French in TEX

Alonzo Gariepy

Abstract

This paper describes a method of producing French documents with TEX that is much simpler than the other available alternatives. No preprocessing of `tex` files is required and the system operates with standard versions of the TEX program, the Computer Modern pixel files, and available device drivers. For IBM PC systems, accented letters can be directly entered from the 8 bit graphics character set. The preloaded version of TEX encompassing all of these changes is called FTEX. It fully hyphenates and kerns French text containing lowercase accented letters. I have used the work of M.J. Ferguson and J. Désarménien where applicable.

1 Introduction

TEX is particularly well known for the variety of its symbols, the accuracy of its hyphenation, the beauty of its output, and the standardization of its various implementations. All of these things suffer when TEX is used for French language typography.

The Computer Modern fonts are as much a part of standard TEX as the program itself. They do not contain accented letters, but provide separate letters and accent symbols that can be combined using TEX's `\accent` primitive. Computer Modern does not contain French quote characters (guillemets) or flattened accents for use with uppercase letters. So the symbol set is not really suitable for typesetting French.

Rules for hyphenating the French language can be formulated much more simply than can those for English. The comprehensive algorithm used by TEX handles French with a minimal set of hyphenation patterns. But the `\accent` primitive inserts *explicit kerns* when forming accented letters, and TEX has been specifically designed not to hyphenate in the vicinity of an explicit kern.

One of the things that makes TEX's output so beautiful is the automatic kerning of letter pairs. But the manner in which the `\accent` primitive operates, prevents this automatic kerning within French text.

Editor's note: The techniques described here seem to be architecture-dependent. Production of this article was not possible on either a TOPS-20 or a VAX/VMS system, but required that TEX be run on an IBM PC compatible, and the `.DVI` file transferred to the VAX for printing.

As a result of these three factors, any attempts at using TEX for French language typesetting have necessitated major deviations from the standard system. Three approaches to French language TEX are described in the following paragraphs.

One approach to this problem, implemented by W. Appelt, involves changing the `tfm` files so that accent/letter pairs become ligature characters with positions above '177 in the fonts.[1] Such ligatures have no character pattern, but are assigned attributes identifying the associated accent and letter. The `tfm` file can be modified so that automatic kerning takes place around the ligature. A special device driver must be created to output this dummy ligature using the character patterns for the accent and letter. The `dvi` file cannot be printed without the special driver and modified `tfm` files.

J. Désarménien came up with another approach that relies on Computer Modern based French fonts incorporating already-accented letters.[2] Some room can be found for these letters at the beginning of a Computer Modern text font by eliminating all the uppercase Greek letters. The vowels à and ù occur only in the words 'à' and 'où' and therefore need not be included in the French fonts for the purposes of hyphenation. These vowels may still be accented with the `\accent` primitive. The problem with this approach is the necessity to maintain a complete set of French Computer Modern `tfm` and pixel files in all needed sizes, resolutions, and magnifications. Such maintenance demands the possession and use of METAFONT and the storage and distribution of large amounts of data.

More recently, M.J. Ferguson has made changes to the TEX program itself to produce a multilingual version called TÊX.[3] This program circumvents the restrictions within TEX that prevent hyphenation of words containing accented letters. To this it adds a facility for loading multiple sets of hyphenation patterns and switching between them. There is no indication that TÊX will automatically kern accented letters from standard Computer Modern. Alas, TÊX is not really TEX.

[1] W. Appelt: *The hyphenation of non-english words with TEX*. Proceedings of the First European Conference on TEX for Scientific Documentation, Addison-Wesley, 1985, pp. 61–65.

[2] J. Désarménien: *How to run TEX in a French environment: hyphenation, fonts, typography*. TUGboat, Vol. 5 (1984), No. 2, pp. 91–102.

[3] M.J. Ferguson: *A Multilingual TÊX*. TUGboat, Vol. 6 (1985), No. 2, pp. 57-58.

2 French TEX for the IBM AT

At my installation we are using Addison-Wesley's MicroTEX (written by David Fuchs). We produce large quantities of unilingual French and English documents. The software is run on IBM ATs. MicroTEX and its French partner, FTEX, operate as components of an interactive text management system (TMS).

Computer editing French language documents that contain normal TEX accent macros (such as \^, \c, \' and \") is very awkward. With the aid of a macrokey program and the IBM PC graphics character set, accented letters can be typed directly into the word processing facilities of the TMS.

With limited disk space on both production and development machines, it is crucial that the TMS and TEX leave enough room for a large quantity of data. The space used by the files and programs needed to develop, distribute, operate, and maintain FTEX must be minimized.

The system maintenance and user support will be performed by a single individual. This will involve a large amount of user training and some TEX macro writing, leaving little time for anything else.

We expect the materials created with this system to have a long lifespan and eventually to be made available for interactive retrieval.

Thus, the constraints for our French language version of TEX:

- a minimum of software maintenance
- minimal storage requirements
- no maintenance of fonts
- easy distribution
- high quality output
- direct editing of accented text on screen
- portability of `tex` and `dvi` files

The three approaches to a French TEX already described were ruled out because of the requirement for developing and maintaining modified versions of device drivers, pixel files, or the TEX program itself.

2.1 Accenting with FTEX

FTEX produces words that can be hyphenated, by forming accented letters in a way that makes use of implicit kerns instead of explicit ones. This kerning is specified by the ligature/kern table in a `tfm` file modified for French.

To produce an accented letter, FTEX inserts three characters in the form: ⟨letter⟩⟨accent⟩⟨letter⟩. For example, on encountering `\^e` in the input file, FTEX will insert `e^e` into the `dvi` file, including the new implicit kerns from the modified `tfm` file.

This works in the following way:

- accent characters in Computer Modern (CM) are already at the right height to go above the lowercase letters
- words that contain uppercase letters will not often need hyphenation and can be accented in the conventional way
- letters that come before and after the `\^e` will automatically kern with the sequence `e^e` in the same way they would with a single `e`
- the three characters `e`, `^`, and `e` can be centered on one another by interposing, between the `^` and each `e`, identical kerns of value $-(\text{width}(\texttt{e}) + \text{width}(\texttt{\^{}}))/2$
- the sum of the widths of these three characters and the two kerns is equal to that of a single `e`
- font substitution (widely available on TEX's device drivers) allows the standard pixel files to be used in concert with FTEX's `tfm` files.
- the three characters, when superimposed on an output device, appear as the printed character pattern ê.

2.2 Modifying `tfm` files

I have written a program that produces a French `tfm` file (eg., `fmr10.tfm`) from the standard Computer Modern `tfm` file (eg., `cmr10.tfm`). The program, called `fkern`, actually works with `pl` files, which are easier to parse and modify, and leaves the translation between `tfm` and `pl` formats to the utilities `tftopl` and `pltotf`.

`Fkern` reads the widths of all of the accents and accentable letters, and adds new entries to the kern table for accent/letter pairs that occur in French. These kerns are negative and equal to the average of the widths of the two characters involved.

There are three strategies for integrating French `tfm` files into your TEX system. The first, already mentioned, is to use FM fonts in your TEX file, but substitute CM fonts when you run the device driver. A second way, requiring more disk storage, would be to make exact copies of all the CM pixel files, naming them FM instead.

The third alternative is to directly modify the CM `tfm` files and use them both for TEX and FTEX. The unusual kerns for French will not cause problems for most applications and could be used to create accented letters in TEX as well. This approach avoids the necessity of redefining the various fonts used by `plain.tex` or LATEX, and setting up font substitutions, but then the names of the files would not distinguish them from the standard set.

2.3 Inputting accented letters

Accented characters from the graphics portion of the IBM PC character set can be directly entered using many IBM PC editors (sometimes with the help of a macrokey program). The TMS we use for entering and managing our documents, allows us to do this.

One solution to the problem of converting these extended (8 bit) characters to a form that TEX can work with would exploit the ability of some editors to substitute any string of characters during output. The substitution facility of such an editor could be customized to convert the IBM graphic character ê to the TEX sequence `\^e` (or `e^e`). Unfortunately, the printer driver for our TMS and many editors does not allow this kind of thing.

When MicroTEX inputs an extended character, it entirely ignores the eighth bit, effectively subtracting 128 from the value of the character. Serendipitously, all French lowercase accented letters in the graphics character set fall into the range from 1 to 23 when the eighth bit has been stripped. These characters may be made active and defined to substitute the ⟨letter⟩⟨accent⟩⟨letter⟩ sequence described above.

One must be careful that the extended characters moved into this range do not conflict with character codes in use by TEX. Fortunately, none of these conflicts with ⟨return⟩. The characters ü, ë, ï and î conflict with ⟨control A⟩, ⟨tab⟩, ⟨control K⟩ and ⟨form feed⟩ defined by `plain.tex` to be equivalent to ⟨Subscript⟩, ⟨Space⟩, ⟨Superscript⟩ and ⟨`\par`⟩. The codes ⟨control A⟩ and ⟨control K⟩ are assigned this way because they correspond to the characters ↓ and ↑ on some non-PC keyboards. FTEX supersedes these definitions. The ë and î may safely be used as long as the input file contains no ⟨tab⟩ or ⟨form feed⟩ characters. These four accented vowels are used rarely enough in French that one could do without the direct entering of them, in order to avoid conflicts.

There are two French uppercase accented letters in the graphics character set. The É can be declared active but, due to its height, must be accented by TEX in the conventional manner. The Ç conflicts with the ASCII NULL when the eighth bit is stripped. FTEX currently uses no uppercase accented letters from the graphics character set.

The guillemets in the IBM graphics character set conflict with alphabetic characters when the eighth bit is stripped. The < and > characters can be used for this purpose in text input while retaining their meanings as relational operators in math mode. There are several ways, using Computer Modern, to create guillemets ≪ of a sort ≫.

The direct entry of 8 bit characters is very system dependent. This feature can be removed from FTEX for non IBM PC systems.

2.4 FTEX and hyphenation

FTEX's French hyphenation patterns are translated from those developed by M.J. Ferguson based on work by J. Désarménien. Examples of the kind of pattern that FTEX needs to perform its hyphenation include `.de´e3s2e´e3gr` and `1c¸c`. Direct entry of accented characters has been made to apply to the hyphenation patterns as well, so that if you edit `ftex.tex` you will actually see lines that look like `.dé3s2é3gr` and `1ç`. Hyphenation exceptions can also be entered in this fashion.

3 Putting it all together

The files that are required to create the preloaded version of FTEX include `ftex.tex`, `plain.tex` and the French Modern `tfm` files created by `fkern`. Two TEX primitives must be temporarily modified when inputting `plain.tex`. The `\patterns` primitive is redefined so that the English hyphenation patterns will be ignored, and the `\font` primitive is redefined to substitute `fm` fonts for `cm` or `am` fonts. A French version of LATEX can be generated this way if there is enough memory for FTEX's larger `tfm` files.

With the FM versions of all the necessary fonts preloaded, FTEX can be distributed as a single file (either `ftex.fmt` or `ftex.exe`). Alternatively, the utilities `fkern`, `tftopl`, and `pltotf` can be included for conversion of `tfm` files on site.

3.1 Limitations

FTEX formats French text only. Words containing uppercase accented letters cannot be hyphenated.

In this version of FTEX, none of the accented characters has been given a `\uccode`. For the `\uppercase` operation to have any effect it should be performed before the active characters have been expanded. Two approaches to setting `\uccodes` are demonstrated below:

```
%
%   if you want the uppercase letter accented
%
{\catcode`A=13 \gdefA{\'E}} \uccode`\é=`A
{\catcode`B=13 \gdefB{\^E}} \uccode`\ê=`B
%   etc.
```

```
%
%   if you do not want it accented
%
{\catcode`A=13 \gdefA{E}}   \uccode`\é=`A
\uccode`\ê=`A \uccode`\ë=`A \uccode`\è=`A
%   etc.
```

TEX device drivers must be careful how they correct the incremental roundoff errors accumulated while setting the letters in a word. The algorithms used ensure identical spacing within a word wherever it is used, while maintaining a correspondence between dvi coordinates and actual pixels. But, in the short run, one cannot be sure that two characters with the same dvi coordinates will be rounded to the same pixel. The obvious impact of this on FTEX is that identical superimposed letters may end up one pixel out of alignment, creating a slightly thicker letter. I believe that a minor improvement to the way that the device drivers are written will remove this imperfection.

4 The source of ftex.tex

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                              %
%  FTEX, Copyright 1987 by Alonzo M. Gariepy   %
%                                              %
%  NOTICE!  This file contains IBM graphics    %
%           characters                         %
%                                              %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
\catcode`\{=1 % begin-group character
\catcode`\}=2 % end-group character
\catcode`\$=3 % math shift
\catcode`\&=4 % alignment tab
\catcode`\#=6 % macro parameter character
%
%
\let\fpatterns=\patterns  % save primitive
\def\patterns#1{}    % disable for English
%
\let\ffnt=\font      % save primitive
\def\font#1=#2#3{\ffnt#1=\ifx#3mfm\else
   #2#3\fi}
%
     \input plain
%
%    \input lplain
%
\let\patterns=\fpatterns    % enable
\let\font=\ffnt
%
\def\multi#1#2#3{\def\multI##1{\ifx\end##1\else
   #1##1#3\expandafter\multI\fi}\multI#2\end}
```

```
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%   Define macros to represent the accents as
%   category 12.  Assign appropriate lccodes.
%
\begingroup    % so category changes are local
%
\catcode`+=7     % we'll be using ^ temporarily
\catcode127=12   % make DELETE valid character
%
\multi{\catcode`}
    {++S ^ ++? ++R ++X ++P ++[ ++^}{=12}
%
\def\fdef#1=#2{{\global\def#1{#2}
    \global\lccode`#2=`#2}}
%
\fdef\Aa=++S
\fdef\Ac=^
\fdef\Ad=++?
\fdef\Ag=++R
\fdef\Cc=++X
\fdef\i =++P
\fdef\oe=++[
\fdef\OE=++^  \lccode`++^=`++[
\fdef\Ap='
%
\endgroup
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%  Activate the IBM PC graphics characters that
%  correspond to the lowercase French accented
%  letters. Comment out this section if you are
%  not using an IBM PC.
%
%  Let ^^A and ^^K be active in verbatim modes
%
\def\dospecials{\do\ \do\\\do\{\do\}\do\$\do\&%
    \do\#\do\^\do\_\do\%\do\~}
%
\multi{\catcode`}
    {\ü \é \â \à \ç \ê \ë \è \ï \î \ô \ö \û \ù}
    {=13}
%
\def ü{u\Ad u}          % 1      129
\def é{e\Aa e}          % 2      130
\def â{a\Ac a}          % 3      131
\def à{a\Ag a}          % 5      133
\def ç{c\Cc c}          % 7      135
\def ê{e\Ac e}          % 8      136
\def ë{e\Ad e}          % 9      137
\def è{e\Ag e}          % 10     138
\def ï{\i\Ad\i}         % 11     139
\def î{\i\Ac\i}         % 12     140
\def ô{o\Ac o}          % 19     147
\def ö{o\Ad o}          % 20     148
\def û{u\Ac u}          % 22     150
\def ù{u\Ag u}          % 23     151
```

```
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%   Modify plain's accent macros so they make
%   lowercase French accented letters with FTEX
%
\begingroup
%
\def\ftxset#1#2{\expandafter\gdef
    \csname#1#2\endcsname{#1}}
%
\ftxset\Aa e
\ftxset\Ac a
\ftxset\Ac e
\ftxset\Ac\i
\ftxset\Ac o
\ftxset\Ac u
\ftxset\Ad e
\ftxset\Ad\i
\ftxset\Ad o
\ftxset\Ad u
\ftxset\Ag a
\ftxset\Ag e
\ftxset\Ag u
\ftxset\Cc c
\endgroup
%
\def\ftxacc#1#2{\if#1\csname#1#2\endcsname
    #2#1#2\else {\accent\expandafter`#1#2}\fi}
%
\def\`{\ftxacc\Ag}
\def\'{\ftxacc\Aa}
\def\^{\ftxacc\Ac}
\def\"{\ftxacc\Ad}
\let\ftexc=\c           % save cedilla macro
\def\c#1{\if\Cc\csname\Cc#1\endcsname
    #1\Cc#1\else\ftexc#1\fi}
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%   French spacing macros by J. Désarménien
%   given in TUGBoat, Volume 5 (1984), No. 2
%
\frenchspacing
\catcode`\;=13
\catcode`\:=13
\catcode`\!=13
\catcode`\?=13
\def;{\relax\ifhmode\ifdim\lastskip>0pt\unskip
  \kern\fontdimen2\font
  \kern-1.2\fontdimen3\font\fi\fi\string;}
\def:{\relax\ifhmode\ifdim\lastskip>0pt
  \unskip\nobreak\ \fi\fi\string:}
\def!{\relax\ifhmode\ifdim\lastskip>0pt
  \unskip\kern\fontdimen2\font
  \kern-1.2\fontdimen3\font\fi\fi\string!}
\def?{\relax\ifhmode\ifdim\lastskip>0pt
  \unskip\kern\fontdimen2\font
  \kern-1.2\fontdimen3\font\fi\fi\string?}
```

```
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Hyphenation patterns for French.  Accented
% letters in this table can be represented in
% any of three forms for FTEX:
%
%     1) e\Ac e   or  \i\Ad\i
%     2) \^e      or  \"\i
%     3) ê        or  ï
%
% On non-IBM PC systems where you are not
% using the graphics characters, you will
% need to use grep or an editor to put the
% patterns into one of the other forms.
%
\patterns{
2'2 'a2 'é2 'e2 'o2 'ö2 'u2 'i2 .é2 1ba 1bâ 1be
1bé 1bè 1bê 1bi 1bî 1bo 1bô 1bu 1bû 1by 4be.

    :

en1i2vr .en1o2 .eu2r1a2 extra1 extra2c extra2i
hémi1é hémo1p2t hypera2 hyperé2 hyper\oe2
hyperi2 hypero2 hypers2 hyperu2 hype4r1 hypo1a2

    :

archi1é2pis moye2n1â2g po1astre uni1o2v
uni1a2x vélo1s2ki vol2t1amp tachy1a2 tchin3t2
chlo2r3a2c chlo2r3é2t n3s2at. n3s2ats.
}
```