TEXing in Hungarian

Gyöngyi Bujdosó & Ferenc Wettl

1. 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 which have been introduced in the last distribution of TEX, as well as those which are not in the distributions but useful for typesetting in Hungarian. We are looking for the problems which have not been solved, and try 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.

2. Grammar

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

2.1. 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{`⟨*arg*⟩`}` is equivalent either to `\az~`⟨*arg*⟩ or to `\a~` depending on the first letter of ⟨*arg*⟩. 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' is beginning with a vowel. For example 'az F. fejezet' (chapter F), 'az x1 változó' (variable x1) is the correct form.

As there were several contributors to the Hungarian part of babel, let us highlight the work of József Bérces and of course that of Johannes Braams [1].

2.2. 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', 'cs' are two different consonants, so 'cukor', 'cuppant', 'csalit', 'csata' are in good 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 [3]. This and some of the following problems can be solved with the makeindex package [5] supplemented with the application of an extra script/preprocessor. This script might apply the '@' metacharacter in the index entry for sorting by makeindex.
- 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 behind 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 e.g. 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 doubled two character consonants is considered as two two character consonants, 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 true for the only three character consonant 'dzs', so the substitution is 'ddzs' → 'dzs + dzs'.

2.3. 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 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 (high) vs. back (low) alternatives of suffixes are selected according to the vowel(s) the stem contains [6]. Examples: házból 'from the house', tűzből 'from fire'.

If the suffix has three forms, then 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, 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 introductory example shows, it can be advantageous to solve this problem for numbers. Fortunately the suffix depends only on the last digit. Back harmonic numbers: 0, 3, 6, 8, 100, labial front harmonic numbers: 2, 5, illabial front harmonic numbers: 1, 4, 7, 9, 10, 1000.

The grammatical problems mentioned above are in connection with spell-checking. So far only commercial programs were available, the first Hungarian GNU ispell program has been released a few weeks ago.

3. Typography of Text

We would like to match the layout of texts with the Hungarian typographical traditions (see e.g. [2], [8], [9]) as much as we can. This section deals with the modifications we have done in this field.

3.1. French spacing

In documents written in Hungarian French spacing has to be used for spacing, hence the `\frenchspacing` command should be included in the Hungarian style files.

3.2. Before : ; ! ?

Before some punctuation marks, such as colon, semicolon, exclamation mark and question mark, one third normal spacing should be applied. In our solution, the chategory codes of these characters are active, and in their definitions, there is a non-breaking small space before the character:

```
\newdimen\@nosp    \@nosp  =0.33333em
\newdimen\ot@nosp  \ot@nosp=0.33333\@nosp
\catcode`\?=\active
\catcode`\!=\active
\catcode`\:=\active
\catcode`\;=\active
\def?{\penalty\@M\kern\ot@nosp\char`\?{}}
\def;{\penalty\@M\kern\ot@nosp\char`\;{}}
\def:{\ifmmode\char`\:\else
      \penalty\@M\kern\ot@nosp\char`\:{}\fi}
\def!{\ifmmode\char`\!\else
      \penalty\@M\kern\ot@nosp\char`\!{}\fi}
```

These definitions produce the following layout:

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

We have to handle the exclamation mark and the colon in a different way because of their special mathematical meanings (see later).

3.3. 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. Although, 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. The definition of the en-dash is as follows:

```
\def\kotojel{\unskip\nobreak
     \hskip\ot@nosp plus0pt\endash
     \hskip\ot@nosp plus0pt\ignorespaces}
```

3.4. Unnumbered lists

Marks of items in unnumbered lists have to be modified (or set) to a layout fitting better our traditions. The mostly used character for this is the en-dash, but we can also use the '$\cdot$', the '$\circ$' and the '$*$' characters, and also the '$\bullet$'. These latter marks should be small and on a raised position:

```
\def\kor{\leavevmode\raise.65ex%
     \hbox{$\scriptscriptstyle\circ$}}
\def\csillag{\leavevmode\raise.65ex%
     \hbox{$\scriptscriptstyle\ast$}}
\def\potty{\leavevmode\raise.55ex%
     \hbox{$\cdot$}}
\def\npotty{\leavevmode\raise.65ex%
     \hbox{$\scriptscriptstyle\bullet$}}
```

Using these definitions, we have

- Tudományterületek
  - Informatika
    - Mesterséges intelligencia
      - Ágensek
        - Mobil ágensek

In this LaTeX list, the values of \topsep, \partopsep, \parsep and \itemsep are set to be zero point, \leftmarginn to \parindent, and \labelsep to two thirds normal spacing (to 2\ot@nosp) whereas \itemindent=0.

3.5. Numbered lists

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

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

The standards of LaTeX document classes have to be changed to conform to these settings.

4. Typography of Math

The Hungarian typographical 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. [8]). Therefore, most of these changes shoud be included in the Hungarian style files.

In this section, we would like to show the modifications we have introduced in order to make the new style files more complete, as well as those that needs to be done in the near future.

4.1. 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.) Applying the default values, the layout will be

$$a+b-c/d*e\circ f=g, \tag{1}$$

and after the modifications

$$a+b-c/d*e\circ f=g. \tag{2}$$

4.2. Exclamation mark and colon

These signs have special mathematical meanings, which forces us to handle them 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. [4], [7]) by adding it to the style file `\mathcode‘\!="6021`, the problem has been solved in most cases. From the source code `{{k!n!(n-1)!} \choose {h!m!}} =1` we have

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

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

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

$$n!+k!+5+6$$

In this example, the spacing around the first + sign has to be corrected, the spacing around the second + is corrected, and the third one gives the new standard layout. Its source code: `n!+k{!}+5+6`. In the first case, TEX puts the + close to the next character reducing it to the class zero, and takes 0pt for spacing according to the table on page 170 of [4]. In the second case we reduced the class of ! to zero, so the class of the + sign has not been changed.

Colon gives a lesser problem. Because we have made the colon active (see above), we have to reset its `\mathcode` to its original class (*Relation*, class number 3): `\mathcode‘\:="303A`

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

We show the source code of the math expression

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

before the modification:

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

and after it:

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

5. What next?

At the time of writing, these modifications have been carried out by continuing the nationalization of TEX in Hungary. Hovewer, more work is needed to add other national specialties to (LA)TEX.

For LATEX documents, we plan to modify, for example, the measure of white spaces effectuated by `\chapter`, `\section` etc. commands, and changing the dot or the spacing written after paragraph titles to a small typographical character. We shall adapt the appearance of footnotes (indentation, marks and rules) to our traditions, and that of captions of tables and figures according to font sizes and types.

It is necessary to write style files for plain TEX whose layout keeps our typographical traditions and uses all of the modifications we have developed.

For (LA)TEX, we would like to propose some additinal modifications in math typing. We also plan to design Hungarian special ligatures (for the pairs *gy*, *gj*, *gz*.

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

[3] Kehr, Roger. “xindy, A Flexible Indexing System.” available from CTAN, `/indexing/xindy`, 1998.

[4] Knuth, Donald E. *The TEXbook.* Reading, MA: Addison-Wesley, 1988.

[5] Lamport, Leslie. “MakeIndex, An Index Processor For LATEX.” available from CTAN, `/indexing/makeindex`, 1987.

[6] Megyesi, Beáta. “The Hungarian Language, A Short Desriptive Grammar.” `http://www.speech.kth.se/~bea/hungarian.pdf`

[7] Salomon, David. *NTG’s Advanced TEX Course: Insights & Hindsight.* Groningen: NTG, 1992.

[8] Timkó, György (ed.). *Helyesírási és tipográfiai tanácsadó.* (Orthographical and Typographical Guide) Budapest: Nyomdaipari Egyesülés, 1971.

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

Gyöngyi Bujdosó
ludens@math.klte.hu
Ferenc Wettl
wettl@math.bme.hu