AN UPDATE ON PHONETIC SYMBOLS IN UNICODE

John Wells

Dept. of Phonetics and Linguistics, University College London
j.wells@ucl.ac.uk

ABSTRACT

The problem of including phonetic symbols in popular computer applications such as word-processing, email, presentation graphics, and web pages has by now been largely, though not entirely, solved through the implementation of the Unicode standard. This paper traces the advances made in this field since the last ICPhS and assesses the current position. With the general availability of Unicode, the various unstandardized custom fonts that phoneticians previously used must now be treated as 'legacy fonts'. A remaining issue is that of the input of special characters: but in this area, too, satisfactory solutions are now readily available.

Keywords: Unicode, phonetic symbols, IPA.

1. INTRODUCTION

As reported in [4], the previously chaotic situation regarding phonetic symbols for computers was by 2003 well on the way to resolution through the adoption of the multi-byte Unicode encoding in place of the unstandardized single-byte custom character sets and fonts that had previously been in use. This process has continued. Nevertheless, there is still some way to go.

2. EARLY CODING SYSTEMS AND LEGACY FONTS

In the early eighties it was not possible to use IPA symbols at all in computer applications. The only available solution was ASCIIization, in which phonetic symbols were replaced by ASCII surrogates: e.g. [θɪŋk] could be represented as [TINk]. This robust method is still in use, despite its obvious drawbacks.

By the nineties a number of customized phonetic fonts had become available, some from commercial companies and others from organizations such as the Summer Institute of Linguistics. With these fonts it was necessary to switch into the special font each time a symbol was required and then switch out of it for the surrounding text. The fonts were of necessity single-byte (8-bit) fonts and therefore restricted to a small character set. They had to encode the phonetic symbols in the same code positions otherwise used for standard characters, e.g. the schwa at the same code point otherwise used for @ or &.

This solution is still used by some phoneticians. Its great disadvantage is that the lack of standardization means that different fonts use different coding and different keyboard layouts. Material encoded for one font appears as gibberish if viewed in another font. Furthermore, there are built-in incompatibilities between Windows and Mac versions of the same fonts [7].

Such fonts are now designated *legacy fonts.* Their use is deprecated. Despite this, the author guidelines for this very conference, ICPhS 2007, recommended legacy fonts, namely the IPA-SAM fonts available from UCL [12] or the older SIL fonts — with the afterthought that "Unicode is accepted". What is not yet taken for granted, as it surely must soon be, is that for scholarly material that is intended to be archived Unicode is now the only serious option.

3. UNICODE

Unicode constitutes an agreed standard for the encoding of written characters for all the world's languages, as well as for specialized disciplines such as, in our case, phonetics. The coding is multibyte and allows for very large character sets.

The members of the Unicode consortium include such industry leaders as Adobe, Apple, Google, HP, IBM, Microsoft, Monotype and Yahoo!. Unicode has been adopted as a standard for Windows, Mac, and Linux systems, and is part of the specification of XML and HTML.

The Unicode standard is set out in the book of the same name [2], now in version 5.0, and is also available on CD-ROM and on the web [3]. It consists in an extensive specification of technical and rendering standards, together with the assignment of code positions to some 100,000 different written characters (the vast majority of which are those required for Chinese). These code

positions, identified by hexadecimal numbering, are organized into blocks of related characters.

Most of the IPA symbols in Unicode are in a block designated IPA Extensions, numbered from 0250 to 02AF. There are two other, adjacent, blocks: (i) Spacing Modifier Letters (02B0-02FF), which includes characters such as the length mark [ː] and the aspiration diacritic [ʰ], and (ii) Combining Diacritical Marks (0300-036F), where we find for example the devoicing and nasalization diacritics, as in [z̥, ã].

However, by no means all IPA characters are in these dedicated blocks. Some are scattered around in other blocks, depending (a) on whether a given symbol is also used in the orthography of some language, and (b) on the date at which the symbol was incorporated into the standard. Thus the symbols [æ, ð, ø], because they are needed for the orthography of Scandinavian languages, are in the Latin-1 Supplement block (code positions 00A1-00FF), along with [ç] and the other diacritic-bearing letters needed for French and other West European languages. The letters [œ], optionally used in French, [ħ], needed for Maltese, and [ŋ], needed for Sámi and various African languages, are in the Latin Extended-A block (0100-017F) along with the diacritic-bearing letters needed for Polish, Hungarian and Czech. As a result, the symbols [æ, ð, ø, œ, ħ, ŋ] are available in many non-phonetic fonts, including Times New Roman, as are the symbols [β] and [θ], which are part of the Greek alphabet and therefore found in the Greek and Coptic block (0370-03FF). The click symbols [ǀ, ǁ, ǂ, ǃ], used in the orthography of Nama, are located in the Latin Extended-B block (0180-024F) along with such exotica as the upper-case [Ɛ, Ɔ, Ə], the ex-IPA [ƙ] and the non-IPA [ǰ].

These facts sometimes come as a surprise to students of phonetics, who expect to find all the phonetic symbols in the IPA Extensions block. It can also make it difficult to locating particular symbols using such accessories as Character Map.

As additional phonetic symbols have been added in the various revisions and extensions of Unicode, some have been placed in yet other blocks. Various non-IPA phonetic symbols are to be found in the Phonetic Extensions (1D00-1D48) and Phonetic Extensions Supplement (1D80-1DBF) blocks. These include the IPA-inspired [ɨ,ʉ] and characters such as [đ,ɖ] from which IPA approval has been withdrawn.

The symbol for the labial flap, [ⱱ], recently approved by the IPA, is not yet part of the Unicode standard, but is in the pipeline for recognition.

4. UNICODE PHONETIC FONTS

With a Unicode font it is possible to use the same font for ordinary orthography and for IPA symbols.

The font Lucida Sans Unicode, routinely supplied with Windows in many locales, contains a wide range of phonetic symbols, namely all those that were recognized in Unicode version 2.0 (1996). It has not been revised to incorporate those added since.

The font MS Mincho, supplied with Japanese and some other versions of Windows, also contains a range of phonetic symbols, but with no stress marks or diacritics available. Some symbols are typographically unsatisfactory.

Both these fonts have been available for a decade or so. More recent fonts naturally reflect subsequent revisions of the Unicode standard, with a larger range of phonetic characters. There are several of these, of varying quality and comprehensiveness. Outstanding are those available free of charge from SIL: Charis SIL, Doulos SIL (replacing the now obsolete non-Unicode SIL IPA93 Doulos) and Gentium. Another is TITUS Cyberbit, which includes a number of potentially useful precomposed symbols such as [t̪]; but they are in the Private Use area rather than at standard Unicode code points. To compare the appearance of these fonts, see Table 1.

Lucida Sans Unicode	ˈθɪŋz ə ˈwʌndəfl
Charis SIL	ˈθɪŋz ə ˈwʌndəfl
Doulos SIL	ˈθɪŋz ə ˈwʌndəfl
Gentium	ˈθɪŋz ə ˈwʌndəfl
Titus Cyberbit Basic	ˈ θɪŋz ə ˈ wʌndəfl

Table 1: Phonetic transcription of the phrase *Things are wonderful* in five Unicode phonetic fonts.

In order for a computer to display particular Unicode characters, the relevant font must first be installed. All those who make use of phonetic symbols are strongly recommended to download Charis SIL, Doulos SIL and Gentium from sil.org and to install them.

5. APPLICATIONS

5.1. Word processing

Microsoft Word has in principle been Unicode-compliant since Word 97. With the introduction of Word XP the inputting of occasional Unicode characters has been greatly simplified by the Alt-X convention (see below). Various earlier glitches have been resolved, and there is now no difficulty in using Unicode characters, including phonetic symbols, in Word.

More basic Microsoft applications, for instance the MS Works word processor, have more limited Unicode functionality. With some other applications, such as the spreadsheet Excel, it may be necessary to compose the wording in Word and then cut-and-paste.

5.2. Email

The adoption of Unicode in email has been very slow. Four years ago probably no one was sending or receiving email with Unicode phonetic symbols. Personal experience suggests that even now the number of phoneticians using Unicode phonetic symbols in email remains very small, even though email software with Unicode capability is now readily available free of charge. Even in China and Japan, where local scripts require very extensive character sets, local encodings such as Shift-JIS (which does not make provision for IPA symbols) are still widely used, and many commonly used e-mail programs cannot handle Unicode data correctly or indeed at all. The prospects for a change in this situation seem meagre in the immediate future.

MS Outlook and Outlook Express, on the other hand, can handle Unicode in both incoming and outgoing messages, as long as the appropriate Options are selected (Send | International Settings | Unicode) and provided the inputting problem can be solved (see below).

Generally considered a better choice is the open-source Mozilla Thunderbird email program [11], which is available for Windows, Macintosh and Unix platforms and in a wide range of languages, free of charge. It has excellent spam filtering and can display Unicode symbols in both incoming and outgoing messages.

With these programs it is possible to use Unicode phonetic symbols not only in the body of the message but also in the Subject line.

The popular email client Eudora is not yet Unicode-compliant, although in 2006 Qualcomm announced that "future versions of Eudora will be based upon the same technology platform as … Mozilla Thunderbird" [10]. This is due to be implemented sometime this year (2007).

5.3. Presentation graphics

MS Powerpoint was slow to acquire reliable Unicode functionality. Recent versions, however, handle Unicode phonetic symbols satisfactorily. Lecturers may still encounter difficulty when using borrowed computers, because of the possible absence of the required fonts. Experience shows that even some speakers at academic conferences can be caught out by this, finding that their carefully-prepared slides display with blanks or gibberish. One work-around is to save each slide (or the part of it containing the symbols) as a graphic, and then to insert the graphic in place of, or as part of, the slide in question. Anything saved as a graphic ought to display correctly under all circumstances.

5.4. Web pages

A valid HTML or XHTML document is a sequence of Unicode characters. Nevertheless, many web pages in practice use older legacy encodings such as Windows-1252 or other local standards.

Web browsers have supported Unicode for many years now. Display problems result primarily from font-related issues. In particular, MS Internet Explorer does not render many code points unless it is explicitly told to use a font that contains them. Other browsers — Mozilla Firefox, Opera, Safari — can intelligently choose an appropriate font if the character required is not in the current font. Naturally, such a font must be present in the operating system: if none has been installed on the computer, obviously no browser can render the character correctly.

Non-ASCII characters are usually incorporated in HTML source files either by using a mathematical transformation such as UTF-8, or (more robustly) by using a numeric character reference. The latter consists of the Unicode number followed by a semicolon and preceded

either by &# (if the number is expressed decimally) or &#x (if it is expressed in hex form).

Thus the symbol [ŋ] can be represented in the source code for a web page as ŋ or as ŋ. The symbol [ʊ] can be entered as ʊ or as ʊ.

6. INPUT METHODS

Other than for web pages, how can we input characters that are not available directly from the keyboard? Apart from the basic Latin a-z, few standard computer keyboards have any phonetic symbols available with a single keystroke.

6.1. The Alt-number method

On Windows desktop computers, characters in the Latin-1 Supplement block (0080-00FF, the single-byte characters) can be entered by holding down the Alt key and typing the decimal code value, with a leading zero, on the numerical keypad. There are only four IPA symbols for which this is relevant: æ (entered as Alt-0230), ç (Alt-0231), ð (Alt-0240), and ø (Alt-0248). There are also Word shortcuts for these symbols: shift-ctrl-& a for æ, ctrl-comma c for ç, ctrl-' d for ð, and ctrl-/ o for ø.

6.2. The Character Map method

The Windows 'Character Map' (since Windows 2000) offers a scroll-down box of all the characters present in a given font. Any desired character can be selected, copied, and pasted into some other Unicode-compliant program. MS Word (since Word 2000) has a similar facility available by accessing Insert | Symbol; this is more user-friendly than Character Map, showing for example the last few special characters that have been used. On the Mac, OS X (since version 10.2) has a similar facility, 'Character Palette', which allows the user to select any Unicode character from a table organized numerically, by Unicode block, or by a selected font's available characters.

6.3. The Alt-X method

In Word (since XP) any Unicode character up to FFFF can be input by typing its hex number and then doing Alt-x. So the letter ʊ, for example, is entered as 028A followed by Alt-x. Conversely, any Unicode character in a document can be converted into its Unicode number in the same way, which may facilitate the identification of unfamiliar characters in a document received from elsewhere. This powerful convention represents a major step in making Unicode characters accessible to all users of word processing.

6.4. The special keyboard method

A much more convenient option for the frequent user of phonetic symbols is the installation of a dedicated virtual keyboard. The UCL Phonetics website [12] offers the Unicode Phonetic Keyboard, an installable keyboard for Windows PCs that provides a convenient keyboard layout for the word-processing of phonetic transcription. Key assignment is based on the SAMPA ASCIIization familiar to many phoneticians, using normal, Shift and AltGr combinations (on keyboards that have an AltGr key; otherwise, Ctrl+Alt). It allows for 107 different special IPA characters, which covers almost everything most users need. A tab on the Language Bar enables the user to switch instantly between the standard and the phonetic keyboard while using the same font. The present author has used this facility with success for inputting extensive material for publication.

SIL offers a similar facility in its Keyman Keyboard [8]. This is designed for US keyboards, and may not work correctly for a non-US keyboard with an AltGr key.

Other, more limited, input methods include CharWrite® [6], Phonemic Script Typewriter [10], and Unicode Phonemic Typewriter [6], the last two covering only symbols for British English.

7. REFERENCES

[1] IPA transcription with SIL fonts. http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=IPAhome. Visited 2-Mar-07.
[2] Unicode Consortium, The 2006. *The Unicode Standard 5.0*. Boston, etc.: Addison-Wesley.
[3] Unicode Consortium, The 2006. The Unicode Standard. http://www.unicode.org/. Visited 2-Mar-07.
[4] Wells, J. 2003. Phonetic symbols in word processing and on the web. *Proc. 15th ICPhS* Barcelona, xxx-xxx.
[5] http://davidbrett.uniss.it/phonemicTypewriter/phonemicTypewriter.html. Visited 5-Mar-07.
[6] http://emeld.org/tools/charwrite.cfm. Visited 5-Mar-07.
[7] http://en.wikipedia.org/wiki/Unicode. Visited 5-Mar-07.
[8] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UniIPAKeyboard. Visited 5-Mar-27.
[9] http://www.e-lang.co.uk/mackichan/call/pron/type.html. Visited 5-Mar-07.
[10] http://www.eudora.com/press/2006/eudora-mozilla_final_10.11.06.html. Visited 5-Mar-07.
[11] http://www.mozilla.com/en-US/thunderbird/. Visited 5-Mar-07.
[12] http://www.phon.ucl.ac.uk/resource/phonetics/. Visited 5-Mar-07.