hunspell le glouton

avec les scripts minimax et dicomax (Linux).

V2. février 2013.

Remerciements à pingouinux, nesthib et Watael.

Une idée différente.

La technique de correction orthographique utilisée avec OpenOffice a fait ses preuves. Elle est même très éprouvée. Elle se caractérise par le recours systématique (certains diront processionnel) à une fenêtre de dialogue. Dans le cas d'un texte contenant de nombreux mots propres nouveaux, la véritable *correction* orthographique est plus une exception qu'une règle. Et plus le texte est soigné, plus le livre est gros (c'est-à-dire + il y a de mots nouveaux), et plus cette technique de correction peut causer de frustration.

Cette manière de procéder n'est pas due au programme *hunspell*. Celui-ci est en effet capable de performances étonnantes, « gloutonnes », à condition d'être utilisé différemment.

Il est proposé une autre technique de correction, plus rapide, basée sur l'idée de *liste*, où le recours à cette fenêtre de dialogue ne se fera plus que sur votre demande. C'est le but des deux mini-scripts suivants :

Utilisation

- *minimax* est un procédé de correction orthographique rapide.
- *dicomax* permet de peupler par lot le dictionnaire utilisateur.

Les deux mini-scripts ci-dessus fonctionnent sous Linux. Veillez à leur donner des droits d'exécution.

Dépendances : *zenity*, *odt2txt, hunspell, hunspell-fr* (ou autre langue) et sa bibliothèque partagée – qui doit déjà être installée.

Le script à utiliser en premier pour la correction du texte est le script *minimax.*

Le script *dicomax* utilisera la liste standard créée par le premier pour peupler par lot le dictionnaire utilisateur.

Nouveautés de la v2.

minimax vous informe sur le nombre des occurrences :

- de chaque mot du texte (fichier **...2.txt**)
- de chaque mot nouveau et/ou inconnu (liste standard **...6.txt**)

minimax
La correction orthographique rapide.

Comment ça marche ?

Le script *minimax* analyse les fichiers odt. Pour un livre standard, il fournit *en moins de cinq secondes*[1] deux listes alphabétiques au format txt :

- une **liste standard (…6.txt)** contenant tous les mots uniques, inconnus et/ou mal orthographiés selon le dictionnaire *hunspell* (elle peut comprendre de 100 à 500 mots selon les livres[2]) et le nombre de leurs occurrences.
- une liste des formes élidées (**…5.txt**) (environ quatre fois plus grosse que la précédente).

L'examen de la seule liste standard suffit. La liste des formes élidées n'est fournie qu'à titre d'information – voir *Le cas des apostrophes* plus bas.

— Il va vraiment avaler tout ça ?
— Voui, en entrée…

À quoi servent ces fichiers ?

En dehors des deux listes numérotées **5** et **6** citées ci-dessus, le script *minimax* produit aussi des fichiers intermédiaires permettant, le cas échéant, d'en vérifier le fonctionnement.

À titre d'exemple, pour un fichier source nommé test.odt, vous trouverez au total :

- test.txt qui est le texte converti au format txt par *odt2txt*
- test2.txt, produite par *sed* et *grep* à partir de test.txt, est la liste des mots du texte classé par nombre décroissant d'occurrence).
- test3.txt, produite par *hunspell* à partir de test2 est une liste intermédiaire.
- test4.txt, produite par *grep* à partir de test.txt est une liste intermédiaire.
- test**5.txt** est la **liste alphabétique des formes élidées** produite par *hunspell* à partir de test4.
- test**6.txt** est la **liste alphabétique standard** des mots inconnus et/ou fautifs avec le nombre de leurs occurrences.

---

1 LMDE 64 bits – Processeur i3.

2 D'ordinaire de un à deux mots par page.

Avantages de la liste alphabétique.

- La vérification des mots inconnus et/ou fautifs s'effectue visuellement de façon fluide au parcours de la **liste standard**. Les graphies multiples éventuelles des mots nouveaux sont mises en évidence par juxtaposition. Le nombre de leurs occurrences est indiqué.
- L'appel à la fenêtre de dialogue F7 n'est plus systématique mais ponctuel. Il est suggéré de travailler en plaçant côte à côte la liste standard et le fichier odt et d'utiliser la recherche par glisser-déposer (voir copie d'écran **[A]** et **[B]**).
- Après validation, la liste standard peut être entrée par lot dans votre dictionnaire utilisateur, ce qui accélérera les corrections ultérieures (voir script *dicomax).*

Il ne paraît pas utile, sauf cas particuliers, de rentrer dans le dictionnaire utilisateur la liste des formes élidées : vous risqueriez de l'engorger rapidement alors que le script *minimax* continuera à traiter les apostrophes de la même façon…

Le cas des apostrophes.

On trouve en français deux types d'apostrophe. L'apostrophe droite, utilisée par défaut dans les traitements de texte et d'autres logiciels, et l'apostrophe courbe ou typographique utilisée si vous le voulez bien.

Cela influe évidemment sur la recherche des formes élidées. Comprenez qu'une recherche incluant *l'autre* apostrophe ne donnera rien. Si l'on n'y prête pas attention, c'est ce qui arrive par exemple avec Sigil qui ne connaît par défaut que l'apostrophe droite.

La liste standard et les apostrophes

La liste standard (**…6txt**) ne tient pas compte de l'apostrophe. *hunspell* va analyser les mots de part et d'autre de celle-ci. Il ne vérifiera pas la forme élidée en totalité, ce qui présente un – petit – inconvénient. Ainsi, l'expression **qu'elle** sera analysée en deux fois, exactement comme l'expression **qu elle** et la seconde graphie, pourtant fautive, ne sera pas signalée.

En attendant qu'un logiciel de correction ne le propose sous forme d'option automatique, vous pallier cet inconvénient en créant une expression régulière[3] destinée à signaler les apostrophes manquantes.

La liste ds formes élidées et les apostrophes.

La liste des formes élidées (**…5txt)** vous fournit à titre d'information la liste complète des formes élidées. Celles-ci sont très nombreuses en raison notamment de la présence de formes conjuguées. Il est conseillé de n'utiliser cette liste que pour vérifier ponctuellement une forme élidée.

Pour créer cette deuxième liste, le script *minimax* recherche par défaut l'apostrophe typographique. Pour la version anglaise jointe (minimax_en_US), il recherche par défaut l'apostrophe droite.

3 Par exemple, à partir de ces racines : _(aujourd|Aujourd|jusqu|Jusqu|lle|lorsqu|Lorsqu|presqu|Presqu|quelqu|Quelqu|qu|Qu|quoiqu|Quoiqu|c|C|d|D|j|J|l|L|m|M|n|N|s|S|t|T)_

Un exemple possible de correction[4]

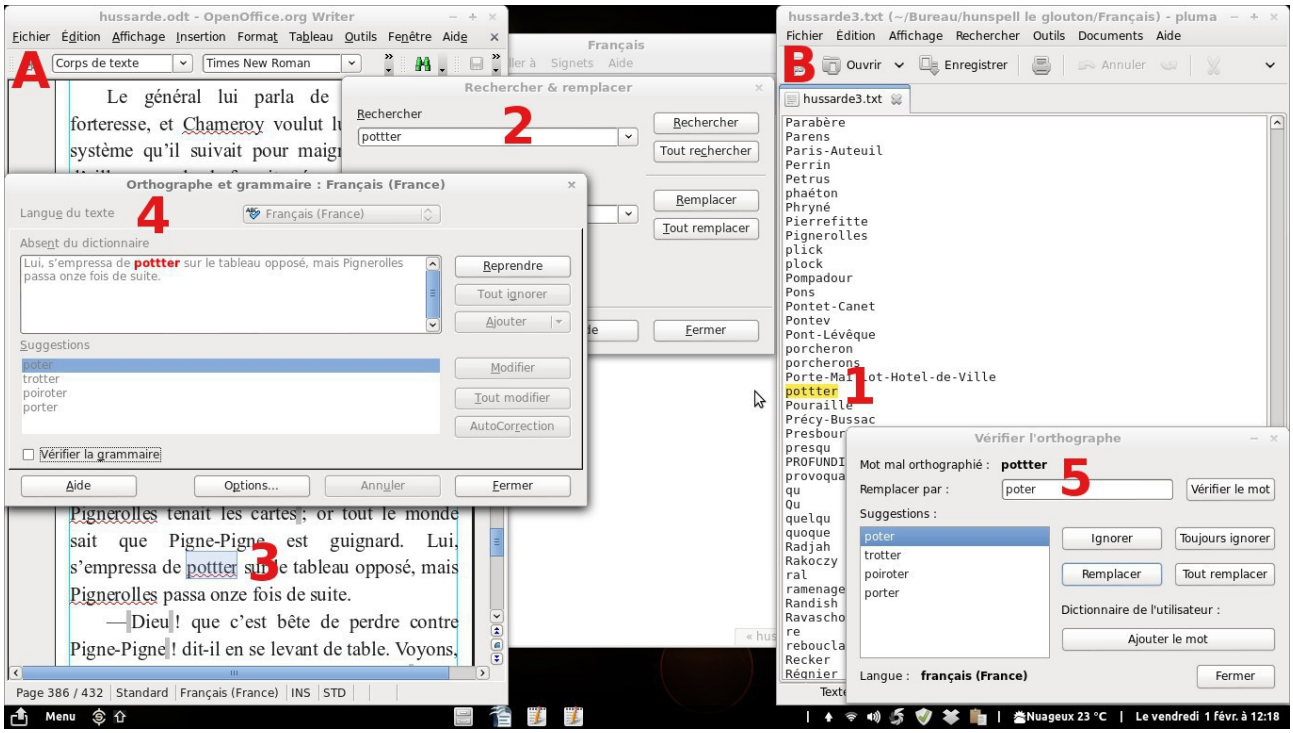
Une erreur a été repérée :[1] il y a deux possibilités :

- Si vous connaissez la réponse, utilisez le glisser-déposer vers la fenêtre chercher/remplacer d'OpenOffice [2] et veillez à son remplacement.
- Si vous ne la connaissez pas : il vous est proposé de faire l'essai suivant :

Prenons un mot qui se trouve sur la liste *hussarde3.txt* : **pottter** [1]
Ce n'est pas un nom propre et il ne veut rien dire, il y a donc une faute à corriger.

- Utilisez le glisser-déposer vers la fenêtre chercher/remplacer d'OpenOffice[2] et regardez le mot sélectionné dans son contexte [3].
- Si vous ne savez toujours pas, cliquer sur F7 pour voir ce que propose *hunspell.* [4] Rien de bien terrible n'est proposé…

En fait, la bonne réponse est **ponter** et il vous faudra la trouver par d'autres moyens.
C'est la vie…



Nota : Vous pouvez vérifier un mot soit à partir d'OpenOffice (F7) soit à partir de l'éditeur de texte (Maj+F7)[5]. Les résultats seront les mêmes puisqu'ils seront fournis tous deux par *hunspell.*

4 Cette image a déjà été utilisée pour la v1. Utilisez désormais le fichier **….6.txt** au lieu du …3.txt.

dicomax
Peuplement par lot du dictionnaire utilisateur.

L'utilisation de dicomax

Le peuplement mot par mot du dictionnaire utilisateur requiert une patience de bénédictin.

dicomax est un mini-script qui permet, après validation de la liste standard …**3.txt** fournie par le script *minimax.,* de transférer en une seule fois, livre par livre, les données de cette liste vers le dictionnaire utilisateur *hunspell.*

Cette injection massive de mots ne devrait pas manquer d'avoir un effet cumulatif rapide sur les capacités de reconnaissance de *hunspell.*

On peut imaginer divers usages. Le vocabulaire d'un lexique spécialisé peut ainsi être extrait *(minimax)* puis intégré *(dicomax)* en quelques secondes. Celui d'une bibliothèque ou des œuvres d'un auteur peut l'être à raison de quelques livres par minute.

La seule limitation – de bon sens – est qu'il convient de travailler sur des fichiers odt entièrement validés.

Si tous les hunspell du monde...

hunspell est installé par défaut sur toutes les distributions Linux (et au-delà). Bien qu'il puisse être installé directement comme il vous l'est proposé ici, il est le plus souvent installé par le biais de logiciels comme OpenOffice, Sigil, etc. Pour cette raison, il peut y avoir plusieurs instances de *hunspell* installées sur le même système d'exploitation, utilisant parfois mais pas toujours la même bibliothèque partagée. Tous les dictionnaires principaux *hunspell* sont classés par langue.

À propos du ou des dictionnaire (s) utilisateur

hunspell-fr installera par défaut le dictionnaire utilisateur à ~/.hunspell_fr_FR

Vérifiez avant de lancer le script que ce fichier existe bien. Sinon, créez-le (même vide).

Il s'agit d'un fichier texte brut (*text-plain*) contenant une liste de mots – sans forme fléchie – qui peut être directement modifiée. On peut y entrer par exemple : *Mussipontin, Mussipontins, Mussipontine, Mussipontines…* Ce fichier permettra à *hunspell* d'apprendre un vocabulaire spécialisé.

Compte tenu de ce qui a été dit au paragraphe précédent, il peut malheureusement exister plusieurs fichiers de dictionnaire utilisateur, ce qui ne contribue pas à l'efficacité d'ensemble. [voir man *hunspell* (-p dict) et fichiers d'aide spécifiques]. Les fichiers texte *standard.dic* ci-après sont des dictionnaires utilisateur qui peuvent aussi servir pour plusieurs langues.

- LibreOffice : ~/.config/libreoffice/4/user/wordbook/standard.dic.
- OpenOffice : ~/openoffice.org/3/ user/wordbook/standard.dic.