Cahiers GUTenberg

FRANÇAIS-GUTENBERG : UN NOUVEAU DICTIONNAIRE FRANÇAIS POUR ISPELL. PROBLÈMES RÉSOLUS ET INTÉGRATION DE CONTRIBUTIONS EXTÉRIEURES

Christophe Pythoud

Cahiers GUTenberg, nº 28-29 (1998), p. 252-275.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_1998___28-29_252_0>

Français-GUTenberg : un nouveau dictionnaire français pour ISPELL. Problèmes résolus et intégration de contributions extérieures

Christophe Pythoud

Université de Lausanne — Section de linguistique
Bâtiment des Facultés des Sciences Humaines 2
CH-1015 Lausanne, Suisse
Christophe.Pythoud@ling.unil.ch

Résumé. Cet article présente les choix qui ont été faits dans l'élaboration d'un nouveau dictionnaire français pour le correcteur orthographique ISPELL. Il y est également expliqué comment augmenter le dictionnaire à l'aide des outils *ad hoc* qui l'accompagnent.

Abstract. *This paper presents choices made in elaborating a new French dictionnary for the ISPELL spell checker. How to augment the dictionnary is also explained. The* ad hoc *tools to do this are demonstrated.*

Keywords: ISPELL, français, orthographe

1. Introduction

Cet article présente *Français-GUTenberg*, un dictionnaire français, qui se veut ouvert et extensible, utilisable avec le correcteur orthographique ISPELL.

1.1. Hommage à Martin Boyer

Je tiens à saluer l'immense travail accompli par Martin Boyer sur *Français-IREQ*, sans lequel je ne me serais peut-être jamais intéressé au problème de la correction de l'orthographe française à l'aide de ISPELL, et sans lequel *Français-GUTenberg* n'aurait probablement pas vu le jour.

1.2. Quelle version de ISPELL et comment la configurer ?

La version 3.1.20 de ISPELL, dernière en date, a été utilisée pour le développement de *Français-GUTenberg*. Il est probable que *Français-GUTenberg* puisse fonctionner avec des versions antérieures, mais ce n'est pas garanti.

Dans tous les cas, votre version de ISPELL doit avoir été compilée avec les options suivantes : `NO8BIT` ne doit pas être défini (i.e. les caractères codés sur huit bits doivent pouvoir être utilisés) et il faut fixer `MASKBIT=64` dans `local.h`.

Ce projet est réalisé sur Unix (Linux) et ne fonctionnera pas sans adaptation sur d'autres plates-formes. Cependant, pour les systèmes d'exploitation où existe une version de ISPELL (DOS, OS/2), les modifications *devraient* se limiter au changement de l'encodage des caractères. Si ISPELL n'est pas disponible, il sera toujours possible de réutiliser le dictionnaire de *Français-GUTenberg* avec un autre correcteur (je pense notamment à *Excalibur* sur Macintosh).

1.3. Remerciements

Je tiens à remercier l'association GUTenberg pour le soutien, notamment financier, qu'elle a apporté à ce projet. Je remercie tout particulièrement les membres du Conseil d'Administration, pour m'avoir accordé leur confiance et pour m'avoir, par la suite, accueilli parmi eux.

2. ISPELL et *Français-IREQ* : problèmes observés et solutions proposées

Cette section présente les problèmes rencontrés lors de l'utilisation du couple ISPELL/*Français-IREQ*, ainsi que la manière dont *Français-GUTenberg* tente de les résoudre.

2.1. Fonctionnement de ISPELL

Cette section a pour objet de présenter les différents aspects du fonctionnement de ISPELL.

2.1.1. Qu'est-ce qu'un correcteur orthographique ?

On appelle généralement « correcteur orthographique » tout programme qui compare les mots d'un texte à ceux présents dans une liste de mots, appelée

« dictionnaire » (procédure de vérification, cf. 2.1.4). Lorsqu'un mot du texte en vérification ne figure pas dans le dictionnaire, un algorithme calcule à partir du mot (potentiellement) mal orthographié un certain nombre (qui peut être nul) de corrections possibles qui sont des mots graphiquement proches figurant dans le dictionnaire (procédure de suggestion, cf. 2.1.5).

Certains logiciels plus ambitieux tentent de corriger, en plus de l'orthographe d'usage, l'orthographe d'accord ou « grammaticale ». ISPELL n'entre pas dans cette seconde catégorie et fonctionne selon le premier modèle évoqué : vérification puis, éventuellement, suggestion de graphies alternatives.

2.1.2. Le format du dictionnaire

Un dictionnaire ISPELL prêt à l'emploi se présente sous la forme de deux fichiers : `langue.aff` et `langue.hash`. (Ils s'appellent en principe `francais.aff` et `francais.hash` pour le français.) Le premier contient la liste des caractères pouvant licitement apparaître dans un mot, ainsi qu'un ensemble de règles d'affixation dont il va être question un peu plus loin. Le second est une table de hachage obtenue à partir du premier et d'un troisième fichier, que l'on peut appeler `langue.dico` (pour le français : `francais.dico`). Il n'existe pas d'extension standard pour ce troisième fichier qui est d'ailleurs généralement le produit de l'agrégation de plusieurs listes de mots.

Un fichier `langue.dico` présente deux types d'entrées :

(1) `mot1`

(2) `mot2/drapeau(x)`

Dans le premier cas[1], on a simplement affaire à un mot faisant partie du dictionnaire. Dans le second, le mot est accompagné d'un ou plusieurs drapeaux renvoyant aux règles d'affixation définies dans `langue.aff` : à la ligne (2) vont correspondre plusieurs mots du dictionnaire. Par exemple, dans le fichier `francais.dico` de *Français-GUTenberg*, on peut trouver :

(3) `coordinatrice/G`

(3) indique que le mot *coordinatrice* est présent dans le dictionnaire, de même que les mots *coordinatrices*, *coordinateur* et *coordinateurs* engendrés à partir de certaines règles d'affixation appartenant au drapeau `G` :

(4) `[^SXZ]  >  S`

1. Les numéros entre parenthèses servent à identifier les différents exemples donnés et ne font pas partie des entrées du dictionnaire.

(5) `T R I C E > -RICE,EUR`

(6) `T R I C E > -RICE,EURS`

La règle (4) engendre un pluriel en « s » à partir de n'importe quel mot qui ne s'achève ni par « s », ni par « x » et ni par « z ». La règle (5) transforme tout mot s'achevant en *-trice* en mot s'achevant en *-eur* en soustrayant les lettres « r », « i », « c » et « e » (dans cet ordre) à la fin du mot de départ et en lui ajoutant les lettres « e », « u » et « r » (dans cet ordre). La règle (6) fait de même, mais ajoute encore la lettre « s » pour créer le pluriel du masculin.

2.1.3. Le choix des lemmes

En lexicographie, on appelle « lemme »[2] un mot choisi pour représenter toute une famille de formes apparentées. Par exemple, si je trouve la forme « soliloquions » dans un texte et que je désire en connaître le sens précis à l'aide d'un dictionnaire courant, je vais chercher dans celui-ci la forme « soliloquer » et non « soliloquions ». Nous sommes en effet habitués à chercher les verbes sous leur forme infinitive dans les dictionnaires et les ouvrages de grammaire. L'infinitif constitue donc le lemme ou « mot-vedette », pour reprendre la terminologie de Nina Catach [2], pour représenter les verbes. Parallèlement, pour un nom ou un adjectif, je me référerais à la forme du masculin singulier.

Par analogie, on peut dire que ISPELL fonctionne de cette manière : au travers des règles d'affixation dont il dispose, il va chercher à trouver le « mot-vedette », le lemme, correspondant à la graphie qu'il est en train d'analyser (cf. 2.1.4). Ici une question se pose : quels lemmes allons-nous utiliser dans la composition de notre dictionnaire ? l'infinitif pour les verbes et le masculin singulier pour les noms et les adjectifs ? ou autre chose ?

Même pour un dictionnaire sur papier, l'emploi de l'infinitif et du masculin singulier n'est qu'une convention et rien ne nous empêcherait, sinon la tradition, de choisir comme lemme, pour un verbe, la première personne du présent de l'indicatif. C'est d'ailleurs cette forme qui est utilisée dans la plupart des dictionnaires grecs et latins. Est conventionnel aussi le fait de n'avoir qu'un lemme pour une famille de mots, convention d'ailleurs très relative : lorsqu'un nom ou un adjectif présente un féminin ou un pluriel irrégulier, cela est généralement indiqué.

Or, le choix des lemmes n'est justement pas un problème trivial au niveau de la composition d'un dictionnaire français pour ISPELL. *Français-IREQ* adopte

2. Je préfère recourir à la terminologie des lexicographes plutôt que d'utiliser le terme de « racine », traduction de l'anglais « `root` », présent dans la documentation de ISPELL, et qui est contraire à la tradition francophone.

la solution traditionnelle : infinitif et masculin singulier. On peut faire mieux en étant plus audacieux.

Par exemple, comment traiter la féminisation[3] et la mise au pluriel des mots : « pleureur », « puériculteur », « chasseur » et « docteur » ? Si on choisit de partir du masculin singulier, il faudra quatre règles différentes pour obtenir les féminins (et il vaudrait mieux ne pas se tromper en attribuant les drapeaux) : « pleureuse », « puéricultrice », « chasseresse » et « doctoresse ». Cela donnerait quelque chose comme ça :

```
flag *A:  # pleureur, pleureuse
        E U R                   >       -R,SE
        E U R                   >       -R,SES

flag *B:  # puériculteur, puéricultrice
        E U R                   >       -EUR,RICE
        E U R                   >       -EUR,RICES

flag *C   # chasseur, chasseresse
        E U R                   >       -UR,RESSE
        E U R                   >       -UR,RESSES

flag *D   # docteur, doctoresse
        E U R                   >       -EUR,ORESSE
        E U R                   >       -EUR,ORESSES
```

avec `pleureur/AS`[4], `puériculteur/BS`, `chasseur/CS` et `docteur/DS` dans le dictionnaire.

Maintenant, si on observe les terminaisons des mots féminins, on s'aperçoit qu'elles sont toutes différentes et qu'il est donc possible, en utilisant le féminin singulier comme lemme, de n'utiliser qu'un seul drapeau. C'est ce qui est fait dans *Français-GUTenberg* et le drapeau `G` est consacré aux mots en *-eur* :

```
flag *G:                        # mots en -eur
        E                       >       S

        # prometteuse
        E U S E                 >       -SE,R
```

3. Les remarques qui suivent ne concernent bien entendu que les mots possédant une forme masculine *et* féminine. Le traitement des termes épicènes est beaucoup plus simple.

4. En partant du principe qu'il existe un drapeau `S` qui renvoie à un traitement général des pluriels.

```
E U S E                 >       -SE,RS

# opératrice
T R I C E               >       -RICE,EUR
T R I C E               >       -RICE,EURS

# chasseresse
E R E S S E             >       -RESSE,UR
E R E S S E             >       -RESSE,URS

# doctoresse
O R E S S E             >       -ORESSE,EUR
O R E S S E             >       -ORESSE,EURS
```

avec `pleureuse/G`, `puéricultrice/G`, `chasseresse/G` et `doctoresse/G` dans le dictionnaire.

Le drapeau `G` est réservé aux mots en *-eur*. Les autres termes non épicènes sont tous traités par le drapeau `F`, dont voici un extrait :

```
# Consonne(s) (t)t + e
T T E                   >       -TTE,T  # coquette, sotte
T T E                   >       -TE,S
[^èT] T E               >       -TE,T   # idiote
[^èT] T E               >       -E,S
è T E                   >       -èTE,ET # discrète
è T E                   >       -èTE,ETS
```

qui présente les mêmes caractéristiques d'économie et dont je laisse l'analyse comme exercice au lecteur.

Quand on s'attaque aux verbes, un problème du même ordre se présente : *Français-IREQ* utilise comme lemme l'infinitif. Or, une des difficultés de la conjugaison en français est qu'un verbe peut avoir plusieurs radicaux : *manger* en a deux, *mang-* et *mange-* ; envoyer en a trois, *envoi-*, *envoy-* et *enverr-* ; tenir en a cinq, *tien-*, *tienn- ten-*, *tin-* et *tiend-*. Ceci a pour conséquences, si on veut traiter exhaustivement du verbe français dans le cadre de *Français-IREQ*, les problèmes suivants :

- une multiplication des drapeaux ;
- des règles d'affixation extrêmement complexes ;
- l'*impossibilité* de traiter de tous les verbes du français, d'où l'absence de nombreuses formes du troisième groupe dans *Français-IREQ*.

La solution consiste ici à augmenter le nombre de lemmes, ce qui amène une diminution, apparemment paradoxale, des drapeaux, réduit la complexité des règles et permet un traitement exhaustif des verbes des trois groupes. Les cinq lemmes suivants sont utilisés par *Français-GUTenberg* : les premières personnes du singulier du présent, de l'imparfait, du passé simple et du futur de l'indicatif, ainsi que l'infinitif. Cette stratégie permet également de résoudre l'élision du pronom personnel *je* (cf. 3.2.3).

La démonstration est un peu courte et incomplète sur ce dernier point : le lecteur intéressé pourra se reporter à [5].

2.1.4. Vérification

ISPELL considère qu'un mot est correctement orthographié s'il s'agit d'un lemme du dictionnaire ou s'il peut être engendré par une règle d'affixation (cf. 2.1.2 : la présence de `coordinatrice/G` dans le dictionnaire implique que *coordinatrice*, *coordinatrices*, *coordinateur* et *coordinateurs* sont corrects).

Maintenant, si le fichier `langue.aff`[5] contient la mention `allaffixes on` sur une ligne, ISPELL essayera alors *tous* les drapeaux sur le mot pour tenter d'aboutir à un lemme. Par exemple, supposons que vous tapiez *l'éléphant* et que le dictionnaire contienne `éléphant/S`, mais pas `éléphant/LS` (avec `S` gérant le pluriel et `L` l'élision du déterminant *le*) : ISPELL va s'apercevoir que *l'éléphant* serait correct si `éléphant/S` était accompagné du drapeau `L` ; il va donc le signaler à l'utilisateur et lui demander de valider (ou non) l'affixation `l'` + `éléphant`.

A priori, ce comportement peut sembler adéquat ; il ne l'est pas, car il ne sert qu'à récupérer une erreur commise dans la conception du dictionnaire : l'absence d'un drapeau. ISPELL produit donc une fausse alerte qui gène l'utilisateur : si de nombreux drapeaux manquent, comme c'est malheureusement le cas dans *Français-IREQ*, les fausses alertes de ce type seront nombreuses et le confort d'utilisation s'en trouvera très diminué (cf. 2.2.1).

Le problème est résolu dans *Français-GUTenberg* en s'assurant que tous les mots du dictionnaire sont accompagnés des drapeaux adéquats et en précisant `allaffixes off` dans `francais.aff`.

2.1.5. Suggestion

Lorsque ISPELL considère un mot comme erroné, il va suggérer quelques graphies proches qui figurent dans son dictionnaire (comme lemmes ou engendrées

5. Dans notre cas `francais.aff`.

par les règles d'affixation). Une graphie sera considérée comme proche si elle remplit une (et une seule) des conditions suivantes :

- elle peut être obtenue en permutant deux lettres voisines (**cmaembert* → *camembert*) ;
- elle peut être obtenue en changeant une lettre (**cjaleur* → *chaleur*) ;
- elle peut être obtenue en supprimant une lettre (**beaucoups* → *beaucoup*) ;
- elle peut être obtenue en ajoutant une lettre (**corp* → *corps*) ;
- ou si l'insertion d'un espace dans la chaîne donne deux mots reconnus par le dictionnaire (**nonlinéaire* → *non linéaire*).

2.2. Inconvénients contournables dans le cadre de ISPELL

Dans les points qui suivent sont présentés les inconvénients liés à l'utilisation de ISPELL et *Français-IREQ*, ainsi que les mesures qui ont été prises dans *Français-GUTenberg* afin de les éviter ou de les atténuer. Ces éléments, déjà largement présentés dans les sections précédentes, constituent la principale motivation de *Français-GUTenberg*.

2.2.1. Les fausses alertes et le problème de l'élision

Le problème majeur de l'utilisation du couple ISPELL/*Français-IREQ* est le nombre de fausses alertes engendrées. En particulier lorsque le programme demande à l'utilisateur de valider, par exemple, `l'atmosphère` comme une combinaison licite de « `l'` » + « `atmosphère` » (cf. 2.1.4).

Comme déjà indiqué, cette fâcheuse situation provient du fait que de nombreux lemmes dans le dictionnaire de *Français-IREQ* ne sont pas accompagnés des drapeaux adéquats.

L'élision des pronoms personnels et des articles pose des problèmes complexes : il en sera continuellement question au fil de la Section 3. Cela peut même engendrer des problèmes au niveau de la procédure de suggestion : reprenons notre exemple de *l'éléphant*. Supposons que le dictionnaire contienne l'entrée `éléphant/S`, qui engendre *éléphant* et *éléphants*, mais pas `éléphant/LS` qui engendrerait, en plus de ce qui précède, *l'éléphant*. Supposons encore que, victime d'une fatigue soudaine, je tape `éléphent` : ISPELL n'aura aucune peine à me suggérer de remplacer cette graphie fautive par *éléphant*. Mais, si j'ai tapé `l'éléphent`, ISPELL ne pourra que me signaler que cette graphie est probablement mauvaise, mais en aucun cas il ne parviendra à suggérer *l'éléphant* comme remplacement. Pourquoi ? Parce que le drapeau `L` manque et que par conséquent, la forme *l'éléphant* est absente du dictionnaire. Voilà une

raison supplémentaire de veiller à ce que chaque entrée du dictionnaire soit accompagnée des drapeaux adéquats.

2.2.2. La gestion des verbes

Français-IREQ contient de nombreux verbes, Martin Boyer ayant basé son travail sur l'excellent, quoique parfois incomplet et contenant quelques coquilles, répertoire Bescherelle [1]. Cependant, quelques verbes du troisième groupe sont absents. D'autre part, le système de lemmatisation choisi par Martin Boyer (l'infinitif comme lemme unique) n'est pas très productif et relativement complexe, comme déjà indiqué à la fin de la Section 2.1.3 (page 257). Les formes qui ne peuvent être traitées par le système d'affixation sont entrées directement dans le dictionnaire, ce qui n'en facilite par la relecture et la mise à jour. Il est, de plus, difficile de ne pas commettre d'erreurs en entrant « manuellement » de nombreuses formes verbales.

Ces problèmes se trouvent résolus par l'introduction du système à cinq lemmes de *Français-GUTenberg*.

2.2.3. Écrire les nombres en toutes lettres

Les règles de l'orthographe française exigent que, dans certains contextes, on écrive les nombres en toutes lettres. Or, beaucoup de nombres manquent dans *Français-IREQ*. Vous les trouverez, sous leurs formes cardinale et ordinale, dans *Français-GUTenberg*.

2.2.4. Le vocabulaire spécialisé

Un reproche souvent fait à *Français-IREQ* est de ne pas connaître le vocabulaire de telle ou telle discipline, en particulier les mathématiques. Idéalement, lorsqu'on corrige un texte relevant d'une discipline particulière, on devrait pouvoir invoquer, en plus du dictionnaire principal, un dictionnaire spécialisé contenant les termes propres à la discipline considérée. ISPELL ne dispose pas de ce mode de fonctionnement (cf. 2.3.2). La seule solution est donc d'inclure le vocabulaire propre à la discipline qui nous intéresse directement dans le dictionnaire. Mais si le dictionnaire contient des termes mathématiques, les physiciens vont alors tout à fait légitimement demander qu'y figure également le vocabulaire de la physique, les juristes celui du droit, les linguistes celui de la linguistique, etc. Inclure le jargon de tous ces différents domaines dans le dictionnaire n'est pas envisageable, cela pour deux raisons :

– la taille excessive du dictionnaire finirait par affecter les performances du correcteur ;

– certaines fautes risqueraient de passer inaperçues à cause de termes spécialisés qui les masqueraient[6].

Au vu de ce qui précède, la solution envisagée est la suivante : l'utilisateur a la possibilité de choisir, lors de la compilation du dictionnaire, quelles listes de mots il compte y faire figurer. La procédure de compilation, ainsi que la création de dictionnaires spécialisés, sont décrites à la Section 3.

2.2.5. Le caractère « œ »

ISPELL utilise l'encodage *Isolatin-1* qui ne comprend pas le caractère « œ ». La solution la plus simple consiste à entrer ce caractère sous la forme « `oe` » dans le dictionnaire. C'est ce qui est adopté par *Français-IREQ* et également, par défaut, par *Français-GUTenberg*. Toutefois, pour ceux qui n'utiliseraient que (LA)TEX, il sera possible d'affecter un autre caractère du jeu *Isolatin-1* au « œ ». Cette astuce sera expliquée dans la documentation qui accompagnera *Français-GUTenberg*.

2.3. Problèmes insolubles dans le cadre de ISPELL

Cette section présente brièvement ce que je considère comme les deux inconvénients majeurs d'ISPELL et qui en eux-mêmes font qu'il est légitime d'envisager la création d'un nouveau correcteur qui ne souffre pas de ces problèmes (cf. Section 4).

2.3.1. Limites intrinsèques de l'algorithme de suggestion

Le but de *Français-GUTenberg* est de rendre l'usage de ISPELL pour corriger des textes en français efficace et agréable, notamment en supprimant le plus possible de fausses alertes. *Français-GUTenberg* ne saurait cependant pallier aux insuffisances de la procédure de suggestion de ISPELL. En se reportant à la description de celle-ci (cf. 2.1.5), on peut immédiatement constater qu'aucune suggestion valide ne pourra être émise pour un mot comportant plusieurs fautes ou pour une faute portant sur plusieurs lettres.

6. Par exemple, lorsqu'on s'intéresse à la logique de Montague [4], on est souvent amené à parler d'*intension* (par opposition à *extension*). On peut imaginer que ce terme soit contenu dans le fichier `logique.dico` inutilement inclus dans le dictionnaire principal. Si un utilisateur, qui n'a jamais entendu parler des théories de Montague, écrit **intension* à la place de *intention*, la faute ne sera pas détectée.

Prenons deux exemples :

- deux fautes dans le même mot : ISPELL est incapable de suggérer *notamment* à partir de **nottament* ;
- une faute portant sur deux lettres : à partir de **précisemment*, ISPELL n'arrive pas à trouver *précisément*.

Que faire lorsque ISPELL ne suggère rien ? Le mieux est probablement d'essayer différentes graphies, jusqu'à ce que l'une d'entre elles soit acceptée ou que ISPELL produise une suggestion intéressante.

Et lorsque cet algorithme simpliste se trouve combiné avec un dictionnaire dans lequel des drapeaux manquent (cf. 2.1.4), on aboutit également à une absence de suggestions.

2.3.2. Unicité du dictionnaire

Le fait de ne pouvoir invoquer qu'un seul dictionnaire est un inconvénient majeur à mon sens. Cela représente une grande perte de flexibilité, comme déjà évoqué plus haut (cf. 2.2.4).

3. Comment augmenter le dictionnaire

Cette section expose brièvement les étapes à suivre pour créer un dictionnaire spécialisé contenant des noms, des adjectifs et des verbes. Celui-ci peut bien entendu aussi contenir des adverbes et d'autres mots invariables. Pour ces derniers, aucune précaution particulière n'est requise, il suffit de les faire figurer dans le dictionnaire tels quels.

Les informations qui suivent constituent davantage un survol qu'une description exhaustive de la procédure. *Français-GUTenberg*, dans sa version finale, sera accompagné d'une documentation plus complète.

3.1. Ajouter des noms et des adjectifs

3.1.1. Le cas trivial

Le cas le plus simple est celui d'un nom ou d'un adjectif épicène ne commençant pas par une voyelle ou un « h » muet. Il suffit de l'entrer dans le dictionnaire et de lui adjoindre le drapeau `S`[7] s'il prend un simple « s » au pluriel, ou le

7. La casse du drapeau est importante et donc `S` n'est pas équivalent à `s` (cf. 1.2).

drapeau `X` dans tous les autres cas. Si le pluriel est identique au singulier, le drapeau n'est pas nécessaire. Par exemple :

(7) `célébration/S`

(8) `bateau/X`

(7) permettra de vérifier *célébration* et *célébrations*. (8) sera la caution de *bateau* et *bateaux*.

3.1.2. Féminin et masculin

Certains noms et adjectifs possèdent des formes féminines et masculines. Dans le cas d'un mot qui ne commence pas par une voyelle ou un « h » muet, il suffit d'entrer la forme au féminin singulier, accompagnée du drapeau `F`, sauf si on a affaire à un mot qui finit en *-eur* au masculin (cf. 2.1.3), auquel cas il faut utiliser le drapeau `G`. Quelques exemples :

(9) `nominative/F`

(10) `malheureuse/F`

(11) `boudeuse/G`

(9) représente *nominative*, *nominatives*, *nominatif* et *nominatifs*. (10) représente *malheureuse*, *malheureuses* et *malheureux*. (11) représente *boudeuse*, *boudeuses*, *boudeur* et *boudeurs*.

Les drapeaux `F` et `G` couvrent la quasi-totalité des correspondances masculin/féminin du français : vous pouvez en principe vous y fiez et, dans le doute, vérifiez ce qui est produit par le drapeau (cf. 3.2.6 pour ce faire). Seuls les cas les plus irréguliers ne sont pas traités : *andalou/andalouse*, *beau/belle*, *favori/favorite*, *fou/folle*, *tiers/tierce*. Il y a peu de chances qu'un dictionnaire spécialisé doive contenir des mots présentant de telles caractéristiques.

3.1.3. Le problème de l'apostrophe

Comment gérer l'élision dans le cas d'un mot commençant par une voyelle ou un « h » muet ? Il faut distinguer deux cas :

- le mot est strictement masculin ou féminin, ou possède une forme identique au masculin et au féminin ;
- le mot possède des formes différentes au masculin et au féminin.

Dans le premier cas, on entre le mot au singulier dans le dictionnaire, accompagné des drapeaux `LMS` ou `LMX` (selon le pluriel du mot, cf. 3.1.1). Par exemple :

(12) `éclaircissement/LMS`

qui représentera *éclaircissement*, *éclaircissements*, *l'éclaircissement*, *d'éclaircissement* et *d'éclaircissements*.

Dans le second cas, l'opération est un peu plus délicate : il faut entrer la forme féminine singulier du mot, accompagnée des drapeaux `LMF` ou `LMG` (cf. 3.1.2), *et* le masculin singulier précédé de « `l'` ». Par exemple :

(13) `ancienne/LMF`

(14) `l'ancien`

où (13) engendre *ancienne*, *anciennes*, *ancien*, *anciens*, *l'ancienne*, *d'ancienne*, *d'anciennes*, *d'ancien* et *d'anciens*, mais pas *l'ancien* qui doit être spécifié séparément. Pourquoi cela ? Regardons le « code » des drapeaux `L` et `M` :

```
flag  L:
        [aàâeèéêiîoôuh]          >       L'

flag *M:
        [aàâeèéêiîoôuh]          >       D'
```

Ces deux règles indiquent que tout mot commençant par une voyelle (accentuée ou non) ou par un « h » peut être précédé de *l'* ou de *d'*. La différence entre elles se situe au niveau de l'astérisque (`*`) qui précède `M` mais pas `L`. Cet astérisque signifie que le préfixe est libre de se combiner avec n'importe quel suffixe également doté de l'astérisque (et réciproquement). Ceci a pour conséquence que (15) est autorisé mais pas (16) :

(15) `d'` + `ancienne` + `s`

(16) `l'` + `ancienne` + `s`

Cette restriction interdit au système d'accepter ou de produire les formes erronées **l'anciens* et **l'anciennes*, mais nécessite alors que la forme correcte *l'ancien* soit spécifiée séparément. Le problème ne se pose pas pour le drapeau `D`.

Vous avez sans doute remarqué que je n'opère pas de distinction entre noms et adjectifs pour l'attribution des drapeaux. On pourrait cependant se demander

si les drapeaux `L` et `M`, représentant respectivement l'élision des déterminants *le de*, doivent vraiment être attribués aux adjectifs. Il ne faut pas oublier qu'en français la plupart des adjectifs peuvent se trouver antéposés au nom qu'ils qualifient, comme dans (17).

(17) L'ignoble individu.

Certains adjectifs, appartenant à des catégories bien précises, ne sont toutefois jamais antéposés. Pour ceux-ci, et seulement pour eux, on peut se passer des drapeaux `L` et `M`. Les catégories d'adjectifs concernées sont les suivantes :

- les adjectifs de couleur ;
- les adjectifs d'appartenance (nationalité, origine géographique, religion, etc.) ;
- les adjectifs de forme (*oval* par exemple) ;
- les adjectifs verbaux formés sur le participe passé (cf. en fin de Section 3.2.3, page 271).

3.1.4. Utilisation des autres préfixes

Le fichier `francais.aff` de *Français-GUTenberg* permet de gérer les préfixes latins suivants : *a(d)-*, *con-*, *dé-*, *dis-*, *ex-*, *in-*, *pré-*, *re-* et *sub-*. Leur utilisation peut faciliter et accélérer la composition d'un dictionnaire.

3.2. Ajouter des verbes

Dans les pages qui suivent, l'adjonction de verbes est examinée. Comme *Français-GUTenberg* contient en principe tous les verbes du troisième groupe, je pars du principe que les verbes qui seront ajoutés appartiennent soit au premier soit au deuxième groupe. (Si vous deviez néanmoins faire figurer un verbe du troisième groupe dans votre dictionnaire, référez-vous à la documentation qui accompagnera *Français-GUTenberg* ou contactez-moi.)

3.2.1. Le cas le plus simple

Dans le meilleur cas de figure, nous aurons affaire à un verbe qui ne commence ni par une voyelle, ni par un « h » muet, pour éviter le problème de l'élision (cf. 3.2.3). Il faudra entrer les premières personnes des temps de l'indicatif suivants : présent, imparfait, passé simple et futur, accompagnés respectivement des drapeaux `u`, `v`, `w` et `x`. Prenons l'exemple du verbe *charmer* :

(18) `charme/u`

(19) `charmais/v`

(20) `charmai/w`

(21) `charmerai/x`

Ces quatre lemmes permettent d'engendrer toutes les formes du verbe *charmer* à l'exception de l'infinitif et du participe passé. Voici les règles, extraites de `francais.aff`, qui leur sont appliquées (ne figurent que les lignes pertinentes pour le verbe *charmer*) :

```
flag *u:
        E                       >       S
        E                       >       NT

flag *v:
        A I S                   >       -S,T
        [^çE] A I S             >       -AIS,IONS
        [^çE] A I S             >       -AIS,IEZ
        A I S                   >       -S,ENT
        A I S                   >       -AIS,ONS
        [^çE] A I S             >       -AIS,EZ
        A I S                   >       -IS,NT
flag *w:
        A I                     >       -I,S
        A I                     >       -AI,A   # = -I
        A I                     >       -AI,âMES
        A I                     >       -AI,âTES
        [^çE] A I               >       -AI,èRENT
        A I                     >       -I,SSE
        A I                     >       -I,SSES
        A I                     >       -AI,âT
        A I                     >       -I,SSIONS
        A I                     >       -I,SSIEZ
        A I                     >       -I,SSENT

flag *x:
        A I                     >       -I,S
        A I                     >       -AI,A   # = -I
        A I                     >       -AI,ONS
        A I                     >       -AI,EZ
```

```
        A I                     >       -AI,ONT
        A I                     >       S
        A I                     >       T
        A I                     >       -AI,IONS
        A I                     >       -AI,IEZ
        A I                     >       ENT
```

Leur interprétation est triviale et je la laisse comme exercice au lecteur.

3.2.2. Infinitif, participe passé et adjectif verbal formé sur le participe présent

Concernant le participe passé, il faut distinguer deux cas :

- il ne s'accorde jamais (ce qui concerne principalement les verbes intransitifs se conjuguant avec l'auxiliaire *avoir*), par exemple *bifurquer* ou *frissonner* ;
- il est susceptible de s'accorder en genre et nombre (ce qui est le cas de la majorité des verbes français).

Dans le premier cas, on entre simplement l'infinitif et le participe passé dans le dictionnaire. Par exemple :

(22) `bifurquer`

(23) `bifurqué`

Dans le second cas, on entre l'infinitif suivi du drapeau `y`. Par exemple :

(24) `charmer/y`

Le drapeau `y` se compose, entre autres, des règles suivantes :

```
flag *y:
        E R                     >       -ER,é
        E R                     >       -ER,éS
        E R                     >       -ER,éE
        E R                     >       -ER,éES
```

Une question du même type se pose pour le participe présent qui est engendré par le drapeau `v` (*charmant* est créé à partir de (19) dans notre exemple). En

effet, bien souvent, au participe présent correspond un adjectif verbal (et parfois un substantif) susceptible de s'accorder en genre et en nombre (*charmants*, *charmante*, *charmantes*). Quand c'est le cas, on peut ajouter le drapeau `c` au lemme à l'imparfait. Dans notre exemple, (19) deviendrait (25).

(25) `charmais/vc`

Les règles du drapeau `c` sont :

```
flag *c:                                    # adjectif verbal accordé
        A I S                       >       -IS,NTS
        A I S                       >       -IS,NTE
        A I S                       >       -IS,NTES
```

Ce dernier mécanisme est très pratique, mais il faut savoir que l'orthographe française recèle quelques pièges dans ce domaine : il faut dans certains cas s'abstenir d'utiliser le drapeau *c*. En effet, les graphies du participe présent et de l'adjectif verbal peuvent parfois être différentes. Quelques exemples :

(26) a. Pierre, *négligeant* toute prudence, appuya sur le champignon.
b. Pierre se montre toujours très *négligent* au volant.

(27) a. Paul asséna son dernier argument à Jean, le *convainquant* ainsi du bien-fondé de son point de vue.
b. Paul a été très *convaincant* face à Jean.

(28) a. Le journaliste continua de poser ses questions, *fatiguant* les coureurs.
b. Les questions du journaliste étaient *fatigantes*.

Au vu de ce qui précède, il faut tenir compte des points suivants :

- à certains participes présents en *-quant* correspondent des adjectifs en *-cant* ;
- à certains participes présents en *-guant* correspondent des adjectifs en *-gant* ;
- à certains autres participes présents (en *-ant*) correspondent des adjectifs verbaux en *-ent*.

Grévisse [3] donne au §1887 une liste d'adjectifs « verbaux » qui ne se forment pas sur le participe présent.

3.2.3. Le problème de l'élision

Lorsqu'on est confronté à un verbe commençant par une voyelle ou un « h » muet, la situation est un peu plus compliquée et il faut être extrêmement prudent par rapport à deux choses :

- les types d'élision possibles devant le verbe considéré (« `j'` », « `n'` », « `m'` », « `t'` », « `s'` » et « `l'` ») ;
- le traitement d'un éventuel adjectif verbal et des participes passés (leur ajouter des drapeaux sans réfléchir entraînerait des effets de bord catastrophiques pour la validité du dictionnaire).

Les pronoms susceptibles de s'élider relèvent de différentes catégories :

- le pronom *je* est toujours susceptible de s'élider en « `j'` » ;
- il en va de même pour la négation : *ne* est toujours susceptible de s'élider en « `n'` » ;
- certains pronoms qui font fonction d'objet direct peuvent s'élider (*me*, *te*, *le*) ;
- certains pronoms qui font fonction d'objet indirect également (*me*, *te*) ;
- de même que certains pronoms réfléchis (*me*, *te*, *se*).

Table 1 – *Les pronoms du français et l'élision*

Sujet	*Obj. dir.*	*Obj. indir.*	*Réflex.*
je → j'	me → m'	me → m'	me → m'
tu → ∅	te → t'	te → t'	te → t'
il/elle/on → ∅	le → l'	lui → ∅	se → s'
nous → ∅	nous → ∅	nous → ∅	nous → ∅
vous → ∅	vous → ∅	vous → ∅	vous → ∅
ils/elles → ∅	les → ∅	leur → ∅	se → s'
Négation : ne → n'			

La Table 1 synthétise les points ci-dessus. On observe un grand syncrétisme entre les différentes formes, ce qui n'est pas de nature à nous faciliter la tâche. Aux différentes formes d'élisions correspondent les drapeaux `j`, `n`, `m`, `t`, `l` :

```
flag  j:        # pour accompagner le système à 5 lemmes
        [aàâeèéêiîoôuh]         >       J'


flag *n:                        # élision de la négation
```

```
        [aàâeèéêiîoôuh]         >       N'

flag *m:                        # élision du pronom 'me'
        [aàâeèéêiîoôuh]         >       M'      # tu m'aides

flag *t:                        # élision du pronom 'te'
        [aàâeèéêiîoôuh]         >       T'      # je t'énerve

flag *l:
        [aàâeèéêiîoôuh]         >       L'      # je l'attends
```

Il ne faut *jamais* combiner ces drapeaux avec `c` et `y` (voir plus bas pour les détails). L'élision de *se* n'est pas prise en charge par le système : voir les Sections 3.2.4 et 3.2.5 pour ce qui concerne les emplois réflexifs et les verbes essentiellement pronominaux. Nous ne nous en occuperons plus dans cette section.

Quels drapeaux employer et dans quels cas ? Le drapeau `j` doit, dans tous les cas, précéder les quatre premiers lemmes du verbe, mais pas l'infinitif. Tout comme le drapeau `L` (cf. 3.1.3), il n'est pas précédé de l'astérisque (`*`), ce qui l'empêche de se combiner avec un suffixe pour donner des résultats aberrants.

Le drapeau `n` concerne tous les lemmes du verbe.

Si le verbe accepte un objet direct, chaque lemme devrait être en principe accompagné des drapeaux `m`, `t` et `l`. En fait, il faut bien réfléchir au sens du verbe : si l'objet direct ne peut pas être une personne les drapeaux `m` et `t` ne devraient en principe pas être spécifiés. On peut prendre comme exemples les verbes *accélérer* ou *ébruiter*.

Si le verbe n'accepte pas d'objet direct, mais seulement un objet indirect, seuls les drapeaux `m` et `t` entrent en ligne de compte (cf. Table 1).

Le critère sémantique invoqué pour les compléments d'objet direct s'applique également aux compléments indirects. Mais il ne faut cependant pas surestimer son importance : le français dispose d'une construction appelée datif éthique, dont les grammaires parlent peu mais qui est très usitée. Rentrent dans cette catégorie toutes les constructions du type :

(29) Je peux *t'*aménager de la place dans mon laboratoire.

(30) Il faudra *m'*améliorer la conclusion de cet article.

Elles indiquent, la plupart du temps, le bénéficiaire de l'action accomplie. Je ne crois pas qu'on puisse sérieusement considérer toutes ces formes comme incorrectes et il faut donc en tenir compte dans l'attribution des drapeaux.

Si le verbe est intransitif, les choses deviennent fort simples : on ignore les trois drapeaux `m`, `t` et `l`.

Avant d'aller plus loin, une remarque complémentaire sur les drapeaux `m` et `t` s'impose : ils autorisent des formes erronées comme **m'absolvons* et **t'énerviez*. Malheureusement, la seule solution pour éviter cela consisterait à supprimer ces drapeaux et à entrer « à la main » toutes les élisions des pronoms *me* et *te* dans le dictionnaire : vu leur nombre, ce n'est pas envisageable.

Concernant le cinquième lemme, l'infinitif, on le fera figurer accompagné des drapeaux `nM`[8], dans tous les cas, et des drapeaux `m`, `t` et `l`, quand il y a lieu (cf. ci-dessus).

Que faire si le participe passé peut s'accorder ? On fait figurer la forme féminine du participe passé accompagnée du drapeau `F` (cf. 3.1.2). Si ce participe passé est à l'origine d'un substantif, on utilise la forme féminine accompagnée des drapeaux `LMF`, suivie de la forme masculine précédée de « `l'` » (cf. 3.1.3). Mais on ne fera pas de même pour un adjectif : contrairement aux adjectifs formés sur le participe présent, les adjectifs formés sur le participe passé ne peuvent jamais être antéposés en français, *i.e.*, utilisés comme épithètes[9]. Ils ne seront donc jamais précédés d'un déterminant. On peut s'en rendre compte empiriquement, en comparant les quatres exemples ci-dessous :

(31) L'homme assailli.

(32) *L'assailli homme.

(33) La question assaillante.

(34) L'assaillante question.

Que faire si le participe présent peut être mis en rapport avec un adjectif verbal ? Au vu de ce qui précède, il convient d'utiliser la forme féminine accompagnée des drapeaux `LMF`, suivie de la forme masculine précédée de « `l'` ».

La Table 2 présente quelques exemples pour clarifier les idées.

8. Je vous rappelle que le drapeau `M` gère le préfixe *d'*.

9. Ils partagent d'ailleurs cette intéressante propriété, comme déjà signalé à la fin de la Section 3.1.3 (page 264), avec les adjectifs de couleur, d'appartenance et de forme.

TABLE 2 – *Verbes et élision : quelques exemples*

`charme/u`	`alunis/jnu`	`ébruite/jnlu`
`charmais/vc`	`alunissais/jnv`	`ébruitais/jnlv`
`charmai/w`	`alunis/jnw`[a]	`ébruitai/jnlw`
`charmerai/x`	`alunirai/jnx`	`ébruiterai/jnlx`
`charmer/y`	`alunir/nM`	`ébruiter/nlM`
	`aluni`	`ébruitée/F`
`archaïse/jnu`	`outrepasse/jnlu`	`assiège/jnmtlu`
`archaïsais/jnv`	`outrepassais/jnlv`	`assiégeais/jnmtlv`
`archaïsante/LMF`		`assiégeante/LMF`
`l'archaïsant`		`l'assiégeant`
`archaïsai/jnw`	`outrepassai/jnlw`	`assiégeai/jnmtlw`
`archaïserai/jnx`	`outrepasserai/jnlx`	`assiégerai/jnmtlx`
`archaïser/nM`	`outrepasser/nlM`	`assiéger/nmtlM`
`archaïsé`	`outrepassée/F`	`assiégée/LMF`
		`l'assiégé`

[a] En principe, on peut utiliser une seule entrée, `alunis/jnuw`, en lieu et place de `alunis/jnu` et `alunis/jnw`. L'utilitaire `makedic` (cf. 3.3.2) peut s'en charger automatiquement.

3.2.4. Élision et emplois réflexifs

Le système développé ci-dessus ne permet pas de régler le problème de l'élision du pronom *se* (dans *il s'aime* ou *ils s'aiment*). Ce problème présente deux caractéristiques :

- il ne concerne que la troisième personne, singulier ou pluriel, des différents temps du verbe ;
- il ne concerne pas tous les verbes (certains ne sont pas, ou rarement, utilisés de manière réflexive).

On peut adopter deux attitudes :

- entrer manuellement dans le dictionnaire toutes les formes contenant une élision du pronom *se* ;
- estimer que ce phénomène est marginal (beaucoup moins fréquent que l'élision du pronom *je* par exemple) et accepter que l'absence de ces formes provoque quelques fausses alertes.

Pour *Français-GUTenberg*, c'est la première solution qui a été choisie, mais la seconde peut être acceptable dans le cadre d'un petit dictionnaire spécialisé. Une solution intermédiaire, où les cas les plus fréquents seraient privilégiés, est aussi envisageable.

3.2.5. Élision et verbes « essentiellement pronominaux »

Un problème de même nature, mais de plus grande ampleur, se pose avec les verbes dits « essentiellement pronominaux » comme *s'absenter*. La seule solution avec de tels verbes est d'en entrer manuellement toutes les formes dans le dictionnaire. Ils sont heureusement peu nombreux (et seuls les verbes commençant par une voyelle ou un « h » muet posent problème) et il est peu probable que de tels cas apparaissent dans la composition d'un dictionnaire spécialisé. *Français-GUTenberg* en contient un certain nombre.

3.2.6. Vérification de la validité des entrées

Comment vérifier la validité d'une entrée de votre dictionnaire spécialisé ? Il suffit d'invoquer ISPELL de la manière suivante :

```
echo lemme/drapeau | ispell -d francais -e
```

et de vérifier que les mots engendrés ne comportent pas de monstruosités.

Si vous désirez valider une liste de mots, taper :

```
cat fichier.dico | ispell -d francais -e
```

3.3. Compiler le dictionnaire

Cette section expose les dernières étapes de la mise en œuvre d'un nouveau dictionnaire.

3.3.1. Étape 1 : créer les listes de mots complémentaires

Tout d'abord, il faut disposer de plusieurs listes de mots : les listes standards de *Français-GUTenberg* et celles que vous voulez ajouter au dictionnaire. Ces dernières devront avoir été composées, par vous ou par un tiers, selon les principes présentés au fil des Sections 3.1 et 3.2.

3.3.2. Étape 2 : `makedic`

Une fois la première étape effectuée, il faut invoquer `makedic`, un programme fourni avec *Français-GUTenberg*, pour créer `francais.dico`. Cette commande s'utilise de la manière suivante :

```
makedic <liste de dictionnaires> <nom du dictionnaire à produire>
```

par exemple :

```
makedic dicos.lst francais.dico
```

où `dicos.lst` est composé d'une série de lignes contenant chacune le nom d'un dictionnaire à inclure pour engendrer `francais.dico`.

Le programme `makedic` ne se contente pas de concaténer les fichiers ; il effectue également les opérations suivantes :

- il ordonne les entrées des dictionnaires (nécessaire pour `buildhash`) ;
- il supprime les doublons éventuels ;
- il regroupe les lemmes identiques (`compatis/u` et `compatis/w` deviennent `compatis/uw`) ;
- (optionnellement) il supprime les lemmes qui produisent des formes qui sont déjà engendrées par une autre entrée.

3.3.3. Étape 3 : `buildhash`

La dernière étape utilise le programme `buildhash`, fourni avec ISPELL, pour créer le fichier `francais.hash` qui sera directement utilisé par le correcteur. La commande type est :

```
buildhash francais.dico francais.aff francais.hash
```

Ceci fait, vous n'avez plus qu'à placer `francais.aff` et `francais.hash` à l'endroit où ISPELL va chercher ses dictionnaires[10]. Reportez-vous à la documentation du programme pour plus d'informations.

10. Généralement, c'est `/usr/local/lib`.

4. Conclusion et musique d'avenir

La plupart de mes travaux portent directement ou indirectement sur la correction automatique de l'orthographe française et l'amélioration des techniques existant dans ce domaine. J'espère pouvoir en faire profiter la communauté des utilisateurs francophones de (LA)TEX, sans doute avec le soutien de GUTenberg, en tenant à jour et en augmentant *Français-GUTenberg*. Aussi en développant, à plus long terme, un nouveau correcteur qui résoudrait les problèmes posés à la Section 2.3, qui permettrait une gestion plus souple de l'élision et serait plus apte que ISPELL à être mis en œuvre sous différents systèmes d'exploitation.

Bibliographie

[1] Bescherelle, *La conjugaison*, (Hatier, Paris, 1990).

[2] Nina Catach, *L'orthographe*, Nathan Université, Paris 3e édition, 1995.

[3] Maurice Grevisse, *Le bon usage*, Duculot, Paris/Gembloux, 12e édition, 1988.

[4] Richard Montague, The proper treatment of quantification in ordinary English, In R. H. Thomason, editor ; *Formal Philosophy: Selected Papers of Richard Montague*, New Haven, 1974, Yale University Press.

[5] Pythoud, Christophe. Problèmes de la correction automatique de l'orthographe lexicale du français à travers une étude de cas : le correcteur orthographique ISPELL et le dictionnaire *Français-IREQ*, Université de Lausanne, 1996, mémoire de licence.