
Francisation d'un format \LaTeX : nouveautés

Daniel FLIPO

Université des Sciences et Technologies de Lille
daniel.flipo@univ-lille1.fr

Résumé. Les distributions \TeX basées sur Web2C v7.x (te \TeX pour Unix, fp \TeX pour Windows, CMac \TeX pour Mac) ainsi que la nouvelle extension `mltex.sty` développée par Bernd RAICHLE ont considérablement simplifié la création et l'utilisation de formats \LaTeX adaptés au français. Cette notice vise à exposer ces nouvelles possibilités.

1. Quelques rappels

Par défaut, \LaTeX est configuré pour travailler en anglais, ou plus exactement en américain. Composer des documents dans d'autres langues nécessite une double adaptation, au niveau dynamique (ajout d'extensions) et au niveau statique (dans le *format*).

1. La traduction des chaînes de caractères générées automatiquement par les commandes comme `\chapter`, `\tableofcontents...`, et l'adaptation aux règles typographiques propres à chaque langue, se font très simplement par appel à des extensions dans le préambule du document (`babel` ou `french` par exemple).
2. En revanche, les règles de césure propres à chaque langue doivent être intégrées dans le *format*.

La présence des motifs de césure adéquats dans le format, ne suffit pas à assurer la coupure correcte des mots contenant des caractères accentués, le codage des fontes¹ utilisées pour visualiser le document joue un rôle crucial.

Pour les textes ne faisant appel qu'à des langues ouest-européennes, les deux codages courants sont

- OT1, codage sur 7 bits (128 caractères) utilisé pour les fontes CM,
- T1, codage sur 8 bits (256 caractères) utilisé pour les fontes EC.

1. Ce codage, dit *de sortie*, est indépendant du codage du texte source appelé *codage d'entrée*, ce point sera développé à la fin de cette section, p. 65.

Deux conditions doivent être remplies pour assurer des césures correctes :

- a) tous les caractères doivent être présents dans la fonte, qu'elle soit réelle ou virtuelle ;
- b) le codage interne utilisé pour les motifs de césure doit être le même que celui des fontes utilisées en sortie. Lorsque plusieurs codages peuvent être utilisés (T1 et OT1 par exemple), le recours à certaines astuces, dues à Bernd RAICHLE, permettent de n'avoir qu'un fichier pour les deux codages : c'est le cas de `frhyph.tex` qui a remplacé pour le français `f7hyph.tex` (codage OT1) et `f8hyph.tex` (codage T1).

Par défaut, \LaTeX fait la composition en codage OT1, avec les fontes CM qui ne contiennent aucun caractère accentué ; ceux-ci sont fabriqués par la commande `\accent` qui superpose l'accent et la lettre. La première condition n'est pas remplie, cette solution a pour effet d'inhiber toute coupure du mot après le premier caractère ainsi fabriqué.

C'est ici qu'intervient \MLTeX : les formats compilés avec l'option \MLTeX disposent d'un mécanisme de substitution (basé sur la primitive `\charsubdef`) pour les caractères absents de la fonte de sortie, ce qui permet d'étendre un codage limité au départ à 128 caractères : LO1, codage prolongeant OT1, est défini dans le fichier `lo1enc.def`, il n'est pas inclus dans le format. L'ajout de la commande `\usepackage{mltex}` dans le préambule d'un document compilé avec un format \MLTeX , fait appel à LO1² et utilise pour l'impression une sorte de fonte virtuelle dont les glyphes sont ceux de la fonte CM. Le codage LO1 étant compatible avec le codage interne des motifs de césure français, la coupure des mots français comportant des signes diacritiques est assurée normalement.

En pratique, pour composer des textes contenant des caractères accentués, il faut proscrire le codage par défaut (OT1), et opter

- soit pour T1, en ajoutant `\usepackage[T1]{fontenc}` dans le préambule du document,
- soit pour LO1, en ajoutant `\usepackage{mltex}`, mais cette seconde solution n'est possible que si le format utilisé intègre le support \MLTeX .

En ce qui concerne les documents utilisant les fontes PostScript standard d'Adobe (Palatino, Times etc.), le plus simple est d'ajouter à la déclaration de changement de fonte, `\usepackage{palatino}` par exemple, `\usepackage[T1]{fontenc}` ; ceci indique à \LaTeX que l'on dispose de fontes à 256 caractères (ce qui est le cas) et permet la césure correcte des mots accentués. L'utilisation d'un format \MLTeX associé à l'extension `mltex` donne le même résultat final sur le plan des césures, mais le positionnement des

2. La commande `\usepackage[LO1]{fontenc}` produit le même effet (changement du codage OT1 en LO1), nous n'utiliserons dans la suite que la forme `\usepackage{mltex}`.

accents, surtout sur les majuscules est différent voire vilain, comparez avec Utopia :

à, é, è, À, É, È (T1) et à, é, è, À, É, È (OT1/LO1).

Pour la production de textes scientifiques, le choix extrêmement restreint de fontes mathématiques disponibles (cf. [1]), fait qu'on se cantonne le plus souvent, par souci de portabilité ou par paresse, aux fontes CM ou EC.

La disponibilité des fontes EC (codage T1) dont l'esthétique est très voisine de celle des CM³ a réduit l'intérêt de la solution $\text{M}^{\text{L}}\text{T}_{\text{E}}\text{X}$. Cependant, un regain d'intérêt pour les fontes CM est venu, d'une part de la mise dans le domaine public⁴ d'une version PostScript type 1 d'excellente qualité, et du développement de $\text{p}^{\text{d}}\text{f}^{\text{t}}\text{e}^{\text{x}}$ d'autre part. Le rendu des documents au format PDF faisant appel à des fontes « bitmap » étant médiocre, on utilise des fontes PostScript pour leur composition. Dès lors qu'on ne dispose pas de fontes EC PostScript d'une qualité équivalente à celle des CM, l'utilisation d'un format $\text{M}^{\text{L}}\text{T}_{\text{E}}\text{X}$, de l'extension $\text{m}^{\text{L}}\text{T}_{\text{E}}\text{X}$ et de fontes CM type 1 offre une solution parfaitement adaptée à la production de documents PDF à caractère scientifique.

La présence dans un format des modifications $\text{M}^{\text{L}}\text{T}_{\text{E}}\text{X}$ ne présente *aucun inconvénient*, ces modifications n'ont aucun effet tant que l'extension $\text{m}^{\text{L}}\text{T}_{\text{E}}\text{X}$ n'est pas chargée dans le préambule du document. Aussi, je recommande de compiler les formats destinés à la composition du français *systématiquement avec le support* $\text{M}^{\text{L}}\text{T}_{\text{E}}\text{X}$.

Terminons ces généralités en mettant en garde contre une confusion possible : l'extension fontenc gère le codage *de sortie*, c'est-à-dire le codage des fontes utilisées pour la visualisation finale du document, tandis que l'extension inputenc gère le codage *d'entrée*, c'est-à-dire la conversion des caractères non-ASCII tapés au clavier, en séquences compréhensibles par $\text{T}_{\text{E}}\text{X}$. Un « ç » entré sur le clavier d'un Mac ne correspond pas au même code que le « ç » entré sur un PC sous Linux ou sous Windows, $\backslash\text{usepackage}[\text{apple}]{\text{inputenc}}$ transforme le premier « ç » en « \c c » tandis que $\backslash\text{usepackage}[\text{latin}]{\text{inputenc}}$ ⁵ transforme le second en la même séquence « \c c » ; ensuite le choix du codage de sortie décidera de la façon de reproduire sur papier ou à l'écran cette séquence : $\backslash\text{usepackage}[\text{T1}]{\text{fontenc}}$ fera appel au caractère « ç » tout fait de la fonte EC, tandis que $\backslash\text{usepackage}[\text{OT1}]{\text{fontenc}}$ (ou une absence de

3. La principale différence est le placement et la forme des accents, plus « plats » sur les fontes EC, notamment sur les capitales.

4. Par Blue Sky et Y&Y.

5. Le codage par défaut étant latin , cette déclaration n'est pas indispensable, il est cependant conseillé de l'inclure lorsque le texte source contient des signes diacritiques non transcrits en séquences $\text{T}_{\text{E}}\text{X}$.

déclaration car OT1 est le codage par défaut) fera composer ce caractère par superposition des deux caractères « c » et « , » de la fonte CM.

Pour un format L^AT_EX, la séquence « \c c » est un caractère, de code 231 pour LO1 (comme pour T1). Il suffit de choisir ce codage pour l'impression (par la commande `\usepackage{mltex}` dans le préambule d'un document) pour obtenir des césures correctes avec les glyphes des fontes CM.

2. Création d'un format

La création d'un nouveau format est souvent considérée comme une opération difficile réservée aux spécialistes. Ce n'est plus du tout le cas avec les distributions modernes basés sur Web2C v7.x.

2.1. Généralités

Voyons d'abord quels fichiers déterminent les langues supportées par le format, il vous faut :

- les fichiers contenant les motifs de césure pour chacune des langues susceptibles d'être utilisées sur votre installation ; ils sont sur le CD-ROM T_EX Live et sur CTAN, par exemple
 - pour le anglais américain : `ushyph.tex` (identique au fichier `hyphen.tex` de Don KNUTH),
 - pour l'anglais britannique : `ukhyph.tex` (appelé aussi `gbhyph.tex`),
 - pour le français : `frhyph.tex` (version 2.4 ou supérieure),
 - pour l'allemand : `dehypht.tex` pour l'orthographe traditionnelle ou `dehyphn.tex` pour l'orthographe « réformée » ;
- un fichier `language.dat` où les langues utilisables sont déclarées. Chaque ligne contient un nom de langue et le nom du fichier de motifs à charger pour cette langue. Vérifier par `kpsewhich language.dat` quel est le fichier `language.dat` pris en compte (il est normalement dans `texmf/tex/generic/config/`). Pour les quatre langues ci-dessus, `language.dat` pourra contenir les lignes⁶ :

```
english      ushyph.tex
=usenglish
british      ukhyphen.tex
french       frhyph.tex
=français
```

6. Il est vivement recommandé de laisser l'anglais américain comme première langue (langue par défaut). Les lignes commençant par = sont optionnelles, elles permettent de définir des alias pour les noms de langues.

```
german      dehyphn.tex
```

– un fichier `hyphen.cfg`, le répertoire `babel` normalement présent dans toute installation \LaTeX en contient un qui fait parfaitement l'affaire⁷.

La création d'un format \LaTeX est réalisée par la commande⁸

```
tex -ini latex.ltx
```

ou si on désire inclure le support $\text{ML}\text{\TeX}$ ⁹

```
tex -mltex -ini latex.ltx
```

ceci crée un fichier `latex.fmt` qu'il faut placer dans un répertoire où \TeX pourra le trouver (en général `texmf/web2c/`).

Si on désire conserver le format `latex.fmt` d'origine, et en créer un nouveau portant le nom `mllatex.fmt`, il suffit de créer un fichier `mllatex.ini`, contenant une seule ligne

```
\input latex.ltx
```

et de lancer la commande :

```
tex -mltex -ini mllatex.ini
```

Voici ce qui se passe lors de la compilation du format pour le chargement des motifs de césures : le fichier `latex.ltx` cherche en priorité un fichier `hyphen.cfg`, ou à défaut il fait appel à `hyphen.tex` qui contient les motifs originaux de Don KNUTH pour l'américain. `hyphen.cfg` va lire dans `language.dat` la liste des langues utilisables et charge pour chacune d'elles le fichier de motifs de césure correspondant.

Pour créer un format `pdflatex` incluant le support $\text{ML}\text{\TeX}$ la commande serait :

```
pdftex -mltex -ini latex.ltx
```

2.2. Cas de $\text{te}\text{\TeX}$

Le script `texconfig`, qui fait partie de $\text{te}\text{\TeX}$ automatise les opérations de création de formats ainsi que toutes les autres opérations de configuration.

Il est recommandé de conserver intacte la distribution d'origine dans un répertoire spécifique (désigné par la variable `$TEXMFMAIN` dans le fichier `texmf.cnf`, par exemple `/usr/local/TeX`), et de faire toutes les adaptations locales dans un autre répertoire (désigné par la variable `$TEXMFLOCAL` dans

7. La démarche présentée ici produit des formats tout à fait utilisables par l'extension `french` de Bernard GAULLE, noter toutefois que la distribution $\text{te}\text{\TeX}$ contient un format `frlatex` spécifiquement adapté à `french`.

8. Les distributions basées sur `Web2C v7.x` font appel à `tex -ini` au lieu de `initex` ou `virtex`.

9. Cette syntaxe est également spécifique à `Web2C v7.x`.

le fichier `texmf.cnf`, par exemple `/usr/local/texmf-localconfig`). On recopiera en particulier le fichier `texmf.cnf` original dans `$TEXMFLOCAL/web2c` et on adaptera cette copie aux besoins locaux. Il faudra bien sûr positionner la variable d'environnement `$TEXMFCNF` à la valeur adéquate (`/usr/local/texmf-localconfig/web2c` dans notre exemple).

En supposant retenue la configuration ci-dessus, le premier lancement du script `texconfig` exécute les opérations suivantes :

- création dans le répertoire `$TEXMFLOCAL` d'une arborescence contenant des copies des fichiers originaux utiles, en particulier `fmtutil.cnf` est recopié dans le sous-répertoire `$TEXMFLOCAL/web2c`, et `language.dat` est recopié dans `$TEXMFLOCAL/tex/generic/config`;
- ensuite création dans le répertoire `$TEXMFLOCAL/web2c` des formats décrits dans `fmtutil.cnf`;
- à la fin des opérations une boîte de dialogue apparaît, elle permet d'éditer les fichiers `fmtutil.cnf` (option `FORMATS`) et `language.dat` (option `HYPHEN`), toute modification faite dans l'un de ces deux fichiers provoque la régénération automatique du ou des formats concernés.

Les formats de la distribution standard qui n'ont pas d'homonyme dans le répertoire local `$TEXMFLOCAL` restent utilisables, en effet la recherche se fait d'abord dans `$TEXMFLOCAL` puis dans `$TEXMFMAIN`. La version « locale » de `fmtutil.cnf` ne devrait contenir que les formats ajoutés ou modifiés.

Voici les lignes actives de mon fichier `fmtutil.cnf` :

```
bplain tex language.dat -mltex bplain.ini
mllatex tex language.dat -mltex mllatex.ini
pdflatex pdftex language.dat -mltex pdflatex.ini
```

Elles créent trois formats multilingues `bplain`, `mllatex` et `pdflatex` adaptés aux langues définies dans `language.dat`, et intégrant tous les spécificités `ML \TeX` . Les deux premiers formats ne sont pas présents dans la distribution de base, le dernier est une variante du format standard `pdflatex` (langues différentes et option `ML \TeX` ajoutée).

Le script `texconfig` peut également être lancé avec un argument :

- `texconfig help` affiche toutes les options possibles ;
- `texconfig init` (ou `fmtutil --all`) recrée tous les formats en prenant en compte les fichiers `fmtutil.cnf` et `language.dat` ; cette commande est bien pratique pour mettre à jour les formats lorsqu'on a modifié l'un ou l'autre de ces deux fichiers ;
- `texconfig rehash` (équivalent de `mktexlsr`) met à jour les fichiers `ls-R` (nécessaire après chaque ajout de fichier ou modification du chemin de recherche).

2.3. Cas de $\text{f}^{\text{p}}\text{T}_{\text{E}}\text{X}$

Le programme `fmtutil.exe` permet de créer ou de mettre à jour les formats. Après avoir édité les fichiers `fmtutil.cnf` et `language.dat`¹⁰ sous WinEdt par exemple, il suffit de lancer la commande `fmtutil.exe` pour régénérer les formats. Ceci peut se faire à partir du menu de $\text{f}^{\text{p}}\text{T}_{\text{E}}\text{X}$ (« **Rebuild formats** ») ou bien à partir de du menu de WinEdt (« **Accessories** → **Rebuild formats** »).

2.4. Cas de $\text{CMacT}_{\text{E}}\text{X}$

La distribution $\text{CMacT}_{\text{E}}\text{X}$ ne propose pas d'équivalent aux scripts `texconfig` et `fmtutil`, chaque format devra être créé comme indiqué à la section 2.1. Noter que tout ajout de fichier ou toute modification des chemins d'accès nécessite de relancer l'utilitaire `setup`.

3. Vérifications, utilisation pratique

Pour vérifier qu'un format produit des césures correctes dans une langue donnée on dispose de la commande `\showhyphens{liste_de_mots}` qui affiche dans le fichier `.log` tous les points de coupure possibles des mots de la liste.

Voici un exemple de fichier test, le paramètre *codage* est à remplacer par *latin1* sous Unix et sous Windows et par *applemac* sur Mac :

```
\documentclass{article}
\usepackage[codage]{inputenc}
\usepackage[T1]{fontenc}
\usepackage[english,français]{babel}
\begin{document}
\showhyphens{signal container \ev\enement alg\ebre}
\showhyphens{événement algèbre dés\oeuvrement naïvetés}
\selectlanguage{english}
\showhyphens{signal container}
\end{document}
```

et le résultat qui *doit être obtenu* si le format est correct (seule compte la *position* des tirets indiquant tous les points de césure possibles, les caractères diacritiques peuvent apparaître sous forme hexadécimale, `^e9` pour é, etc.) :

```
Underfull \hbox (badness 10000) in paragraph at lines 6--6
[] \T1/cmr/m/n/10 si-gnal contai-ner évé-ne-ment al-gèbre
```

10. Ils se trouvent respectivement sous `texmf/web2c` et `texmf/tex/generic/config`.

```
Underfull \hbox (badness 10000) in paragraph at lines 7--7
[] \T1/cmr/m/n/10 évé-ne-ment al-gèbre dés-œu-vre-ment naï-ve-tés
```

```
Underfull \hbox (badness 10000) in paragraph at lines 9--9
[] \T1/cmr/m/n/10 sig-nal con-tainer
```

Remarquer les différences de césures en français (ligne 6) et en anglais (ligne 9) pour les mots « signal » et « container » présents dans les deux langues.

Pour tester un format $\text{ML}\text{T}\text{E}\text{X}$ avec les fontes CM, on remplace la ligne `\usepackage[T1]{fontenc}` par `\usepackage{mltex}`, les points de césure doivent être les mêmes que ci-dessus. Si on tente d'ajouter `\usepackage{mltex}` sans que le format ait été compilé avec l'option `-mltex`, l'extension `mltex.sty` n'aura aucun effet et le signalera dans le fichier `.log`.

4. Conclusion

Pour assurer des césures correctes en présence de signes diacritiques, il importe de prendre l'habitude de déclarer dans le préambule du document, à la fois le codage d'entrée utilisé (`\usepackage[applemac]{inputenc}`, `\usepackage[latin1]{inputenc}` etc.) et un codage de sortie adapté : T1 (`\usepackage[T1]{fontenc}`) ou codage LO1 (`\usepackage{mltex}`), le codage LO1 ne fonctionnant que si le format a été compilé avec les modifications $\text{ML}\text{T}\text{E}\text{X}$.

Dans ces conditions la commande `\hyphenation{liste_de_mots}` fonctionne même si la liste contient des signes diacritiques, par exemple : `\hyphenation{Lurçat anté-diluvien anti-dérapant}` empêche toute coupure du nom propre « Lurçat » et n'autorise la coupure des deux suivants qu'entre le préfixe et la racine.

Remerciements. — Je tiens à remercier Michel BOVANI et Fabrice POPINEAU pour leurs indications concernant $\text{C}\text{M}\text{a}\text{c}\text{T}\text{E}\text{X}$ et $\text{f}\text{p}\text{T}\text{E}\text{X}$, distributions que je ne pratique pas, et surtout Bernd RAICHLE et Thierry BOUCHE dont les remarques judicieuses m'ont permis de préciser et de compléter le contenu de cet article.

Bibliographie

- [1] Th. BOUCHE. — « Sur la diversité des fontes mathématiques », *Cahiers GUTenberg*, 25, 1996.
- [2] D. FLIPO, B. GAULLE & K. VANCAUWENBERGHE. — « Motifs français de césure typographique », *Cahiers GUTenberg*, 18, 1994.