The use of Phonetic and other Symbols in Dictionaries: A brief survey

May 08, 2006 Asmus Freytag, Ph.D.

Summary

This Unicode Technical Note presents the result of a brief survey about the use of special symbols to represent phonetic and other information in dictionaries. The survey intends to document specific examples of typical usage, rather than provide a complete summary of existing practices. Many dictionaries use The International Phonetic Alphabet [IPA], which is fully described elsewhere. A few of the special symbols mentioned in this document are not encoded, but would have to be realized with special fonts or ligatures.

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.

For this survey, several dictionaries were researched and their notational systems are compared here 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, however, in some cases, recent editions of the Unicode Standard have added some of the characters discussed here. The Unicode Consortium continues to add phonetic symbols and general symbols to the Unicode Standard, whenever they meet the criteria for character encoding.

Phonetic symbols in widely used American dictionaries

The following two excerpts (Sample 1 and 2) are from an American dictionary for college use, showing a variation of the phonetic transcription system for which the character U+1D7A LATIN SMALL LETTER TH WITH STRIKETHROUGH was added in Unicode 4.1. Instead of strikethrough's, ligatures are used.

ch chin, catcher, arch sh she, cushion, dash th thin, nothing, truth th then, father, lathe zh azure, leisure 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:

oo ooze, tool, crew
oo look, pull, moor
yoo use, cute, few
yoo united, cure, globule
oi oil, point, toy
ou out, crowd, plow

Sample 2

Note that Sample 2 shows an oi and an ou ligature, as well as an oo ligature.

Not all dictionaries use either the TH with strike through or a even a ligated th. Sample 3 below is from a dictionary that uses an unligated digraph, but with italics to indicate voiced pronunciation.

A' I. A or to a so a or of	L	1.
tight, stopped		0
thin	th	θ
this	th	ð
cut	ŭ	Λ
urge, term, firm, word, heard	ûr	3°, 3r
valve	v	V
with	W	W
yes	У	j
zebra, xylem	Z	Z
vision, pleasure, garage	zh	3
about, item, edible, gallop	Э	Э
circus		· · · · · · · · · · · · · · · · · · ·
butter	ər	Dr.
FOREIGN	AHD	IPA
French feu	œ	œ
German schön		
French tu	ü	У
German über	•	
German ich	KH	ç, x
Scottish loch		
French bon***	N	õ, æ, ã, œ

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 (\hat{o}) and (\check{o}). In these words the (\hat{o} r) pronunciation is given first: **forest** ($\hat{f}\hat{o}r'$ ist, $\hat{f}\hat{o}r'$ -).

***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.

Sample 3

Glyph representation in online reference works

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

oo oo th oy ɔy au

The ligated and accented digraphs $\overline{00}$ and $\overline{00}$ 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 the TH WITH STRIKE THROUGH, but here realized as an incomplete horizontal strikethrough.

The two forms QY and QY are equivalent to some forms of oi, depending on the precise

phonetic value, while **QU** 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:

ts ks pf ui tf aı ai

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 excerpt in Sample 4 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 [$\frac{a}{l}d^{e}rm^{e}n$]
- 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 [...chton]
- ' ist das schwache e, z.B. Blamage [...maseh']
- ng bedeutet, daß der Vokal davor durch die Nase (nasal) gesprochen wird, z.B. Arrondissement [arongdißemang]
- ' ist das nur angedeutete r, z.B. Girl $\lceil g\ddot{\varphi}^r l \rceil$
- ist das nur angedeutete i, z. B. Lady $[\underline{le}^i di]$
- s ist das stimmhafte (weiche) s, z.B. Diseuse $[dis \ddot{\varrho}s^e]$
- β ist das stimmlose (harte) s, z.B. Malice [... $li\beta^e$]
- sch ist das stimmhafte (weiche) sch, z.B. Genie [sche...]
- th ist der mit der Zungenspitze hinter den oberen Vorderzähnen erzeugte stimmlose Reibelaut, z.B. Commonwealth [komen allth]
- dh ist der mit der Zungenspitze hinter den oberen Vorderzähnen erzeugte stimmhafte Reibelaut, z.B. Rutherford [radherferd]
- * ist das nur angedeutete u, z.B. Paraguay $[...g^{u}ai]$.

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)mang, schweiz. auch: ...münt]

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 [ägräte]; Plateau [...to].

Sample 4

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** (muth '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 'muth '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.)

Use of symbols for subject classification in dictionaries

Dictionaries often need a shorthand notation to classify terms by subject matter or by other usage. A system of using iconic symbols for subject matter classification is fairly widespread,

especially on the European continent. However, there are both differences in elaboration as well as some differences in particular symbols chosen.

The symbols used for this classification are dingbats, but their function is to serve directly as a shorter and more easily recognizable stand-in for one or more words or abbreviations. Where glyphs are equivalent depictions of the same object, and can stand for each other in more than one context, unification to a set of generic dingbats would be appropriate. However, where the representative object for a category is different, for example a hypothetical use of both *ship* and *anchor* to indicate nautical terms, both symbols should remain distinct. In other words, what would be encoded in this hypothetical situation would be a symbol for a ship and a symbol for an anchor. The fact that they are both used to indicate 'nautical term' is an externally applied convention.

Note: some of the symbols are more traditional and are both used, and readily recognized outside the context of the notation for a given publication. The symbols are arranged in that order in the table of proposed symbols.

Examples of subject classification symbols

The following examples show sets of symbols actually used for classification of subject matter in dictionaries.

1. Symbols — Bildliche Zeichen

- siehe Seite 13: Fette und einfache Tilde;
- 2) see page 13: Bold-faced and Simple Tilde.
- F familiar; Umgangssprache, colloquial.
- vulgär, vulgar; unanständig, indecent.
- Botanik, botany.
- Handwerk, handicraft; Technik, engineering.
- Bergbau, mining.
- * militärisch, military term.
- Schiffahrt, nautical term.
- + Handel u. Wirtschaft, commercial term.

- Eisenbahn, railway.
- Flugwesen, aviation.
- W Postwesen, postal affairs.
- J Musik, musical term.
- A Architektur, architecture.
- Elektrotechnik, electrical engineering.
- Rechtswissenschaft, legal term.
- A Mathematik, mathematics.
- Landwirtschaft, agriculture.
- Chemie, chemistry.
- Medizin, medicine.
- → siehe Seite 14: Verweiszeichen; see page 14: Mark of Reference.

Sample 5

A nearly identical set of symbols has been identified in an English-Turkish dictionary.

Symbols — Semboller

F	colloquial language	konuşma dili
P	provincialism	taşra dili
$\cdot \mathbf{v}$	vulgar	kaba konuşma
♥	botany	botanik
\oplus	mechanics	mekanik
*	mining	madencilik
×	military term	askeri terim
Ţ	nautical term	denizcilik terimi
CS	railway, railroad	demiryol u
+	aviation	havacılık
•	postal affairs	postacılık
3	musical term	müzik terimi
£	electrical engineering	elektrik mühendisliği
***	legal te rm	hukuk terimi
A A	mathematics	matematik
•	agriculture	ziraat
	chemistry	kimya
\$	medicine	ιφ

Sample 6

A slightly more elaborate set of symbols was located by in an Icelandic dictionary. See Sample 7 below. All three samples (5-7) show a different design for the symbol for aviation.

afleidd eða huglæg merking

Ásatrú, norræn goðafræði byggingarlist B bókmenntafræði endurgerò oròmynd (oftast D dýrafræði frumnorræn) ⊕ sjaldgæ.. E) eòlisfræði og efnafræði 🕶 flugmál sjómennska (H) heimspeki † fornt eða úrelt mál, óbundið skáldamál, gamalt og nýtt jarðfræði grasafræði slangur eða óstaðfest nýyrði M málfræði, þar með bragfræði guðfræði og trúarbrögð staðbundið málfar Sálfræði £ hagfræði og verslun 🖋 stjörnufræði U uppeldisfræði, skólamál iðnaður stærðfræði V veðurfræði iþróttir)) timatal þjóðtrú landbúnaður tónlist ? vont mál, orð eða merking sem f líffræði, lífeðlisfræði, læknistækni forðast ber í íslensku (yfirleitt fræði og fleiri skyldar greinar → verður að aðeins sett þar sem betra orð 🌃 lögfræði og önnur samfélags-← komið af er sýnt í skýringu) fræði sjá, samanber skýring, það er 🖪 rafmagnsfræði

Sample 7

1 sagnfræði, sögulegt efni (við-

Sample 8 is a sample from a Danish-Islandic dictionary, showing a glyph variant on the anchor.

$^{\odot}$	Bókmenntir, bragfræði, stílfræði	×	Hernaður
(D)	Dýrafræði	*	Iðnaður, handverk
Œ	Eðlis- og efnafræði	%~	Íþróttir
E	Félagsfræði, stjórnmál	\$	Líffræði, læknisfræði, lyfjafræði
\oplus	Heimspeki		og fleiri skyldar greinar
①	Jarðfræði	2.9	Lögfræði
(L)	Listir (nema tónlist og bókmenntir)	۴	Sagnfræði
M	Málfræði, hljóðfræði	*>	Sjómennnska
S	Sálar- og uppeldisfræði, kennsla	::	Spil
1	Tölvufræði	☆	Stjörnufræði
63	Byggingarlist	π	Stærðfræði
⊛	Grasafræði, garðyrkja	J.	Tónlist
*	Guðfræði		
£	Hagfræði, verslun, viðskipti	*	fyrir framan orð táknar nýyrði

Sample 8

Sometimes, only a subset of these symbols is used, but as can be seen, with very consistent representation. The following excerpt in Sample 9 is from the list of symbols used in an English-Swedish dictionary. In addition to the symbols shown here, that dictionary also uses the sansserif capital letters F and V with the same meanings as in Sample 7 above. We may consider this

as pointing towards an overall pattern of usage that transcends individual editors' or publishers' style, but is fairly common.

- * sällsynt.
- × militärterm.
- & sjöterm.

Sample 9

Here is a sample from a typical page of the English-German dictionary showing the symbols as they are used in an actual entry:

worries von Sorgen bedrückt; ab'sorb·ent [-bənt] I. adj. absorbierend, aufsaugend: ~ cotton & Verbandwatte; II. s. Absorptionsmittel n; ab'sorb·ing [-bin] adj.

□ 1. aufsaugend; fig. fesselnd, pakkend; 2. ⊕, biol. Absorptions...,
Aufnahme... (a. †); ab·sorp·tion
[əb'səːpʃən] s. 1. a. ź, &, ⊕, biol.,
phys. Auf-, Einsaugung f, Aufnahme f, Absorption f; Vereinigung f;

Sample 10

Three additional subject classification symbols, which are commonly found, are not documented here. They include the mask for *theater*, the film clip for *cinematic*, or *cinematography* and use of a stylized palette to signify the *visual arts*.

Other symbols used in dictionaries

The use of some form of tilde to repeat terms is common. Where the capitalization changes in the new context, a circle is added:

Wenn sich die Anfangsbuchstaben ändern (groß zu klein oder umgekehrt), steht statt der Tilde das Zeichen 2 oder 2.

Beispiele:

Brit|e m; \sim in f; \sim isch adj.; falt|bar adj.; \sim -blatt n; \sim boot n; höchst...: \sim wert m; \sim zahl f; \sim zulässig adj.; Güterstand m: ehelicher \sim matrimonial regime; heilig adj. holy etc.; $der \sim$ e Geist the Holy Ghost; hängen v/i. hang etc.; \sim n hanging; harsch adj. harsh etc.; \sim m crust.

Sample 11

As can be seen in Sample 11, the glyph chosen for the tilde is not the ordinary mid-line tilde, but a low tilde that rests on the baseline. The symbol that combines the tilde with the circle is only topologically similar to the combination of U+007E TILDE and U+030A COMBINING CIRCLE ABOVE, but when typeset with common fonts such a combination ($^{\circ}$) looks noticeably different (and the circle, which is a distinguishing feature no longer appears emphasized). Note that while the text mentions the tilde by name, it does not describe it as a tilde with a circle added, rather treating it as a symbol of its own.

Sample 12 is a sample from Duden Rechtschreibung, a work usually considered the authority on German orthography, containing three further symbols not yet encoded in Unicode.

- Die senkrechte punktierte Linie dient zur Angabe der Nottrennung, z. B. Nation (vgl. S. 50, R 176).
- Als Warenzeichen geschützte Wörter sind durch das Zeichen wekenntlich gemacht. Etwaiges Fehlen dieses Zeichens bietet keine Gewähr dafür, daß es sich hier um ein Freiwort handelt, das von jedermann benutzt werden darf.
- Der waagerechte Strich vertritt das Stichwort buchstäblich, z. B. ab; und zu; oder: Brett s; -[e]s, -er; oder: Allerleis; -s, -s; Leipziger -.
- ... Drei Punkte stehen bei Auslassung von Teilen eines Wortes, z. B. Buntdruck (*Mehrz.* ...drucke); oder: Streß m; ...sses, ...sse. Über die Punkte bei Aussprachebezeichnungen vgl S. 13, VI.
- Der Bogen steht innerhalb einer Ableitung oder Zusammensetzung, um anzuzeigen, daß der vor ihm stehende Wortteil bei den folgenden Wörtern an Stelle der drei Punkte zu setzen ist, z. B. Augen braue, ...diagnose.

Sample 12

- The first is a series of four dots, spanning the height of the entire line and used in the cited work to indicate places for emergency word breaks.
- The second is based on the German word for trademark and is used where an English language work might use ® or TM. [Indirect evidence for the use of this symbol in online reference works for Microsoft Office 2000 comes from the fact that the 'Verdana Ref' font that shipped with that product contains a glyph for this Wz symbol]
- The third is used to split a term into two parts, one of which can then be repeated by an ellipsis. (...). Other dictionaries use a vertical bar and low tilde for the same purpose.

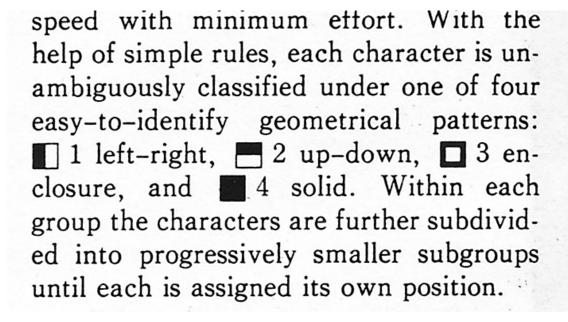
Use of geometrical shapes

Many dictionaries use geometrical shapes as symbols; cf. the use of \square in Sample 10 above. The following two samples also show the use of a symbol that appears to be the character U+25EB \square WHITE SQUARE WITH VERTICAL BISECTING LINE, but whose shape here deliberately emphasizes the dividing line, rather than the two rectangular halves of the square.

COUNT STROKES OF SHADED PART DETERMINE TOTAL STROKE-COUNT Count the strokes of the shaded Determine the total stroke-count of part to get the second part of the your character to get the second part of the SKIP number. ▶ 下 ■ 4-3-SKIP number. ▶相 ■ 1-4-REFERENCE: Appendix 2. How to Count Strokes REFERENCE: Appendix 2. How to Count Strokes COUNT STROKES OF BLANK PART IDENTIFY SOLID SUBPATTERN Count the strokes of the blank Determine to which of the four solpart to get the third part of the id subpatterns your character be-SKIP number. ▶相 1-4-5 longs to get the third part of the SKIP number. Select from: 1. □ 2, □ 3, or □ 4. ▶下 ■ 4-3-1

Sample 13

Sample 14 is a second sample from the same source, completing the set of symbols used.



Sample 14

Special uses of Punctuation and Stress Symbols

Dictionaries follow specific conventions that guide their use of special characters to indicate features of the terms they list. Marks used for some of these conventions may occur near line break opportunities and therefore interact with line breaking, for example, in one dictionary a natural hyphen in a word becomes a tilde dash when the word is split.

Examples of conventions used in several dictionaries were investigated by looking up the noun "syllable" in eight dictionaries:

Dictionary of the English Language, Samuel Johnson, 1843 **SY'LLABLE** where ' is an over-sized U+02B9 and follows the vowel of the main syllable (not the syllable itself).

Oxford English Dictionary (1st Edition) $si \cdot l \ddot{a}' b l$ where \cdot is a slightly raised middle dot indicating the vowel of the stressed syllable (similar to Johnson's acute). The letter \ddot{a} is U+0103. The ' is an apostrophe.

Oxford English Dictionary (2nd Edition) has gone to IPA 'sIləb(ə)l where ' is U+02C8, I is U+026A, ə is U+0259 (both times). The ' comes before the stressed syllable. The () indicate the schwa may be omitted.

Chambers English Dictionary (7th Edition) sil'a-bl where the stressed syllable is followed by 'U+02B9, a is U+0259, and - is a hyphen. When splitting a word like abate'- ment the stress mark 'goes after stressed syllable followed by the hyphen. No special convention is used, when splitting at hyphen.

BBC English Dictionary slləbl where \underline{I} is <U+026A, U+0332>, ϑ is U+0259. The vowel of the stressed syllable is underlined.

Collins Cobuild English Language Dictionary slləbə'l where I is <U+026A, U+0332> and has the same meaning as in the BBC English dictionary. The $\mathfrak a$ is U+0259 (both times). The $\mathfrak a$ is u+2070 and indicates the schwa may be omitted.

Readers Digest Great Illustrated Dictionary. syl·la·ble (sílləb'l) The spelling of the word has hyphenation points (· is a U+2027) followed by phonetic spelling. The vowel of the stressed syllable is given an accent, rather than being followed by an accent. The ' is an apostrophe.

Webster's 3rd New International Dictionary. syl·la·ble /'siləbəl/ The spelling of the word has hyphenation points (• is a U+2027) and is followed by phonetic spelling. The stressed syllable is preceded by 'U+02C8. The ə's are schwas as usual. Webster splits words at the end of a line with a normal hyphen. A U+2E17 DOUBLE OBLIQUE HYPHEN indicates that a hyphenated word is split at the hyphen.

This survey was originally published in Unicode Standard Annex #14, "Line Breaking Properties", which can be accessed at http://www.unicode.org/reports/tr14/.

About the use of oblique double hyphen in dictionaries

A lot has been written recently about the use of Fraktur hyphen ("oblique double hyphen"). See for example document N2639. For completeness, here is a citation of the use of this symbol in a dictionary context.

Webster's Ninth New Collegiate Dictionary, ISBN 0-87779-508-8 Copyright 1989 by Merriam-Webster Inc.

Explanatory Notes, End-of-line division (p11)

A double hyphen at the end of a line in this dictionary (as in the definition at indexation) stands for a hyphen that belongs at that point in a hyphenated word and that is retained when the word is written as a unit on one line.

And the definition of indexation reads (line breaks indicated by |):

a system of economic control | in which certain variables (as wages and interest) are tied to a cost-of<double-> | living index so that both rise of fall...

Graphically, the double hyphen is slanted upward.

Acknowledgements

Michael Everson contributed to the collection of samples from dictionaries, especially on symbols. Tim Partridge contributed the set of samples of stress marks and punctuation characters (originally published as part of UAX#14 Line Breaking Properties, but re-edited and incorporated here for completeness),

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.

Akdikmen, Resuhi. 1992. Langenscheidt's pocket Turkish dictionary: Turkish-English English-Turkish. Berlin & München: Langenscheidt. ISBN 0-245-60405-7

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

Árni Böðvarsson. 1992. Íslensk orðabók. Reykjavík: Mál og menning. ISBN 9979-3-0446-4

Engelsk-svensk ordbok, skoluplaga, Esselte Herzogs, Nacka 1976, ISBN 91-24-19070-2

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

Langenscheidt's New College German Dictionary, German-English © 1973 Langensheidt KG, Berlin and Munich, ISBN:0-88729-018-3

NTC's New Japanese-English Character Dictionary, Jack Halpern ed., Tokyo 1993.

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

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 Copyright 1989 by Merriam-Webster Inc.

The dictionaries used for the survey on representing 'syllable' are not separately cited as references. Sufficient information to identify them is incorporated in the text.

Web sites referenced

http://www.wikipedia.org/wiki/Caduceus

http://www.wikipedia.org/wiki/Rod_of_Asclepius