## APPENDIX: UNICODE and UTF-8

One very common pitfall is the UNICODE representation of letters – different representations may look just the same on the screen. In Toolbox, it is important to enter ALL possible representations of a letter in the sort order in the language configurations. This holds in particular for accentuated letter, as for example ***ã***, the Latin lowercase letter 'a' with a tilde. There are two UNICODE representations of this, which in the UTF-8 encoding correspond to two bytes (a combined, complex character), or to three bytes (a combination of two UNICODE characters, one being represented by one byte, the other by two bytes). What does this all mean?

Each letter is in binary terms a sequence of *bits* – zeros and ones. The letter "***a***" is the sequence 01100001 (of eight bits, one *byte*), in decimal notation ***97*** (= 1*64 + 1*32 + 1). In *hexadecimal* notation (with 16 instead of 10 digits: 0 1 2 3 4 5 6 7 8 9 A B C D E F), this number is represented as "61" (6*16 + 1 = 96 + 1 = 97), usually written ***0x61***, to avoid confusing hexadecimal with decimal representations if no letter-digit occurs, as in this case.

With one *byte*, one can represent $2^8$ = 256 different letters or other symbols. This is enough for upper and lower letters, the numbers, interpunctuation, and a selection of letters with accents, tilde etc. The problem is: each language needs different special letters, and some need more than 256 – think of Chinese! So the widely used encoding Latin1 is just ***one*** manner to assign the numbers 0 to 255 to certain letters or symbols – there are many other *encodings*, which is why sometimes emails or documents look strange, with the wrong symbols.

*UNICODE* is not much more than an assignment of *one unique name* and *one unique number* to ANY letter or symbol in ANY language. For instance, the phonetic symbol "***ɔ***" is the UNICODE character ***U+0254*** (UNICODE uses the prefix ***U+*** and hexadecimal notation, the decimal number of this character is ***596***), called "*LATIN LETTER SMALL OPEN O*". Luckily, the basic letters (ASCII) are the same as in Latin1: ***a*** = ***U+0061*** (=***97***), named *"LATIN SMALL LETTER A*".

*UTF-8***.** In order to represent all the UNICODE characters, one would need at least three bytes for each character – that is not practical. Therefore, different UNICODE-encodings exist. A very popular and practical one is *UTF-8*. It is a "compromise" character encoding that can be as compact as ASCII (if the file is just plain English text) but can also contain any UNICODE characters. Often used characters have one or two bytes, and some characters have up to four.

In UNICODE, the letter **ã** ("a with tilde") can be represented as a combination of character "*a*" (***U+0061***, see above) with the UNICODE character "*COMBINING TILDE*", ***U+0303*** (=771). The latter character needs two bytes in UTF-8 representation, 204 and 131 (0xCC and 0x83). In the old Latin1 encoding, these two numbers correspond to the letters ***Ì*** and ***ƒ***, which is why the whole combination of three bytes, 97 204 131, or 0x61 0xCC 0x83, or 01100001 11001100 10000011, looks, in Latin1, as *"**aÌƒ**"*. (The letter ***ẽ*** corresponds to the three bytes *"**eÌƒ**"*.)

But there is also a *combined* UNICODE character for the "a with tilde". It is called "*LATIN SMALL LETTER A WITH TILDE*", has the UNICODE number ***U+00E3*** (=227), and in UTF-8, it is represented by a sequence of two bytes, ***0xC3 0xA3***, = 195 160, = 11000011 10100011, which in Latin1 renders as *"**Ã£**"*. (The accidental Latin1 *Ã=0xC3* has nothing to do with the represented character *ã=U+00E3*.)

Sometimes, there are also two **functionally** different UNICODE characters that *look the same* (in some fonts). This holds in particular for the **apostrophe-like** characters:

*APOSTROPHE* = **U+0027** = 39. Looks like this: **'**. In UTF-8: 0x27 (one byte): equally: **'** (is ASCII).

In word processors, it is often ***automatically replaced*** by ‘, ’, ‚ or similar.

*MODIFIER LETTER APOSTROPHE* = **U+02BC** = 700. Looks like this: **ʼ**. ***Use this character to represent ejectives and the glottal stop*** (as in Awetí)‼ In UTF-8: 0xCA 0xBC (two bytes): "**Ê¼**".

*RIGHT SINGLE QUOTATION MARK* = **U+2019** = 8217. Looks like this: **’** (identical to the previous U+02BC in most fonts, or very similar). In UTF-8: 0xE2 0x80 0x99 (three bytes): "**â€™**".

There are others, such as *PRIME* ′ and the *ʻOKAINA* ‵, see the Wikipedia s.v. "Apostrophe".