A REFLEXION ON UNICODE

Letterjuice

The Unicode consortium, as they describe themselves is a «*non-profit corporation devoted to developing, maintaining, and promoting software internationalization standards and data, particularly the Unicode Standard, which specifies the representation of text in all modern software products and standards*»[1].

In plain words, and regarding what concern type designers, they assign a code called Unicode to glyphs allowing the software to identify a particular glyph when typed by the user. By giving a code to a glyph they make a much deeper decision than what the people involved may realise. They have the power of giving voice to many people and to silent the voice of many others. In lots of cases that has even a political implication. Therefore, they have an enormous responsibility and they should do enough research to make the right decisions on how they assign those codes and also how they organise them in different code charts.

These decisions affect specially minority languages, since they have no official status, and thus they are very weak. Unicode Consortium at the moment is far from helping the preservation of languages which are endangered but rather helping them to disappear by ignoring their needs. Many languages have fascinating stories about human history, they are our cultural patrimony and they deserve respect and effort to be preserved.

In this article we give a few examples of what we consider wrong decisions that the Consortium has made regarding minority languages. There are three common mistakes we will talk about in order to explain their consequences. However, they are mere examples we use for illustrating those problems, we have not analysed the entire Unicode Charts. The issues shown here affect the Latin part of Unicode, further investigation should be carried out regarding other scripts.

[1] http://www.unicode.org/consortium/consort.html

MISS-REPRESENTATION OF SHAPES

There is a glyph called "ldot" (UNI013F and UNI0140), which is completely miss-represented, even the name is miss-leading. The letter coded is just the representation in half of its actual shape, which is formed by an l a period (centered between the two) and another l (l·l, it is used in Catalan). The name ldot responds to its visual representation in Unicode l·, which unfortunately has become the standard way in which type designers draw this letter. The letter has a different sound than a single l or than a double l, it is an l germination, so a "long l" sound. This is a fairly common letter used in a language that though it has official status of minority language it is used by 11,530,160 speakers according to Wikipedia (data from ethnologue.com, 2009), and that has become important enough to be covered by numerous software companies such as Adobe.

The two "l" are too far appart

l·l Il·lustració

L·L COL·LEGI

wrong

"l" are closer and the dot is centered between them

l·l Il·lustració

L·L COL·LEGI

right

The problems this create is obvious, we end up with lots of badly design letters which deprives readers of typographic quality.

There are two ways this can be sorted by an informed designer:

1. Designing l• as it is now represented but taking special care with the spacing of this glyph followed by another l.

2. Designing the right shape l•l with the right placement of the dot. We recommend this solution, which is more faithful to reality.

This are, of course, "patch" solutions, the real solution is to have the right shape and name attached to its Unicode number, and that only can be sorted by the Unicode Consortium. Another issue surrounding this glyph is the fact of not having the letter itself in the keyboard layout, but this issue has to do with economic reasons, which are much harder to sort out, and out of the scope of this article. As things are now a contextual alternate OT feature is needed in order to obtain the right glyph when typing it.[2]

[2] Please note there are two ways in which people will type this: l period l or l periodcentered l.

LANGUAGES WITH GLYPHS WITHOUT CODE

Guaraní is a South American language spoken by 4,850,000 people[3]. It is official language in Corrrientes (Argentina), Paraguay and Bolivia, and also spoken in Brazil and Uruguay. Unfortunately, it has three missing glyphs in the Unicode chards. Gtilde, gtilde and Puso, which represents a glottal interruption, the shape of this glyph is similar to apostrophe though it is is located lower and has slightly different shape (according to Paraguayan type designer Juan Heilborn).

[3] http://en.wikipedia.org/wiki/Guarani_language

G̃ g̃ ´

Guaraní glyphs without unicode

This makes it very difficult to typeset text since the computer is not capable to recognise those glyphs. The only way to access these glyphs is through the glyph palette in professional typesetting and graphic design software and through insert symbol option in text editors. Professional typesetters might know that but for the majority of people this is a big barrier which they probably do not know how to trespass. The consequence of this is that people choose to use the big language which has no problem to typeset. Young generations do not have the possibility of using their own language in digital environment. That helps the gradual abandoning of the use of the affected language, and it puts it in a position of danger of disappearance.

LANGUAGES HAVING HALF OF THE LANGUAGE IN ONE CHART AND THE REST IN AN OBSCURE ONE

This is the case of some of the Sámi languages, Sámi languages are the languages used by the indigenous people of Sampi, which is the northern part of the European continent in across Norway, Sweden, Finland and Russia. There are approximately 25,000 speakers[4] of different Sámi languages and decreasing.

The problem some of the Sámi languages suffer is that they have one of their specific glyphs covered in the Latin Extended-A Chart which is covered by plenty of typefaces. However, they have missing glyphs to be fully supported. For instance Skolt Sámi, has some glyphs with an assigned code but in Charts that are not so commonly covered by type designers. The letters g, ǩ, ǧ, ʒ and ǯǯ are in the Latin Extended-B Chart. To make matters worse essential punctuation mark ʹ is in a more remote location, in the spacing modifier letters. This limits the amount of typefaces covering the language drastically reducing their typographic choice to a few user interface typefaces which design is specific for this use alone. Meaning they do not have typefaces for children's books, newspapers, magazines, etc.

a â b c č ʒ ǯ d đ f g ǧ g h i j k ǩ
l m n ŋ o õ p r s š t u v z ž å ä ʹ

Glyphs on Latin B unicode chart

Glyphs on spacing modifier unicode chart

Skolt Sámi alphabet

Evidently this little attention paid to so-called minority languages[5] is due mostly to economic reasons. Software developers, phone makers, computer manufacturers, etc. have no interest in satisfying the needs of what for them is just a small niche in a global market. It is then the responsibility of the Unicode Consortium, to fully represent all languages without discriminating minorities. It is not tolerable that a smiley face (or any other unnecessary or little important dingbat) has a code, and glyphs needed to help survive endangered languages do not have one. It shows either a lack of respect, ignorance or conscious discriminating policy on their part. On the other hand, type designers and foundries have also a part of responsibility when deciding character sets and when designing glyphs which they are not native users. Designers need to do research on unknown characters and diacritics, if possible consulting native users. We also need to stop following blindly Unicode Charts. Why not to ignore charts and misleading nomenclature in order to fully cover languages which need our compromise regardless of commercial intrests?

www.letterjuice.cat