

Motifs français de césure typographique

Daniel FLIPO^α

Bernard GAULLE^β

Karine VANCAUWENBERGHE^γ

Résumé. Cet article a pour objectif de faire un bilan des différentes versions de fichiers de motifs de césure français et de proposer une nouvelle version entièrement revue et corrigée. Une introduction à la division des mots y est faite ainsi qu'au processus de division employé par T_EX.

Abstract. *The aim of this article is to compare the various current versions of the French hyphenation files and to propose a completely new updated and corrected version. A short introduction is given to French hyphenation as well as to T_EX word-splitting mechanisms.*

1. La méthode LIANG

Avant les années 80, la division des mots se faisait souvent encore à la main selon les habitudes acquises de l'expérience ou de la sagesse des anciens. L'auteur d'un document n'intervenait que rarement dans ces décisions qui regardaient surtout les typographes. L'informatisation de la chaîne éditoriale a changé progressivement les méthodes et les professions. Les meilleurs programmes de division de mots ne s'intéressaient qu'aux règles classiques de syllabation du français, sur lesquelles nous reviendrons par la suite. Il était alors souvent nécessaire de se constituer des dictionnaires d'exceptions personnels. Par ailleurs, l'aptitude de ces programmes à traiter des documents multilingues était quasiment nulle. C'est de l'université de Stanford (EUA) que nous est arrivé encore une fois, la solution à ce problème.

F.M. LIANG [18] a publié une thèse en 1983 dans laquelle il propose un algorithme capable de prendre en compte plusieurs langues, pourvu que

α. Maître de conférences, université de Lille 1.

β. Ingénieur de recherche, IDRIS-CNRS, Orsay.

γ. Agrégée de lettres modernes, Bordeaux.

lui soient fournis des données linguistiques adaptées. C'est cet algorithme que Donald KNUTH a justement choisi d'implémenter dans \TeX et c'est grâce à lui que J. DÉSARMÉNIEN a pu commencer en 1983 son travail [4] sur les césures françaises dans \TeX (toujours à Stanford). Nous verrons que cet algorithme n'utilise pas de dictionnaire mais des *motifs* spécifiques à chaque langue. Cette méthode, efficace et économique, a malgré tout sa contre-partie : toute altération même minime d'un seul motif peut *casser* entièrement la validité de l'algorithme pour la langue en question. C'est pourquoi nous souhaitons que les motifs français continuent d'être *certifiés* par l'association GUTenberg, en faveur de laquelle nous avons réalisé ce travail.

Pour la petite histoire il faut dire que cet algorithme a été *repris* par Ronald MCINTOSH¹ qui fournit le monde entier et en exclusivité, en algorithmes de coupure de mots pour photocomposeuses. Ses fichiers de motifs sont bien entendu très confidentiels, nous ne les avons pas approchés ; mais les milieux bien informés disent qu'une révision est déjà annoncée pour le français...

2. La division des mots : une opération délicate

La division des mots a lieu normalement aux frontières des syllabes graphiques : la coupe en syllabes graphiques répond à un certain nombre de règles (voir plus loin page 38), la syllabe graphique se distinguant par quelques points de la syllabe phonique. En français, la syllabe est une unité phonique comprenant obligatoirement une voyelle, accompagnée éventuellement de consonnes ou de semi-consonnes (qui la précèdent et/ou la suivent).

Les principales divergences entre syllabe graphique et syllabe phonique sont les suivantes :

- la syllabe graphique conserve tout « *e* » muet placé entre une consonne et une autre consonne ou une fin de mot ;
- la syllabe graphique sépare les consonnes doubles même si elles sont prononcées simples.

1. Voir 1° R. MCINTOSH, *Hyphenation*, Ronald MCINTOSH publication, ISBN 1 872757-00-6, London, 1990 ; 2° D. FAWTHROP, *Hyphenation by Algorithm of English/American and other Languages*, Computer Hyphenation Ltd, 1990, ISBN 1 872757-02-2.

On peut voir, par exemple, que l'on distingue graphiquement trois syllabes dans **pu-re-té**, même si l'on prononce [pyr-te] (deux syllabes phoniques).

Il est coutume de dire que la division des mots se fait normalement entre syllabes mais aussi conformément à l'étymologie lorsque celle-ci est « *nettement perceptible* » [12]. Il n'existe malheureusement pas de traité ni de dictionnaire français sur ce sujet. Contrairement au domaine orthographique géré par l'Académie, il n'existe pas d'organisme détenteur de droits sur la typographie française. Du côté de la linguistique, domaine qui nous occupe aussi tout naturellement et où les *traités, grammaires* ou *dictionnaires* foisonnent, rien n'est établi non plus concernant la division des mots. Seuls quelques auteurs, en nombre restreint, ont publié des éléments de réponse en donnant des exemples, mais jamais on ne trouvera une liste vraiment exhaustive des règles à appliquer. Notre objectif n'est toutefois pas d'y remédier car nous pensons qu'il faut laisser la place aux différentes écoles, tout comme à l'innovation. Les règles que nous proposons plus loin permettront à chacun de disposer d'éléments de décision (lorsque l'une d'elles est difficilement mise en œuvre dans T_EX nous l'indiquons en note de bas de page).

Pour poser clairement le problème de la division des mots en français, citons Jacques DÉSARMÉNIEN [5] :

« Comme l'écrit THIMONNIER [19] : « ceux dont le français est la langue maternelle, appliquent spontanément les règles de la syllabation parlée. Ils n'ont qu'à se laisser guider par la prononciation usuelle. » dans la pratique typographique, les choses ne sont pas aussi simples : on sait bien que, en français, la correspondance n'est pas parfaite entre l'écriture et la prononciation. De fait, les problèmes rencontrés pour la détermination de la coupure d'un mot proviennent de la confusion faite généralement entre syllabation parlée et syllabation écrite. »

On ne saurait donc être trop prudent avant d'énoncer quelques règles complémentaires. Pourtant, il nous est apparu qu'il était important de diffuser quelques rudiments que l'on trouve éparés dans quelques livres d'auteurs [2, 3, 6, 12, 19, 15, 10], ou dans le *Code typographique* [7]. Ces références divergent parfois sur quelques points de vue ; nous avons fait notre propre choix.

3. Définitions

Le terme de *césure* (*typographique*) est souvent jugé équivalent aux termes plus simples de *coupe*, *coupure*, ou *division*. Il nous faut, cependant, expliquer les distinctions que nous allons leur appliquer par la suite :

- la *division d'un mot* est une sorte d'algorithme, non formel, qui doit être employé lorsqu'un mot arrive en bout de ligne et déborde dans la marge, de façon à obtenir l'endroit approprié de sa coupure (marquée par un trait d'union en fin de ligne) ;
- la *césure* est une autre forme d'algorithme, qui peut être tout à fait formel comme dans le cas de T_EX et qui sert à déterminer *tous* les points possibles de coupure dans un mot donné ; il faut alors parler de *points de césure* ; l'algorithme de T_EX est basé sur un ensemble de *motifs de césure* dont nous parlerons plus loin ;
- le terme *coupure* sera souvent employé par la suite en lieu et place de *division*.

Dans ce qui suit nous allons aussi adopter les notations suivantes :

-	:	trait d'union ou point de césure
/	:	division interdite à cet endroit
-/	:	division interdite après le trait d'union
-[:	division proposée
[]	:	notation phonétique (API ²)

4. Quelques règles

Avant de faire état des règles de coupure syllabique et étymologique, voyons d'abord quelques règles générales puis typographiques. Le terme de *règle* employé ici et par la suite est probablement un peu fort s'il est lu avec sa connotation de *règlement*. Il s'agit plutôt pour nous de règles de conduite, voire de conseils, d'autant que ces règles évoluent dans le temps, parfois très fortement, nous en verrons des exemples par la suite : aussi ce que nous écrivons aujourd'hui doit être considéré comme un recueil (non exhaustif) de ce que nous considérons comme de *bonnes habitudes* actuelles.

². Alphabet Phonétique International.

Les points de césure figurant dans les exemples à venir sont aussi ceux exactement produits par T_EX après application des éléments fournis dans cet article.

4.1. Règles générales

- g-a) Il faut éviter au maximum de couper les mots. Cela étant dit, nous devrions nous en tenir là, mais comme rien n'est parfait... et qu'il faut, malgré tout, parfois couper, voyons les autres règles.
- g-b) Il est important de suivre les mêmes règles tout au long du document.
- g-c) On n'isole pas en fin de ligne une seule lettre et on en rejette au moins trois (ex. : **obéi** n'est pas coupé, tout comme **agréé**, voir [6]).
- g-d) « Les mots étrangers employés dans un texte en français doivent être divisés, le cas échéant, suivant les règles de la langue étrangère » (extrait de [7]). Il faut donc que l'utilisateur de logiciel puisse indiquer de quelle langue il s'agit, et ce au niveau le plus bas c'est-à-dire au niveau du mot.

Il est précisé, par exemple, dans [2] que : « Les règles d'emploi du trait d'union varient d'une langue à l'autre. En français, on ne coupe pas un mot avant ou après un *x* ou un *y* placé entre deux voyelles. En anglais, on peut le faire. »

Cette idée est reprise par [3] qui donne les exemples suivants : en français, *fixer*, *moyen* ne peuvent pas être coupés, à l'inverse de *mixture*, *tex-tile* et *cy-près*.

- g-e) « On évitera les coupures malsonnantes :
 con/trôle **cul/ture**
 « [...] De ce fait on ne peut couper certains mots comme **connexion** »
 (extrait de [7]).
- g-f) On ne coupe pas les noms propres (on ne coupe pas non plus entre initiale et nom³), les abréviations, les acronymes, les nombres et les dates exprimés en chiffres⁴.
- g-g) On ne sépare pas un nombre de l'unité qui lui est rattachée⁵ (m, kg, %, F, *etc.*).

3. Nous n'aborderons pas, volontairement, le problème de la coupure entre les mots qui n'est pas l'objet de cet article mais qui, à lui tout seul, mériterait une longue discussion (on en trouvera un aperçu avec la règle suivante).

4. En (L^A)T_EX l'option de style **french** apporte quelques commandes pour effectuer ce travail.

5. Pour éviter ce problème, les logiciels disposent en général d'un espace insécable (~ en T_EX).

- g-h) On coupe les mots composés au trait d'union qu'ils possèdent déjà ;
- dans le cas du *t* euphonique, on coupe naturellement au premier trait d'union [6] : `dira-[t-il`
 - par contre, « pour la même raison d'euphonie, on divisera : `c'est-à-[dire` et *non* `c'est-/à-dire` » (extrait de [7]).

4.2. Règles typographiques

Traitant, ici encore, de règles très générales, nous n'entrerons pas dans le débat *pour* ou *contre* la justification à droite dans certaines conditions, ceci étant un préalable à la décision de couper ou de ne pas couper les mots. Nous ne parlerons pas non plus de la taille des justifications qui modifient parfois les données de la coupure ; on se reportera à [7], par exemple, pour de plus amples détails. On retiendra toutefois les quelques points suivants, applicables dans la grande majorité des cas :

- t-a) Aucune césure ne doit intervenir en bas de page (et spécialement sur les pages impaires) pour ne pas rendre la lecture plus difficile⁶.
- t-b) Un paragraphe ne doit pas être terminé, du fait d'une division de mot, par une ligne de longueur inférieure au renforcement⁷.
- t-c) On ne doit jamais trouver 3 lignes consécutives terminées par un trait d'union⁸. Mais au temps des DIDOT on n'hésitait pas à couper 6 ou 7 fois de suite !

4.3. Règles de coupure en syllabes graphiques

Les règles générales et typographiques étant énoncées, il est maintenant possible de savoir si un mot donné, placé à un endroit précis du document, peut être candidat à la division ou non. En supposant qu'il le soit, il faut

6. T_EX choisit l'endroit où il doit changer de page en maximisant une « note d'aspect » pour la page en cours. Malheureusement aucun paramètre n'est prévu pour pénaliser de telles coupures de mots, et il arrive que T_EX produise effectivement des pages dont le dernier mot est coupé (voir [17] et [13]). Dans ce cas il est nécessaire d'intervenir manuellement.

7. Cette règle n'est pas mise en œuvre directement en T_EX sous cette forme. Dans son calcul de note d'aspect T_EX prend en compte le paramètre `\finalhyphendemerits` dont la valeur standard (5 000) le dissuade fortement de terminer par une césure l'avant dernière ligne d'un paragraphe (cf. [17] p. 98).

8. T_EX prend en compte un paramètre `\doublehyphendemerits` dont la valeur standard (10 000) pénalise très fortement toute paire de lignes consécutives terminées par une césure (cf. [17] p. 98). Dans le cas d'un triplet de telles lignes la pénalité serait doublée, rendant la disposition du paragraphe inacceptable.

alors s'attacher, en premier lieu, à le diviser « normalement » aux frontières des syllabes graphiques. Voici quelques règles à cet effet :

s-a) la coupure est possible entre voyelle et consonne unique [6] :

co-[lon

s-b) la coupure peut être réalisée entre les consonnes lorsqu'il y en a deux :

col-[lant fac-[teur

al-[bi-nos par-[tir

La coupure entre deux consonnes identiques a lieu même si ces consonnes doublées dans l'écriture sont presque toujours prononcées simples. Ainsi on coupe **souf**-[fler et **col**-[lê-gue, bien que l'on prononce généralement [su-fle] et [ko-lɛg].

- la coupure s'effectue devant la plus forte de deux consonnes formant un groupe décroissant. Un tel groupe est en général constitué d'une consonne que l'on peut qualifier de forte – en particulier une occlusive, sourde ([p], [t], [k]) ou sonore ([b], [d], [g]), ou parfois une constrictive, notamment [f] et [v] – suivie d'un phonème plus *faible* : une semi-consonne⁹ ou le plus souvent une sonante, telle la latérale [l] et la vibrante [r]. D'où de nombreux groupes indissociables notés **cr**, **dr**, **bl**, **fr**...

câ-[bler nau-[frage ta-[bleau

- mais on peut aussi tenir compte de l'étymologie (voir 4.4) :

sub-[stan-tif¹⁰

- *ch*, *gn*, *ph*, *th*, ne sont jamais coupés lorsqu'ils ne transcrivent qu'un son, ce qui n'est pas toujours le cas. En effet, on relève quelques termes, qui font figure d'exceptions, pour lesquels les deux consonnes transcrivent deux phonèmes, et sont donc susceptibles d'être coupés. J. DÉSARMÉNIEN avait parfaitement noté cela pour les mots suivants :

stag-[na-tion wag-[né-rien.

s-c) Des voyelles consécutives en hiatus (ex. **thé/âtre**) constituent des syllabes entre lesquelles on ne va pas à la ligne, sauf s'il s'agit de préfixes [6] :

pré-[avis pro-[émi-nent

9. Les trois semi-consonnes du français sont :

[j], comme dans *ciel*, *faïlle*

[ɥ], comme dans *lui*, *nuit*

[w], comme dans *loi*, *Louis*

10. [15] préfère la coupure syllabique ordinaire : **subs-tan-tif**.

s-d) On ne rejette pas à la ligne, pour la commodité de la lecture, une syllabe muette [6] :

es-[pè/rent

s-e) On ne divise jamais après une apostrophe, ni immédiatement après une élision (comme dans s'en-tr'/égor-ger).

4.4. Règles de coupure étymologique

Le *Code typographique* [7] fait grandement abstraction des conditions à appliquer pour respecter l'étymologie :

« La division étymologique est si rarement et si difficilement observée, même par l'Académie et les lexicographes, qu'il est préférable de suivre l'avis des techniciens du Livre tels que TASSIS, FREY, Théotiste LEFÈVRE, etc. qui recommandent la division d'après l'épellation française.

« Nous reconnaissons néanmoins que certains auteurs de travaux scientifiques préfèrent la division étymologique, qui fait ressortir la racine grecque ou latine. »

Nous avons vu en g-h qu'il fallait couper les mots composés au trait d'union qu'ils possèdent déjà.

e-a) Il en est de même pour les mots composés sans trait d'union¹¹ :

baise-[main chèvre-[feuille contre-[maître

Par ailleurs, J. DÉSARMÉNIEN donne *ses* bons exemples mais aussi les cas difficiles, voire insolubles de façon informatique usuelle. Pour schématiser à l'extrême, disons :

e-b) les préfixes tels que **re-**, **ré-**, **co-**, **trans-** (mais attention ici aux faux amis, on coupera **trans**-[at-lan-tique mais **tran**-[sac-tion) et **inter-** peuvent être considérés comme étymologiques (on peut alors couper **inter**-[action¹²);

e-c) ceux tels que **sur-**, **psy-**, **pro-**, **ex-** et **sub-** sont aussi à examiner avec attention, mais à l'exception de ce dernier la coupure syllabique coïncide en général avec la séparation préfixe/base ;

11. Le *Code typographique* [7] dit improprement « les mots composés dont le trait d'union a disparu ». En effet, le mot *chèvrefeuille* existait déjà en un seul mot au Moyen Age ; il semble donc qu'il n'ait jamais comporté de trait d'union.

12. Cette coupure n'est pas effectuée automatiquement par T_EX avec les motifs actuels.

- e-d) d'autres comme **para-** sont plus difficilement décidables : *parabole*, *paravent* et *parachever* n'ont pas la même racine étymologique, ainsi on devrait couper **pa/ra-[chro-nisme**¹² mais aussi **par-[ache-ver**.
- e-f) et parfois on s'interroge, sans fin, sur des préfixes comme **poly-** ou **télé-** : faut-il couper ou conserver intact ces préfixes ?

J. DÉSARMÉNIEN précise dans [5] que 96,8 % des mots peuvent être traités de façon syllabique classique. Alors on est porté à croire que la méthode et les quelques règles que nous venons d'exposer conviendront dans la grande majorité des cas pour les textes usuels.

5. Un peu de T_EXnique

Le procédé utilisé par T_EX pour choisir les points de césure est exposé dans le T_EXbook [17] appendice H p. 449.

Rappelons ici quelques notions nécessaires à la lecture d'un fichier de césures. Un tel fichier se compose essentiellement d'une macro `\patterns{}` dont le contenu est une suite de lignes contenant chacune un ou plusieurs motifs séparés par un espace.

Ce fichier principal peut éventuellement être accompagné d'un fichier d'exceptions (par exemple `enhyphex.tex` contient les exceptions au fichier de motifs anglo-américains, `f7hyphex.tex`¹³ contient les exceptions françaises en format 7-bits).

5.1. `\patterns` : un motif de discussion...

Pour préciser l'interprétation des motifs de césure, nous allons considérer quelques lignes extraites du fichier `f7hyph.tex`¹⁴:

```
\patterns{...
        il2l
        vil3l
        chevil4l
```

13. Il existe aussi `f8hyphex.tex` qui lui est équivalent mais sous une forme 8-bits. Dans la pratique c'est le fichier `frhyphex.tex` qui est appelé et qui décide quel est le fichier à charger en mémoire, soit le 7-bits, soit le 8-bits. Tous ces fichiers font partie intégrante de la distribution des *fichiers du style french* sur le serveur d'archives GUTenberg : `ftp.univ-rennes1.fr` dans le répertoire `pub/GUTenberg/french`.

14. C'est le code contenu dans le fichier `frhyph.tex` qui est exécuté et qui décide quel est le fichier à charger en mémoire, soit le 7-bits `f7hyph.tex`, soit le 8-bits `f8hyph.tex`.

```
...
4che.
4ches.
.ch4
...}
```

Le principe est le suivant : les chiffres impairs indiquent les coupures possibles, les chiffres pairs non nuls les coupures interdites. Considérons les mots en *ill* : lorsque ce groupe se prononce [ij] (grillage, fille...), la coupure *il-l* est interdite par le motif *il2l*, mais cette même coupure entre les deux *l* est permise dans *vil-lage* pour le son [il] : on ajoute pour ce faire le motif *vil3l* ; ensuite il faut apporter une correction pour que le mot *cheviller* (phonème [j]) ne soit pas coupé entre les deux *l* : c'est l'objet du motif *chevil4l*. T_EX prend en compte comme valeur paire ou impaire le maximum des chiffres rencontrés dans tous les motifs concernés : si une coupure est interdite par un 2 on peut l'autoriser dans certains cas particuliers par un 3, ensuite l'interdire de nouveau par un 4 pour certaines exceptions et ainsi de suite.

On remarque dans la macro `\patterns` des lignes commençant ou se terminant par des points : en début de ligne un point signifie que le motif de la ligne ne s'applique qu'aux mots *commençant* par lui, le point en fin de ligne indique que le motif concerne seulement les mots *se terminant* par lui. En français, comme nous l'avons rappelé précédemment, on ne renvoie jamais à la ligne suivant une terminaison de mot muette : par exemple les coupures devant la syllabe terminale *che(s)* des mots *pervenche(s) cravache(s) fantoche(s)*... sont interdites par les motifs *4che.* et *4ches.* De même, sans le motif *.ch4* la coupure *ch/rysanthème* serait possible...

5.2. Mettre l'accent : comment ?

Pour les langues utilisant des caractères accentués se pose le problème de la représentation de ces caractères dans les motifs de césure. Il existe deux représentations possibles, le choix dépend du moteur T_EX utilisé :

- si on travaille avec un moteur M_IT_EX (cas de la distribution UNIX de GUTenberg) il est possible d'utiliser une représentation des caractères accentués sous la forme `\'e` pour *é* comme cela est fait dans les fichiers *f7hyph.tex* et *f7hyphex.tex* pour les césures françaises ;

- si on travaille avec `TeX V3.xx` et des caractères 8-bits, soit avec des fontes DC/EC (cas de la distribution MAC de GUTenberg), soit avec des fontes virtuelles (cas de la distribution PC de GUTenberg, à base `emTeX`) il faut utiliser une représentation des caractères accentués sous la forme hexadécimale `^^xx` décrite dans [17] p.73 (`^^e9` pour é); c'est ce qui est fait dans les fichiers `f8hyph.tex` et `f8hyphex.tex` pour les césures françaises.

Rappelons enfin que pour obtenir toutes les coupures possibles des mots accentués, il faut¹⁵ impérativement utiliser soit `MTex`, soit un `TeX V3.xx` associé à des fontes 8-bits (virtuelles ou réelles comme les DC). En effet, si les caractères accentués sont fabriqués à partir de la macro `\accent` (qui superpose le caractère et l'accent), toute coupure est inhibée après le premier caractère accentué du mot.

6. Petite histoire de la césure française dans `TeX`

Les premiers travaux sur l'adaptation de `TeX` aux césures françaises ont été faits par J. DÉARMÉNIEN qui a commencé son travail à Stanford au début des années 80. Une première publication a été faite dans le *TUGboat* en 1984 [4]. Mais l'essentiel du travail de J. DÉARMÉNIEN sur les césures est exposé dans la revue *TSI* [5].

La solution technique utilisée par J. DÉARMÉNIEN pour parvenir à des coupures correctes des mots accentués reposait sur l'utilisation de fontes francisées: les caractères spécifiques du français (â, ê, î, ô, û, é, ç, ë, è, î) remplaçaient les caractères grecs placés par Don KNUTH en position 01 à 10 des fontes CMR¹⁶.

Dans le même temps M.-J. FERGUSON apportait une solution intéressante aux problèmes multilingues avec son `MTex`. Il devenait possible de couper des mots contenant des lettres accentuées sans recourir à des fontes francisées. Le fichier des motifs de césure de J. DÉARMÉNIEN a donc été revu et adapté au codage 7-bits standard de `TeX`. `MTex` a commencé à cette époque à être très largement diffusé dans les milieux francophones (aidé en cela par les distributions `PCTex` de Personal `TeX` Inc.). Les der-

15. Ces conditions ne sont pas entièrement suffisantes, les `\lccode` et `\catcode` des caractères utilisés dans le texte étant, notamment, très importants.

16. Le « à » n'a pas pu trouver place dans les dix positions réservées aux caractères grecs, mais cette lettre n'intervient qu'en fin de mot (*voilà*) ou dans des mots composés: (*tout-à-l'égout*).

nières « *marques* » de M.-J. FERGUSON dans les fichiers semblent dater de juin 88.

Ensuite est sorti T_EX V3, reprenant l'essentiel (mais pas la totalité) des idées de M.-J. FERGUSON. Cette différence a amené M.-J. FERGUSON à faire son M_TE_X V3.

Une autre façon de traiter le problème des césures de mots accentués, qui généralise l'approche de J. DÉSARMÉNIEN, consiste à disposer dans chaque fonte de tous les caractères accentués nécessaires. L'adoption à Cork en 1990 [8] du codage de fontes EC¹⁷ à 256 caractères résout donc définitivement le problème pour le français comme pour toutes les autres langues européennes, à condition toutefois de disposer de fontes réelles (DC¹⁸ par exemple) ou virtuelles.

L'apparition de cette norme a motivé le transcodage des fichiers de césures prévus pour M_TE_X en fichiers 8-bits. Le transcodage naïf consiste à remplacer les séquences du type \'*e* par le caractère 8-bits correspondant... Mais le code de ce caractère dépend bien sûr de la machine utilisée (les codages PC, Mac, ISO-Latin1 et PostScript sont *tous* différents). Il faudrait donc un fichier 8-bits spécial pour chaque machine... Il existe malheureusement encore aujourd'hui sur les réseaux, y compris dans les toutes nouvelles archives CTAN, des versions utilisant ce type de codage (cf. `f8hyph.tex` récupéré le 01/12/93 sur `ftp.tex.ac.uk`).

La seule façon raisonnable de coder un fichier de motifs de césure en 8-bits est d'utiliser la notation hexadécimale $\wedge^{\wedge}xy$ de T_EX V3 et de respecter le codage le plus standard, c'est-à-dire le codage DC. Cette notation interne à T_EX a deux qualités essentielles : elle est indépendante de la machine utilisée¹⁹ et elle résiste (en principe) aux transferts par les réseaux puisqu'elle ne fait appel à aucun caractère de code supérieur à 127.

C'est ce codage qu'a fort judicieusement choisi Y. HARALAMBOUS pour créer son fichier de motifs adapté aux fontes DC²⁰ ; malheureusement la liste des motifs de césure dont il est parti était altérée. Ce fichier est encore présent dans les archives sous le nom de `fr8hyph.dc` ou `fr8hyph.tex` (version 1.0 du 10 juillet 1991)...

17. Ce codage est désormais appelé T1 avec L^AT_EX2 ϵ .

18. Le créateur de ces fontes considère que leur nom est temporaire ; le nom définitif doit être EC.

19. Sous réserve d'utiliser des fontes avec le même codage.

20. Il est nécessaire de disposer du fichier de motifs adapté au codage de la police utilisée à la composition du document.

Au moment où nous avons effectivement commencé notre travail (fin 1991) de nombreuses versions différentes des motifs de césure français étaient disponibles sur les réseaux. Celles que nous possédions nous-mêmes étaient-elles correctes? Les fichiers originaux de J. DÉSARMÉNIEN avaient de toute évidence subi des ajouts et des modifications, certaines judicieuses et d'autres moins: comment faire le tri? Dans un premier temps nous avons décidé de travailler uniquement sur les fichiers de motifs prévus pour $\text{M}\text{T}\text{E}\text{X}$ et de transcoder ceux-ci, lorsqu'ils seraient au point, en version 8-bits.

7. Notre travail

Nous avons commencé par une comparaison, à la main, des différents fichiers de césure française à notre disposition, dans le domaine public, en vue de déterminer lequel pouvait être considéré comme le meilleur point de départ pour notre étude. Très vite il a fallu se rendre à l'évidence et renoncer devant l'énormité du travail. Nous disposions de 5 fichiers d'origines diverses mais tous étaient vraiment différents.

Dans le doute, nous avons pris comme référence le fichier de césure distribué par GUTENBERG depuis près de trois ans, pour le comparer avec les motifs publiés par J. DÉSARMÉNIEN. Nos comptages manuels font apparaître 259 différences (tableau 1).

TABLE 1 - *Divergences survenues depuis le travail de J. DÉSARMÉNIEN*

Motifs de césure	modifiés	disparus	nouveaux	total
étymologiques	25	63	69	157
phonétiques	0	60	42	102
	25	123	111	259

Notre fichier de référence n'était visiblement pas bon. Il fallait donc refaire le décompte sur d'autres fichiers mais sans avoir la garantie que celui qui se rapproche le plus de celui de J. DÉSARMÉNIEN soit vraiment le meilleur. Alors nous avons décidé, non plus de comparer les motifs de césure, mais de comparer les résultats obtenus avec la liste des 499 mots publiée en annexe de l'article de J. DÉSARMÉNIEN [5] (nous avons ajouté en fait un 500^e mot, *correctement*, souvent mal coupé: **cor-re/cte-ment** au lieu de **cor-rec-[te-ment]**). Il fallait alors automatiser ce processus de comparaison pour pouvoir le répéter souvent.

Nous avons donc créé les outils permettant d'effectuer les comparaisons : un premier fichier de code T_EX lit le fichier des 500 mots tests correctement *hyphénés*²¹ en ignorant les points de césure présents et crée un fichier `.log` comportant toutes les coupures proposées par le système. Un deuxième relit ce fichier `.log` et compare les points de césure trouvés à ceux du fichier test.

Dès les premiers essais, des anomalies graves apparurent dans la majorité des fichiers de motifs de césure disponibles :

- les points en début de ligne étaient souvent remplacés par des doubles points, ce qui a pour effet pratique d'invalider totalement le motif correspondant ; probablement est-ce dû à des problèmes de transmission sur les réseaux...
- le plus souvent, tous les motifs concernant *ill* (liste allant des motifs `il21` à `xi31` chez J. DÉARMÉNIEN) avaient disparu... Est-ce le résultat d'une fausse manœuvre d'un contributeur anonyme ?
- les motifs `.con4` et `.cons4` (règle g-e) présents chez J. DÉARMÉNIEN étaient absents dans plusieurs fichiers.
- un motif `1ct` a été ajouté, par qui ? Pourquoi ? Il a pour effet fâcheux de faire couper *correctement* en *cor-re/cte-ment* !

Inversement, d'autres personnes semblent être intervenues (judicieusement) sur les transcriptions du fichier original en vue de son utilisation par M_IT_EX, comme celles de M.-J. FERGUSON (juin 88) pour ajouter un 0 derrière les `\i` et les `\oe` de façon à permettre effectivement la coupure après les motifs concernés.

Comme on le voit, ces fichiers de motifs mis à la disposition de tous, dans le domaine public, ont pu être utilisés par tous mais aussi modifiés avec plus ou moins de bonheur. Nous étions alors encore plus confortés dans l'idée qu'il était indispensable de mettre un peu d'ordre et de laisser ensuite l'association GUTenberg seul maître et garant de leur pérenité.

Après toutes ces corrections, notre même fichier de motifs (qui comportait 259 anomalies) donna, *seulement*, 118 mots mal divisés sur les 500. Ce n'était pas encore un exploit !

21. Cet anglicisme signifie que chaque mot comportait les points de césure (sous forme de trait d'union) proposés par J. DÉARMÉNIEN.

Le meilleur fichier de motifs disponible s'est avéré être celui de la machine Gould de l'ex-CICB de Rennes²² qui donnait seulement 7 différences significatives²³ par rapport à la référence. Le tableau 2 récapitule les 7 différences observées. Nous avons choisi d'en faire notre référence et d'essayer

TABLE 2 - *Dernières différences observées.*

Original de J.D.	Notre résultat	Index
désami-don-ner	dés-ami-don-ner	(1)
désen-flure	dés-en-flure	(2)
dés-hy-dro-gé-ner	dés-hy-dro-gé-ner	(3)
dés-rien-ter	dés-orien-ter	(4)
pé-réqua-tion	pér-équa-tion	(5)
phy-topph-thora	phy-to-phthora	(6)
pontet	pon-tet	(7)

de supprimer les différences observées en le modifiant le moins possible. Le premier type de différences concerne les coupures étymologiques des mots commençant par `d\'es`. Les motifs concernant ces mots semblent avoir été profondément modifiés par rapport à la version de J. DÉARMÉNIEN. Nous ne disposons d'aucune indication ni sur l'auteur de ses modifications ni sur ses intentions.

Nous avons réintroduit le motif `.d\'e2s` qui était présent dans le fichier original de J. DÉARMÉNIEN, et supprimé les motifs suivants :

- `.d\'e2s1a2` responsable de la différence (1)
- `.d\'e2s1e2` responsable de la différence (2)
- `.d\'e2s3h` responsable de la différence (3)
- `.d\'e2s1o2` responsable de la différence (4)

De plus, nous avons ajouté le motif `.d\'es2a3m` pour obtenir la division `désa-mi-don-ner` et non `désami-don-ner`. Ce choix n'est pas celui de J. DÉARMÉNIEN, mais il nous semble préférable de traiter de la même

22. Actuellement appelé CRI de l'université de Rennes 1.

23. Le premier essai nous donna en fait 6 erreurs de plus avec `MITEX V3`, ce qui était dû à des fautes de frappe. Mais nous avons surtout été troublés un temps par les 33 erreurs produites par `MITEX V2`. M.-J. FERGUSON nous expliqua alors que les accents perturbaient le décompte des caractères en fin de mot. De fait, tous les mots supplémentaires en erreur se terminaient par un e accent aigu.

manière désa-mi-don-ner, désa-mor-cer et désa-bu-ser. Enfin, la ré-introduction du motif `.d\'e2s` nous a obligés à augmenter la force de la césure des motifs présents dans le fichier original `.d\'e1s2c` `.d\'e1s2p` et `.d\'e1s2t` en `.d\'e3s2c` `.d\'e3s2p` et `.d\'e3s2t`.

La différence (5) provenait vraisemblablement d'une faute de frappe (`p\'e2r1\'e2q` au lieu de `p\'e1r2\'e2q`) d'un auteur inconnu car ce motif n'existait pas chez J. DÉSARMÉNIEN.

Le motif `.phyto3ph2` (ajouté par qui?) n'était pas bon, il donnait `phy-to-phthora` (6) au lieu de `phy-toph-[thora`; nous l'avons supprimé.

Enfin le cas du mot `pon/tet` (7) peut être réglé en ajoutant le motif `.pon2tet`.

Nous avons remplacé le motif `p\'e2n1u1` par `p\'e2nul` conformément au fichier original de J. DÉSARMÉNIEN (pour `an-té-pénul-[tième`).

Autres modifications introduites :

- rétablissement des motifs `.re2s3cisi` et `.re2s3ciso`, présents chez J. DÉSARMÉNIEN mais absents du fichier de Rennes;
- les exceptions précisées en e-a ont aussi été ajoutées²⁴;
- rétablissement du motif `1\c c` présent chez J. DÉSARMÉNIEN mais absent du fichier de Rennes;
- ajout du motif `.cul4` (même raison que `.con4` et `.cons4`);
- le mot `ré/union` était mal coupé (règle s-c); nous avons remplacé le motif en cause (`.r\'e1u2`) par `.r\'eu2`;
- nous n'avons pas voulu retirer les motifs de début de mot à 1 caractère ainsi que ceux de fin de mot à 2 caractères qui n'ont pas de sens avec `TeX`, puisque par défaut il ne coupera jamais de la sorte; mais les deux paramètres `\righthyphenmin` et `\lefthyphenmin` qui permettent, depuis la version 3.00 de `TeX`, d'imposer le minimum de lettres en fin de ligne et le minimum à rejeter sur la ligne suivante, peuvent avoir été malencontreusement changées; et puis il reste aussi quelques installations `MITeX V2` pour lesquelles l'algorithme de dénombrement des caractères de début et de fin de mot n'est par sûr à 100 %;

24. Nous les avons introduites dans un premier temps dans les fichiers d'exceptions, mais il est beaucoup plus performant, en terme de gestion mémoire, de les mettre sous forme de motifs de césure.

- nous avons complètement revu l'ordre des motifs de césure afin d'essayer de faciliter la lecture et la maintenance ultérieure du fichier : les motifs étymologiques regroupés en deuxième partie du fichier par J. DÉARMÉNIEN ont été interclassés avec les motifs phonétiques²⁵. L'ordre retenu est l'ordre alphabétique, à ceci près que nous avons regroupé les motifs agissant sur le même groupe de lettres. Par exemple, le motif générique *il2l* est à sa place alphabétique et les motifs complémentaires concernant le *ill* (*ci13l*, *rcil4l*, ... *xi13l*) sont rassemblés à la suite de celui-ci et indentés de manière à aligner verticalement le groupe *ill* afin d'améliorer la lisibilité. De même, nous avons préféré ne pas séparer les terminaisons muettes : par exemple les motifs *4de.*, *4des.*, *2dent.* sont dans cet ordre, les exceptions étymologiques *.d\'e1a2*, *.d\'e1io*, *.d\'e1o2*, étant reléguées après, en (légère) violation de l'ordre alphabétique.

Le test des 500 mots demeurant pour nous trop réduit par rapport à la richesse de notre vocabulaire français, nous avons souhaité vérifier les césures obtenues sur plusieurs milliers de mots. Il aurait été souhaitable de reprendre le fichier du « vocabulaire informatisé du laboratoire d'automatique et de linguistique » de Paris VII fourni par le professeur GROSS comme base de travail à J. DÉARMÉNIEN mais nous ne l'avions pas. Par contre nous avons pu bénéficier d'une convention de recherche avec le GDR-PRC « Communication Homme-Machine » du CNRS, ce qui nous a permis d'utiliser la *base de données lexicales du français écrit et parlé* (BDLEX), soit environ 330 000 formes fléchies du français.

De ce volumineux fichier nous avons extrait aléatoirement 3 000 mots et produit la liste des mots avec les césures proposées par T_EX. Nous avons relevé quelques problèmes, notamment les terminaisons muettes de pluriel de verbes (règle s-d), telles que :

at-té-nuè/rent fric-tion/nent se-ri-nas/sent

Aussi avons-nous décidé d'essayer de régler le problème des terminaisons en **-ent**, celles-ci n'ayant pas été traitées par J. DÉARMÉNIEN. Le problème n'est pas simple car ces terminaisons peuvent être de trois types :

- muettes : les verbes à la troisième personne du pluriel (cf. ci-dessus), les plus nombreux et de loin, étant en **-sent** et **-rent** (imparfait du subjonctif et passé simple...),

25. Dans les fichiers **f7hyph.tex** et **f8hyph.tex**, les motifs étymologiques sont indentés de 20 caractères de façon à les distinguer des motifs phonétiques. Dans la liste publiée en appendice, ils sont précédés du signe |.

- sonores : principalement les adverbes en **-ment** (*vraiment, sûrement*) et d'autres mots (*comment, élément, parent, présent*),
- mixtes, sonores ou muettes selon le cas : **président** (*ils président, le président*), **pressent** (*ils pressent, il pressent*), **équivalent, ferment** etc.

T_EX n'analyse pas le contexte d'un mot et ne sait pas couper *ils pré-sident* dans un cas et *le pré-si-dent* dans l'autre. Nous avons donc choisi de traiter les terminaisons mixtes comme si elles étaient muettes et donc *insé-cables* : nous avons préféré interdire la coupure possible *le prési-dent* plutôt que d'accepter la césure incorrecte *ils prési-dent* ; ce mot sera dans tous les cas coupé par T_EX en **pré-[si/dent**. L'auteur a toujours la possibilité d'infléchir *localement* ce choix en codant dans le texte **pré\si\dent** au lieu de **président**.

La base BDLEX nous a été d'une grande utilité pour faire le tri des terminaisons : grâce aux utilitaires standard d'UNIX, nous avons extrait *tous* les mots se terminant par **ent**, puis nous les avons triés par ordre alphabétique mais *de droite à gauche* afin de dégager des règles de classement en motifs sonores ou muets. Une lecture attentive du fichier produit nous a permis d'établir la liste de motifs à ajouter. Le fichier BDLEX nous a semblé assez complet pour permettre une classification satisfaisante.

Le tableau 3 présente le bilan de toutes les terminaisons en **ent**. présentes dans le fichier BDLEX.

L'ajout des 182 motifs proposés permet de corriger 18 331 coupures incorrectes parmi les mots du fichier BDLEX.

Prenons, pour expliquer notre démarche, l'exemple des terminaisons en **-cent** : elles sont en général muettes (115 contre 47, cf. tableau 3), on ajoute donc le motif de base **2cent**. pour interdire en général la coupure devant le **c**. Mais lorsque la syllabe **-cent** est précédée de **-ja-** (*adjacent, subja-cent*) ou de **-é-** (*décent, indécent, récent*) elle est sonore et on corrige en ajoutant **ja3cent**. et **\'e3cent**. pour autoriser la coupure devant le **c** dans ce cas. Les terminaisons en **-escent** (*évanescent, incandescent, sénes-cent...*) sont également sonores sauf *acquiescent*, on ajoute donc **es3cent**. et **acquies4cent**.

On comprend sur cet exemple que l'ordre pertinent pour l'analyse des *terminaisons* de mots est l'ordre alphabétique *de droite à gauche*. C'est dans cet ordre qu'on été rangés les motifs corrigeant chaque motif de base

TABLE 3 - *Fréquences des terminaisons en -ent dans BDLEX.*

	muettes	sonores	motifs ajoutés		muettes	sonores	motifs ajoutés
-bent	55	0	1	-kent	1	0	1
-blent	35	0	1	-lent	647	22	18
-brent	29	0	1	-ment	164	2247	58
-cent	115	47	12	-nent	698	14	8
-chent	179	0	1	-pent	140	3	4
-ckent	2	0	1	-phent	5	0	1
-clent	16	0	1	-plent	14	0	1
-crent	9	0	1	-prent	3	0	1
-dent	313	16	12	-quent	183	6	4
-dlent	4	0	1	-rent	6767	21	7
-drent	12	0	1	-sent	7307	4	4
-fent	44	0	1	-shent	1	0	1
-flent	27	0	1	-tent	877	16	9
-frent	16	0	1	-trent	70	0	1
-gent	203	18	14	-vent	162	2	3
-glent	18	0	1	-vrent	21	0	1
-gnent	86	0	1	-went	1	0	1
-grent	12	0	1	-xent	19	0	1
-guent	70	1	2	-zent	5	1	2
-jent	1	0	1	Total	18 331	2 418	182

dans les fichiers `f7hyph.tex` et `f8hyph.tex`: ainsi `2cent.` est placé à la lettre `c`, après les deux motifs `4ce.` et `4ces.` qui ont la même fonction que lui, et les exceptions (`ja3cent ... is3cent.`, `immis4cent.`) sont classées par ordre alphabétique *de droite à gauche* à la suite, tous les motifs du groupe ayant été alignés sur le point final afin d'améliorer la lisibilité.

8. Conclusion provisoire et perspectives

Le présent travail constitue une remise en ordre des fichiers de césures françaises de J. DÉARMÉNIEN. Nous avons essayé de faire le tri des améliorations apportées et des détériorations survenues. Nous ne prétendons pas que les fichiers de motifs (`f7hyph.tex`, `f8hyph.tex`) soient parfaits, il y a encore beaucoup de travail à faire sur ce sujet : si les motifs phonétiques nous semblent bien au point, les motifs étymologiques mériteraient eux d'être testés plus avant, mais aucun des trois auteurs n'a eu le temps, à ce jour, de se mettre à ce travail. Toutes les propositions de collaboration dans ce domaine sont évidemment les bienvenues.

Les nouveaux fichiers de césure se trouvent dans la distribution des *fichiers de style french*, ils seront inclus dans les distributions GUTenberg. Des versions bien antérieures qui sont encore actuellement diffusées sur les réseaux, sont loin d'être satisfaisantes. Nous comptons faire en sorte que toutes les versions altérées soit retirées des archives CTAN et remplacées par les seuls fichiers :

- **frhyph.tex** (V1.06 du 20/4/94) : code T_EX effectuant le chargement en mémoire des motifs de césure sous leur forme 7 ou 8-bits selon le cas le plus approprié ;
- **f7hyph.tex** (V2.0 du 20/5/94) motifs de césure 7-bits ;
- **f8hyph.tex** (V2.0 du 20/5/94) motifs de césure 8-bits (codage DC).

Toutes les suggestions et améliorations peuvent nous être transmises à l'adresse électronique suivante : **cesure-l@ens.fr** ou directement aux auteurs.

Remerciements Nous tenons à remercier les diverses personnes qui ont bien voulu critiquer une version provisoire de cet article, et en particulier Jacques ANDRÉ.

Références bibliographiques

- [1] Association suisse des compositeurs à la machine : *Guide du typographe romand*, 5^e édition, 1994
- [2] M. ARRIVE, F. GADET et M. GALMICHE, *La grammaire d'aujourd'hui : guide alphabétique de linguistique française*, Flammarion, Paris, 1986.
- [3] H. BONNARD : *Code du français courant*, Magnard, 1981.
- [4] J. DÉSARMÉNIEN : « How to run T_EX in a French environment : Hyphenation, Fonts, Typography », *TUGboat*, vol. 5 no. 2, 1984, pp. 91–102.
- [5] J. DÉSARMÉNIEN : « La division par ordinateur des mots français : application à T_EX », *Techniques et Sciences Informatiques*, vol. 5 no. 4, 1986, pp. 251–265.
- [6] B. DUPRIEZ, *Gradus : les procédés littéraires (dictionnaire)*, Union générale d'Éditions, Paris, “10-18”, 1984.
- [7] Fédération CGC de la communication : *Code typographique, Choix des règles à l'usage des auteurs et professionnels du livre*, 16^e édition, 1989.

- [8] M.-J. FERGUSON : « Fontes latines européennes et T_EX 3.0 », *Cahiers GUTenberg* N° 7, 1990.
- [9] D. FLIPO and L. SIEBENMANN : « Hyphenation in presence of accents and diacritics. An easy and low-cost solution » *Cahiers GUTenberg* N° 15 (EuroT_EX 92), 1992.
- [10] A. FREY : *Manuel nouveau de typographie*, réed. Léonce Laget, 1979, *épuisé*.
- [11] B. GAULLE : *Manuel d'utilisation du style french*, document électronique fourni avec la distribution logicielle, 6^e édition, 1994.
- [12] J. GIRODET : *Dictionnaire du bon français*, Bordas, 1981.
- [13] M. GOOSSENS, F. MITTELBACH and A. SAMARIN : *The L^AT_EX Companion*, Addison-Wesley, New York, 1994.
- [14] C. GOURIOU : *Mémento typographique*, Hachette, 1973.
- [15] M. GRÉVISSE : *Le bon usage*, 13^e édition Duculot, 1994.
- [16] Imprimerie nationale : *Lexique des règles typographiques en usage à l'Imprimerie nationale*, 3^e édition, 1990.
- [17] D. E. KNUTH : *The T_EXbook*, Addison Wesley, 1991.
- [18] F. M. LIANG : *Word Hy-phen-a-tion by Comput-er*, Ph. D. Thesis, Departement of Computer Science, Stanford University, report N° STANCS-83-977, 1983.
- [19] R. THIMONNIER : *Code orthographique et grammatical*, Marabout, 1978, *épuisé*.

Annexe Liste des motifs de césure français

Les lettres accentuées des motifs de césure sont exprimées en notation \TeX de base (ASCII c'est-à-dire 7-bits). Chaque macro-instruction terminale est suivie du chiffre 0 pour clore effectivement le motif. Rappelons que les chiffres impairs à l'intérieur des motifs autorisent la coupure des mots calqués sur ce motif (à condition qu'aucun autre motif ne le contredise, voir à ce sujet le paragraphe 5.1). Les motifs étymologiques sont ici précédés du caractère |. Le classement est d'ordre alphabétique, de gauche à droite pour l'essentiel, mais aussi de droite à gauche pour les exceptions (on les remarquera par leur alignement à droite).

```
\patterns{
2'2      | .bi1a2c      4che.      | co1assur
.a4      | .bi1a2t      4ches.      | co1au
'a4      | .bi1au      2chent.      | co1ax
.\^a4    | .bio1a2      | .ch\^e2vre3feuille 1c\oe0
'\^a4    | .bi2s1a2    2chg      | co1\^e2
| .ab3r\^ea | .biu2      4ch2l      | co1ef
| 'ab3r\^ea | 1b2l      4chle.      | co1en
| a1\^e2dre | 4ble.      | 4chles.      | co1ex
| .ae3s4ch | 4bles.      | chlo2r3a2c    | .con4
| 'ae3s4ch | 2blent.      | | chlo2r3\^e2t    | .cons4
| 1alcool  | 1bo        2chm      | .contre1s2c
| a2l1algi | 1b\^o      2chn      | .contre3ma\^i0tre
| .amino1a2c | 1b2r      2chp      | .co2nurb
| 'amino1a2c | 4bre.      ch2r      | .co1o2
| .ana3s4tr | 4bres.      4chre.      | .co2o3lie
| 'ana3s4tr | 2brent.      4chres.      | 1c2r
| 1a2nesth\^esi | 1bu      2chs      | 4cre.
| .anti1a2  | 1by        2cht      | 4cres.
| 'anti1a2  | 1\c c      2chw      | 2crent.
| .anti1e2  | 1ca        1ci      | 1cu
| 'anti1e2  | 1c\^a      1c\^i0      | 1c\^u
| .anti1\^e2 | ca3ou3t2   | .ci2s1alp     | 1cy
| .anti2enne | 1ce        1c2k      | .cul4
| 'anti2enne | 1c\^e      4ck.      | 1d'
| .anti1s2   | 1c\^e      2ckb      | 1da
| 'anti1s2   | 1c\^e      4cke.      | 1d\^a
| .apo2s3ta  | 4ce.       4ckes.      | | .dacryola2
| 'apo2s3ta  | 4ces.      2ckent.      | did2h
| apo2s3tr   |           2ckf      | 1de
| archi1\^e2pis |           2ckg      | 1d\^e
| .as2ta     |           2ck3h     | 1d\^e
| 'as2ta     |           2ckp      | 4de.
| a2s3tro    |           2cks      | 4des.
| 1ba        |           2ckt      |
| 1b\^a      |           1c2l      | 2dent.
| .bai2se3main |           4cle.      | d\^eca3dent.
| 1be        |           4cles.      | \^e3dent.
| 1b\^e      |           2clent.      | cci3dent.
| 1b\^e      |           1co        | inci3dent.
| 1b\^e      |           1c\^o      | confi3dent.
| 1b\^e      |           | colacc      | tri3dent.
| 4be.       |           | colacq      | dissi3dent.
| 4bes.      |           | co1a2d      | chien3dent.
| 2bent.     |           | co1ap       | .ar3dent.
| 1bi        |           | co1ar       | impu3dent.
|           |           | colassoc    | pru3dent.
```

Motifs de césure français

.d\'e1a2	\'^e4	tan3gent.	hyperu2
.d\'e1io	.\'e4	rin3gent.	hypo1a2
.d\'e1o2	\'e4	contin3gent.	hypo1e2
.d\'e2s	.\'e4	.ar3gent.	hypo1\'e2
.d\'e3s2a3cr	\'e4	\'ar3gent.	hypo1i2
.d\'es2a3m	1\'e2drie	ser3gent.	hypo1o2
.d\'e3s2a3tell	1\'e2drique	ter3gent.	hypo1s2
.d\'e3s2astr	1\'e2lectr	r\'esur3gent.	hypo1u2
.d\'e3s2c	1\'e2l\'ement	1g2ha	.i4
.d\'e2s1\'e2	.en1a2	1g2he	\'i4
.d\'e3s2\'e3gr	\'en1a2	1g2hi	.\'^i4
.d\'e3s2ensib	1\'e2nerg	1g2ho	\'^i4
.d\'e3s2ert	e2n1i2vr	1g2hy	i1algi
.d\'e3s2exu	.en1o2	1gi	i1arthr
.d\'e2s1i2	\'en1o2	1g\'^i0	i1\'e2dre
.d\'e3s2i3d	\'epi2s3cop	1g2l	i12l
.d\'e3s2i3gn	\'epi3s4cope	4gle.	cil3l
.d\'e3s2i3li	e2s3cop	4gles.	rcil4l
.d\'e3s2i3nen	.eu2r1a2	2glent.	ucil4l
.d\'e3s2invo	\'eu2r1a2	1g2n	vacil4l
.d\'e3s2i3r	eu1s2tat	.sta2g3n	gil3l
.d\'e3s2ist	extra1	wa2g3n	hil3l
.d\'e3s2o3d\'e	extra2c	4gne.	lil3l
.d\'e2s1\'oe0	extra2i	4gnes.	l3lion
.d\'e3s2o3l	1fa	2gnent.	mil3l
.d\'e3s2o3pil	1f\'^a	1go	mil4let
.d\'e3s2orm	1fe	1g\'^o	\'emil4l
.d\'e3s2orp	1f\'e	1g2r	semil4l
.d\'e3s2oufr	1f\'e	4gre.	rmil4l
.d\'e3s2p	1f\'e	4gres.	armil5l
.d\'e3s2t	4fe.	2grent.	capil3l
.d\'e2s1u2n	4fes.	1gu	papil3la
3d2hal	2fent.	1g\'^u	papil3le
3d2houd	1fi	gis2	papil3li
1di	1f\'^i0	4gue.	papil3lom
1d\'^i0	1f2l	4gues.	pupil3l
di2s3cop	4fle.	2guent.	piril3l
.di1a2c\'e	4fles.	.on3guent.	thril3l
.di1a2cid	2flent.	\'on3guent.	cyril3l
dia2g3n	1fo	1gy	ibril3l
.di1ald	1f\'^o	1ha	pusil3l
.di1a2mi	1f2r	1h\'^a	.stil3l
.di1a2tom	4fre.	1he	distil3l
.di1e2n	4fres.	1h\'e	instil3l
.di2s3h	2frent.	1h\'e	fritil3l
2dlent.	f1s2	1h\'e	boutil3l
1do	1fu	h\'emi1\'e	vanil3lin
1d\'^o	1f\'u	h\'emo1p2t	vanil3lis
1d2r	1fy	4he.	vil3l
4dre.	1ga	4hes.	avil4l
4dres.	1g\'^a	1hi	chevil4l
2drent.	1ge	1h\'^i0	uevil4l
d1s2	1g\'e	1ho	uvil4l
1du	1g\'e	1h\'^o	xil3l
1d\'u	1g\'e	1hu	1informat
1dy	4ge.	1h\'u	.in1a2
.dy2s3	4ges.	1hy	\'in1a2
.dy2s1a2	2gent.	hypera2	.in2a3nit
.dy2s1i2	r\'e3gent.	hypere2	\'in2a3nit
.dy2s1o2	entre3gent.	hyper\'e2	.in2augur
.dy2s1u2	indi3gent.	hyperi2	\'in2augur
.e4	dili3gent.	hypero2	.in1e2
\'e4	intelli3gent.	hypers2	\'in1e2
.\'e4	indul3gent.	hype4r1	.in1\'e2

'in1\`e2	1k\`e	.ma2r1x	fu2ment.
.in2effab	4ke.	1me	hu2ment.
'in2effab	4kes.	1m\`e	fichu3ment.
.in2\`e3lucta	2kent.	1m\`e	llu2ment.
'in2\`e3lucta	1k2h	1m\`e	plu2ment.
.in2\`e3narra	4kh.	.m\`e2g1oh	bou2ment.
'in2\`e3narra	.kh4	.m\`e2sa	bru2ment.
.in2ept	1ki	.m\`e3san	su2ment.
'in2ept	1k\`i0	.m\`e2sies	tu2ment.
.in2er	1ko	.m\`e2s1i	1mi
'in2er	1k\`o	.m\`e2s1u2s	1m\`i0
.in2exora	1k2r	.m\`eta1s2ta	.milli1am
'in2exora	1ku	4me.	1m2n\`emo
.in1i2	1k\`u	4mes.	1m2n\`es
'in1i2	1ky	\^a2ment.	1m2n\`esi
.in2i3miti	1la	da2ment.	1mo
'in2i3miti	1l\`a	fa2ment.	1m\`o
.in2i3q	1l\`a	amalg2ment.	1m\`oe0
'in2i3q	la2w3re	cla2ment.	.mono1a2
.in2i3t	1le	ra2ment.	.mono1e2
'in2i3t	1l\`e	temp\`era3ment.	.mono1\`e2
.in1o2	1l\`e	ta2ment.	.mono1i2
'in1o2	1l\`e	testa3ment.	.mono1\`i2d\`e
.in2o3cul	4le.	qua2ment.	.mono1o2
'in2o3cul	4les.	\`e2ment.	.mono1u2
.in2ond	2lent.	car\`e2ment.	.mono1s2
'in2ond	.ta3lent.	diaphrag2ment.	mon2t3r\`eal
.in1s2tab	iva3lent.	ryth2ment.	m1s2
'in1s2tab	\`equiva4lent.	ai2ment.	1mu
'inte4r3	monova3lent.	rai3ment.	1m\`u
.intera2	polyva3lent.	ab\`i2ment.	1my
'intera2	re3lent.	\`eci2ment.	moye2n1\`a2g
.intere2	.do3lent.	vidi2ment.	1na
'intere2	indo3lent.	subli2ment.	1n\`a
.inter\`e2	inso3lent.	\`eli2ment.	1ne
'inter\`e2	turbu3lent.	reli2ment.	1n\`e
.interi2	succu3lent.	mi2ment.	1n\`e
'interi2	f\`ecu3lent.	ani2ment.	1n\`e
.intero2	trucu3lent.	veni2ment.	4ne.
'intero2	opu3lent.	ri2ment.	4nes.
.inte4r3	corpu3lent.	d\`etri3ment.	2nent.
.interu2	ru3lent.	nutri3ment.	r\`ema3nent.
'interu2	sporu4lent.	inti2ment.	imma3nent.
.inters2	1li	esti2ment.	perma3nent.
'inters2	1l\`i0	l2ment.	.\`emi3nent.
.in1u2	1lo	flam2ment.	pr\`e\`emi3nent.
'in1u2	1l\`o	gram2ment.	pro\`emi3nent.
.in2uit	1is2t	.gem2ment.	sur\`emi3nent.
'in2uit	1lu	om2ment.	immi3nent.
.in2u3l	1l\`u	.com3ment.	conti3nent.
'in2u3l	1ly	\`o2ment.	perti3nent.
.io1a2ct	1ma	slalo2ment.	absti3nent.
.i1oxy	1m\`a	chro2ment.	1ni
.i1s2tat	.ma2c3k	to2ment.	1n\`i0
1j	.macro1s2c	ar2ment.	1no
2jk	.ma2l1a2dres	.sar3ment.	1n\`o
4je.	.ma2l1a2dro	er2ment.	1n\`oe0
4jes.	.ma2l1a1s\`e	antifer3ment.	.no2n1obs
2jent.	.ma2l1ap	.ser3ment.	1nu
1ka	.ma2l1a2v	fir2ment.	1n\`u
1k\`a	.ma2l1en	or2ment.	n3s2at.
1ke	.ma2l1int	as2ment.	n3s2ats.
1k\`e	.ma2l1oc	au2ment.	n1x
1k\`e	.ma2l1o2d	\`ecu2ment.	1ny

Motifs de césure français

.o4	1p2h	1p2t\`er	.r\`etro1a2
'o4	.ph4	1p2t\`er	4re.
'\~o4	4ph.	1pu	4res.
.\~o4	.phalan3s2t	.pud1d2l	2rent.
o2b3long	4phe.	1p\~u	.pa3rent.
loctet	4phes.	1py	appa3rent.
oid2l	2phent.	1q	transpa3rent.
o1\`e2dre	ph2l	4que.	\`e3rent.
olioni	4phle.	4ques.	tor3rent.
ombud2s3	4phles.	2quent.	cur3rent.
omni1s2	2phn	\`e3quent.	1r2h
o1s2tas	photo1s2	\`elo3quent.	4rhe.
o1s2tat	ph2r	grandilo3quent.	4rhes.
o1s2t\`ero	4phre.	1ra	2r3heur
o1s2tim	4phres.	1r\~a	2r3hydr
o1s2tom	2phs	radio1a2	1ri
o1s2trad	2pht	1re	1r\~i0
o1s2tratu	3ph2tal\`e	1r\`e	1ro
o1s2triction	3ph2tis	1r\`e	1r\~o
.oua1ou	1pi	1r\~e	1ru
'oua1ou	1p\~i0	.r\`e1a2	1r\~u
.ovi1s2c	1p2l	.r\`e2a3le	1ry
'ovi1s2c	4ple.	.r\`e2a3lis	1sa
oxy1a2	4ples.	.r\`e2a3lit	1s\~a
1pa	2plent.	.r\`e2aux	.sch4
1p\~a	.pluri1a	.r\`e1\`e2	1s2caph
pal\`eo1\`e2	1p2n\`e	.r\`e1e2	1s2cl\`er
.pa2n1a2f	1p2neu	.r\`e2el	1s2cop
.pa2n1a2m\`e	1po	.r\`e2er	1s2ch
.pa2n1a2ra	1p\~o	.r\`e2\`er	e2s3ch
.pa2n1is	pol1astre	.r\`e1i2	i2s3ch\`e
.pa2n1o2ph	poly1a2	.r\`e2i3fi	i2s3chia
.pa2n1opt	poly1e2	.r\`e1o2	i2s3chio
.pa2r1a2che	poly1\`e2	.re1s2	4sch.
.pa2r1a2ch\`e	poly1\`e2	.re2s3cap	4sche.
.para1s2	poly1i2	.re2s3cisi	4sches.
.pa2r3h\`e	poly1o2	.re2s3ciso	2schs
1pe	poly1s2	.re2s3cou	1se
1p\`e	poly1u2	.re2s3cri	1s\`e
1p\`e	.pon2tet	.re2s3pect	1s\`e
1p\~e	.pos2t3h	.re2s3pir	1s\~e
4pe.	.pos2t1in	.re2s3plend	sesqui1a2
4pes.	.pos2t1o2	.re2s3pons	4se.
2pent.	.pos2t3r	.re2s3quil	4ses.
re3pent.	.post1s2	.re2s3s	2sent.
.ar3pent.	1p2r	.re2s3t	ab3sent.
'ar3pent.	4pre.	.re3s4tab	pr\`e3sent.
ser3pent.	4pres.	.re3s4tag	.res3sent.
.pen2ta	2prent.	.re3s4tand	.seu2le
per3h	.pr\`e1a2	.re3s4tat	.sh4
p\`e2nul	.pr\`e2a3la	.re3s4t\`en	1s2h
.pe4r	.pr\`e2au	.re3s4t\`er	4sh.
.per1a2	.pr\`e1\`e2	.re3s4tim	4she.
.per1e2	.pr\`e1e2	.re3s4tip	4shes.
.per1\`e2	.pr\`e1i2	.re3s4toc	2shent.
.per1i2	.pr\`e1o2	.re3s4top	2shm
.per1o2	.pr\`e1u2	.re3s4tr	2s3hom
.per1u2	.pr\`e1s2	.re4s5trein	2shr
p\`e1r2\`e2q	.pro1\`e2	.re4s5strict	2shs
.p\`eri1os	.pro1s2c\`e	.re4s5trin	1si
.p\`eri1s2	pro2s3tat	.re3s4tu	1s\~i0
.p\`eri2s3s	.prou3d2h	.re3s4ty	1s2lav
.p\`eri2s3ta	1p2syh	.r\`eu2	1s2lov
.p\`eri1u2	.psycho1a2n	.r\`e2uss	1so

1s\`o	.su2r1inf	4thre.	2vent.
1s\oe0	.su2r1int	4thres.	conni3vent.
1s2patia	.su2r1of	2ths	.sou3vent.
1s2perm	.su2r1ox	1ti	1vi
1s2por	1sy	1t\`~\i0	1v\`~\i0
1s2ph\`er	1ta	1to	1vo
1s2ph\`er	1t\`a	1t\`o	1v\`o
1s2piel	1t\`a	1t2r	vol2t1amp
1s2piros	tachy1a2	tran2s1a2	1v2r
1s2tandard	tchin3t2	tran3s2act	4vre.
1s2tein	1te	tran3s2ats	4vres.
st\`er\`eo1s2	1t\`e	tran2s3h	2vrent.
1s2tigm	1t\`e	tran2s1o2	1vu
1s2tock	1t\`e	tran2s3p	1v\`u
1s2tomos	t\`el\`e1e2	tran2s1u2	1vy
1s2troph	t\`el\`e1i2	4tre.	1wa
1s2tructu	t\`el\`e1o2b	4tres.	1we
1s2tyle	t\`el\`e1o2p	2trent.	4we.
1su	t\`el\`e1s2	.tri1a2c	4wes.
1s\`u	4te.	.tri1a2n	2went.
.su2b1a2	4tes.	.tri1a2t	1wi
.su3b2alt	2tent.	.tri1o2n	1wo
.su2b1\`e2	.la3tent.	t1t2l	1wu
.su3b2\`e3r	.pa3tent.	1tu	1w2r
.su2b1in	comp\`e3tent.	1t\`u	2xent.
.su2b3limin	\`eni3tent.	tung2s3	.y4
.su2b3lin	m\`econ3tent.	1ty	'y4
.su2b3lu	omnipo3tent.	.u4	y1asth
sub1s2	ventriipo3tent.	'u4	y1s2tom
.su2b1ur	\`equipo3tent.	.\`u4	y1algi
supero2	impo3tent.	'\`u4	1za
supe4r1	mit3tent.	unilo2v	1ze
supers2	.th4	unila2x	1z\`e
.su2r1a2	1t2h	u2s3tr	1z\`e
su3r2ah	4th.	1va	4ze.
.su3r2a3t	4the.	1v\`a	4zes.
.su2r1e2	4thes.	1ve	2zent.
.su3r2eau	thermo1s2	1v\`e	privatdo3zent.
.su3r2ell	2t3heur	1v\`e	1zi
.su3r2et	2thl	1v\`e	1zo
.su2r1\`e2	2thm	v\`elo1s2ki	1zu
.su2r3h	2thn	4ve.	1zy
.su2r1i2m	th2r	4ves.	}