Manuel

PRESTO

Version alpha

Interface développée conjointement par Virginie Lethier (Maître de Conférences) et Nicole Salzard (Ingénieur en développement et déploiement d'applications) au laboratoire ELLIADD – pôle DTEPS (Université de Franche-Comté) avec l'appui des travaux réalisés par les Professeurs Jean-Marie Viprey et Philippe Schepens
le 24 avril 2017

SOMMAIRE

PRESTO

1. Présentation de Presto

Chercheurs et étudiants qui pratiquent l'analyse de données textuelles assistée par l'informatique et la statistique sont bien souvent confrontés à des données numériques « bruitées », qui rendent particulièrement chronophage l'étape de constitution et de documentation du corpus.

Le caractère « bruité » des données textuelles peut résulter de l'océrisation (ou « OCR ») d'un document dégradé (exemple : presse antérieure à 1950) : les sorties caractérisées par un faible taux de reconnaissance optique des caractères impliquent une laborieuse face de **correction** qui vise restituer une version textuelle fidèle à celle inscrite sur le document d'origine.

D'autres données numériques sont pour leur part « nativement bruitées » : c'est par exemple le cas des données textuelles issues des réseaux sociaux numériques (*Facebook, Twitter*) où abondent acronymes, sigles et/ou fautes d'orthographe et qui impliquent une nécessaire phase de **normalisation des graphies.**

Si **correction** et **normalisation** des données textuelles sont deux moments bien distincts (et parfois complémentaires) de l'étape de l'établissement des données, toutes deux sont véritables choix heuristiques, qui ne peuvent être totalement automatisés sous peine de créer davantage de bruits dans les données. Parce qu'ils constituent autant de choix placés sous la responsabilité chercheur, dont il s'agit de garder trace, ces moments impliquent également de pouvoir être archivés, facilement documentés et partageables.

Le logiciel Presto se propose de faciliter les opérations de prétraitement de données textuelles « bruitées » en offrant à l'utilisateur une interface conviviale, destinée à **l'assister** dans ses opérations de correction et de normalisation. Le logiciel Presto propose à l'utilisateur **différents parcours de correction,** permettant par exemple à l'utilisateur de corriger des formes fautives par lot de 25, tout en ayant une vue sur leur contexte d'emploi. L'utilisateur effectue alors la correction en une seule saisie en ayant la certitude de ne pas générer de nouvelles « coquilles ». Le **logiciel Presto** peut également soumettre à l'attention de l'utilisateur différentes formes soupçonnées fautives, car elles ne sont pas connues par le dictionnaire ou parce qu'elles correspondent à des erreurs fréquemment observées dans les sorties OCR. L'utilisateur peut également parcourir et corriger son corpus en procédant à des requêtes à partir d'une dizaine de critères, qui lui permettent par exemple de rechercher l'ensemble des mots contenant certaines lettres.

Le logiciel Presto invite l'utilisateur à **valider** systématiquement chacun de ses choix et à **en garder trace** dans des rapports archivés automatiquement. Ce faisant, le **logiciel Presto** vise à aider le chercheur à **décrire** et **historiciser** son pré-traitement des données.

Les données traitées dans Presto peuvent correspondre à des simples textes non-formatés (format .txt ou .xml) ou à des corpus respectant les normes attendues par l'un des logiciels suivants : TXM,

Iramuteq, Astartex, hyperbase, hyperbaseweb, lexico,Trameur, Coocs, Ictena. L'exportation du corpus corrigé suivra au choix l'un des formats des logiciels proposés en entrée ou sera un texte simple. Ce faisant, **Presto** peut donc être utilisé pour changer le format d'un corpus.

2. Installation de Presto

L'application a été testée sur Windows Vista, Windows 7 et la version **5.1** et 5.6 de MySQL.

a. Copie du répertoire de PRESTO

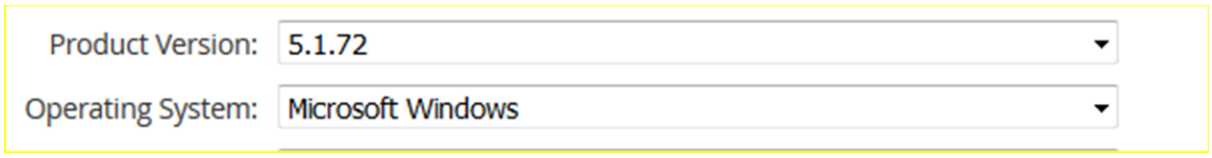
Télécharger le répertoire Presto.zip ici, puis le décompresser dans un répertoire que vous placerez où vous voulez sur votre ordinateur et que vous nommerez Presto.
Ne jamais modifier ou déplacer une partie du contenu de ce répertoire.

b. Installation de MySQL *5.1*

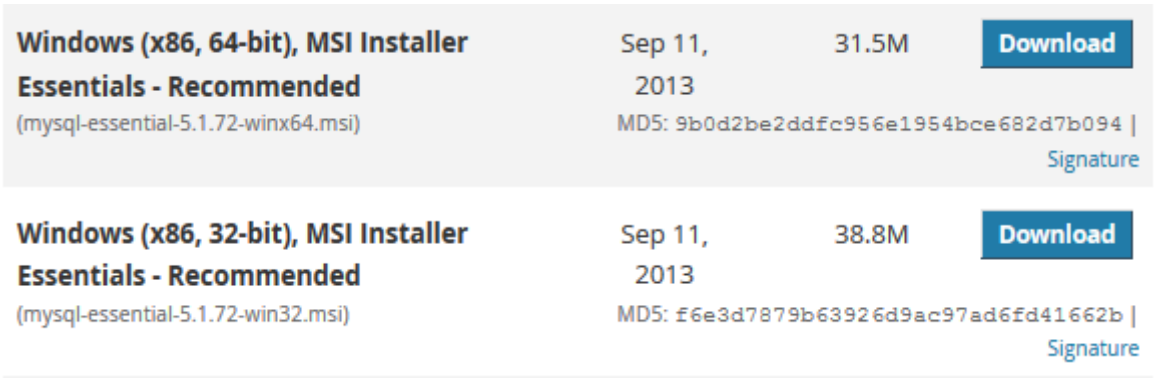
Il faut ici impérativement respecter les indications, entourées en rouge : toute erreur exigerait une désinstallation complexe.

Se rendre sur la page : https://downloads.mysql.com/archives/community/

Compléter le formulaire ainsi :

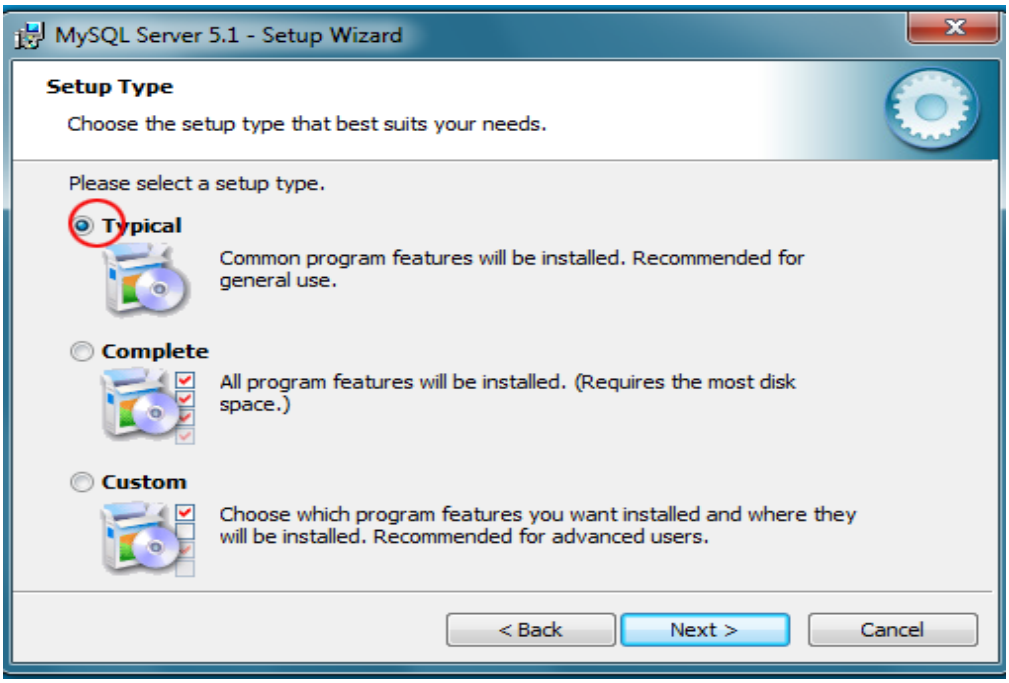


Puis télécharger la version qui convient à votre ordinateur :

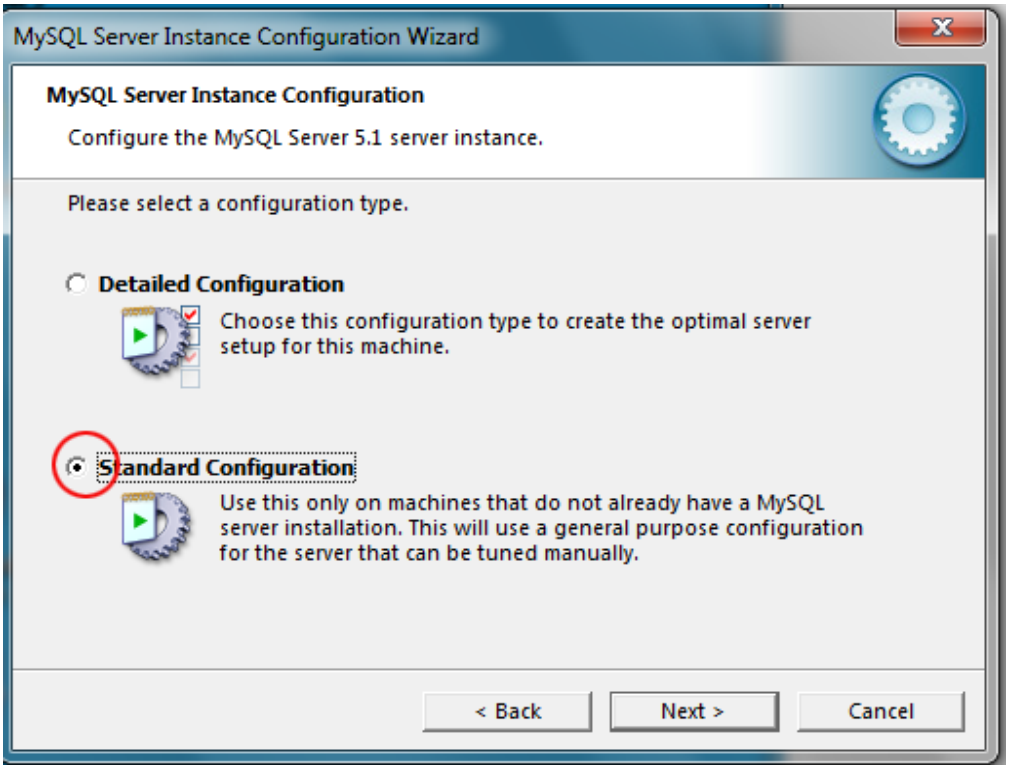


Double-cliquer sur le fichier téléchargé.

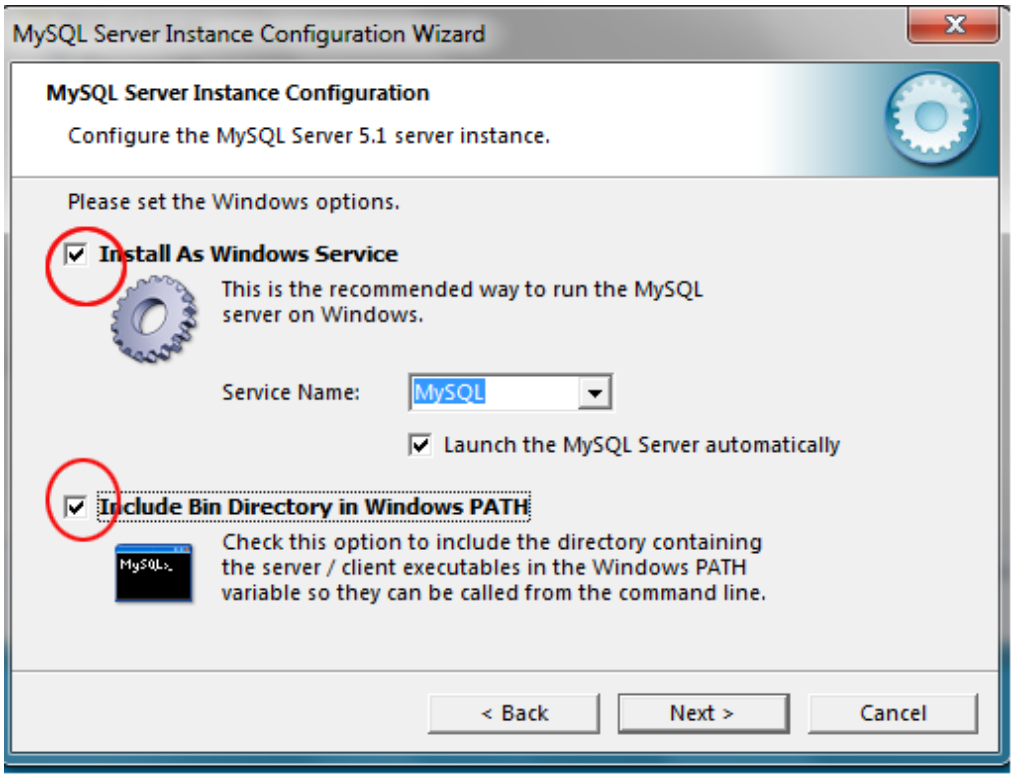
Choisir l'installation "typical ".



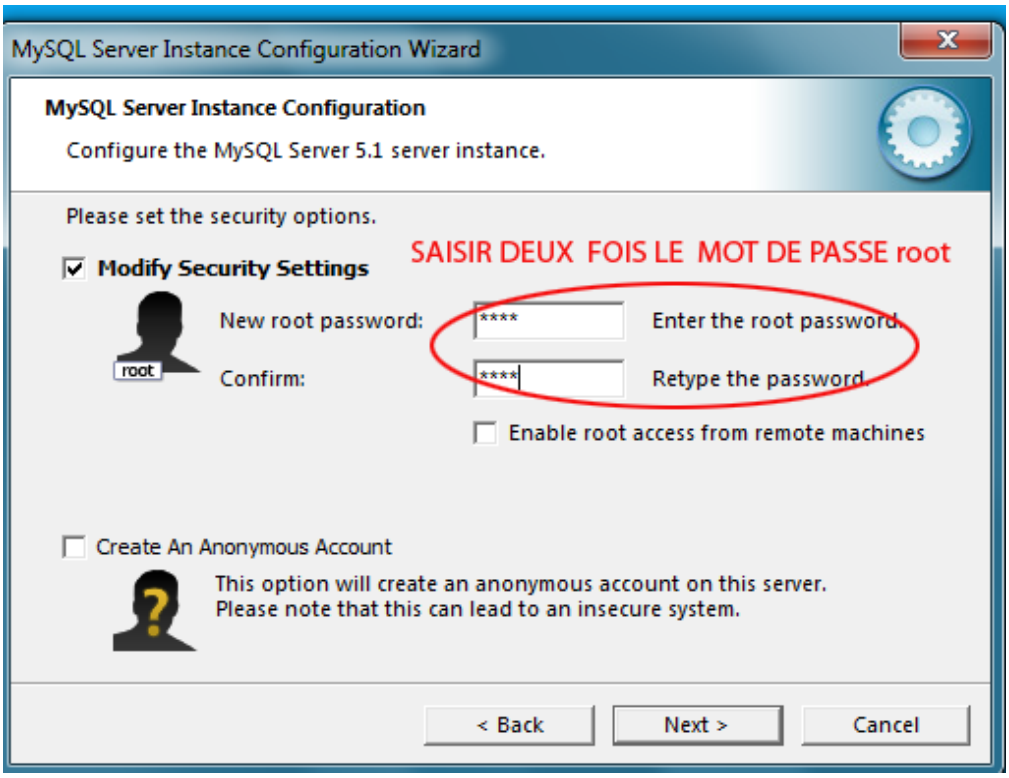
Choisir la configuration standard



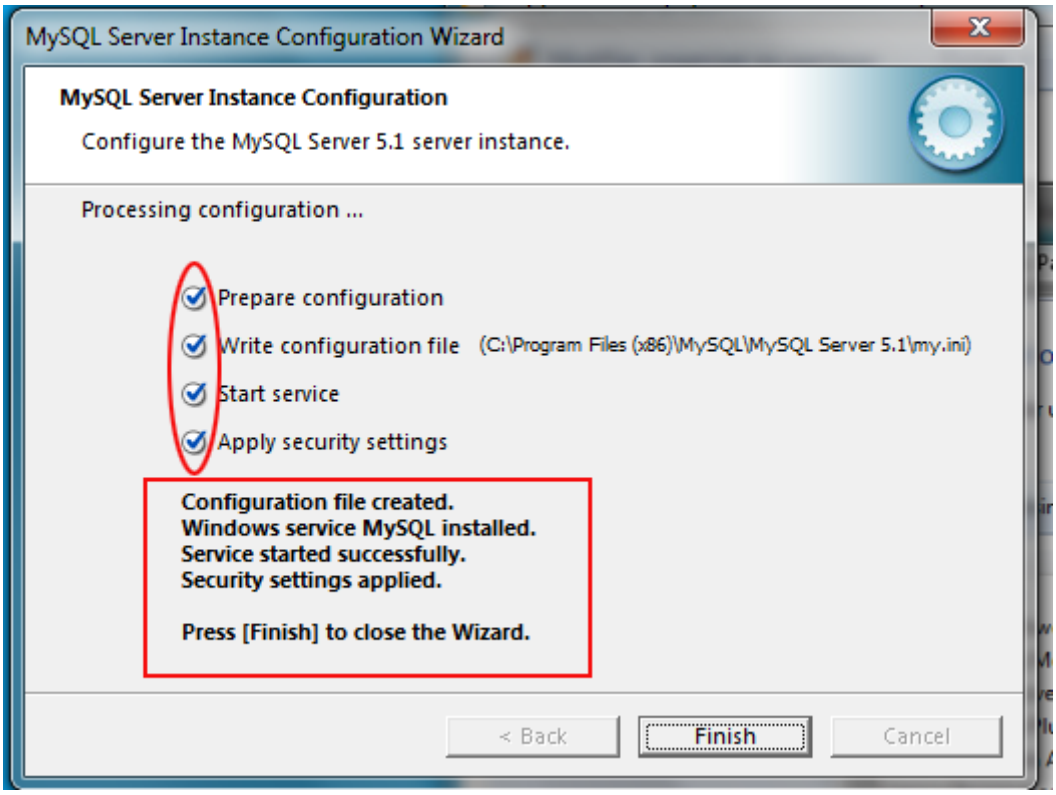
Cocher les deux cases : “ Install As Windows Service “ et “ Include Bin Directory in Windows Path “



Choisir “Modify Security Settings” :
new root password : **root**
confirm : **root**



L'installation est terminée quand vous obtenez l'interface suivante.



c. Installation de java (version supérieure ou égale à 7)

Télécharger Java à cette adresse : http://www.java.com/fr/download/.
Double-cliquer sur le fichier "java.exe" et suivre les explications.

d. Lancement de l'application

Pour lancer l'application, double-cliquer sur le fichier "PRESTO" qui se trouve dans le répertoire PRESTO. Le chargement des données prendra alors **quelques minutes**.

3. Formats d'entrée des données

Presto accepte plusieurs formats d'entrée de vos données textuelles.
Quel que soit le format d'entrée de vos données, celles-ci doivent être regroupées dans un dossier source, du nom de votre choix.
Tous les fichiers respectent un encodage UTF-8 ou cp1252 (encodage des textes par défaut de windows).

Cependant quelques restrictions sont imposées par Presto pour des raisons techniques ou de simplification. Il remplace en effet quelques caractères par d'autres :
les ¤ et les $ par un espace, tous les types de tirets par – , tous les types d'apostrophes par ' et un espace insécable par un espace sécable.

a. Textes non-formatés sans métadonnées

Le corpus contient un ou plusieurs fichiers dans un ou plusieurs répertoires au format txt.

b. Corpus "Astartex"

La configuration d'un corpus accepté par Astartex est la suivante :
Les données textuelles sont regroupées dans un (ou plusieurs fichiers) enregistré(s) dans un dossier-répertoire. Le format des fichiers correspond à du texte brut (.txt).
Chaque texte doit être précédé de métadonnées selon le modèle suivant :

<ASTX_ATTR SOURCE=Petit_Comtois DATE=1883_08_13 NUMERO=0013 PAGE=1 LARGEUR=1>

Les principes de rédaction des métadonnées sont les suivants :
Les métadonnées doivent être introduites par la séquence <ASTX_ATTR et se succèdent jusqu'à la fermeture du chevron. Les métadonnées sont séparées entre elles par un espace. Le nom d'une métadonnée et sa valeur sont séparées par un signe « égal ». Le nom et la valeur d'une métadonnée ne comportent pas de signe « égal » ni d'espace.

c. Corpus "ICTeNA"

La configuration d'un corpus accepté par ICTeNA est la suivante :
Les articles sont répartis dans un ou plusieurs fichiers dans un ou plusieurs répertoires. Chaque texte doit être précédé de métadonnées selon le modèle suivant :

*DATE: 27/01/2011
*SOURCE: LIBERATION
*AUTEUR: VINCENT GIRET
*RUBRIQUE: EDITORIAL
*TITRE: Soif de liberté
*TEXTE-ARTICLE:
Le peuple avance comme …

Les principes de rédaction des métadonnées sont les suivants :
Le groupe des métadonnées commence obligatoirement par **DATE* : et se termine par **TEXTE-ARTICLE :*
Chaque nom de métadonnées est placé sur une nouvelle ligne et est précédé de * .
Le nom d'une métadonnée et sa valeur sont séparées par le signe :

d. Corpus "Iramuteq"

La configuration d'un corpus accepté par Iramuteq est la suivante :
Les données textuelles sont enregistrées dans un seul fichier mettant bout-à-bout tous les textes que l'on souhaite analyser.
Chaque texte doit être précédé de métadonnées selon le modèle suivant :
**** *an_1998 *mois_1 *code_01-98 *jour_16 *loc_Duisenberg *stat_Pres *sex_Hom
Les principes de rédaction des métadonnées sont les suivants :
Les métadonnées sont sur une ligne qui commencent par ****. Chaque nom de métadonnée est précédé d'une étoile. Le nom et la valeur d'une métadonnée sont séparés par le signe underscore (_).Le nom et la valeur ne doivent contenir que des caractères parmi a-z, A-Z, 1-9.
Pour plus d'information sur les principes de rédaction des métadonnées Iramuteq : http://www.iramuteq.org/documentation

e. Normes d'un corpus "Hyperbase (version logiciel)"

La configuration d'un corpus accepté par la version logicielle d'Hyperbase est la suivante : vos données peuvent être enregistrées dans un seul fichier ASCII en format *.TXT.* Si vos données textuelles sont enregistrées dans plusieurs fichiers, le corpus les considérera comme autant de textes, constitutifs du corpus.

Chaque texte doit être précédé de métadonnées selon le modèle suivant :

&&&La vie en rose, Rose, VR&&&

Pour plus d'information sur les principes de rédaction des métadonnées Hyperbase logiciel : http://ancilla.unice.fr/bases/manuel.pdf

f. Normes d'un corpus "Hyperbase (web)"

Les configurations d'un corpus accepté par Hyperbase Web (http://hyperbase.unice.fr/) sont au nombre de trois.
La première consiste à soumettre à l'interface un fichier de type .txt par texte : l'interface Hyperbase Web vous proposera alors de renseigner les métadonnées dans un tableau.
La deuxième consiste également à envoyer un fichier txt par texte dont les métadonnées sont reportées dans le nom des fichiers et séparées par un underscore :

auteur1_annee1_genre1.txt
auteur2_annee2_genre2.txt
auteur3_annee3_genre3.txt

La troisième demande de suivre les normes d'Iramuteq.

Hyperbase admet tous les encodages de caractères mais il est préférable d'utiliser UTF-8.

Pour plus d'information sur Hyperbase Web : http://hyperbase.unice.fr/hyperbase/

g. Normes d'un corpus "Lexico3-Trameur-Coocs"

La configuration d'un corpus accepté par Lexico3 est la suivante :

Les données textuelles sont enregistrées dans un seul fichier mettant bout-à-bout tous les textes que l'on souhaite analyser. Le texte doit être enregistré sous la forme d'un fichier texte seulement (.txt).
Chaque texte doit être précédé de métadonnées selon le modèle suivant :

<source=monjournal> <date=2015> <rubrique=actualités>

Pour plus d'information sur les métadonnées Lexico3 : http://lexi-co.com/ressources/manuel-3.41.pdf

h. Normes d'un corpus "TXM-XML/w+csv"

La configuration d'un corpus XML accepté par TXM est la suivante :
Chaque texte est enregistré dans un fichier de type .xml dont le prologue est : <?xml version="1.0" encoding="UTF-8"?>
Nous suivrons le modèle précédent pour décrire ses fichiers sans se reporter aux normes xml. Le groupe des métadonnées commencent par la balise : <text et se termine par ><body>
Les métadonnées sont séparées par un espace, on place un signe égal entre le nom et la valeur d'une

métadonnée. Les valeurs des métadonnées sont encadrées par des guillemets anglais ". Le texte se termine par </body></text>

Chaque texte doit être précédé de métadonnées selon le modèle suivant :

```
<?xml version="1.0" encoding="UTF-8"?>
<text catDesc="Non" date="1884" fulldate="1884-03-23" author="Aristarque" genre="Article de fond">
<body>
M. Oudet était déjà républicain en 1848. Je n'examine …
</body>
</text>
```

i. Normes d'un corpus "TXM-TXT+métadonnées.cvs"

La configuration d'un corpus "TXM-TXT+métadonnées.cvs » accepté par TXM est la suivante :

Chaque texte est dans un fichier d'extension .txt qui ne contient pas de métadonnées. En effet celles-ci sont collectées dans un tableur qui doit être enregistré sous format cvs.

Les normes de ce type de corpus sont détaillées sur Internet :
https://groupes.renater.fr/wiki/txm-users/public/tutoriel_import_txt_csv

4. Importer mes données textuelles dans Presto

Créer un dossier dans lequel vous placerez vos données textuelles en format .xml ou .txt. Ces données peuvent être réunies en un seul fichier ou être composées de plusieurs fichiers.

Il vous est possible d'afficher des images dans Presto pour faciliter vos opérations de correction des données textuelles. Si vous souhaitez afficher dans PRESTO des documents sources au format image, il vous faudra créer un dossier pour les données textuelles et un pour les images. A l'intérieur de ces répertoires, le nom du fichier texte et de l'image correspondante sans extension seront identiques

Lancer Presto en cliquant sur le fichier PRESTO.jar dans le dossier PRESTO.

Dans le menu « PROJET », choisir « NOUVEAU ». La fenêtre de dialogue suivante s'ouvre :

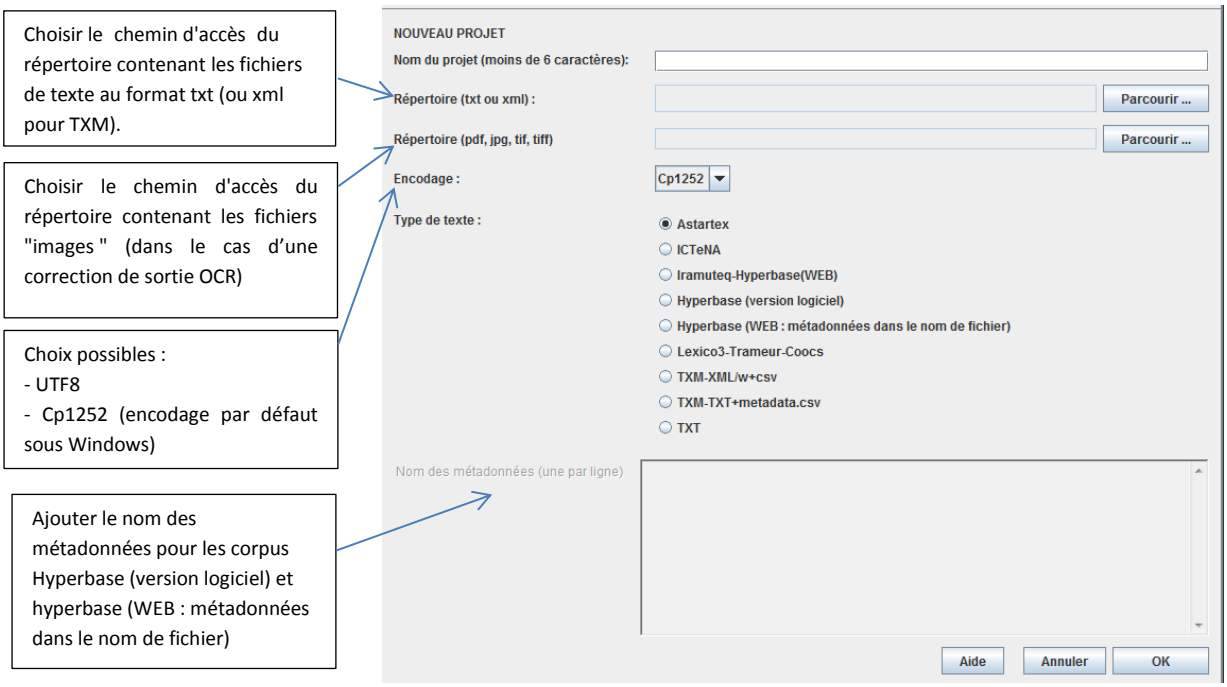


Illustration 1 : Fenêtre de dialogue de création d'un nouveau projet

Valider vos choix en cliquant sur OK, puis patienter: le traitement de vos données peut prendre quelques minutes.

5. Vérifier / modifier les métadonnées.

Choisir le menu CORRECTIONS/Métadonnées

Le tableau des noms et des valeurs des métadonnées s'affiche. (Illustration 2)

Fichier	catDesc	date	fulldate	author	macroca...	genre
1883\1...	AVIS	1883	1883-0...	NON	articl...	AVIS
1883\1...	NONE	1883	1883-0...	NON	articl...	ARTICL...
1883\1...	Chambr...	1883	1883-0...	SYLVIN	articl...	compte...
1883\1...	NONE	1883	1883-0...	SYLVIN	articl...	COMPTE...
1883\1...	SENAT	1883	1883-0...	NON	articl...	BREVE

Illustration 2 : Tableau de métadonnées pour vérification et modification

a. Pour trier les colonnes

Faire un clic gauche sur les entêtes

b. Renommer une colonne

Faire un clic droit sur son nom *n* pour ouvrir une boîte de dialogue dans laquelle la saisie du nouveau nom *nv* sera effectuée.

Si le nom *nv* est déjà donné, l'application essaiera de fusionner les colonnes c'est-à-dire de reporter les valeurs de la colonne n dans la colonne nv. En cas de doublon le processus de fusion n'aura pas lieu.

c. Déplacer les colonnes

Les faire glisser avec la souris

d. Supprimer les fichiers

Sélectionner la ligne et cliquer sur le bouton "Supprimer les fichiers sélectionnés"
Attention cette action supprime définitivement toutes les données concernant ce fichier

e. Ajouter une colonne / supprimer une colonne

Cliquer sur le bouton adéquat

f. Agréger des contenus de colonnes

Cliquer sur le bouton "Agréger les colonnes".
Ceci consiste à réunir les valeurs de plusieurs colonnes dans la première en les séparant par _ dans l'ordre d'affichage du tableau. Si cet ordre ne vous convient pas déplacer au préalable les colonnes.

g. Réinitialiser

Cliquer sur le bouton "Réinitialiser".
Le tableau reprend le contenu qu'il avait lors de l'ouverture de la fenêtre à **l'exception des fichiers supprimés**.

h. Le nom d'un fichier ne peut pas être modifié

i. Valider

Les valeurs affichées à l'écran sont enregistrées dans un fichier.
Nous garderons toujours le fichier créé lors de l'initialisation du corpus. Par contre, chaque validation entraînera l'écrasement du fichier précédent des données corrigées s'il existe.

6. Corriger des données textuelles dans Presto

Le texte du corpus est segmenté en Unités textuelles Atomiques (UTA) qui seront la base de notre correction. Il s'agira globalement de remplacer une ou plusieurs UTA entières par une ou plusieurs UTA.
Ainsi par exemple l'UTA *latable* sera remplacée par les deux UTA : *la* et *table* ou encore *lata ble* par *la table.*
A noter : Presto ne gère pas la suppression ou l'ajout de sauts de ligne dans les textes.

a. Quelques principes généraux

A partir du menu Corrections/Texte, une interface de correction s'affiche ayant la configuration suivante.

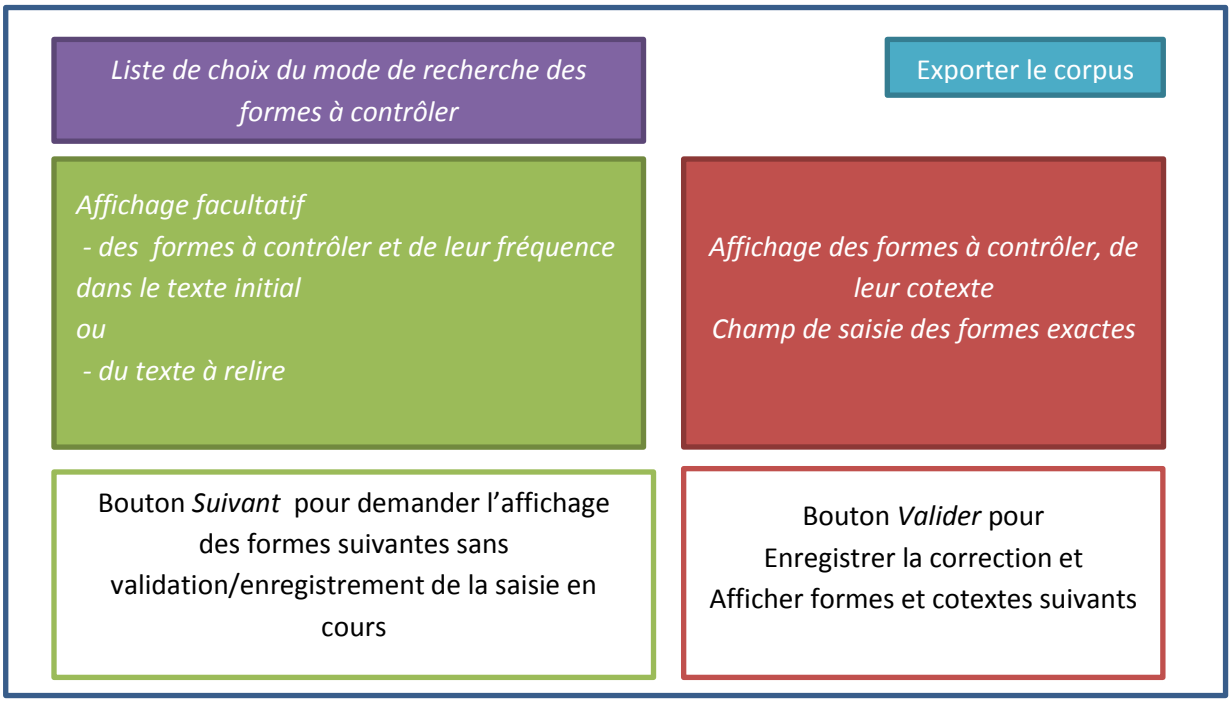


Illustration 3 : Schéma de l'interface

Le panneau en haut violet permet de choisir un des 10 **modes de recherche des formes à contrôler** et de le paramétrer.

Le vert est facultatif et modulable en fonction du **mode de recherche des formes à contrôler**. Il affiche tantôt la forme ou les formes à contrôler, tantôt l'arborescence des articles et le texte à relire. Dans le cas de l'affichage des formes à contrôler cliquer sur le bouton *Suivant* pour passer aux formes suivantes sans faire d'enregistrement de formes corrigées.

Dans le panneau rouge s'affiche toujours de la même façon un tableau de 4 colonnes. La colonne *Forme à contrôler* est entourée de cotextes gauche et droit constitués au maximum de 10 UTA. La première colonne permet de situer l'emplacement de la *Forme à contrôler* dans le corpus en indiquant le nom de son fichier et son numéro de ligne. Cliquer sur le bouton *Valider* pour enregistrer la correction éventulle et passer aux formes suivantes

b. 10 modes de recherche des formes à contrôler

Il existe en tout 10 modes de recherche des formes à contrôler, proposés dans une liste de choix qui se trouve dans le panneau violet (Illustration 4).

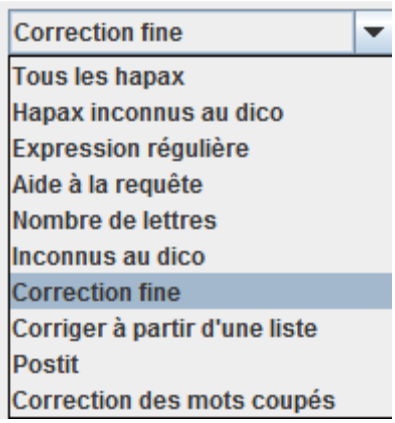


Illustration 4 : Liste de choix des modes de recherche des formes à contrôler

Pour chacun de ces modes de correction, la démarche est la même : choisir un mode, le paramétrer éventuellement et cliquer sur *OK* pour lancer la recherche sur tout le corpus.

Illustration 5: Choix et paramétrage du mode de recherche des formes à contrôler

i. *Le mode « correction fine »*

Corriger un texte peut se faire en le lisant du début à la fin et en corrigeant les erreurs au fur et à mesure. C'est ce que nous appellerons une *correction fine.* Choisir cet item dans la liste déroulante. *(Illustration 6)*

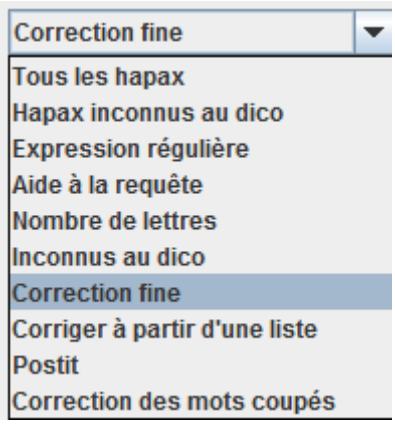


Illustration 6 : Choix du mode correction fine

Dans ce cas, le logiciel affiche la liste des fichiers des corpus ainsi que le premier texte.
La navigation d'un texte à l'autre se fait en cliquant sur le nom d'un fichier. L'utilisateur lit le texte et sélectionne à l'aide de la souris les formes à contrôler : une ou plusieurs UTA. (Illustration 7)

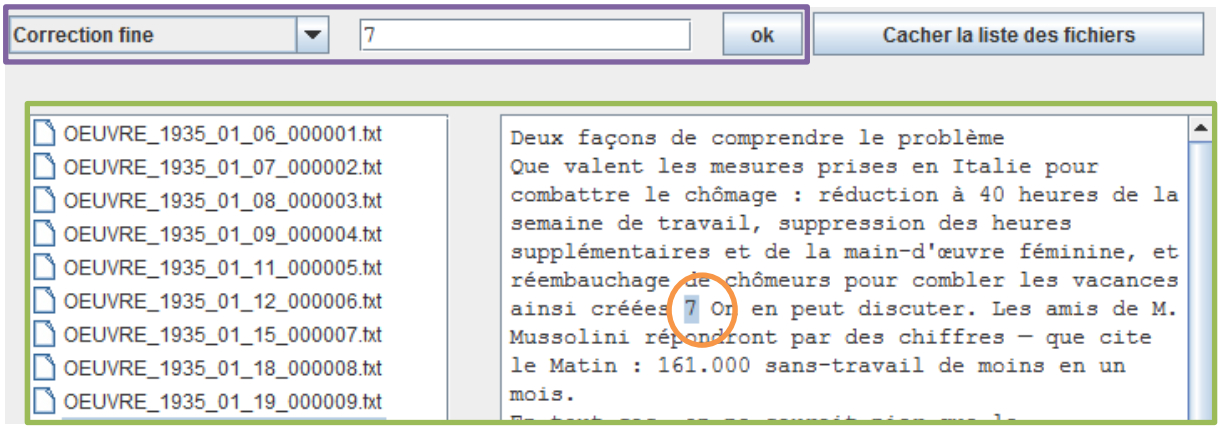


Illustration 7 : Choix d'une ou plusieurs UTA consécutives

La recherche de la forme sélectionnée ainsi que leurs deux cotextes et leur emplacement dans les fichiers est lancée sur tout le corpus. Le résultat s'affiche dans un tableau de 25 lignes maximum. (Illustration 8)

Fichier	Cotexte gauche	Forme à contrôler	Cotexte droit
OEUVR...	le flot trompeur de l'optimisme. Où allons-nous	7	»
OEUVR...	haines, ni faveurs, ni clans, ni dynasties	7	L'avancement sera s...
OEUVR...	L'avancement sera soustrait à l'influence du pouvoir législatif	7	II le sera à l'acti...
OEUVR...	et réembauchage de chômeurs pour combler les vacances ainsi créées	7	On en peut discuter...
OEUVR...	Vingt-cinq milliards, monsieur	7	Mais n'y a-t-il pas...

Illustration 8 : Tableau affiché suite à la recherche de la forme à contrôler *7*

Cliquer une fois sur la forme à contrôler pour saisir la correction.

Fichier	Cotexte gauche	Forme à contrôler	Cotexte droit
OEUVR...	le flot trompeur de l'optimisme. Où allons-nous	7	»
OEUVR...	haines, ni faveurs, ni clans, ni dynasties	7	L'avancement sera s...
OEUVR...	L'avancement sera soustrait à l'influence du pouvoir législatif	7	II le sera à l'acti...
OEUVR...	et réembauchage de chômeurs pour combler les vacances ainsi créées	7	On en peut discuter...
OEUVR...	Vingt-cinq milliards, monsieur	7	Mais n'y a-t-il pas...

Illustration 9 : Correction de la forme à contrôler *7*

Pour faire la correction de plusieurs formes à contrôler en seule saisie sur plusieurs lignes, cliquer sur les formes à retenir avec la souris tout en maintenant
- la touche Contrôle enfoncée dans cas de lignes non consécutives,
- la touche Majuscule dans le cas de lignes consécutives.

Noter qu'il faut appuyer sur la touche *Entrée* pour que le tableau affiche la nouvelle saisie.

Quand toutes les formes sont revues, cliquer sur le bouton *Valider* pour enregistrer les nouvelles données du tableau dans la base. L'application recherchera alors les formes à contrôler suivantes. Si vous ne désirez modifier aucune forme, cliquer sur le bouton *Valider* pour afficher les formes suivantes.

A chaque validation l'application mémorise les caractéristiques de la dernière ligne du tableau. Ainsi en cas d'arrêt l'utilisateur pourra reprendre son travail là où il l'avait laissé.

Cas particuliers :

1. Si vous désirez **supprimer une forme** il faut la remplacer par $. Ce signe sera supprimé lors de la restitution du corpus corrigé et permet de garder une trace de la modification pendant le travail de correction.
2. Pour supprimer **un ou des espaces** les sélectionner dans le texte et se laisser guider par les boîtes de dialogue.
3. Pour **ajouter un espace**, sélectionnez l'UTA précédente et ajouter l'espace dans la cellule du tableau de correction.
4. Il se peut qu'une partie de la **forme à contrôler se trouve non seulement dans la colonne prévue mais aussi dans celles des cotextes** comme dans la ligne 2 du tableau ci-dessous.

Cotexte gauche	Forme à contr...	Cotexte droit
aucoup moins nombreuses à Amiens qu'à	j	Chantilly, et elles s...
n en plein air,	j	mais au Cirque. Elles...
r but, puisqu'elles	j	ont provoqué, comme i...
mme un événement d'une	j	rare gravité, d'où po...

Illustration 10 : Exemple d'une forme à contrôler sur deux colonnes

La sélection de la forme erronée dans la colonne cotexte provoque l'affiche de la boîte de dialogue suivante (Illustration 11).

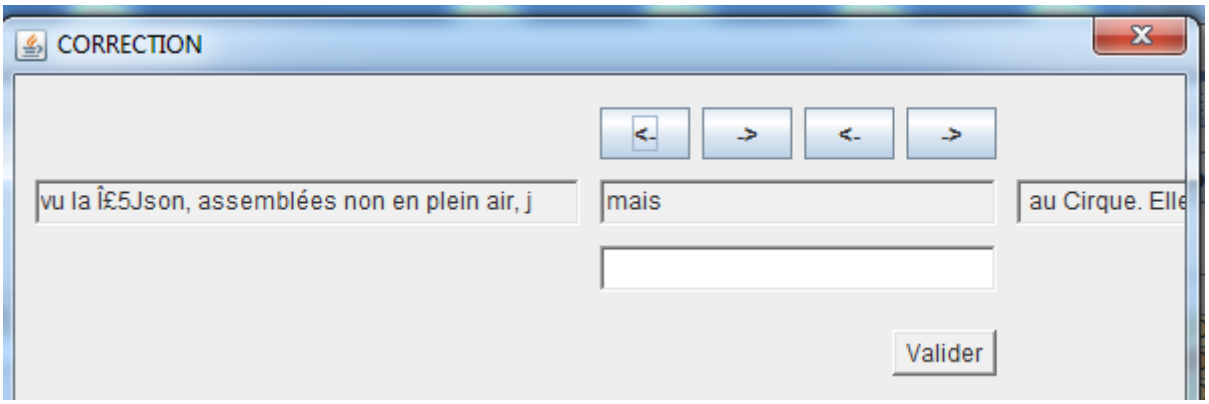


Illustration 11 : Dialogue pour regrouper deux formes graphiques

A l'aide des flèches ajouter le *j* au *mais* dans le champ central et écrire la forme correcte dans le champ texte en bas vide. (Illustation 12)

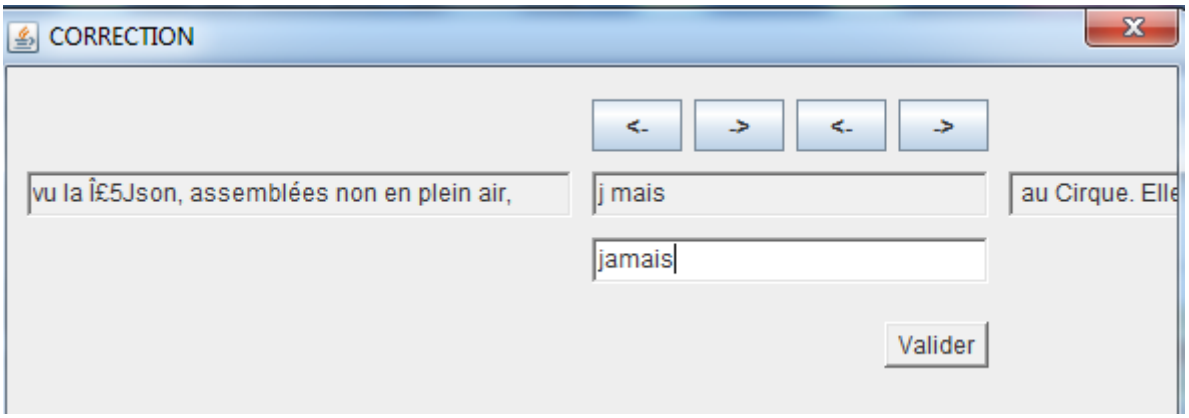


Illustration 12 : Dialogue pour réécrire une forme graphique à partir de deux formes graphiques

Ainsi *j mais* sera corrigé en *jamais*. Il se peut que la forme ainsi corrigée se trouve ailleurs dans le corpus. Si c'est le cas, la forme *j mais* ainsi que sa correction seront enregistrées et nous pourrons y revenir à partir du mode de choix de correction *Postit* . (cf paragraphe 6.b.ix ci-dessous)
Ce mode *Postit* autorise aussi de sélectionner une forme à corriger
- sur un seul cotexte
- sur la forme à contrôler et un de ses cotextes
- sur la forme à contrôler et ses deux cotextes

5. En cliquant sur le nom d'un fichier **l'image du texte issu d'un scan s'affichera.**
   Attention : il faut que l'adresse relative des fichiers images et des textes correspondent.
6. Pour **rechercher une forme dans le texte lu** lors de la correction fine, appuyer sur la touche *Ctrl*

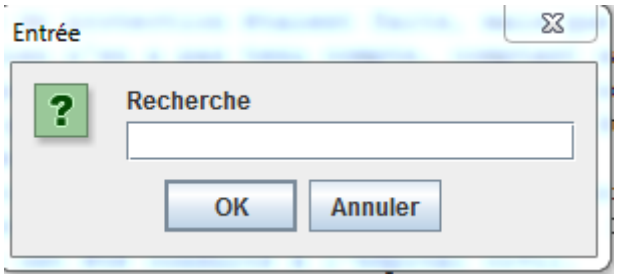
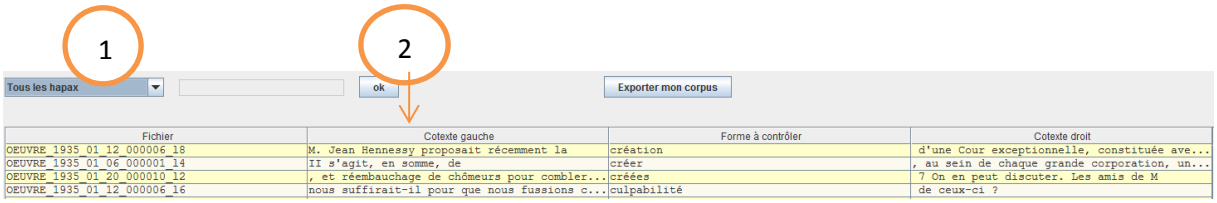


Illustration 13 : Dialogue pour recherche une suite graphique dans un texte lu lors de la *Correction fine*

Pour rechercher les suivantes appuyer sur la touche *F3*.

ii. Mode « Par hapax »

Choisir l'item « Tous les hapax ». Cliquer sur le bouton OK pour lancer une recherche, sur tout le corpus, des UTA qui ne sont présentes qu'une seule fois dans le corpus et qui n'ont jamais été corrigées. (Illustration 14)
Cette recherche respecte les majuscules et les minuscules, ainsi que les accents.



Fichier	Cotexte gauche	Forme à contrôler	Cotexte droit
OEUVRE_1935_01_12_000006_18	M. Jean Hennessy proposait récemment la	création	d'une Cour exceptionnelle, constituée ave...
OEUVRE_1935_01_06_000001_14	II s'agit, en somme, de	créer	, au sein de chaque grande corporation, un...
OEUVRE_1935_01_20_000010_12	, et réembauchage de chômeurs pour combler...	créées	7 On en peut discuter. Les amis de M
OEUVRE_1935_01_12_000006_16	nous suffirait-il pour que nous fussions c...	culpabilité	de ceux-ci ?

Illustration 14 : Dialogue pour rechercher tous les hapax

Effectuer la correction comme précédemment au paragraphe 6.b.i.

iii. Mode « Par hapax inconnus au dictionnaire »

Ce mode de correction soumet les hapax inconnus au dictionnaire annexé à Presto et non corrigées. Ce dictionnaire est une ressource construite par Jean-Marie Viprey. Il contient plus de 470 000 formes graphiques de français contemporain.

iv. Mode « Par expression régulière »

Choisir le mode de recherche *Expression régulière*, saisir l'expression dans le champ texte et lancer la recherche en cliquant sur « OK ». Elle se fera sur les formes qui ne sont pas encore corrigées.
La liste des formes à contrôler s'affichera par ordre de fréquence décroissant.
Le tableau des formes à contrôler et leurs cotextes sera également édité. (Illustration 15)
La correction et la validation s'effectuent comme précédemment au paragraphe 6.b.i.

Par exemple, lors d'un OCR, les caractères *7* et *?* sont souvent confondus. Recherchons donc les mots contenant un 7.

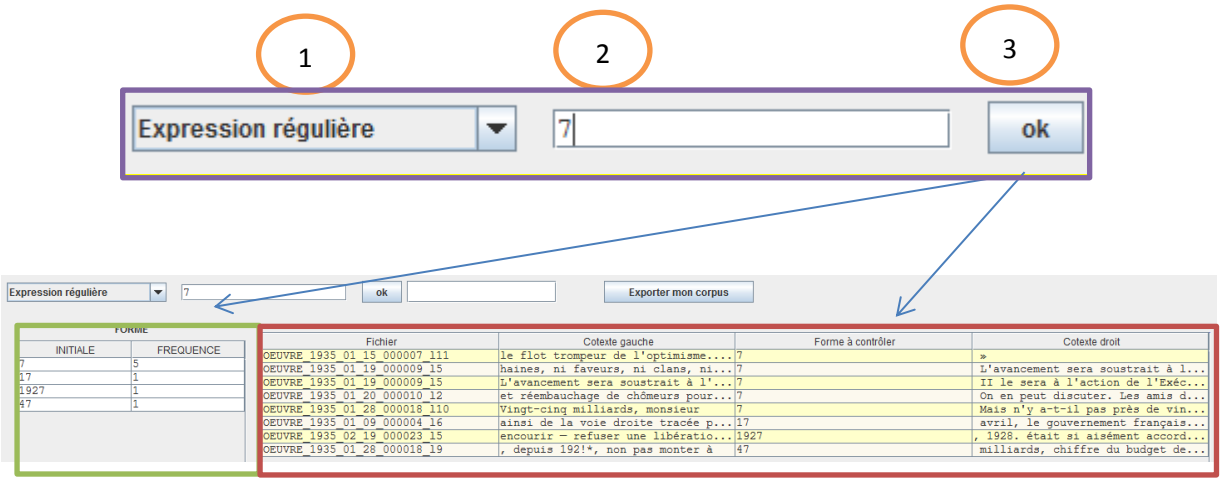


Illustration 15 : Contrôle à partir d'une expression régulière

Certains caractères # ! ^ $ ( ) [ ] { } | ? + * .>< ne sont reconnus que s'ils sont précédés d'un antislash : ce sont les métacaractères.

Exemples :

- Pour rechercher la forme <p> il faudra saisir \<p\>

- L'expression régulière ^[0-9]{1,2}$ recherche tous les nombres de un ou deux chiffres. Si les formes trouvées n'ont aucun cotexte, il s'agit d'un nombre seul sur une ligne et nous avons de forte chance que soit un numéro de page.

Variantes :

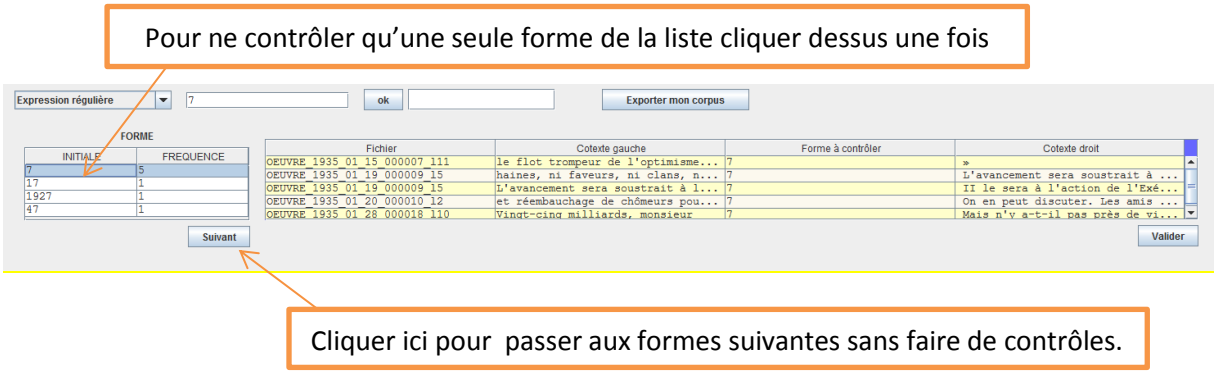


Illustration 16 : Astuces pour affiner la recherche des formes à contrôler

v. *Mode « Aide à la requête »*

Cette aide permet de rechercher des mots commençant, terminant, égal ou contenant la forme saisie dans le champ texte tout en respectant casse et accents.

Choisir l'item *Aide à la recherche* puis le sous-menu *Mot commençant par …*, enfin saisir la suite graphique à rechercher. Cliquer sur le bouton *OK* pour lancer une recherche qui aboutira à l'affichage de la liste des formes à contrôler et du tableau des cotextes. (Illustration 17)

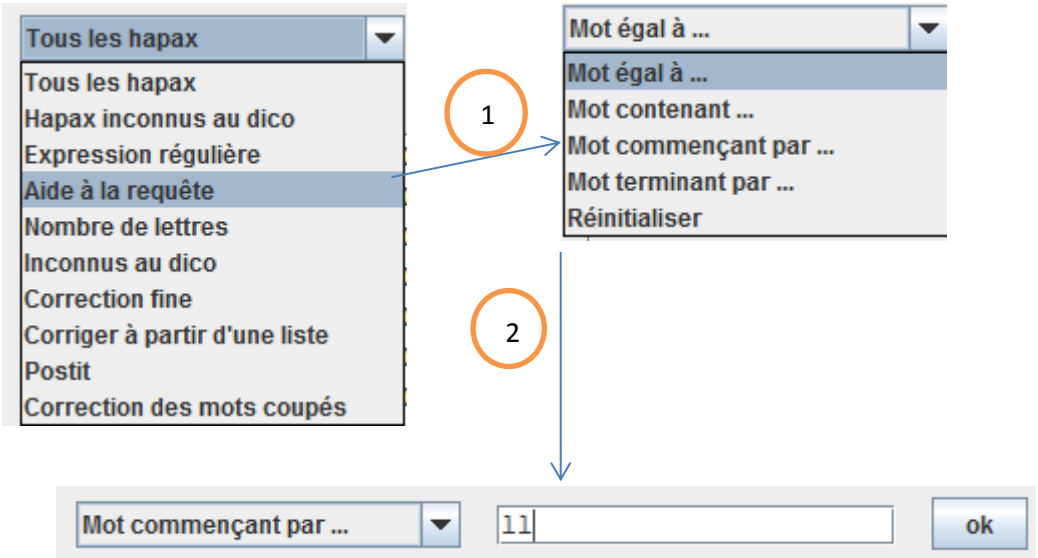


Illustration 17 : Dialogue pour rechercher les mots commençant par les caractères inscrits dans le champ texte

Effectuer la correction comme précédemment au paragraphe 6.b.i.

vi. *Mode « Nombre de lettres »*

Les formes à corriger auront un nombre de lettres donné par l'utilisateur. (Illustration 18)

Illustration 18 : Dialogue pour rechercher les formes ayant un nombre donné de lettres

Effectuer la correction comme précédemment au paragraphe 6.b.i.

vii. Mode « Inconnus au dictionnaire »

Ce mode de correction soumet les formes qui ne sont pas référencées dans le dictionnaire annexé à Presto.

Cette ressource de 470 000 formes graphiques de français contemporain a été construite par Jean-Marie Viprey

La démarche de correction est identique à celle du paragraphe 6.b.i (Illustration 19)

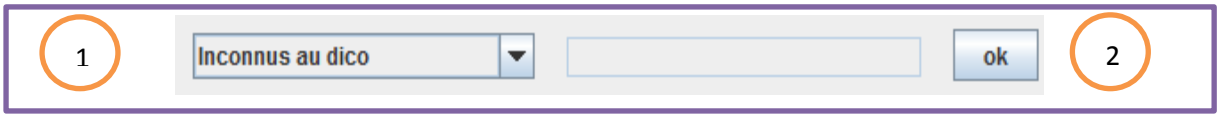


Illustration 19 : Dialogue pour rechercher les formes inconnues au dictionnaire

viii. Mode « Corriger à partir d'une liste »

L'expérience démontre que certaines erreurs d'OCR sont régulières. Presto vous propose de vérifier un certain nombre de formes qui constituent habituellement des erreurs d'OCR dans des données en langue française, qui ont été répertoriées dans un fichier nommé « erreursOCR », que vous trouverez à la racine du dossier PRESTO.

Ce fichier se charge à partir du mode de correction « Corriger à partir d'une liste ». (Illustration20)

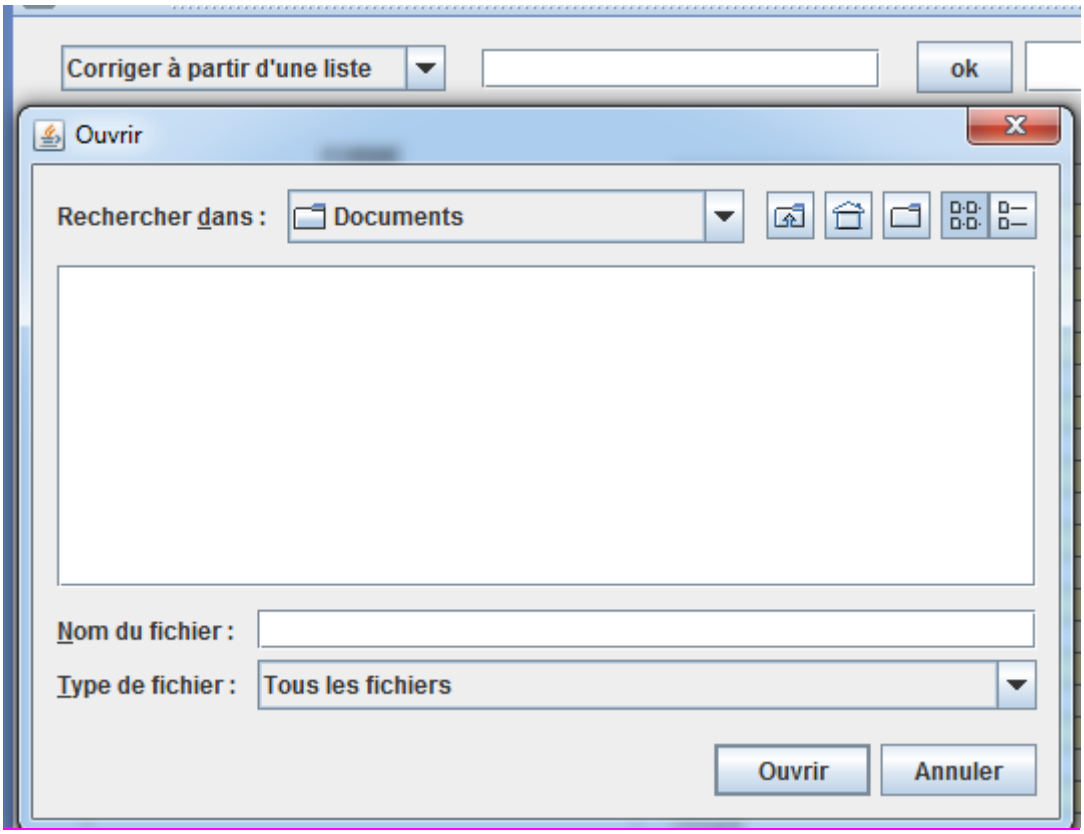


Illustration 20 : Dialogue pour rechercher un fichier dans lequel se trouve une liste de proposition de correction

Le logiciel recherche ensuite les formes à contrôler et procède à un affichage légèrement modifié par rapport aux cas précédents (Illustration 21).

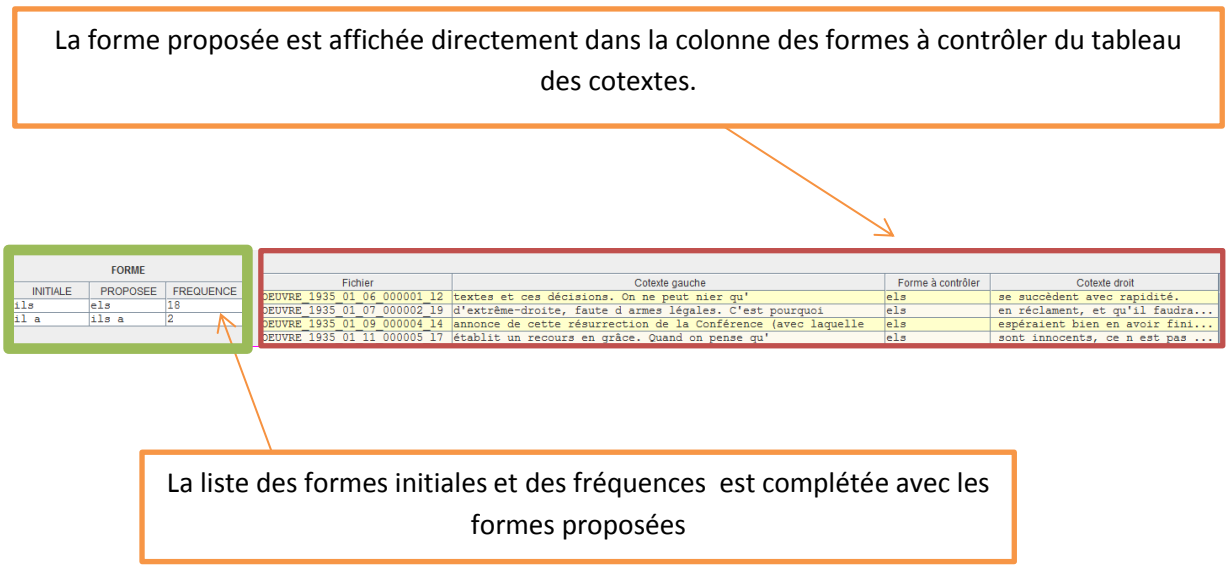


Illustration 21 : Formes initiales et proposées

La liste fournie par Presto n'est pas exhaustive, et vous pouvez la compléter pour l'adapter aux spécificités de votre corpus OU constituer votre propre liste de formes à contrôler.

Pour constituer des listes de formes erronées et de formes correctes dans des fichiers textes la norme suivante est à respecter : inscrire sur chaque ligne la forme initiale et la forme proposée séparées par une tabulation avec un encodage UTF-8. (Illustration 22)

î'	l'
1'	l'
1'	l'
!a	la
la	la
là	la
]a	la
Lé	Le

Illustration 22 : Exemple de liste de correction

Effectuer la correction comme précédemment au paragraphe 6.b.i.

ix. Mode « Postit »

Cet item permet de rappeler les formes à contrôler que l'utilisateur a enregistrées lorsqu'une forme à contrôler déborde sur les cotextes. Ceci est expliqué dans le paragraphe 6.b.i dans le 4° cas particuliers. L'affichage est le même qu'au cas précédent car nous disposons à la fois d'une forme initiale et d'une proposée.

x. Mode « Correction des mots coupés »

L'application recherche dans toutes les lignes se terminant par un tiret, la dernière UTA.
Elle la concatène avec la 1° UTA de la ligne suivante pour former un mot qu'elle tentera d'identifier dans notre dictionnaire. En cas d'échec elle essaie de trouver un mot avec le tiret. Si le mot ainsi formé existe elle fera la correction automatiquement**. C'est le seul cas où Presto n'offre pas à l'utilisateur la possibilité de contrôler les corrections.**

7. Utiliser les Curseurs

A chaque fois que vous validez un tableau corrigé, l'application mémorise les informations de la dernière ligne du tableau que vous venez de contrôler.

Le rappel se fait à partir de la liste des modes de recherche (Illustration 23)

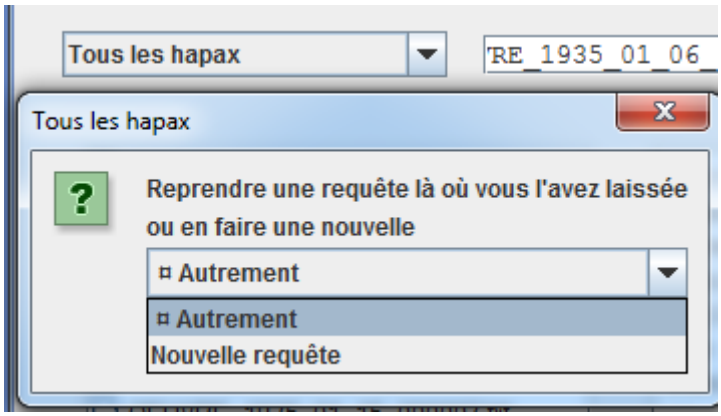


Illustration 23 : Dialogue pour reprendre une requête là où on l'a laissée

Ainsi ce menu déroulant vous rappelle que le dernier hapax relu est *Autrement* . En validant ce choix le contrôle reprendra au premier hapax suivant *Autrement*.
En sélectionnant *Nouvelle requête* PRESTO revient au premier hapax non corrigé.

8. Erreur de correction

Une erreur de correction C se récupère lors de la « correction fine ».
Sélectionner dans le texte la forme C. Le logiciel recherchera toutes les UTA qui ont subi la même correction pour les afficher le tableau de relecture habituel, tout en autorisant une nouvelle saisie et en proposant la forme initiale.

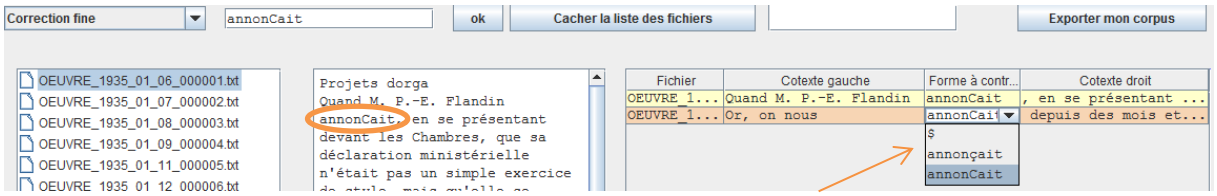


Illustration 24 : Rectification d'une erreur de correction

9. Exporter les données corrigées

Presto propose d'exporter le corpus corrigé en suivant les normes d'un logiciel bien précis, avec un encodage UTF-8 ou cp1252 dans le répertoire de votre choix. (Illustration 25)

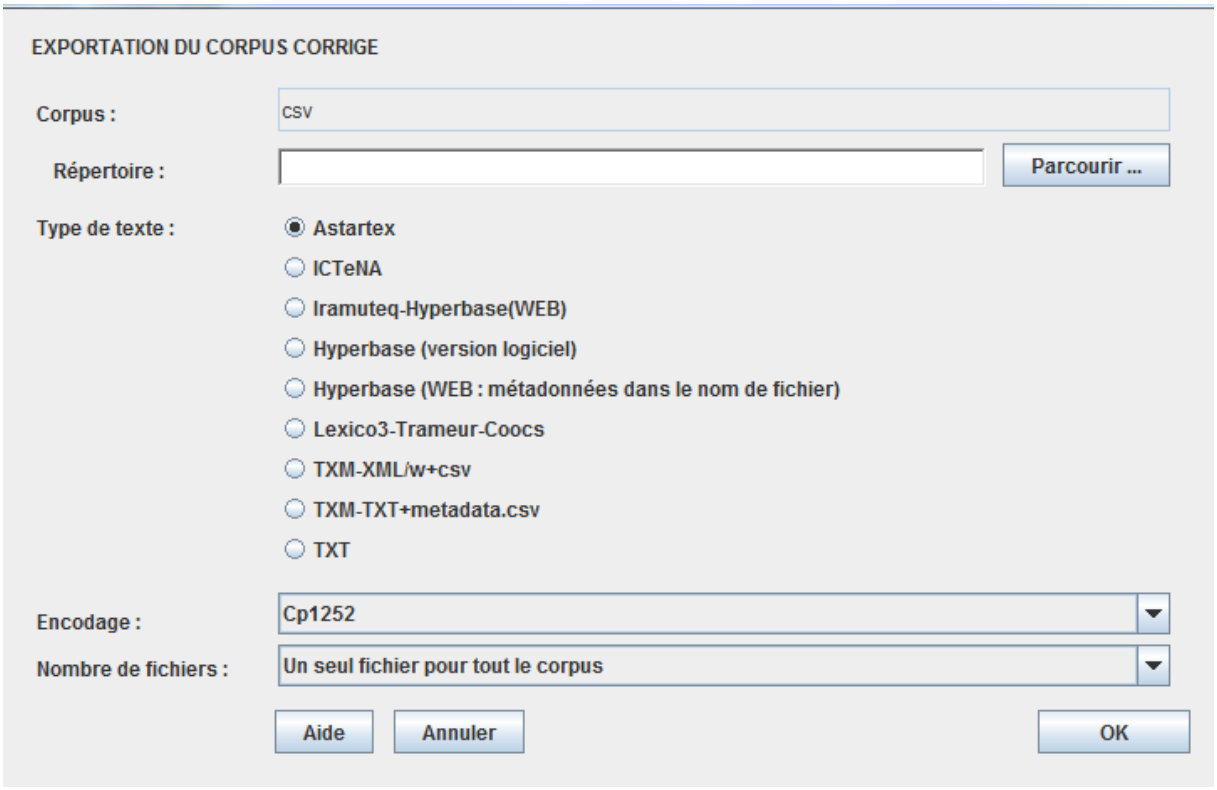


Illustration 25 : Dialogue pour exporter un corpus fini

L'organisation des fichiers et des répertoires du corpus corrigé se fait de trois façons différentes :
- *Un seul fichier pour tout le corpus* : tout le corpus exporté sera réuni dans un seul fichier
- *Un fichier par article dans un seul répertoire* : le corpus sera exporté dans un seul répertoire et chaque article sera enregistré dans un fichier
- *Un fichier par article_arborescence des répertoires initiaux reproduite* : Presto placera le corpus final dans un ou plusieurs répertoires en respectant l'organisation du corpus initial. Par contre il y aura toujours un seul article par fichier.

10. Supprimer/copier/importer un projet

L'illustration 26 présente les sous-menus du menu PROJET.

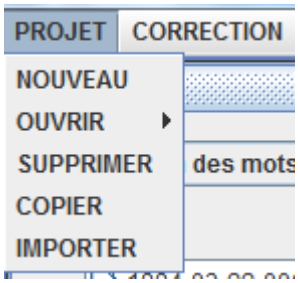


Illustration 26 : Menu projet

La fonctionnalité SUPPRIMER effacera toutes les données du corpus en cours de correction.

COPIER fera une sauvegarde complète des données du corpus en cours de correction. Elle ne permet

pas de générer un corpus corrigé.

Réciproquement, IMPORTER réintroduira un corpus précédemment copié dans l'application