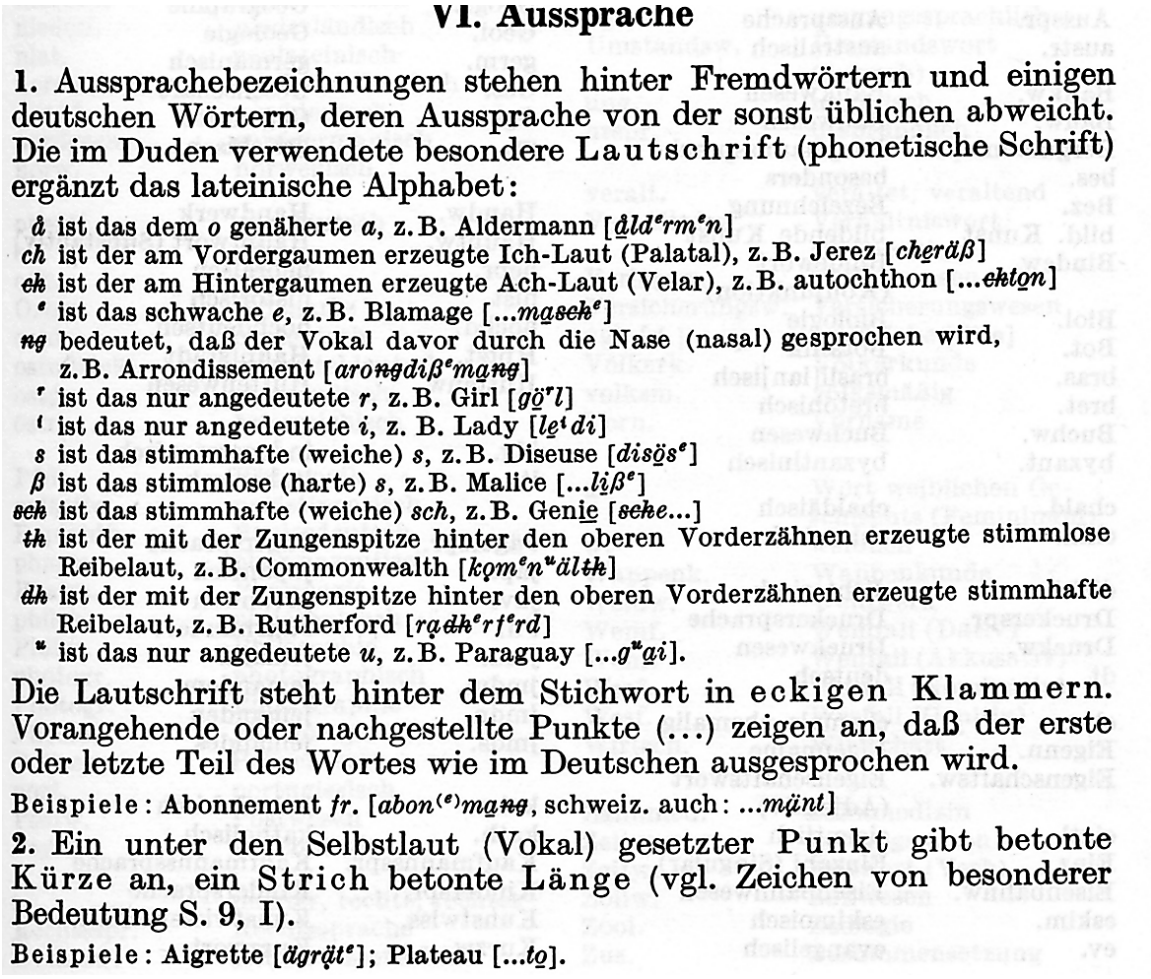
t͜s k͜s p͜f u͜i t͜ʃ a͜ɹ a͜i

The sounds that they intend to represent are immediately understandable from the constituent characters (some of which are from IPA). Nevertheless none of these characters can be represented with existing Unicode characters.

While the sound could be represented by writing just the two base characters, the double diacritic carries the essential information that the letters must be pronounced in an uninterrupted sequence. This document proposes encoding a double wide combining mark for the purpose of indicating the connection.

Non-US dictionaries

The use of such non-IPA systems to indicate pronunciation is not limited to US dictionaries. The following excerpt is from the pronunciation guide used by Duden.



VI. Aussprache

1. Aussprachebezeichnungen stehen hinter Fremdwörtern und einigen deutschen Wörtern, deren Aussprache von der sonst üblichen abweicht. Die im Duden verwendete besondere Lautschrift (phonetische Schrift) ergänzt das lateinische Alphabet:

å ist das dem *o* genäherte *a*, z.B. Aldermann [*å̲ldᵉrmᵉn*]
ch ist der am Vordergaumen erzeugte Ich-Laut (Palatal), z.B. Jerez [*cheräß*]
ch ist der am Hintergaumen erzeugte Ach-Laut (Velar), z.B. autochthon [...*chtǫn*]
ᵉ ist das schwache *e*, z.B. Blamage [...*mą̲schᵉ*]
ng bedeutet, daß der Vokal davor durch die Nase (nasal) gesprochen wird, z.B. Arrondissement [*aroⁿgdiß*ᵉ*mą̲ⁿg*]
ʳ ist das nur angedeutete *r*, z.B. Girl [*gö̲ʳl*]
ⁱ ist das nur angedeutete *i*, z. B. Lady [*le̲ⁱdi*]
s ist das stimmhafte (weiche) *s*, z.B. Diseuse [*disös*ᵉ]
ß ist das stimmlose (harte) *s*, z.B. Malice [...*li̲ß*ᵉ]
sch ist das stimmhafte (weiche) *sch*, z.B. Genie̲ [*schᵉ*...]
th ist der mit der Zungenspitze hinter den oberen Vorderzähnen erzeugte stimmlose Reibelaut, z.B. Commonwealth [*kǫmᵉnʷälth*]
dh ist der mit der Zungenspitze hinter den oberen Vorderzähnen erzeugte stimmhafte Reibelaut, z.B. Rutherford [*rą̲dhᵉrfᵉrd*]
ʷ ist das nur angedeutete *u*, z.B. Paraguay [...*gʷą̲i*].

Die Lautschrift steht hinter dem Stichwort in eckigen Klammern. Vorangehende oder nachgestellte Punkte (...) zeigen an, daß der erste oder letzte Teil des Wortes wie im Deutschen ausgesprochen wird.

Beispiele: Abonnement *fr.* [*abon*⁽ᵉ⁾*mą̲ⁿg*, schweiz. auch: ...*mänt*]

2. Ein unter den Selbstlaut (Vokal) gesetzter Punkt gibt betonte Kürze an, ein Strich betonte Länge (vgl. Zeichen von besonderer Bedeutung S. 9, I).

Beispiele: Aigrette [*ägrą̈tᵉ*]; Plateau [...*to̲*].

Marking Stress

There are many different systems to mark stress. One common system uses oversized *primes* in two different weights to mark primary and secondary stress. See the following sample:

PRONUNCIATION KEY

The symbol (′), as in **moth·er** (muŧħ′ər), **blue′ dev′ils,** is used to mark primary stress; the syllable preceding it is pronounced with greater prominence than the other syllables in the word or phrase. The symbol (′), as in **grand·moth·er** (grand′muŧħ′ər), **buzz′ bomb′**, is used to mark secondary stress; a syllable marked for secondary stress is pronounced with less prominence than one marked (′) but with more prominence than those bearing no stress mark at all.

(This sample also shows one of the symbols used to show the pronunciation of voiceless th.)

Additional information about use of already encoded special characters in dictionaries can also be found in Unicode Standard Annex #14, “Line Breaking Properties”, which can be accessed at http://www.unicode.org/reports/tr14/.

Table of Proposed Symbols

Code	Glyph	Suggested Name
0242	ch	LATIN SMALL LIGATURE CH used for ‘ch’ sound as in English. IPA []
0243	<kh>	LATIN SMALL LIGATURE KH *used for the sound of ch in German ‘ach’* (actual glyph looks like the others in this series,)
0244	sh	LATIN SMALL LIGATURE SH used for voiceless ‘sh’ sound, IPA []
0245	th	LATIN SMALL LIGATURE TH used for voiceless ‘th’ sound, IPA []
0246	*th*	LATIN SMALL LIGATURE ITALIC TH used for voidec ‘th’ sound, IPA[]
0247	zh	LATIN SMALL LIGATURE ZH used for voiced ‘sh’ sound, IPA []
023D	*ch*	LATIN SMALL LETTER CH WITH BAR *ch as in German ‘ach’*
023C	*ng*	LATIN SMALL LETTER NG WITH BAR
023B	*sch*	LATIN SMALL LETTER SCH WITH BAR
0238	*th*	LATIN SMALL LETTER TH WITH BAR
0239	ʯh	LATIN SMALL LETTER TH WITH STROKE this is the form of the letter proposed in N2645
023A	*dh*	LATIN SMALL LETTER DH WITH STROKE
023F	oi	LATIN SMALL LIGATURE OI used for transcribing the ‘oi’ sound in English
0240	oo	LATIN SMALL LIGATURE OO used for transcribing the ‘oo’ sound in English
0241	ou	LATIN SMALL LIGATURE OU used for transcribing the ‘ou’ sound in English
1D7A	H	LATIN LETTER SMALL CAPITAL H
1D7B	N	LATIN LETTER SMALL CAPITAL N This is the un-reversed of 1D0E
2Exx	◌͡◌	COMBINING CONNECTOR BELOW
205D	ˈ	LARGE THICK PRIME *primary stress* This prime is very prominent, it extends from above the top of a parenthesis, to below x height.

205E		LARGE THIN PRIME *secondary stress* same as the large thick prime, except thinner

Suggested Character Properties:

As appropriate for Latin digraphs, double combining marks, and primes, respectively.

References

Note that a large number of additional dictionaries were researched, but since they either use no phonetic symbols, or use IPA and/or other symbols already encoded in Unicode, or simply duplicate the set of proposed symbols they have not been cited here.

American Heritage Dictionary of the English Language, 3rd ed., Houghton Mifflin, Boston 1992, ISBN 0-395-44895-6

Der Große Duden, Band 1, *Rechtschreibung*, Bibliographisches Institut, 1968, Mannheim, Zürich.

The Random House College Dictionary, revised edition, Random House, New York 1975, ISBN 0-394-436008-8

Webster's New World Dictionary, Second College Edition, Williams Collins, Cleveland 1979, ISBN 0-529-05234-1

Webster's Ninth New Collegiate Dictionary, ISBN 0-87779-508-8

PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
(Please read Principles and Procedures Document for guidelines and details before filling this form.)
See http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html **for latest *Form*.**
See http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for latest *Principles and Procedures* document.**
See http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest roadmaps.**

A. Administrative

1. **Title:** **Proposal — Phonetic symbols used in dictionaries**
2. Requester's name: Asmus Freytag
3. Requester type (Member body/Liaison/Individual contribution): ___ Individual Contribution
4. Submission date: 2003-10-15
5. Requester's reference (if applicable):
6. (Choose one of the following:)
This is a complete proposal: X
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: X
      Name of the existing block: Various
2. Number of characters in proposal: 20
3. Proposed category (see section II, Character Categories):
4. Proposed Level of Implementation (1, 2 or 3) (see clause 14, ISO/IEC 10646-1: 2000): 1
   Is a rationale provided for the choice? No
   If Yes, reference:
5. Is a repertoire including character names provided?
   a. If YES, are the names in accordance with the 'character naming guidelines in Annex L of ISO/IEC 10646-1: 2000? ___Yes
   b. Are the character shapes attached in a legible form suitable for review? ___Yes
6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? proposer
   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:
7. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? No
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? yes
8. 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)?
9. Additional Information:
Submitters are invited to provide any additional information about Properties of the proposed Character(s) …

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? No
   If YES explain
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
   If YES, with whom? ___Irish NB, Michael Everson

[1] Form number: N2352-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09)

If YES, available relevant documents: _____

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

Reference: ___Publishers and users of common, widely available dictionaries of several languages_____

4. The context of use for the proposed characters (type of use; common or rare) ___common____

Reference: ___In active use, see references section_____

5. Are the proposed characters in current use by the user community? _____yes_____

If YES, where? Reference: ___see references section_____

6. After giving due considerations to the principles in *Principles and Procedures document* (a WG 2 standing document) must the proposed characters be entirely in the BMP? _____no_____

If YES, is a rationale provided? _____

If YES, reference: _____

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

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? _____no_____

If YES, is a rationale for its inclusion provided? _____

If YES, reference: _____

9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? _____no_____

If YES, is a rationale for its inclusion provided? _____

If YES, reference: _____

10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? _____yes_____

If YES, is a rationale for its inclusion provided? _____yes_____

If YES, reference: _____see attached_____

11. Does the proposal include use of combining characters and/or use of composite sequences (see clauses 4.12 and 4.14 in ISO/IEC 10646-1: 2000)? ____ no_____

If YES, is a rationale for such use provided? _____

If YES, reference: _____

Is a list of composite sequences and their corresponding glyph images (graphic symbols) 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 character(s)? _____no_____

If YES, is the equivalent corresponding unified ideographic character(s) identified? _____

If YES, reference: _____