DIACRITICS IN NATIVE LANGUAGES

CHRIS HARVEY © 2009

CONTENTS

THE RELATION BETWEEN SOUNDS AND LETTERS

The modern Latin script (as used by English, French, Spanish, etc.) has twenty-six letters: A–Z. This means that any language which has more than twenty-six sounds (phonemes) must modify the alphabet in some way to accommodate the full range of phonemes. Similarly, some strategy to change the alphabet is required if the language uses sounds which were absent from the original Latin.

It is generally understood that each letter of the Latin A–Z has some kind of inherent sound, or group of sounds, which writing systems should follow. For example, the letter **e** is inherently a vowel sound, which is pronounced somewhere at the front of the mouth. The letter **b** is a consonant involving the coming together in some way of both lips, while **t** is a consonant using the tongue-tip near the front of the mouth. If twenty people speaking twenty different languages which use the Latin script saw the word **bet**, we can assume that most would all pronounce it with *something* close to the same sound.

Due to historical reasons, some languages diverge from the inherent sounds of the Latin script more than others. English has undergone some major vowel changes, which result in the letter **u,** for example, being pronounced in a rather unusual way when short [ʌ]. In other cases, languages which have recently begun to use the Latin script have matched sounds and letters in atypical ways: e.g. Pinyin Chinese **q** pronounced close to English “ch” [tɕ]. And some letters, like **c** or **x** represent a wide variety of phonemes in different writing systems (orthographies). Generally speaking, though, **q** is still a consonant and **u** is still a vowel. The consonants **c, j, q,** and **x** are among the letters most commonly re-assigned, as their Latin pronunciation values are often superfluous: they could be replaced with **k, y, kw,** and **ks** respectively.

How then, can a language’s Latin-script orthography write sounds that didn’t exist in Latin? As mentioned in the previous paragraph, one could assign novel sound values to letters. Many languages do this to some small degree, but too many changes make knowledge of the script difficult to transfer to other languages:

- English: **j** [dʒ], **long-i** [ɑɪ], **long-a** [eɪ] and a few other vowel sounds.
- Welsh: **y** [ə], **u** [ɨ]/[i]
- Hungarian: **c** [ts], **s** [ʃ]
- Kiowa (McKenzie Orthography): **f** [p], **j** [t], **v** [p’], **x** [tʃ’], **q** [k’]

A second strategy is to combine letters of the Latin script together to represent unique phonemes. While a very popular way to extend the alphabet, this technique runs into problems where a two or more of charaters could be pronounced multiple ways: either as a single sound or a series of sounds, for example: English **sh** in **fishing** [ʃ] and **mishap** [sh].

Some writing systems use a punctuation mark or accent to separate letter combinations: the mid-dot (l·l) in Catalan, the apostrophe (n'g) in Inuktitut, the underline (en) in Mohawk. Many orthographies do not have consistent ways to separate these types of combinations.

- English: **ch** [tʃ], **sh** [ʃ], **th** [θ]/[ð], **ee** [i], **igh** [ɑɪ] and so on
- Welsh: **ch** [x], **dd** [ð], **ll** [ɬ], **rh** [r̥], **th** [θ], and so on
- Hungarian: **cs** [tʃ], **dzs** [dʒ], **gy** [ɟ], **ly** [j], **sz** [s], and so on
- Kiowa (McKenzie Orthography): **ch** [tʃ], **th** [t'], **au** [ɔ]

Yet another method is to introduce completely new or modified letters to the Latin scripts. This is relatively uncommon in orthographies which have been in use for many centuries, however, newly developed spelling systems often contain characters borrowed from various phonetic alphabets:

- Icelandic: **ð** [θ]/[ð], **þ** [θ]/[ð]
- Polish, Dene languages: **ł** [w] in Polish, [ɬ] in Dene
- Halkomelem (Musqueam): **ʔ** [ʔ], **ƛ** [tɬ], **ə** [ə], **ɬ** [ɬ], χ [χ]
- Ktunaxa: **ȼ** [ts], **ɬ** [ɬ], **ʔ** [ʔ]

WHY DIACRITICS?

Diacrics, often called *accents,* are the final way to extend the alphabet that I will discuss. Cross-linguistically this is probably the most popular means (along with letter combinations) to spell out sounds lacking in Latin, though it is not at all common in English.

- Swedish: **ä** [ɛ], **ö** [ø], **å** [o]
- Uummarmiutun Inuvialuktun: **ñ** [ɲ], **ɨ̂** [ɻ]

Diacritics are especially effective as they allow readers to see associations between sets of sounds. Typically a diacritic indicates that the base letter has been modified in some predictable way.

- Welsh: **a** [a], **â** [aː], **e** [ɛ], **ê** [eː] the circumflex accent indicates a long vowel in an unexpected place.
- Italian: **a** [a], **à** ['a], **e** [ɛ], **è** ['e] the grave accent indicates an unusually stressed vowel.
- Nisga'a: **m** [m], **m̓** [m̰], **n** [n], **n̓** [n̰] the apostrophe accent indicates a glottalised consonant. **g̲** [ɢ], **k̲** [q], **x̲** [χ] the low-macron accent indicates a uvular consonant.
- Tłı̨chǫ Yatiì: **a** [a], **à** [a+low tone], **e** [e], **è** [e+low tone] the grave accent indicates low tone. **a** [a] **ą** [ã], **e** [e] **ę** [ẽ], the ogonek accent indicates a nasal vowel.

By using diacritical marks, the relationships between sounds and sound changes are not confused by the addition of new characters. In Mohawk, a vowel can carry three different stress/tones: unstressed (no diacritic), high tone stressed (acúte accent), falling tone stressed (gràve accent). When suffixes are added, stress usually shifts towards the end of the word, meaning what was once a stressed vowel becomes unstressed: **oháha** 'road' > **ohahákta** 'beside the road'. The change in stress is shown by the accent leaving the base characters unchanged. If, hypothetically, Mohawk indicated stressed vowels with a new symbol (such as §), the spelling of the root word would no longer be consistent: ***oh§ha** 'road' > ***ohah§kta** 'beside the road'.

*An asterisk * before a word means that *itz form is incorrect.*

Remembering the correct usage of diaritics can be difficult at first for people who are only familiar with English spelling (which uses accent marks sparingly if at all), and there is often an initial distaste towards these marks. However diacritics are an integral part of most Latin-based orthographies on earth and give a writing system its character and aesthetic: what would French be without é or ç, Spanish without ñ, or Navajo without ę?

Even the ancient Romans used an accent mark in Latin: called an apex

DIACRITICS AND FONTS

It was a fact of life on early computers that most languages could not be displayed properly because the ASCII character set did not contain any accented characters whatsoever. In 1985, the ISO 8859-1 (often called Latin-1) character set was released including a number of pre-composed accented characters for major western European languages, though French and Finnish could not be written correctly as the characters **œ, š,** and **ž** were absent. Proper quotation marks were also lacking.

To display the major central Eurpean languages, one had to install special CE fonts, which would re-arrange the character map, removing western European accented characters and replacing them with those needed for Hungarian, Czech, Slovak, etc. There were similar re-encodings for Baltic Languages (Latin-4), Turkish (Latin-5), and many more. Users of each encoding needed to install special fonts. If one wanted to view Lithuanian, for example, one would need a font based on Latin-4, the language could not be read with a Latin-1 font.

Some encodings were standardised; generally these were all in Europe. For speakers of indigenous languages without their own encodings, speakers had to resort to home-made, ad-hoc fonts with idiosyncratic character mapping. If you have ever used 'Times Navajo', 'WinMac' (for NWT Dene languages), or the Cherokee Nation's 'Cherokee', you are familiar with ad-hoc fonts.

While the myriad different encodings—some standard, some not—enabled one to print out hard copies in many languages, with the arrival of the internet and e-mail, a serious flaw emerged. Here are some commonly encountered situations, even today:

- You want to send me an e-mail in your language which has diacritics which do not exist in Latin-1. You either leave out diacritics altogether, or use type-fudges. For example, the word **Tŝilhqot'in** would be typed either ***Tsilhqot'in** or ***Ts^ilhqot'in**. These fudges amount to spelling mistakes.
- You want to create a web-page in your Native language, which contains diacritics when written. You include a link to download an ad-hoc font which will allow me, the reader, to make sense of the page. Chances are, I don't want to download and install software just to read your page, so I click off somewhere else. Without the font, the text is garbled and illegible. Or you have to upload everything as a PDF.
- In desparation, the local language authority replaces the orthography with a new system devoid of special characters or diacritics. While this solves some technical problems, it is an example of people serving the machine, instead of the machine serving the people.

Ndè Naàwo is an example of a Tłı̨chǫ Yatıì language website using an ad-hoc font. Assuming you don't have the font installed, the text appears full of diaereses, circumflexts, æ's, and å's, none of which belong in the orthography.

With the release of Unicode, and Unicode support becoming standard on all modern computers, the days of requiring ad-hoc fonts had come to an end. Unicode introduced the *combining diacritic*, a character consisting of a floating accent mark which binds to the preceding character. So that **r̂** is made up of two characters: **r** + **combining circumflex**. With a broad selection of combining accents to choose from, virtually any base-letter diacritic combination is possible. There are several complicating factors:

- Not all fonts contain the combining diacritic characters. The system fonts, like Times New Roman or Helvetica do have these diacritics, as do the fonts from

*Here I will start to distinguish a **character** from a **letter**. A letter is a unit of orthography: in many languages combinations like **lh** or **t'** are considered one letter. A character is a unit of computers: the smallest unit of type as the computer understands it. A base letter is a character, a combining diacritic is another character, and*

Languagegeek.

*something like **lh** is two characters, irrespective of how it is used in specific languages.*

- Even when the combining diacritics are present in the font, often the font's designer did not include instructions on how to properly place those diacritics above or below the base characters. In this case, the diacritic will appear too high or not high enough, or too far to the left or right. Languagegeek fonts include instructions for diacritic placement for North American Native languages, and many other languages using the Latin Script around the world.
- The designers of Unicode did not want to make documents using earlier encodings—like Latin-1, Latin-2, etc.—obsolete. The precomposed accented characters found in other encodings: like **ä** or **î**, had to be included in Unicode as precomposed in addition to building these by base character + combining diacritic. Therefore, a letter like **ä** can be either a precomposed character: U+00E4, or a base character (a) followed by the combining diacritic (diaeresis) U+0061 U+0308. Both versions of **ä** should be treated as identical on computers, but not all software is in compliance yet.

*Unicode characters are usually referenced by number, U+0058 is capital **X** and U+0142 is lowercase slash-l **ł**. Unicode characters also have official names, which are typically given in all-caps.*

In the end, it is my advice that everyone should be using Unicode encodings and fonts as Unicode is the global standard which allows all languages to work within the same system no matter whether one is using Windows, Mac, Linux, or whatever. The Languagegeek fonts were specially designed to use combining diacritics to write any indigenous language.

TYPING DIACRITICS

Unicode fonts allow one to read the language, typing it another matter. The computers used by most Native language speakers around the world come with a keyboard for the dominant language of that country. In some cases, this is not a problem, for example: Quechua can be typed on a Spanish keyboard, or Abenaki on a Canadian French keyboard. However, a great many indigenous languages use letters which are not accessible on the Native speakers' computers' keyboards.

- The best solution is to use a keyboard layout specifically designed for your language. If your computer does not have such a keyboard already, please download and install a Languagegeek keyboard layout which will allow you to quickly and easily type all the characters you need.
- The standard keyboard layouts on Macs can type certain diacritic marks by using the option key. This method does not meet the needs of most Native languages, and is not the most efficient way to type. Windows has a similar kind of keyboard called US International.
- You can open the Character Palette or Character Map, and find-and-click the characters you need. This is a reasonable solution if you only need to add one or two characters to your document, but for any amount of typing in the Native language, this technique is frustrating.

THE NAMES OF THE DIACRITICS

Each diacritical mark has a name. Different languages often have different words for accents they use, and in a few cases, different accent names can be used when the same mark has different functions. Often speakers of indigenous languages come up with their own words to describe the diacritics, both in the Native language and in English. These words usually refer to either how the mark affects pronunciation or what it looks like on the page. The **háček** accent (the down-pointing arrow on top of the č pronounced: HA-check) is often called a 'wedge', and many people call the **circumflex** (the up-pointing arrow on top of â) a 'hat'. Some descriptors by pronunciation are: the **acute accent** (as in

é) can be called ‘high tone’ or ‘stress accent’ (depending on the language) and the **ogonek** accent (as in ą) is often referred to as a ‘nasal hook’.

Below is a list of the most commonly seen diacritics in Native languages, along with their standard English name and Unicode encoding number, followed by some other commonly heard words to describe these accents, and a few Native languages which use this diacritic. The mark is shown with the letter ‘a’ as a demonstration, it does not mean that in the languages given, the diacritic is combined specifically with ‘a’.

Mark	Name	Unicode	Other Names	Languages
à	grave	U+0300	low-tone	Tsek’ehne, Kanien’kéha
á	acute	U+0301	high-tone, stress	Dene, Bodéwadminwen
â	circumflex	U+0302	hat, falling-tone	Kaska, Karúk Vahi
ã	tilde	U+0303	squiggle, nasal	Avañe’ẽ, Onoñda’gega’
ā	macron	U+0304	long, above-line	Mvskoke, X̄a''islaka̓la
ă	breve	U+0306	short	Tohono ’O’odham
ȧ	dot accent	U+0307		Lakota, Dakota
ä	diaeresis	U+0308	umlaut, two dots	Onödowága, Hän
å	ring accent	U+030A	whispered	Etsêhesenestse
ǎ	háček	U+030C	wedge, rising tone	ʔayʔaǰuθəm, Nuučaaʕuł
ả	comma above	U+0313	apostrophe, glottal	Nisg̱a’a, Secwepemctsín
ạ	dot below	U+0323	dot	Yokuts, Nłeʔkepmxcin
ą	ogonek	U+0328	nasal hook	Gov̨ogohó:nǫ’, Diné Bizaad
a̱	macron below	U+0331	underline	Kwak’wala, X̱aad Kil
a̲	low line	U+0332	underline	Dakelh, Sosoni’

Home Previous Page