On the localization of TEX in Hungary

Gyöngyi Bujdosó
Department of Computer Graphics and Library and Information Science
Institute of Mathematics and Informatics
University of Debrecen
H-4010 Debrecen, P.O.B. 12
Hungary
ludens@math.klte.hu

Ferenc Wettl
Department of Algebra
Institute of Mathematics
Budapest University of Technology and Economics
Budapest, Műegyetem rakpart 1–3, H. ép. V. 5.
Hungary
wettl@math.bme.hu

Abstract

This paper deals with the present and future of the localization of TEX in Hungary. The authors review some of the necessary tools for preparation Hungarian documents, and especially the improvements needed to make TEX more usable in Hungary. Some of the work has been done, and a short "to do" list will be presented for work to be done in the near future. The problems stemming from the specialities of Hungarian grammar (e.g., hyphenation, handling definite articles and suffixes) will be considered as well as the tasks implied by the heritage of the Hungarian typography (e.g., page layout).

Introduction

The Hungarian Users Group (called MATEX) exists formally since the end of 2001. As the very first activity, we have resumed the localization of (LA)TEX for the Hungarian language. We have started getting together all the developments in the domain of localization, we are looking for the problems which have not been solved, and trying to organize teams for finding out the answers.

This paper deals with the developments which have been achieved and with our aims for the near and distant future.

Grammar

There are some specialties in the Hungarian language which might be interesting in connection with TEX or generally with document preparation. We concentrate on the problems of generated texts.

Definite articles In Hungarian there are two definite articles, 'a' and 'az'. 'a' is used before words beginning with a consonant, and 'az' is used before words beginning with a vowel, just like 'a/an' in English. If any generated text needs an article, it must also be generated. This is the situation with `\ref`, `\pageref`, and `\cite` in LATEX. The babel package nicely solves this problem with a more generally usable command `\az`. This generates the definite article along its argument and the argument itself; that is, the command `\az{`$\langle arg \rangle$`}` is equivalent either to `az~`$\langle arg \rangle$ or to `a~`$\langle arg \rangle$ depending on the first letter of $\langle arg \rangle$. Beside `\ref`, `\pageref`, and `\cite` one may use `\aref`, `\apageref`, and `\acite` with babel/magyar which also generate the appropriate definite articles. These macros use the command `\az`. The rule mentioned above has some consequences, which are satisfied by `\az`:

- a number beginning with 5 or a number beginning with 1 and having $3k + 1$ digits (k is a nonnegative integer) is preceded by 'az', all the other numbers are preceded by 'a'. For example 'az' is before 1, 5, 51, 524, 1020, 1000000, and 'a' is before 2, 3, 4, 6–49, 60–499, 10000, 100000, etc.
- the same rule is applied to roman numerals; that is, 'az' is before I, V, LI, DXXIV, MXX, and 'a' is before II, III, IV, VI, etc.

- if a notation has one letter only, or begins with a letter and is followed by a number or any special character, then the pronunciation of the letter must be considered. The pronunciation of 'f', 'l', 'm', 'n', 'r', 's', 'x', 'y' begins with a vowel. For example 'az F. fejezet' (chapter F), 'az x1 változó' (variable x1) is the correct form.

There were several contributors to the Hungarian part of babel. We would especially like to highlight the work of József Bérces and of course that of Johannes Braams [1].

Alphabetical order The special rules of alphabetical ordering are as follows:

- A one character consonant is handled separately from a two character consonant beginning with the same sign. For example, 'c' and 'cs' are two different consonants, so 'cukor', 'cuppant', 'csalit', 'csata' is the correct order. This rule is not applied for the ancient type of two character letters, which are frequently used in family names, and for the two or more character letters of other languages, like 'sch'. This rule can be well handled by the xindy package [4]. This and some of the following problems can be solved with the makeindex package [6] supplemented with the application of an extra script/preprocessor which applies the '@' metacharacter in the index entry.
- The short and long vowels are equivalent (a=á, e=é, i=í, o=ó, ö=ő, u=ú, ü=ű), although the long ones are after the short ones in the Hungarian alphabet (a, á, b, c, ...). For example 'alma', 'álom', 'alorvos' is the correct order. The only exception is the case when two words differ only in the length of the same vowels. In this case the short vowel comes first (e.g., 'kerek', 'kerék', 'kérek').
- The long digraphs are considered as two short digraphs, so the next substitutions must be applied before the ordering: 'ccs' → 'cs + cs', 'ggy' → 'gy + gy', 'ssz' → 'sz + sz', 'zzs' → 'zs + zs', etc. This is also true for the only trigraph 'dzs', where the substitution is 'ddzs' → 'dzs + dzs'.

Inflexional suffixes There are several things to do in connection with TEXing in Hungarian. The topic of this section is not the most urgent or important, but it is interesting in general as a question of the generated texts in Hungarian documents. The essence of the problem is that there are several suffixes, and most of them have more then one form.

Let us suppose that some equations are numbered from (1) to (7) in a math text. The translation of the English text "adding (1) to (3) and subtracting it from (4) gives (5)" is "(1)-et hozzáadva (3)-hoz, majd azt kivonva (4)-ből az (5)-öt kapjuk." If we change the numbers only, we may get different suffixes as the next example shows: "(3)-at hozzáadva (4)-hez, majd azt kivonva (6)-ból a (7)-et kapjuk." The problem with such a sentence from TEX's point of view is that generating the equation numbers also demands generating the suffixes. The suffixes follow the vowel harmony. This means that suffixes, which may assume two or three different forms, usually agree with the last vowel of the stem. In other words, front vs. back alternatives of suffixes are selected according to the vowel(s)[1] the stem contains [8]. Examples: 'tűzből' (from fire), 'házból' (from house), where 'tűz' and 'ház' are the base words and the front and back forms of the suffix are '-ből' and '-ból'.

If the suffix has three forms, one of them has a back vowel (o), the other has a labial front vowel (ö), and the third has an illabial front vowel (e). If the last vowel of the stem is labial (illabial), the labial (illabial) suffix is used. Harmony causes the following alternations among suffix combinations:

- a/e (-ban/-ben 'in', -nak/-nek 'to'),
- á/é (-nál/-nél 'at'),
- ó/ő (-ból/-ből 'from', -ról/-ről 'about'),
- u/ü (-ul/-ül 'for, by'),
- o/e/ö (-hoz/-hez/-höz 'to').

There are several uncertainties. For example the vowels may show paradigmatic alternations as long and short vowels alternate in some stems (ver*é*b → ver*e*bet 'sparrow', f*a* → f*á*t 'tree'). Another problem is that 'i' and 'í' can be both front and back harmonic (híd → hidat 'bridge' + acc., but szív → szívet 'heart' + acc.). The suffix may have different forms if the word is a compound word. So a good suffix-generator needs a dictionary.

As the first example shows, it can be advantageous to solve the problem for numbers. Fortunately, the suffix depends on the last nonzero digit and on the number of closing zeros only:

- back harmonic numbers: 0, 3, 6, 8, 100;

[1] The vowels created in the front of the oral cavity are called *front* vowels, and those formed in the back of the oral cavity are called *back* vowels. The front/back vowels cause the feeling of high/low sound. In Hungarian the front vowels are e, é, ö, ő, ü, ű, the back vowels are a, á, o, ó, u, ú. The vowels i and í are neutral as they can be either front or back vowels depending on the word. For example í is a front vowel in the word 'víz' (water), but a back vowel in the word 'zsír' (grease). Four of the front vowels are labial (ö, ő, ü, ű), others are illabial (e, é, i, í). A suffix is called front (back) suffix if the vowel it contains is a front (back) vowel.

- labial front harmonic numbers: 2, 5;
- illabial front harmonic numbers: 1, 4, 7, 9, 10, 1000.

These grammatical problems also come up in relation to spell-checking. Until recently, only commercial programs were available; happily, the first Hungarian GNU ispell program [10] and a free Linux and FreeBSD version of a commercial program were released recently [9].

Hyphenation The limitations of TEX in hyphenation of non-English text are well known. Unfortunately some of the problems are still not solved, or in some cases the known solutions cause other problems. The phonetic rules of Hungarian hyphenation are simple and easily programmable (unfortunately in TEX the third doesn't):

- Every syllable has exactly one vowel, so a Hungarian word has as many syllables as vowels (fi-a-i, me-ta-fo-ra, pa-ra-di-csom).
- A chain of consonants between two vowels is cut before the last consonant, so the last one starts the next syllable (csu-por, kap-tár, Ham-burg-ban). In Hungarian 'cs', 'dz', 'gy', 'ly', 'ny', 'sz', 'ty', 'zs' are digraphs, and 'dzs' is a trigraph, that is, two or three letters form one consonant (ki-csi, gú-nya).
- Although only the first letter is doubled in a long digraph (or trigraph), when hyphenated both syllables contain the full digraph (or trigraph) (mennyi – meny-nyi, hosszú – hosz-szú, gallyak – galy-lyak, briddzsel – bridzs-dzsel)

There is a grammatic rule of hyphenation, which overrides the previous phonetic ones and causes real difficulties:

- Compound words or words beginning with verbal or superlative prefixes have to be hyphenated at the morpheme boundaries (ö-reg-asz-szony – öreg+asszony, meg-e-szi – meg+eszi).

To handle this, either a (never complete) exception list or a morphological analysis is needed. At the moment TEX uses the first method (the wordlist is implicitly given in `huhyph.tex`), but only the second can produce a perfect solution. We list some cases when the fourth rule conflicts with the first three ones:

- Two words, a simple and a compound or prefixed, have the same form but different hyphenation (fe-lül – over, fel-ül – sit up, me-gint – again, meg-int warn, gép-e-lem – machine part, gé-pe-lem – I type it).
- Morpheme boundary seems to be a long digraph (villamos-szék – electric chair).
- Either of two hyphenations is acceptable if there is a Latin or Greek morpheme boundary, but it is not clear for the average reader (depresszió dep-resz-szió or de-presz-szió – depression).

Words containing a hyphen may be hyphenated at other points according to the rules. The hyphen may be repeated at the beginning of the next line if it is necessary to show the hyphen, for example in a specialized book: nátrium-<*newline*>-klorid. More difficulties are implied by the typographic rule that no hyphenation can be applied after the first or before the last letter of a word when applied to compound words (in this case, `\lefthyphenmin` and `\righthyphenmin` can not be used).

The present official version of the hyphenation file `huhyph.tex` is made by hand, and not by `patgen`. It fulfills the first two phonetic rules by a simple list of the possible syllable boundaries, and the grammatical rule by an exception list. Recently Gyula Mayer made a big hyphenation dictionary for `patgen` [7], and generated new hyphenation patterns.

In summary, all of these problems show that joining a morphologic analyser to TEX would produce better results.

Typography of text

We would like to match the layout of texts with the Hungarian typographic traditions (see e.g. [3], [14], [15]) as much as we can. This section deals with the modifications we have to do in this field.

Baseline grid In each type of texts written in Hungarian, a baseline grid has to be applied. It is easy to typeset plain texts following this rule, but for texts containing mathematical formulae, the task is particularly difficult.

Titles It is not allowed to put a period after title names. A small typographic character on a raised position should separate the paragraph title and the text of the paragraph with a non-stretchable normal spacing around it.

> CÍMEK ◦ Magyar nyelvű szövegekben a címek után sohasem teszünk pontot.
>
> *Jelek* * A cím betűképéhez illeszkedő bármilyen jel alkalmazható elválasztóként.

In general, the medium series of fonts is used as standard for typesetting titles within documents. The fonts can be upright, italic, in small caps or capitalized.

Application of bold upright and bold italic fonts is also allowed.

Fonts Hungarian typography generally uses bold extended fonts only for typesetting title pages, not for titles within documents (chapter, section, etc.). For these titles within documents, regular bold fonts are used.

Applying slanted fonts is absolutely contraindicated.

Short page Blank pages should not contain page numbers.

Vertical space between paragraphs The standard for every document is setting `\parskip` to 0 pt. Additional vertical space may be applied in short documents only, such as prospectus, if the paragraphs have no first line indentation.

Indentation First line indentation is the standard for typesetting paragraphs in almost every type of documents.

The indentation should be equal to 1 quad if the line length is less than or equal to 24 ciceros ([12], [13], [15]) or 20 ciceros (see e.g., [2]), otherwise the indentation should equal 2 quads. Within a given document, the measurement of the first line indentation of every paragraph (including, e.g., footnotes, references) is fixed. Every left or right indentation is equal to this size or to the multiple of it.

The first paragraph following a chapter or a section title may be typeset with or without first line indentation, but the method is fixed in a given document.

Document design without first line indentation may be applied for typesetting short documents (see e.g., [2]).

Break-line There are some rules for the length of the last line of paragraphs.

In the case of `\parindent` > 0 pt, the length of the break-line has to be longer than the first line indentation, and it has to be shorter than ⟨*line length*⟩ minus ⟨*first line indentation*⟩ or to be equal to the line length exactly.

In a document design with `\parindent` $= 0$ pt, in the case of `\parskip` $= 0$ pt, the break-line has to be longer than 1 or 2 quad and to be shorter than ⟨line length⟩ minus 2 quad (see [12], [13]) or more (see e.g., [15]). If the paragraphs are ragged right, the upper bound of the break-line length is 3/4 line length [15]. If the `\parskip` > 0 pt, the length of the break-line is allowed to be equal to the line length.

French spacing In any document, French spacing is used for spacing, i.e., the `\frenchspacing` command should be included in every Hungarian style file.

Before : ; ? ! Before some punctuation marks, such as colon, semicolon, question mark and exclamation mark, one third normal spacing should be applied.

– Megfeledkeztél a virágról?
– Nem, nem! Hoztam: ibolyát és gyöngyvirágot; kaktuszt és fikuszt; no és persze tulipánt is!

As a result, we have to modify the kerning tables of the fonts, so the standard font names used by TeX have to be changed to new ones.

Setting the spaces around exclamation marks is problematic because of its special mathematical meaning (see later).

Unnumbered lists Marks of items in unnumbered lists have to be set to a layout more closely following our traditions. The mostly used character for item labels is the en-dash. We can also use the '·', '∘', '*' characters, and also the '•'. These latter marks should be small and on a raised position:

– Tudományterületek
  • Informatika
    ∘ Mesterséges intelligencia
      * Ágensek
        · Mobil ágensek

In the typography of lists there is no vertical space between an item and the preceding or the following text (or item), nor between the paragraphs of an item. Labels are separated by two thirds normal spacing from the text.

If there is just one line per item in a list, the distance of the *labels* and the left margin of the main text is $i \times$ `\parindent` ($i = 0, 1, \ldots, 5$).

If the items contain more than one line the distance of the left margin of the *item paragraph(s)* and the main text can be

- equal to $i \times$ `\parindent` (as it is in standard TeX list formats) ($i = 0, 1, \ldots, 5$), or
- set to zero (i.e., just the first line of the paragraph is indented by $i \times$ `\parindent`).

Enumerated lists The numbers of items in enumerated lists have the following order and appearance:

I. Informatika
  1. Mesterséges intelligencia
    a) Ágensek
      α) Mobil ágensek

Numbers are followed by period, letters are followed by parenthesis. In the labels, letters and parentheses are emphasized.

For the rules of indentation see the preceding section.

Footnotes If a document contains relatively few footnotes (i.e., the average is less than 1.5 footnotes per page) we use asterisks (*) for marking them begining with one star on every page. Otherwise we can use numbers (as superscripts) ordinarily.

If the first line of a footnote text is indented, the footnote marks are set flush right in the space of first line indentation or in the labels of the list. The indentation in both cases has the same size as in the main text.

> A lábjegyzet* jeleként a legtöbb esetben** csillagot alkalmazunk.
>
> * Megjegyzés a szóhoz.
> ** Számokat alkalmazzunk, ha a szöveg valamely más részén hivatkozunk bizonyos lábjegyzetekre.

En-dash The normal usage of en-dash is the same as in English, for example, in the 'ld. 12–24. oldal' (see pages 12–24), or in the 'Budapest–Debrecen' expressions. Sometimes we are obliged to put one third (non-stretchable) normal space around the en-dash, for example, using it between names with first names: 'Kiss Előd – Nagy Pál – Tóth Éva'. The space is unbreakable before and breakable after the en-dash.

Typography of math

The Hungarian typographic traditions need some modification in the layout of mathematical formulae, too. TEXers perform some of these corrections on their own, because they see that the standard of (LA)TEX does not fit the Hungarian conventions (see e.g., [12], [13], [14]).

In this section, we show the modifications we have to introduce in order to make the new style files more complete.

Spacing around binary operators and relations Spacing is very important in mathematical expressions. (LA)TEX typeset mathematical formulae in a nice form, however, our standards slightly differ from those used in Hungarian typesetting.

In mathematical typesetting, one third normal spacing is used around binary operators and two thirds normal spacing around relations. (LA)TEX uses three `mglue` parameters called `\thinmuskip`, `\medmuskip` and `\thickmuskip`, for adjusting the spaces around elements of mathematical formulae. For setting these parameters to the required measures, we have given new values to these parameters:

```
\thickmuskip=4mu plus  2mu minus 4mu
\medmuskip=2mu plus 1.5mu minus 2mu
\thinmuskip=3mu
```

where `2mu` equals one third normal spacing. (The strange thing is that after the modification the thin space becomes wider than the medium space.) With the default values, the layout is

$$a + b - c/d * y \circ x = z, \tag{1}$$

and after the modifications

$$a\!+\!b\!-\!c/d\!*\!y\!\circ\!x = z. \tag{2}$$

Line break If TEX breaks a line after a binary operator or a relation in an inline mathematical formula, the sign has to be repeated on the next line.

Formulae cannot be broken at `\cdot` or slash.

Exclamation mark This sign has a special mathematical meaning, which forces us to handle it differently than the others.

For the most part, exclamation mark means the factorial sign in math mode. In this usage, it has to be followed by a small space without glue. Changing the class of this character from closing (number 5) to punctuation (number 6) (see e.g., [5], [11]), the problem has been solved in most cases. From the source code `{{k!n!(b-1)!} \choose {h!m!}} =1` we have

$$\binom{k!\,n!\,(b-1)!}{h!\,m!} = 1$$

which is an acceptable solution of our problem in most cases.

However, when a punctuation mark is followed by a binary operator, the spacing needs correction:

$$n!+k!+5a+6b,$$

its source code: `n!+k{!}+5a+6b`.

Space after commas In the Hungarian language, the decimal character is the comma, so we changed the default class of comma in math mode to 0, i.e., `\mathcode`\,="013B` results in a decimal "point" in math mode, too:

$$F_i(x,y) = y^i + 1{,}3x \quad x,\ y \in A,\ i = 1,\ 2,\ 3,\ldots$$

This modification reduces the number of mistakes in the layout, although, we have to type approximately the same number of characters.

What next?

At the moment there is no standard way of writing text easily in Hungarian using plain TEX. The authors generally make and use their own macros. It is necessary to write style files, whose layout keeps our typographic traditions and uses the modifications mentioned above.

The situation is better with LaTeX as babel offers an acceptable output and even some nice features. With some minor changes it can be improved enough to be closer to the needs of professional publishers: for example, changing the measure of white space produced by `\chapter`, `\section`, etc. commands, changing the dot or the spacing written after paragraph titles to a small typographic character, changing the appearance of footnotes to our traditions, and that of captions of tables and figures according to font sizes and types.

For (LA)TeX, we would like to propose some additional modifications in math typesetting. We also plan to design Hungarian special ligatures (for the pairs *gy*, *gj*, *gz*), and even perhaps new fonts.

Last but not least we mention the different TeX-variants, like ε-TeX or Ω, the usage of which can help our work.

There is a lot to do in the future.

References

[1] Braams, Johannes. "Babel, a multilingual package for use with LaTeX's standard document classes." available from CTAN, `/macros/latex/required/babel`, 2001.

[2] Bardóczy, Irén. *Magasnyomó formakészítés.* (Making frames for letter-press printing), 3rd edn, Textbook, Budapest, Műszaki Könyvkiadó, 1979.

[3] Gyurgyák, János. *Szerkesztők és szerzők kézikönyve.* (Manual for Editors and Authors) Budapest, Osiris, 1998.

[4] Kehr, Roger. "xindy, A Flexible Indexing System." available from CTAN, `/indexing/xindy`, 1998.

[5] Knuth, Donald E. *The TeXbook.* Reading, MA, Addison-Wesley, 1988.

[6] Lamport, Leslie. "MakeIndex, An Index Processor For LaTeX." available from CTAN, `/indexing/makeindex`, 1987.

[7] Mayer, Gyula. "A TeX és LaTeX elválasztási modulja, 2002." (Hyphenation module for TeX and LaTeX) `http://www.typotex.hu/huhydok.pdf`

[8] Megyesi, Beáta. "The Hungarian Language, A Short Descriptive Grammar." `http://www.speech.kth.se/~bea/hungarian.pdf`

[9] Morphologic. "MSPELL." `http://www.morphologic.hu/en/en_mspell.htm`

[10] Németh, László. "Magyar ISPELL." `http://www.szofi.hu/gnu/magyarispell/`

[11] Salomon, David. *NTG's Advanced TeX Course: Insights & Hindsight.* Groningen: NTG, 1992.

[12] Szántó, Tibor. *Könyvnyomtatás – tipográfia.* (Printing – Typography), 2nd edn, Budapest, Műszaki Könyvkiadó, 1964.

[13] Szántó, Tibor. *Könyvtervezés.* (Designing books) Budapest, Kossuth Nyomda, 1988.

[14] Timkó, György (ed.). *Helyesírási és tipográfiai tanácsadó.* (Orthographic and Typographic Guide) Budapest, Nyomdaipari Egyesülés, 1971.

[15] Virágvölgyi, Péter. *A tipográfia mestersége – számítógéppel.* (Craft of Typography – by computer) Budapest, Tölgyfa, 1998.