A Multilingual TêX

Michael J. Ferguson
INRS-Télécommunications
Montréal, Canada

This note details an extension to TEX that allows for multilingual hyphenation using **standard TEX fonts**, including words with accented letters. Switching between languages within a document is sufficiently simple and efficient that it could even be done on a word by word basis. Although it really has not been given a name, perhaps it could be called "TêX".

Briefly the features of the extension are:

- A new **primitive** integer parameter **\language** has been introduced. The value of this parameter controls the set of language **\patterns** and hyphenation exceptions actually in force when hyphenation is attempted.
- Hyphenation exceptions are language dependent.
- Words with accents, such as "l'épicerie", will be hyphenated correctly. These modifications of TEX will work, with accented letters such as "é" built using TEX's accent primitive or resident in the font. In addition they will work whether the letter is accessed as a single character or as a ligature.
- **\lccode** 1 and 2 are now used to indicate which characters are accents. Note that **\lccode** 0 indicates a non letter.
- Theoretically any number of languages could be used. The only problem, at this point is a restriction to 256 of one of TEX's tables (the **trie_op** for those that know). The bilingual (French/English) running at INRS-Télécommunications uses, 238 of 256 of these table locations. French alone uses 108 and English alone 181.
- The changes are upward compatible — a standard **plain.tex** can be built into a format file by a modified **initex**. However, the format file has been modified so that a non-extended **plain.fmt** will cause a "fatal format error" if used with the extended TEX.

Some restrictions are as follows:

- Discretionary hyphenation spellings, as required in German, are not automatically included. However, it is felt that these could be added, in special format to the **\patterns** and handled during hyphenation much as ligatures.
- Accents must be in the same font as the characters in the word to be hyphenated. It is not clear whether this is an important restriction.
- A new value of **\language** determines both a set of hyphenation patterns and exceptions. There is no provision for using an additional set of hyphenation exceptions with an already existing set of patterns. For instance, if it was really important that "Random House" hyphenation be used rather than "Websters", a set of patterns for both would be required.

To change hyphenation rules it is only necessary to change the value of **\language**. However, since accents and certain characters may be legitimate in one language and not others, it may also be desirable to modify certain **\lccode**s. There are checks in the modifications to prevent disasters if **\language** is somehow not within the range allowed.

Modifications for Hyphenating Words with Accented Characters

The basic idea is as follows:

- Designate accents with `\lccode` of 1 or 2.
- Make accent kerns implicit so that they disappear before the word is sent to the hyphenation routine.
- Reconstitute the accents after the hyphens are returned from the hyphenation routine.

The net effect of this is that hyphenation patterns will be applied to words involving accents. This means, for instance, that the word "envôuterions" has the hyphens "en-vôute-ri-ons" if the English patterns in Plain are used, but is hyphenated as "en-vôu-te-rions" if the French patterns are used. Note that the English patterns inserts one incorrect hyphen and misses another. In addition there will never be a hyphen inserted between an accent and its following character since that case has never been given an odd number.

Two `\lccodes` were used to allow for different placement of the same accent symbol – for instance above or below the accented character. Initially it was thought that the cedilla " ¸ " would require such special treatment but that turned out not to be the case. The second value could be used for language dependent accent placements. However this is of limited utility at the moment as there is no way, other than recompiling TEX to modify the accent placement routine.

There are a few restrictions with respect to accents and hyphenations.

- The accent must come from the same font as the accented character.
- It is not possible to accent ligatures.
- It is not possible to put on more than one accent.
- Accents placed by raising or lowering boxes cannot be hyphenated. This means that the cedilla in some fonts may prevent hyphenation.*

Comments and Caveats

There are several extensions possible. The most obvious, and one that is probably necessary, is the introduction of discretionary spellings involving hyphenation ... such as those that occur in German. It appears to this author that the rules could be placed in the hyphenation patterns, and the invocation handled much like the multiple choices involved in ligatures. This was not done because of a lack of precise understanding of what was required.

Finally, there is one data structure in short supply. This is the `trie_op` that is one byte. The combined French and English patterns use 238 of 256 possible values. This suggests that for several languages the trie will have to be modified.

The changes to TEX have been rather simple, involving precise rather than massive surgery. The fact that it could be done at all amazes this author. The credit for this lies with squarely with Don Knuth. There are two major reasons. The first is obvious, to those who know, and is due to the incredible level of documentation possible in WEB. However, WEB by itself does not guarantee a well designed program. TEX, quite simply, is well designed. Although there is ample opportunity for obscure data dependencies, they do not seem to occur. Several times this author was worried about problems that did not exist. Under normal circumstances, they might have been there.

---

* Interestly the umlaut, "ö" appears to be incorrectly placed in the most fonts for (nice) use as an accent, at least in French. It is not clear why this is so.