Diacritics

J.C. Wells, University College London

Diacritics are distinguishing marks attached to letters of the alphabet, for example the acute accent on the **é** in **café**. Most language orthographies based on the Latin alphabet make some use of diacritics, as indeed do those based on other alphabets and writing systems. The focus of this article is on the role of diacritics in the orthography of languages written with the Latin alphabet.

Indeed, the origin of some letters that are now a standard part of the alphabet lies in the use of diacritics. The letter **G** was invented in Roman times as a variant of **C**, distinguished by the crossbar on the upstroke. The letter **J** was not distinguished from **I**, nor **U** from **V**, until the 16th century (Sampson 1985: 110). The new letter **ŋ** is obviously a variant on **n** and so could be seen as incorporating a diacritic tail. Diacritics proper, though, are seen as marks attached to a base letter. In this sense, **j u ŋ** do not involve diacritics.

The extensive use of diacritics to supplement the Latin alphabet in cases where it was seen as inadequate for the sounds of other languages is generally attributed to the religious reformer Jan Hus (1369-1415), who devised a reformed orthography for Czech incorporating 'accented' letters such as **č ě ň ř š**.

Most diacritics are placed above the base letter with which they are associated. A few, however, are placed below it (as **ç**) or through it (as **ł**).

Latin letters come in lower-case and upper-case versions. Diacritics can be awkward with the latter, and there is a convention in French, for example, that under some circumstances the diacritics of **Ê, À** etc. may be omitted.

The principal diacritics are as follows. An example is given of each, with the name of a language in whose orthography it is used. Those placed over the base letter are the acute (◌́ **á**, Spanish), the grave (◌̀ **à**, French), the circumflex (◌̂ **â**, French), the caron (or háček, ◌̌ **č**, Czech), the breve (◌̆ **ă**, Romanian), the macron (◌̄ **ā**, Latvian), the dot (◌̇ **ė**, Lithuanian), the dieresis (◌̈ **ä**, German), the tilde (◌̃ **ã**, Portuguese), the double acute (◌̋ **ő**, Hungarian), the ring (◌̊ **å**, Danish), and the hook (◌̉ **ả**, Vietnamese). Those placed through the base letter are the stroke (**/ ø**, Norwegian), and the bar (**- đ**, Croatian). The horn (◌̛ **ơ**, Vietnamese) is attached alongside. The principal diacritics placed under the base letter are the cedilla (◌̧ **ç**, Turkish) and its variant the comma (◌̦ **ş**, Romanian), the ogonek (◌̨ **ą**, Polish), and the dot (◌̣ **ạ**, Vietnamese). For exhaustive listings of the combinations of alphabetic letters and diacritics to be found in language orthographies, and what they stand for, see Wells (2001). See also The Unicode Consortium (2003).

The same diacritic may be applied to very different purposes in different orthographies. For example, the acute accent on vowel letters can indicate word stress (Spanish), vowel length (Czech), vowel quality (French **é**), a diphthong (Icelandic **á**), high tone (Vietnamese), or rising tone (the Pinyin romanization of Chinese). In Dutch it signals emphasis or the strong form of a word (e.g. **één**). On consonant letters it indicates palatalization (Polish **ń**). Typographically, the acute

accent should properly be at a steeper angle in Polish than in French or Spanish (Twardoch, 1999).

Equally, the same phonetic quality may be indicated by a range of different diacritics in different languages. The palatoalveolar [ʃ] is written **š** (Czech etc.), **ș** (Romanian), **ş** (Turkish), **ŝ** (Esperanto), and **ṣ** (Yoruba).

The dot is not a diacritic in **j** nor, usually, in **i**. In Turkish, however, undotted **Iı** and dotted **İi** are treated as distinct, standing for back and front vowels respectively.

Two or more diacritics may be combined on the same letter. This is frequent in the orthography of Vietnamese, where in **ẫ ắ ậ ẻ ờ**, for example, one diacritic indicates tone, the other vowel quality.

Letters bearing diacritics have always presented something of a problem to typographers. Both on the web and in print, combining a 'floating' one-size-fits-all diacritic with a base letter tends to give unsatisfactory results, since it is difficult to position the diacritic properly. In the days of typewriters, diacritics (if provided) were typically put on 'dead' keys, and the typist had to strike first the diacritic and then the base letter. Nowadays, in Unicode and therefore in HTML, the code for the diacritic must follow the code for the base letter. If possible, however, precomposed (ready-made) combinations are nowadays preferred. The Unicode standard provides separate code points for almost all combinations used in language orthographies (though not necessarily for those needed in phonetic transcription, philological work, or mathematics, logic and physics). Those required for Western European languages

(e.g. **á ç è î õ ü**) are to be found in the Latin-1 Supplement block, U+00A0 to U+00FF. Most of those needed for other Latin-letter orthographies (e.g. **ā ċ ĕ ģ ĥ į ļ ł ő ţ ů ŵ**) are in the Latin Extended-A block, U+0100 to U+017F. Various unusual combinations are in the Latin Extended-B block, U+0180 to U+0236 (e.g. **ǖ ǣ ǰ ǿ ȍ**) or the Latin Extended Additional block U+1E00 to U+1EDE (e.g. **ạ ḍ ế ḥ ầ**).

The localization of computer software and hardware includes the provision of special keyboards (physical or virtual) to facilitate access to the diacritics or diacritic-letter combinations required in local languages. Where a dedicated keyboard is not available, these special characters can be entered in other ways, e.g. by using an application such as Character Map or, in the most recent versions of Word, by typing the Unicode hex code number followed by Alt-X.

References

Alvestrand, H.T., 1995. 'Languages and character sets'. On the web at www.alvestrand.no/ietf/lang-chars.txt.

Korpela, J., n.d. 'A tutorial on character code issues'. On the web at www.cs.tut.fi/~jkorpela/chars.html.

Sampson, G., 1985. *Writing Systems*. London: Hutchinson.

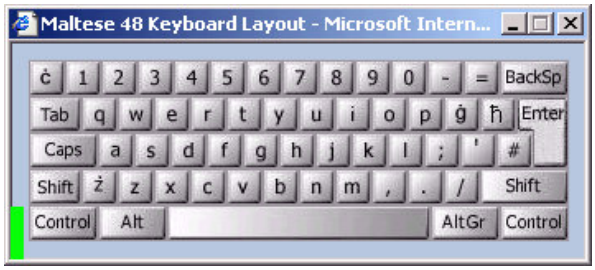
The Unicode Consortium, 2003. *The Unicode Standard, Version 4.0*. Reading, MA: Addison-Wesley. ISBN 0-321-18578-1. Also on the web at www.unicode.org.

Twardoch, Adam, 1999. 'Polish diacritics: how to?'. On the web at studweb.euv-frankfurt-o.de/twardoch/f/en/typo/ogonek

Wells, J.C., 2001. ‘Orthographic diacritics’. *Language problems and language planning* 24, 3. There is a revised and updated version on the web at www.phon.ucl.ac.uk/home/wells/dia/diacritics-revised.htm.

Wood, A., n.d. ‘Combining diacritical marks’. On the web at www.alanwood.net/unicode/combining_diacritical_marks.html.

Wood, A., n.d. ‘Unicode character ranges and the Unicode fonts that support them’. On the web at www.alanwood.net/unicode/fontsbyrange.html.



Microsoft keyboard localization. Layouts for Slovak, Maltese, Latvian and Swedish-with-Sami, including precomposed letter-diacritic combinations. The Latvian keyboard is shown with Shift and Alt-Gr depressed, the Swedish-Sami with Alt-Gr depressed. (from www.microsoft.com/globaldev/reference/keyboards.aspx)

ABSTRACT

Diacritics are distinguishing marks attached to letters of the alphabet, for example the acute accent on the **é** in **café**. Most language orthographies using the Latin alphabet supplement it in this way, using a variety of marks placed over, through, beside or under the base letter. The Unicode standard for computers provides for precomposed versions of most letter-diacritic combinations needed in orthographies. Localized keyboards allow them to be simply accessed.

KEYWORDS

orthography, diacritic, accent, spelling, acute, grave, circumflex, caron, háček, breve, macron, dieresis, tilde, ring, hook, stroke, bar, horn, cedilla, comma, ogonek, dot, Unicode