[9] G.A. Kubba. The Impact of Computers on Arabic Writing, Character Processing, and Teaching. *Information Processing*, 80:961–965, 1980.

[10] Pierre Mackay. Typesetting Problem Scripts. *Byte*, 11(2):201–218, February 1986.

[11] J. Marshall Unger. *The Fifth Generation Fallacy—Why Japan is Betting its Future on Artificial Intelligence.* Oxford University Press, 1987.

[12] X/Open Company, Ltd. *X/Open Portability Guide, Supplementary Definitions*, volume 3. Prentice-Hall, 1989.

◇ Nelson H.F. Beebe
Center for Scientific Computing
and Department of
Mathematics
South Physics Building
University of Utah
Salt Lake City, UT 84112
USA
Tel: (801) 581-5254
Internet: Beebe@science.utah.edu

On Standards for Computer Modern Font Extensions

Janusz S. Bień

Abstract

Haralambous' proposal to standardize the unused part of Computer Modern fonts is discussed, and some modifications and extensions suggested. The idea is pursued by designing the extended CM font layout, and an example is given for one of its possible uses.

1 Introduction

In my note [4] I advocated an old ([15, p. 46], [6, p. 45]) but rarely used idea to place national letters (actually, the Polish ones, but the generalization is obvious) in the unused part of Computer Modern fonts, i.e. as the characters with the codes higher than 127; this approach allows the handling of national languages in a way upward compatible with the standard (American) English TEX. A similar proposal was made independently by Yannis Haralambous [8], who states also that the use of non-English letters of latin alphabets should be coordinated, resulting in a single widely used extension to Computer Modern fonts — I strongly support the principal idea, and I pursue it in the present paper. To organize the discussion in a systematic way, I will use the notions — borrowed from [2] — of text *encoding*, *typing* and *rendering.*

2 Text encoding

In the context of TEX, *encoding* means the character sets of the fonts in question and their layouts. In the present section I will focus my attention on the character sets, as the layouts should be influenced, among others, by *typing* considerations.

In an attempt to obtain a general idea about the use of the latin alphabet worldwide, I looked up the only relevant reference work I am aware of, namely *Languages Identification Guide* [7] (hereafter *LIG*). Apart from the latin scripts used in the Soviet Union and later replaced by Cyrillic ones, it lists 82 languages using the latin alphabet with additional letters (I preserve the original spelling):

> Albanian, Aymara, Basque, Breton, Bui, Catalan, Choctaw, Chuana, Cree, Czech, Danish, Delaware, Dutch, Eskimo, Esperanto, Estonian, Ewe, Faroese (also spelled Faroeish), Fiji, Finnish, French, Frisian, Fulbe, German, Guarani, Hausa, Hungarian, Icelandic, Irish, Italian, Javanese, Juang, Kasubian, Kurdish, Lahu, Lahuli, Latin, Lettish, Lingala, Lithuanian, Lisu, Luba, Madura, Miao, Malagash, Malay, Mandingo, Minankabaw, Mohawk, Mossi, Navaho, Norwegian, Occidental, Ojibway (also spelled Ojibwe), Polish, Portuguese, Quechua, Rhaeto-Romanic (Ladin, Romansh), Rumanian, Samoan, Seneca, Serbo-Croatian, Sioux, Slovak, Slovene, Spanish, Suto, Sundanese, Swahili, Swedish, Tagalog, Turkish, Uolio, Vietnamese, Volapük, Welsh, Wolof, Y, Yoruba, Zulu.

This list includes some languages and dialects with no script at all, for which the information supplied concerns more or less standard transcription. For most of them this fact is noted explicitly, but the exception of Kasubian (usually recognized as a dialect of Polish) suggests that this is not always the case. I noticed some inconsistencies in the numerous indexes to the book, but only one omission (described later) in the proper text. Of course, it is difficult for me to judge the reliability of the work as a whole.

The number of additional letters in the latin alphabets listed in *LIG* — including some variants of shape but excluding upper case letters — is 176.

Hence the total number of lower and upper case letters is definitely over 300. The possible errors and omissions cannot change this estimate significantly, so in general we have to cope with the number of additional letters substantially exceeding the number of free slots in the Computer Modern fonts.

My solution to this problem is to postulate two levels of standards:

Extended Computer Modern fonts, with a small number of slots unassigned.

Full Extended Computer Modern fonts, i.e. national or regional fonts compatible with Extended CM fonts, but having some additional characters assigned.

Of course, both of them will include all the characters of the original CM fonts in their proper places; although teletypewriter layout fonts are much less used, our standards should take them into account, too.

It should be noted now that there are numerous national and international standards for text encoding. The most relevant for us is the ISO 6937 international standard ([12], [13], [14]), described thoroughly in [25] and discussed in [24]. Annex D to the standard [13] is entitled *Use of Latin alphabetic characters*; formally it is not part of the standard, but its goal is to provide

> justification for the composition of the alphabetic part of the graphic character repertoire. It does not attempt to define which characters should, and which ones should not, be used in any language.

The annex contains a table (quoted in [25]) listing the following languages (I again preserve the original spelling):

> Albanian, Basque, Breton, Catalan, Croat, Czech, Danish, Dutch, English, Estonian, Faroese, Finnish, French, Frisian, Galician, German, Greenlandic, Hungarian, Icelandic, Irish, Italian, Lapp, Latvian, Lithuanian, Maltese, Norwegian, Occitan, Polish, Portuguese, Rhaeto-Romanic, Romanian, Scots Gaelic, Slovak, Slovene, Sorbian, Spanish, Swedish, Turkish, Welsh, Afrikaans, Esperanto.

With the exception of the last 2 languages, the list contains 39 living European languages. However, despite the quoted reservation, it seems rather strange that, according to the table, English uses 28 additional letters (namely á Á, à À, æ Æ, ç Ç, é É, è È, ê Ê, ë Ë, î Î, ï Ï, ñ Ñ, ô Ô, ö Ö, œ Œ). The standard associates with all the characters their unique identifications (explained in Annex A to the standard [13]) and names; I will use these names in the sequel when appropriate.

The ISO 6937 character set includes 87 additional letters which exist in both lower and upper case form, 6 letters which have only lower case form and 2 letters which have only upper case form. Additionally, 3 lower case letters and 1 upper case letter have shape variants (I refer here to the shapes of the letters, not to their function in specific languages; although e.g. in Lapp and Latvian Ģ is the upper case equivalent of ǵ, I count them as having no case counterparts). This gives us the total of 186 additional letters. Although 10 of them are already included in the original CM fonts (namely æ Æ, ı, ł Ł, œ Œ, ø Ø, ß), again the number of additional characters exceeds the number of free slots. Moreover, we should not forget the problem of the missing punctuation marks. The most demanded ones seem to be the *angle quotation marks* («,») used also e.g. in French, German and Polish, the “continental” left quotation mark („) used e.g. in German and Polish, and perhaps the German right quotation mark (“); cf. [6], [21], [18].

Let us have now a closer look at the character set proposed by Haralambous. To understand fully its implications, let us discuss first the language list contained in [8]. The ISO standard and *LIG* confirm consistently only 8 items:

Croat (spelled Croatian in [8] and [7]): ć Ć, č Č, đ Đ, š Š, ž Ž.

Hungarian: á Á, é É, í Í, ó Ó, ö Ö, ő Ő, ú Ú, ü Ü, ű Ű.

Polish (in addition, I vouch for its correctness personally): ą Ą, ć Ć, ę Ę, ł Ł, ń Ń, ó Ó, ś Ś, ź Ź, ż Ż.

Romanian (spelled Rumanian in [7]): â Â, ă Ă, î Î, ş Ş, ţ Ţ.

Slovene (spelled Slovenian in [7] and [8]): č Č, š Š, ž Ž.

Spanish: á Á, é É, í Í, ñ Ñ, ó Ó, ú Ú, ü Ü.

Turkish: â Â, ç Ç, ğ Ğ, ı I, i İ, î Î, ö Ö, ş Ş, û Û, ü Ü.

In the case of 7 languages my sources consistently disagree with Haralambous' list:

Albanian. There is *c with cedilla* (ç Ç) instead of *c with caron* (č Č).

Catalan. There is the additional letter *i with diaeresis* (ï); according to ISO 6937, there is an additional letter *l with middle dot*, while *LIG* states

Two successive letters l which do not denote one sound are separated by a point **l.l** (or **l·l**).

Czech. The letter d' is treated as variant of ď; both of them are called in ISO 6937 *small d with caron*; the same holds respectively for t'. *LIG* distinguishes also a variant of ď differing in the placement of the caron. For upper case letters both sources list only Ď and Ť (neither D' nor T').

Faroese (in [7] often spelled Faroeish). Instead of *small d with stroke* and *small thorn* there should be *small eth* (*ð*)[1] and *capital D with stroke* (i.e. the capital eth *Đ*).

Icelandic. Instead of *small d with stroke* there should be *capital thorn* (*Þ*).

Irish. Besides its own alphabet, the language uses the latin script with the following additional letters: á Á, é É, í Í, ó Ó, ú Ú.

Lithuanian. It uses *ogonek* instead of *cedilla*, so there is e.g. ą and ę instead of a̧ and ȩ, etc.

For the remaining 28 languages, 9 languages are not accounted for in the ISO 6937 standard (Corsican, Creole, Gaelic, Guarani, Indonesian, Kurdish, Latin, Qhëshwa, Vietnamese) and 7 languages are not listed in *LIG* (Corsican, Creole, Gaelic, Galician, Maltese, Occitan, Qhëshwa); however, some languages may be called by different names (I happen to know that Latvian is Lettish, but is Scots Gaelic different from Gaelic, is Qhëshwa different from Quechua?). For the rest of them both my sources more or less disagree. Fortunately, with the exception of Slovak and Vietnamese, the differences concern the use of accented letters by specific languages and do not affect the character set itself.

For Slovak (spelled Slovakian by Haralambous), the problem concerns the letter *l with acute accent* included in the ISO standard but not listed at all in *LIG*, and the letter *l with caron*, listed in *LIG* (and by Haralambous) only in its variant shape (l'). I consulted an original Slovak grammar [22], which confirms the existence of ĺ and Ĺ and lists *l with caron* only in the form l' L'.

As for Vietnamese, *LIG* (and also some books published in Poland) uses o' and u' instead of ơ and ư (o and u "with beard"), listed not only by Haralambous but also in [26]; on the other hand, there is no doubt about the correct shape of the accent called question mark in [26], which is given by Haralambous in a simplified form. I intend to consult an expert on this matter (I suspect different usage in North and South Vietnam), but his answer is not relevant for our further discussion — anyway, the Vietnamese letters and accents should be included in a specific Full Extended CM font, not in the Extended CM font.

In my opinion, the Extended CM fonts should contain the following additional letters:

- small and capital *a* with acute (á Á), grave (à À) and circumflex (â Â) accent, with diaeresis (ä Ä), tilde (ã Ã), ring (å Å) and ogonek (ą Ą),
- small and capital *c* with acute (ć Ć) accent, with cedilla (ç Ç) and with caron (č Č),
- small and capital *d* with caron (ď Ď) and with stroke (đ, Đ),
- small *eth* (ð), small and capital *thorn* (þ Þ),
- small and capital *e* with acute (é É), grave (è È) and circumflex (ê Ê) accent, with diaeresis (ë Ë) and ogonek (ę Ę),
- small and capital *g* with breve (ğ Ğ),
- small and capital *i* with acute (í Í), grave (ì Ì) and circumflex (î Î) accent, with diaeresis (ï Ï) and caron (ǐ Ǐ), and capital *I* with dot above (İ),
- small and capital *l* with acute accent (ĺ Ĺ), with caron (ľ Ľ) and with stroke (ł Ł),
- small and capital *n* with acute accent (ń Ń), with tilde (ñ Ñ) and caron (ň Ň),
- small and capital *o* with acute (ó Ó), grave (ò Ò) and circumflex (ô Ô) accent, with diaeresis (ö Ö) and caron (ǒ Ǒ), and with double acute accent (ő Ő),
- small and capital *r* with caron (ř Ř),
- small and capital *s* with acute (ś Ś) accent, with cedilla (ş Ş) and with caron (š Š),
- small and capital *t* with cedilla (ţ Ţ) and with caron (ť Ť),
- small and capital *u* with acute (ú Ú), grave (ù Ù) and circumflex (û Û) accent, with diaeresis (ü Ü), ring (ů Ů) and with double acute accent (ű Ű),
- small and capital *y* with acute (ý Ý) accent and with diaeresis (ÿ Ÿ),
- small and capital *z* with acute (ź Ź) accent, with caron (ž Ž) and with dot above (ż Ż).

and the following additional punctuation marks:

- the left and right angle quotation marks (« »),
- the "continental" left quotation mark („),
- the German right quotation mark (“).

[1] The editors thank Jörgen Pind for supplying his METAFONT sources (see also [23]) to create the eths and thorns in this article.

The proposed character set thus contains 112 additional letters and 4 additional punctuation marks. It includes the Polish letters ł Ł, already present in some CM fonts, because they are needed also in the fonts with the teletypewriter layout (I follow Haralambous in this respect).

The Extended Computer Modern font leaves 12 slots to be assigned in the regional or national Full Extended CM fonts (in particular, for Vietnamese).

3 Text typing

In my note [4] I advocated a novel idea (at least at that time — now cf. [27, p. 335]) to use several `tfm` files to access the same font for different purposes — a Polish font with the layout upward compatible with the original CM font can be accessed by the original `tfm` for standard work, and by a special `tfm` file for typesetting Polish texts. In my opinion, this approach should be applied to the multilingual fonts discussed here — they should be offered with many `tfm` files tailored for specific regions, nations and languages. Therefore in the sequel I will limit my attention to the *default* `tfm` files for Extended CM fonts.

In general, the typing considerations have two aspects

- echo problem,
- sorting problem.

By the echo problem I mean the typing feedback — can the user pressing a key on the keyboard see the proper character shape on the screen without resorting to the graphic mode? As for the sorting problem, many people are not aware that the alphabetic ordering is language dependent, and that it can differ substantially from one language to another. Of course, TEX users are first of all interested in sorting by various TEX utilities, such as BIBTEX or MakeIndex. I hope that the re-implementation of LATEX proposed in [20] will be accompanied by the universal versions of these programs, allowing the sorting algorithm to be controlled by appropriate parameters.

Unfortunately, the echo problem is not an internal affair of the TEX community, but a general problem heavily dependent on hardware and operating systems. As mentioned in [9], over half of TEX users work on IBM compatible computers, so it would not be wise to ignore what IBM intends to do in this domain. Therefore I have done my best to collect the tables of the so called *code pages* designed by IBM (or with its approval).

In [19] I found the following tables:

1. Code page 437 — United States,
2. Code page 850 — Multilingual,
3. Code page 860 — Portuguese,
4. Code page 863 — French-Canadian,
5. Code page 865 — Nordic.

Surprisingly enough, there were mistakes in the tables; I managed to correct them by consulting other sources.

In [10] I found, apart from Cyrillic, the following page

1. Code page 852 — Multilingual Group 2.

In [11] I found, apart from Cyrillic and 22 EBCDIC-based pages, the following code pages (for the curious reader I include also non-latin scripts):

1. Code page 838 — Latin #5, Thailand,
2. Code page 850 — Multinational,
3. Code page 851 — Greece,
4. Code page 857 — Latin #5, Turkey,
5. Code page 860 — Portugal,
6. Code page 861 — Iceland,
7. Code page 862 — Israel,
8. Code page 863 — Canadian French,
9. Code page 864 — Arabic,
10. Code page 865 — Nordic,
11. Code page 891 — Korea,
12. Code page 897 — Japan #1,
13. Code page 903 — Peoples Republic of China (PRC),
14. Code page 904 — Republic of China (ROC).

As you can see, the page names differ slightly in various documents.

My goal was to design the layout of Extended CM fonts in a way as compatible as possible with the above listed code pages. I think that seeing on the screen — instead of a letter — a non-letter character is less confusing than seeing a wrong letter; therefore I looked first of all for those letters which appear in at least two code pages and which conflict only with some non-letter characters. I found 8 such letters, and I included them in the font on the positions identified by their codes in the code pages (the octal values are given in parentheses):

small A with circumflex accent (â) 131 (*'203*),
small c with cedilla (ç) 135 (*'207*),
capital C with cedilla (Ç) 128 (*'200*),
small e with acute accent (é) 130 (*'202*),
small o with acute accent (ó) 162 (*'242*),
small o with circumflex accent (ô) 147 (*'223*),
small u with diaeresis (ü) 129 (*'201*),
capital U with diaeresis (Ü) 154 (*'232*).

I decided also to prefer those code pages which are provided now with MS-DOS and PC-DOS, namely the pages 437 and 850. So the second step was to include those letters which occur in both of them, and those which occur in page 850 and are in conflict only with non-letter characters in page 437. It resulted in the following 49 assignments.

small a with acute accent (á) 160 (*'240*),
capital A with acute accent (Á) 181 (*'265*),
small a with grave accent (à) 133 (*'205*),
capital A with grave accent (À) 183 (*'267*),
capital A with circumflex accent (Â) 182 (*'266*),
small a with diaeresis (ä) 132 (*'204*),
capital A with diaeresis (Ä) 142 (*'216*),
small a with tilde (ã) 198 (*'306*),
capital A with tilde (Ã) 199 (*'307*),
small a with ring (å) 134 (*'206*),
capital A with ring (Å) 143 (*'217*),
small eth (ð) 208 (*'320*),
capital D with stroke (Đ) 209 (*'321*),
small thorn (þ) 232 (*'350*),
capital thorn (Þ) 231 (*'347*),
capital E with acute accent (É) 144 (*'220*),
small e with grave accent (è) 138 (*'212*),
capital E with grave accent (È) 212 (*'324*),
small e with circumflex accent (ê) 136 (*'210*),
capital E with circumflex accent (Ê) 210 (*'322*),
small e with diaeresis (ë) 137 (*'211*),
capital E with diaeresis (Ë) 211 (*'323*),
small i with acute accent (í) 161 (*'241*),
capital I with acute accent (Í) 214 (*'326*),
small i with grave accent (ì) 141 (*'215*),
capital I with grave accent (Ì) 222 (*'336*),
small i with circumflex accent (î) 140 (*'214*),
capital I with circumflex accent (Î) 215 (*'327*),
small i with diaeresis (ï) 139 (*'213*),
capital I with diaeresis (Ï) 216 (*'330*),
small n with tilde (ñ) 164 (*'244*),
capital N with tilde (Ñ) 165 (*'245*),
capital O with acute accent (Ó) 224 (*'340*),
small o with grave accent (ò) 149 (*'225*),
capital O with grave accent (Ò) 227 (*'343*),
capital O with circumflex accent (Ô) 226 (*'342*),
small o with diaeresis (ö) 148 (*'224*),
capital O with diaeresis (Ö) 153 (*'231*),
small o with caron (ǒ) 228 (*'344*),
capital O with caron (Ǒ) 229 (*'345*),
small u with acute accent (ú) 163 (*'243*),
capital U with acute accent (Ú) 233 (*'351*),
small u with grave accent (ù) 151 (*'227*),
capital U with grave accent (Ù) 235 (*'353*),
small u with circumflex accent (û) 150 (*'226*),
capital U with circumflex accent (Û) 234 (*'352*),
small y with acute accent (ý) 236 (*'354*),
capital Y with acute accent (Ý) 237 (*'355*),
small y with diaeresis (ÿ) 152 (*'230*).

This rule applies also to the punctuation marks:

left angle quotation mark («) 174 (*'256*),
right angle quotation mark (») 175 (*'257*).

The next step was to transfer to our font the letters included only in the second multinational page, namely 852, and not in conflict with some letter in other pages, i.e. the following letters:

capital N with caron (Ň) 213 (*'325*),
small r with caron (ř) 253 (*'375*),
capital R with caron (Ř) 252 (*'374*),
capital S with cedilla (Ş) 184 (*'270*),
capital S with caron (Š) 230 (*'346*),
small t with cedilla (ţ) 238 (*'356*),
capital T with cedilla (Ţ) 221 (*'335*),
small t with caron (ť) 156 (*'234*),
small u with double acute accent (ű) 251 (*'373*),
small z with acute accent (ź) 171 (*'253*),
small z with dot above (ż) 190 (*'276*),
capital Z with dot above (Ż) 189 (*'275*).

By this time we have filled in 71 slots in the font; 12 slots are to be left free and 45 characters are still to be assigned. It is the right moment to concentrate on the free slots. I decided to leave free the positions 145 (*'221*), 146 (*'222*), 155 (*'233*) and 157 (*'235*), because in the most used page, 437, they contained the characters æ Æ ø Ø, which can be useful for many TeX users. For similar reasons I left free the position 225 (*'341*), which in the two popular pages 850 and 852 (and also 857) contain the character ß. I decided also to leave free the positions 159 (*'237*), 166 (*'246*), 167 (*'247*), 168 (*'250*), 169 (*'251*), 172 (*'254*) and 173 (*'255*), because I see no simple criterion for solving the letter conflicts among the code pages. There are also serious conflicts on the positions 158 and 170, so I decided to devote them to the punctuation marks:

the "continental" left quotation mark („) 158 (*'236*),
the German right quotation mark (“) 170 (*'252*).

The remaining 43 characters have been assigned in an arbitrary way:

small a with ogonek (ą) 176 (*'260*),
capital A with ogonek (Ą) 177 (*'261*),
small a with breve (ă) 178 (*'262*),
capital A with breve (Ă) 179 (*'263*),
small c with acute accent (ć) 180 (*'264*),
capital C with acute accent (Ć) 185 (*'271*),

small c with caron (č) 186 (*'272*),
capital C with caron (Č) 187 (*'273*),
small d with caron (ď) 188 (*'274*),
capital D with caron (Ď) 191 (*'277*),
small d with stroke (đ) 192 (*'300*),
small e with ogonek (ę) 193 (*'301*),
capital E with ogonek (Ę) 194 (*'302*),
small e with caron (ě) 195 (*'303*),
capital E with caron (Ě) 196 (*'304*),
small g with caron (ǧ) 197 (*'305*),
capital G with caron (Ǧ) 200 (*'310*),
small i with caron (ǐ) 201 (*'311*)
capital I with caron (Ǐ) 202 (*'312*),
capital I with dot above (İ) 203 (*'313*),
small l with caron (ľ) 204 (*'314*),
capital L with caron (Ľ) 205 (*'315*),
small l with acute accent (ĺ) 206 (*'316*),
capital L with acute accent (Ĺ) 207 (*'317*),
small l with stroke (ł) 217 (*'331*),
capital L with stroke (Ł) 218 (*'332*),
small n with acute accent (ń) 219 (*'333*),
capital N with acute accent (Ń) 220 (*'334*),
small n with caron (ň) 223 (*'337*),
small o with double acute accent (ő) 239 (*'357*),
capital O with double acute accent (Ő) 240 (*'360*),
small s with acute accent (ś) 241 (*'361*),
capital S with acute accent (Ś) 242 (*'362*),
small s with cedilla (ş) 243 (*'363*),
small s with caron (š) 244 (*'364*),
capital T with caron (Ť) 245 (*'365*),
small u with ring (ů) 246 (*'366*),
capital U with ring (Ů) 247 (*'367*),
capital U with double acute accent (Ű) 248 (*'370*),
capital Y with diaeresis (Ÿ) 249 (*'371*),
capital Z with acute accent (Ź) 250 (*'372*),
small z with caron (ž) 254 (*'376*),
capital Z with caron (Ž) 255 (*'377*).

Editor's note: The encoding scheme above is presented in a font layout on p. 183.

The default `tfm` files for Extended CM fonts for use with 8-bit TEX should not contain any ligatures except those needed for kerning or inherited from the original CM fonts. However, for 7-bit TEX another *default* `tfm` scheme is to be designed, because in it, ligatures are the only way to access the second half of the fonts without disturbing the hyphenation. I would like to advocate here another idea from my note [4], consisting in using the character with the code 32 (the stroke for the Polish ł) as a part of the ligatures accessing the national letters. The idea is further developed here in two respects:

- The ligatures in question should consist of a letter *followed* by the character 32. The reason is that such representation of national letters affects the alphabetic ordering in a less substantial way and, under some additional conditions, can even preserve the ordering for some languages.
- There should be a general rule saying that the ligature composed of a character with the code x followed by the character with the code 32 accesses the character with the code $x + 128$. The rule can be called a 7-bit equivalent of the double circumflex notation [16, p. 325].

Of course, the character 32 is not directly accessible, because it coincides with the space character in the ASCII code. However, it can be easily assigned to any active character. On the other hand, to preserve the compatibility in case of the teletype layout fonts, the macro for the visible space has to be changed.

I think that the language specific `tfm` files are especially useful for 7-bit TEX. My experience with typesetting Russian texts using the AMS Cyrillic fonts showed that sophisticated multipurpose ligature tables are more a nuisance than a real help. In consequence, Haralambous' ligatures can be accepted only as one of several alternative `tfm` files, and not as a general standard.

4 Text rendering

In the context of TEX, *rendering* means the actual fonts used by the device drivers. Again, in my opinion, there should be a *default* METAFONT definition, not the standard one. First, I am not sure that e.g. French *capital A with acute accent* looks the same as the Hungarian one (my impression — maybe wrong — is that they differ substantially). Secondly, I do not know whether such problems as the actual shape of e.g. Czech *d with caron* can be solved definitively; perhaps both versions are to be used depending on the situation.

Last but not least, it should be remembered that some letters use up the font space only for hyphenation purposes — even in TEX 3.0 an accented letter (i.e. constructed by the `\accent` command) disables hyphenation until the next glue. Sooner or later a standard for *virtual fonts* — i.e. for creating new characters from the elements already present in the fonts — will emerge as a part of the standarization of the device drivers. One of the first virtual font mechanisms was mentioned in the Editor's comment to Haralambous' paper ([8, p. 342]), but the idea of "fooling" the TEX program can be traced down at least to Appelt [1]. Incidentally, the term *virtual fonts* is used in the context of Beebe's drivers in a totally different sense — cf. the 'a' parameter ([3, p. 3]); I hope this confusing use will soon be abandoned.

5 An example

Let us imagine an IBM PC computer equipped with the code page 852 character set (supported by IBM on the Polish market and accepted by some state-owned manufacturers), used to typeset Polish texts with 8-bit TeX and the proposed Extended CM fonts. There are 18 Polish national letters, and only for 4 of them their codes coincide in the code page and the proposed layout. In consequence, some kind of translation is needed for the remaining 14 letters (such a compromise seems necessary to make the proposal acceptable by the users of other languages).

Assuming that the fonts have been set up correctly (by assigning to their characters the proper values of `\catcode`, `\lccode`, `\uccode`, `\sfcode`, `\mathcode` and `\delcode`), the following definitions are sufficient for the compatibility of the echo (when working with a standard 8-bit editor) with the font layout.

```
% 165 small a with ogonek
\catcode^^a5=\active\chardef^^a5=176
% 164 capital A with ogonek
\catcode^^a4=\active\chardef^^a4=177
% 134 small c with acute accent
\catcode^^86=\active\chardef^^86=180
% 143 capital C with acute accent
\catcode^^8f=\active\chardef^^8f=185
% 169 small e with ogonek
\catcode^^a9=\active\chardef^^a9=193
% 168 capital E with ogonek
\catcode^^a8=\active\chardef^^a8=194
% 136 small l with stroke
\catcode^^88=\active\chardef^^88=217
% 157 capital L with stroke
\catcode^^9d=\active\chardef^^9d=218
% 228 small n with acute accent
\catcode^^e4=\active\chardef^^e4=219
% 227 capital N with acute accent
\catcode^^e3=\active\chardef^^e3=220
% 224 capital O with acute accent
\catcode^^e0=\active\chardef^^e0=224
% 152 small s with acute accent
\catcode^^98=\active\chardef^^98=241
% 151 capital S with acute accent
\catcode^^97=\active\chardef^^97=242
% 141 capital Z with acute accent
\catcode^^8d=\active\chardef^^8d=250
% no translation needed for
% 162 small o with acute accent
% 171 small z with acute accent
% 190 small z with dot above
% 189 capital z with dot above
```

After changing the representation of Polish letters in the hyphenation patterns [17], the Polish hyphenation algorithm will operate with no problems.

As for 7-bit TeX, using directly the default 7-bit `tfm` would make the input text completely unintelligible. However, it is not difficult to create a convenient interface, either by means of macro definitions similar to those quoted in [18, p. 5] and [5], or by introducing a special Polish `tfm` file with appropriate ligatures.

In both cases the explicit use of national letters (i.e. echoed on the screen in a reasonable way) in control sequences is severely limited. Unfortunately, we have to live with it till the next change in TeX.

6 Concluding remarks

For a standard to be widely accepted, it has to be fully adequate to actual needs — neither too general nor too specific. I hope that my modifications and extensions of Haralambous' proposal achieve the proper balance.

It should be also noted that a substantial part of actual and potential TeX users who will be affected by the standards are not yet organized into users groups; moreover, most of them have no access to electronic mail. If the standard is to be developed — as proposed by Haralambous — in a democratic way, then the traditional forms of communication should be the primary medium.

References

[1] Wolfgang Appelt. The Hyphenation of Non-English Words with TeX. In Dario Lucarella, editor, *Proceedings of the First European Conference on TeX for Scientific Documentation*, Addison-Wesley, Reading, Massachusetts, 1985, pp. 61–65.

[2] Joseph D. Becker. Multilingual Word Processing. *Scientific American* Vol. 251 No. 1 (July 1984), pp. 82–93.

[3] [Nelson H. F. Beebe]. DVIxxx — Display TeX DVI Files on Assorted Output Devices. Beebe's driver distribution version 2.10.

[4] Janusz S. Bień. Polish Language and TeX. *TeXline* 8 (January 1989), p. 2.

[5] Janusz S. Bień. Co to jest TeX? [What is TeX? In Polish]. *Wiadomości Matematyczne* Vol. 29 No. 1 (to appear).

[6] Jacques Désarménien. The Use of TeX in French: Hyphenation and Typography. In

Dario Lucarella, editor, *Proceedings of the First European Conference on TEX for Scientific Documentation*, Addison-Wesley, Reading, Massachusetts, 1985, pp. 41–59.

[7] R. S. Gilyare and V. S. Grivnin. *Languages Identification Guide.* "NAUKA" Publishing House, Central Department of Oriental Literature, Moscow 1970.

[8] Yannis Haralambous. TEX and latin alphabet languages. *TUGboat* Vol. 10 No. 3 (November 1989), pp. 342–345.

[9] Don Hosek. Guidelines for creating portable METAFONT code. *TUGboat* Vol. 10 No. 2 (July 1989), pp. 173–176.

[10] IBM Corporation. *Personal System/2 Natural Language Supplement.* First edition (February 1988) 07F3226.

[11] IBM Corporation. *Application System/400 Natural Language Support: User's Guide.* First edition (September 1989) GC21-9877-0.

[12] International Organization for Standardization. *Information processing. Coded character sets for text communication — Part 1: General Introduction.* First edition 1983-11-01. Ref. No. ISO 6937/1-1983(E).

[13] International Organization for Standardization. *Information processing. Coded character sets for text communication — Part 2: Latin alphabetic and non-alphabetic graphic characters.* First edition 1983-12-15. Ref. No. ISO 6937/2-1983(E).

[14] International Organization for Standardization. *Information processing. Coded character sets for text communication — Part 2: Latin alphabetic and non-alphabetic graphic characters.* Addendum 1, 1989-05-01. Ref. No. ISO 6937-2-1983/Add1:1989(E).

[15] Donald E. Knuth. *The TEXbook.* Addison-Wesley, Reading, Massachusetts, 1984.

[16] Donald E. Knuth. The new versions of TEX and METAFONT. *TUGboat* Vol. 10 No. 3 (November 1989), pp. 325–328.

[17] Hanna Kołodziejska. Dzielenie wyrazów polskich w systemie TEX [Polish hyphenation patterns for TEX; in Polish]. IInf UW Report 165, Institute of Informatics, Warsaw University, 1987.

[18] Hanna Kołodziejska. Le traitement des textes polonais avec le logiciel TEX. *Cahiers GUTenberg* Numéro zéro (Avril 1988), pp. 3–10.

[19] Microsoft Corporation. *MS-DOS User's Guide and User's Reference* [Version 3.3]. Doc. No. M5123-8806B.

[20] Frank Mittelbach and Rainer Schöpf. With LATEX into the Nineties. *TUGboat* Vol. 10 No. 4 (December 1989), pp. 681–690.

[21] Hubert Partl. German TEX. *TUGboat* Vol. 9 No. 1 (April 1988), pp. 70–72.

[22] Eugen Pauliny. *Krátka gramatika slovenská.* Slovenské Pedagogické Nakladatel'stvo, Bratislava 1963.

[23] Jörgen L. Pind. Lexicography with TEX. *TUGboat* Vol. 10 No. 4 (December 1989), pp. 655–665.

[24] Staffan Romberger and Yngve Sundblad. Adapting TEX to languages that use Latin alphabetic characters. In Dario Lucarella, editor, *Proceedings of the First European Conference on TEX for Scientific Documentation*, Addison-Wesley, Reading, Massachusetts, 1985, pp. 27–40.

[25] Joan M. Smith. Transmitting Text: A Standard Way of Communicated Characters (Part 1). *Association for Literary and Linguistic Computing Bulletin* Vol. 12 (1983) No. 2, pp. 11–38.

[26] Eric Vogel. Printing Vietnamese characters by adding diacritical marks via TEX. *TUGboat* Vol. 10 No. 2 (July 1989), pp. 217–221.

[27] Dimitri Vulis. Notes on Russian TEX. *TUGboat* Vol. 10 No. 3 (November 1989), pp. 332–336.

◇ Janusz S. Bień
Institute of Informatics
Warsaw University
PKiN p.850
00-901 Warszawa, Poland

	´0	´1	´2	´3	´4	´5	´6	´7	
´20x	Ç	ü	é	â	ä	à	å	ç	"8x
´21x	ê	ë	è	ï	î	ì	Ä	Å	
´22x	É	free	free	ô	ö	ò	û	ù	"9x
´23x	ÿ	Ö	Ü	free	ť	free	„	free	
´24x	á	í	ó	ú	ñ	Ñ	free	free	"Ax
´25x	free	free	“	ź	free	free	«	»	
´26x	ą	Ą	ă	Ă	ć	Á	Â	À	"Bx
´27x	Ş	Ć	č	Č	ď	Ż	ż	Ď	
´30x	đ	ę	Ę	ě	Ě	ğ	ã	Ã	"Cx
´31x	Ğ	ĭ	Ĭ	İ	ľ	Ľ	ĺ	Ĺ	
´32x	ð	Đ	Ê	Ë	È	Ň	Í	Î	"Dx
´33x	Ï	ł	Ł	ń	Ń	Ţ	Ì	ň	
´34x	Ó	free	Ô	Ò	ŏ	Ŏ	Š	Þ	"Ex
´35x	þ	Ú	Û	Ù	ý	Ý	ţ	ő	
´36x	Ő	ś	Ś	ş	š	Ť	ů	Ů	"Fx
´37x	Ű	Ÿ	Ź	ű	Ř	ř	ž	Ž	
	"8	"9	"A	"B	"C	"D	"E	"F	

The layout of the proposed CM font extensions

Janusz S. Bień

Fonts

Circular Reasoning: Typesetting on a Circle, and Related Issues

Alan Hoenig

Owing to the generality of both TEX and METAFONT, it's easy to typeset in and on circles. Here's how.

The METAFONT Part

TEX can't actually turn characters on their side; we ask METAFONT to create special fonts where each character in the font is rotated around its reference point (the lower left corner of the bounding box of any character). Then TEX properly positions characters from the rotated fonts to achieve the illusion of circular typesetting. We need one rotated font for each position on the circle.

What does it mean to typeset characters around the circumference of a circle? I imagined a regular

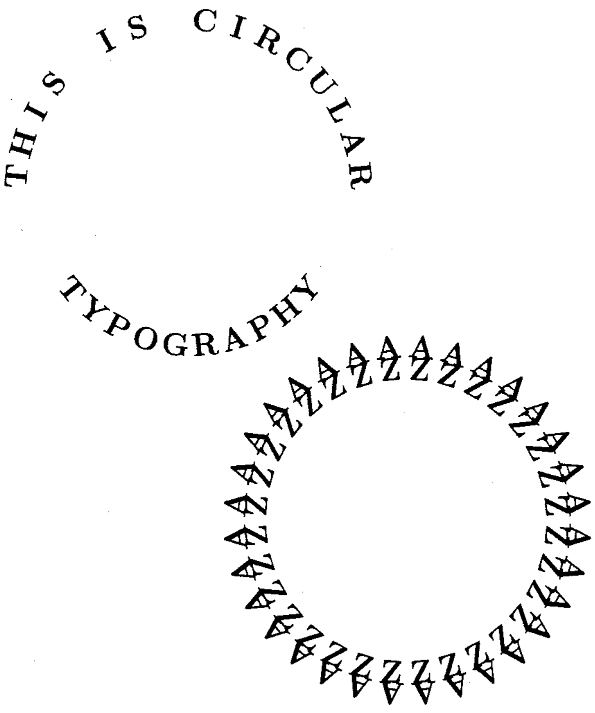

Figure 1. What this article is about.