Manuel de TXM

Version 0.7.8

Août 2017

Dans l'esprit du logiciel open-source, ce manuel vous est fourni gracieusement en échange de retours de corrections et de propositions d'amélioration au bénéfice de la communauté des utilisateurs de TXM. **Nous avons besoin de vous pour améliorer ce manuel**.

Voir la FAQ de TXM pour savoir comment contribuer : https://groupes.renater.fr/wiki/txm-users/public/faq.

Table des mises à jour (lignes plus anciennes à la fin du document)

16/05/2015	SH	Updated macro, import, Groovy and R sections
17/06/2015	SJ	Màj pour TXM 0.7.7 : graphiques
27/07/2015	SH	Finalisation pour livraison 0.7.7, ajouts MD&SJ+màj AL+réglages
09/08/2017	SH	Début mise à jour pour livraison 0.7.8 et traduction EN, ajouts MD, màj section macros

N° d'édition : 1096

Contenu : 239 pp., 54859 occ., 140 ill., 41 tab.

Date d'édition : 30/11/17, 10:52:23

Sommaire

1 Préface

1.1 Pourquoi lire ce manuel ?

Si vous souhaitez utiliser la plateforme TXM, ce manuel vous expliquera, pas à pas, les différents concepts du logiciel ainsi que ses divers outils d'analyse de corpus textuels.

Le logiciel TXM vous aide à construire et à analyser des corpus annotés et structurés :

- Le logiciel vous permet d'importer des ressources textuelles afin de construire un corpus provenant de diverses sources, ou directement, depuis n'importe quel texte copié dans le presse-papier.
- Il construit des sous-corpus à partir de différentes propriétés des unités textuelles
- Il construit des partitions à partir de ces propriétés
- Il construit une édition HTML pour chaque unité textuelle du corpus
- Il calcule le vocabulaire d'ensemble d'un corpus ou la liste des valeurs d'une propriété particulière
- Il construit des tables lexicales à partir de partitions ou d'index.
- Il recherche des motifs lexicaux complexes construits à partir des propriétés des unités lexicales et produit des concordances kwic à partir des résultats. Depuis chaque ligne de concordance, vous pouvez accéder à la page correspondante dans l'édition HTML
- Il calcule des cooccurrents situés autour d'un motif lexical complexe.
- Il calcule les références de motifs lexicaux complexes
- Il calcule le modèle des spécificités de mots ou d'étiquettes situés à l'intérieur d'une partition ou d'un sous-corpus
- Il calcule l'analyse factorielle des correspondances de propriétés des mots sur une partition.
- Il calcule la classification d'une partition

Le logiciel est composé de quatre modules en synergie :

- le moteur de recherche plein texte CQP ;
- le moteur statistique R ;
- le module d'importation de textes pour construire les corpus ;

- l'interpréteur de scripts.

Ce manuel vous présentera chacun de ces modules au travers des diverses commandes disponibles.

1.2 Comment est organisé ce manuel ?

Le manuel décrit dans un premier temps comment installer le logiciel selon les différents systèmes d'exploitation et comment lancer TXM. Il expose ensuite les éléments de l'interface utilisateur et la manière d'importer de nouveaux corpus dans la plateforme. La section qui vient par la suite présente les divers outils et la façon de les utiliser pour analyser un corpus. Enfin, il vous sera introduit le moyen de piloter la plateforme grâce aux scripts. Le document se clôt sur un glossaire et un index de référence.

1.3 Documentation complémentaire

Pour toute information complémentaire à ce manuel, nous vous invitons tout d'abord à consulter le wiki des utilisateurs de TXM qui est la source d'informations les plus récentes et à jour à l'adresse https://listes.cru.fr/wiki/txm-users/index

1.3.1 Le wiki des utilisateurs de TXM

Le wiki est actuellement structuré de la façon suivante :

- F.A.Q : donne des réponses aux questions les plus fréquemment posées sous une forme plus pragmatique - ou à jour - que le manuel de référence ;
- retours de bugs de la version poste de TXM : cette page agrège les retours de bugs des différentes réunions et mails de retours que nous recevons ;
- retours de bugs de la version portail de TXM
- demande de fonctionnalités : recensement des diverses demandes de nouvelles fonctionnalités pour TXM ;
- le wiki vous permet également de participer à l'amélioration de la documentation (y compris ce manuel) ou à sa traduction.

1.3.2 La liste de diffusion des utilisateurs de TXM

Nous vous invitons également à vous inscrire à la liste de diffusion francophone des utilisateurs de TXM à l'adresse https://listes.cru.fr/sympa/subscribe/txm-users.

Non seulement cette inscription sera nécessaire pour pouvoir participer au wiki indiqué précédemment, mais cela vous permettra surtout de dialoguer directement avec des utilisateurs de TXM (expérimentés ou non) ainsi qu'avec les concepteurs du logiciel, tout en recevant régulièrement les informations concernant les nouvelles versions du logiciel.

Nous vous invitons à consulter les archives de cette liste de diffusion pour rechercher si un sujet en rapport avec vos questionnements n'aurait pas déjà été abordé par le passé auquel cas certains éléments de réponse seraient peut-être déjà disponibles. Le moteur de recherche plein texte dans le sujet et le corps des articles est très pratique pour cela. L'archive se trouve à l'adresse : https://listes.cru.fr/sympa/archives/txm-users

1.3.3 Le site web du projet Textométrie

Vous trouverez également sur le site du projet Textométrie tous les documents officiels ayant trait à la plateforme TXM : http://textometrie.ens-lyon.fr/spip.php?article98&lang=fr (un tutoriel vidéo, les manuels disponibles pour les utilisateurs et les développeurs, les documents fondamentaux de la méthodologie textométrique, les documentations sur l'encodage des textes, sur les moteurs de recherches et statistiques et sur les interpréteurs de scripts) ainsi que la liste des publications scientifiques liées au développement et à l'utilisation de la plateforme : http://textometrie.ens-lyon.fr/spip.php?article82&lang=fr

1.3.4 Le site web des développeurs du logiciel TXM

Enfin, les personnes intéressées par le développement du logiciel open-source TXM lui-même peuvent consulter le wiki francophone des développeurs à l'adresse https://groupes.renater.fr/wiki/txm-info

Pour pouvoir participer à l'édition de ce wiki, vous devez :

1) soit disposer d'un compte d'une institution enregistrée auprès de Renater, soit vous créer un compte CRU de Renater : https://cru.renater.fr/sac/faces/casRedirect.jsp?target=create
2) vous inscrire à la liste de diffusion des développeurs « txm-info » avec l'adresse mail de votre compte : https://groupes.renater.fr/sympa/subscribe/txm-info

Nous les invitons également à s'inscrire à la liste de diffusion (anglophone) des développeurs de TXM à l'adresse https://lists.sourceforge.net/lists/listinfo/txm-open

TXM vous est fourni gracieusement. En contre-partie, dans l'esprit du logiciel open-source, vous êtes invité à participer à son adaptation ou à son amélioration. Pour cela, vous n'êtes pas obligé d'être développeur informatique. Par exemple, vous pouvez :

- nous transmettre vos publications ou autres documents de travail montrant votre usage de TXM
- ajouter un lien depuis votre site web vers la page d'accueil du projet Textométrie : http://textometrie.ens-lyon.fr
- nous faire part des dysfonctionnements que vous pourriez constater dans TXM, de préférence directement dans le wiki des utilisateurs : https://groupes.renater.fr/wiki/txm-users/public/retours_de_bugs_logiciel. Vous pouvez bien sûr toujours utiliser la liste de diffusion txm-users pour vos retours : https://groupes.renater.fr/sympa//info/txm-users
- nous aider à traduire son interface utilisateur ou sa documentation dans d'autres langues
- partager avec la communauté des utilisateurs vos supports de cours, ainsi que des corpus exemples
- inviter vos collègues développeurs à adapter TXM à vos propres besoins et partager avec nous ces évolutions. Soit en son coeur (langages Java et C pour développeurs professionnels), soit à sa périphérie (langage de script Groovy beaucoup plus accessible)
- monter un projet mobilisant TXM (e.g. ANR) pour lequel nous pouvons vous conseiller dans la mise en oeuvre ou dans l'adaptation du logiciel
- nous faire des propositions pour améliorer la documentation et sa diffusion

1.3.5 Les plaquettes de présentation de TXM

- La plaquette en français : http://sourceforge.net/projects/txm/files/documentation/TXM%20leaflet%20FR.pdf/download
- La plaquette en anglais : http://sourceforge.net/projects/txm/files/documentation/TXM%20Leaftlet%20EN.pdf/download
- la fiche PLUME : https://www.projet-plume.org/relier/txm
- la page TXM du wiki du consortium TEI (en anglais) : http://wiki.tei-c.org/index.php/TXM

1.3.6 TXM dans les réseaux sociaux

- Twitter : https://twitter.com/txm_
- Facebook : https://www.facebook.com/groups/148388185204997
- Le canal IRC #txm sur le serveur irc.freenode.net

- ou directement sur webchat : http://webchat.freenode.net/?channels=txm

1.3.7 Les Ateliers de formation TXM

Les Ateliers TXM sont des journées de formation ouvertes à tous et gratuites : https://groupes.renater.fr/wiki/txm-users/public/ateliers_txm

1.4 Accéder à la documentation en ligne

Ce manuel est accessible en ligne :

- au format HTML : http://txm.sourceforge.net/doc/manual/manual.xhtml
- sous forme d'un corpus TXM : https://sourceforge.net/projects/txm/files/corpora/refman

Cette version ainsi que sa traduction sont également disponibles à l'adresse :

http://sourceforge.net/projects/txm/files/documentation

1.5 Conventions typographiques

Dans ce manuel, certains éléments sont mis en valeur par une typographie différente :

- les expressions littérales sont en police `Courier` : les chemins d'accès aux dossiers, les noms de fichiers, les exemples de requêtes et de chaîne de caractères ainsi que les liens hypertextes.
- la police **Arial** est réservée aux titres de sections
- la police **Arial** est réservée aux commandes de l'application

2 Installation

2.1 Installer TXM sur sa machine

2.1.1 Prérequis d'installation

Vous aurez besoin des droits d'installation sur votre machine pour pouvoir installer TXM.

Vous aurez besoin d'un accès à Internet.

Cette version du logiciel est compatible avec les systèmes d'exploitation suivants :

- Windows 7 : voir les instructions d'installation à la section 2.1.2 ci-dessous ;
- Mac OS X 10.12 (Sierra) : voir les instructions d'installation à la section 2.1.3 page 13 ;
- Linux Ubuntu 16.04 et supérieur et ses variantes (Xubuntu, Kubuntu, Lubuntu) : voir les instructions d'installation à la section 2.1.4 page 15 .

Autres versions de systèmes pour lesquels TXM fonctionne également.

Espace disque nécessaire pour l'installation de TXM :

- Windows : 250 Mo
- Mac OS X : 335 Mo
- Linux : 200 Mo
- Ajouter 120 Mo pour les corpus exemples DISCOURS et GRAAL quel que soit le système

Ressources nécessaires pour l'usage de TXM :

- disque : prévoir de l'espace pour chaque corpus supplémentaire, exemples :
  - DISCOURS : 50 Mo (100 000 mots, 4 propriétés de mots)
  - Base de français médiéval (BFM) : 1.4 Go pour 5 M mots, 14 propriétés de mots et 60 propriétés de structures
  - etc.
- mémoire : 1 Go (voir la section 2.7 page 32 pour visualiser l'utilisation mémoire courante)

2.1.2 Installation sur Windows

2.1.2.1 Avertissement avant installation (Windows 7 et 8)

L'installeur de TXM n'est pas certifié par un organisme reconnu de Microsoft ce qui provoque l'affichage d'une fenêtre d'avertissement en plein écran. Pour continuer l'installation, il suffit de sélectionner « OK » puis « Exécuter quand même ».

En cas de doute sur l'installeur de TXM téléchargé, le seul site web diffusant TXM officiellement est celui du projet TXM de sourceforge :

https://sourceforge.net/projects/txm/files/software/0.7.7

2.1.2.2 Exécution de l'installeur

- Télécharger le fichier d'installation « `TXM_0.7.7_WinXX.exe` » à l'adresse ci-dessous, XX étant l'architecture de votre machine (32 = 32-bit ou 64 = 64-bit)[1] : https://sourceforge.net/projects/txm/files/software/0.7.7
- Exécuter le fichier d'installation par double-clic sur son icone :

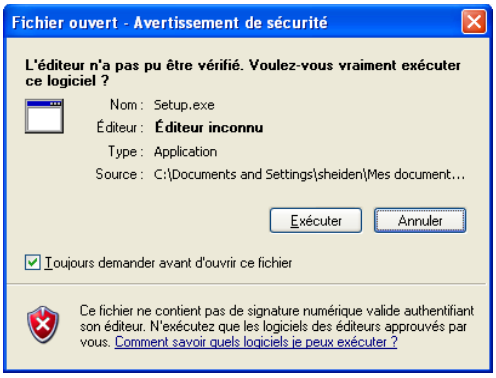


Illustration 2.1: Avertissement de sécurité

  - En fonction du niveau de sécurité de votre version de Windows, la fenêtre 2.1 peut apparaître. Si tel est la cas, veuillez cliquer sur le bouton « Exécuter » .

[1] Pour connaître l'architecture 32 ou 64 bits de votre système Windows : voir la documentation Microsoft.

- Dans la fenêtre 2.2 cliquer sur « `Install` » (si nécessaire, choisir un autre dossier d'installation au préalable).

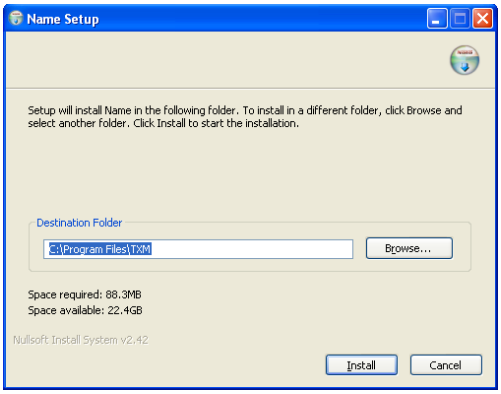


Illustration 2.2: dossier d'installation

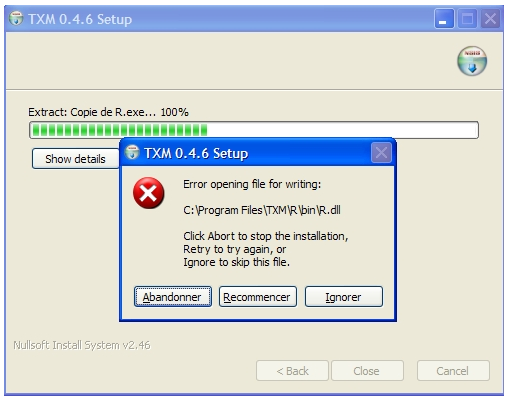


Illustration 2.3: Erreur d'accès au fichier "R.dll"

- L'installation dure environ une minute.

- Si, pendant l'installation, le message suivant apparaît (Illustration 2.3: Erreur d'accès au fichier "R.dll") : Cela signifie que le processus 'Rserve' (le moteur statistique de TXM) est toujours en cours d'exécution sur votre ordinateur et que l'installation ne peut pas mettre à jour ses fichiers binaires. Vous devez d'abord quitter le TXM en cours d'exécution ou terminer le processus Rserve depuis le gestionnaire de tâches de Windows et ensuite cliquer sur « `Recommencer` » pour reprendre l'installation.

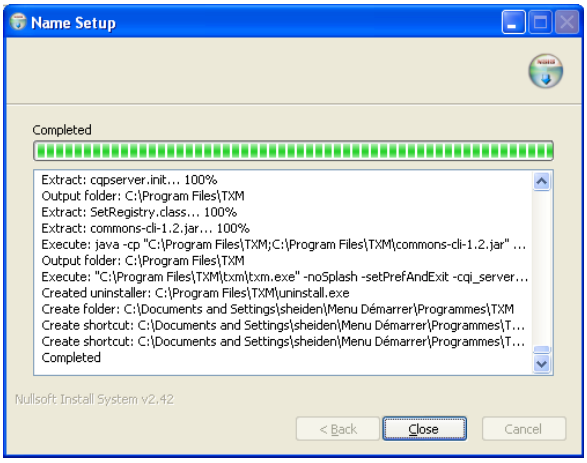


Illustration 2.4: Fin de l'installation

- Si l'écran 2.4 s'affiche, tout s'est bien déroulé et l'installation est terminée. Vous pouvez alors clore la fenêtre de l'installeur en cliquant sur « `Close` ».

Si l'installation n'a pas abouti, veuillez suivre les instructions de retour de problème de la section 2.8 page 32.

2.1.2.3 Premier lancement de TXM

Vous devez lancer TXM une première fois pour finaliser le processus d'installation, voir la section 1 page 34.

Remarque : les mises à jour automatiques et les extensions ne sont disponibles qu'à partir du deuxième lancement de TXM, voir la section 2.3 page 23.

2.1.3 Installation sur Mac OS X

2.1.3.1 Étape 1 : pré-requis

TXM fonctionne sur Mac OS X à partir de la version 10.6 (Snow Leopard)[2].

Particularités des différentes versions :

- à partir de Mac OS X 10.7 (Lion) : il faut modifier les paramètres de “sécurité” pour pouvoir installer correctement TXM (si on ne fait pas cette modification, le problème ne se manifeste pas au moment de l'installation, mais au premier lancement). Pour ce faire, il faut aller dans “Préférences systèmes > Personnel > Sécurité” et autoriser les applications téléchargées depuis Internet ;
- à partir de Mac OS X 10.9 (Maverick) : il faut installer à préalable l'application XQuartz. Depuis la page https://xquartz.macosforge.org/landing, télécharger et installer l'application « XQuartz-x.x.x.dmg ».

[2] Pour connaître la version du système de son Mac, il faut afficher les Informations Système à partir du menu Pomme > "A propos de ce Mac".

2.1.3.2 Étape 2 : Exécution de l'installeur

Procédure d'installation pour tous les Mac OS X :

- Télécharger le fichier d'installation « `TXM_0.7.7_MacOSX.pkg` » à l'adresse : https://sourceforge.net/projects/txm/files/software/0.7.7
- En étant connecté à Internet[3], double-cliquer sur l'icone du fichier « `TXM_0.7.7_MacOSX.pkg` ». Cela lance l'installateur de logiciels Mac qui se déroule sur plusieurs écrans :

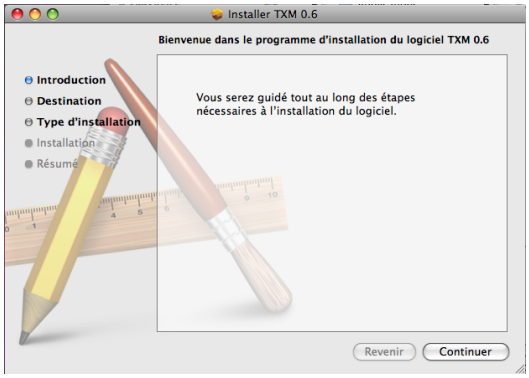


Illustration 2.5: Accueil

Accueil

Au niveau de la page d'accueil (ill. 2.5), vous pouvez cliquer sur « Continuer ».

Disque d'installation

Si nécessaire (ill. 2.6), choisir le disque sur lequel vous souhaitez installer TXM (par défaut, le disque « Macintosh HD » sera utilisé) et cliquer sur « Continuer ».

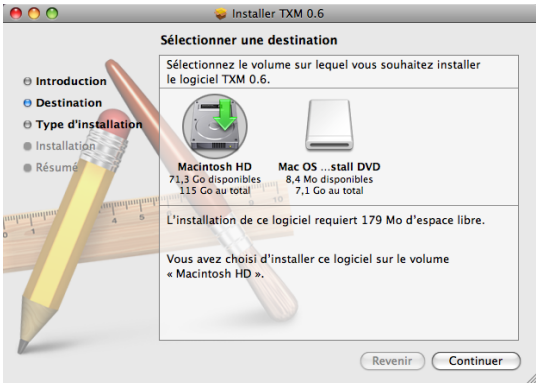


Illustration 2.6: Disque d'installation

Authentification

Pour installer TXM (ill. 2.8), il faut disposer des droits d'administration de votre Mac OS X à l'aide de votre identifiant (login) et de votre mot de passe.

Installation

TXM va maintenant copier ses fichiers dans son dossier d'installation et préparer son environnement de travail (ill. Erreur : source de la référence non trouvée).

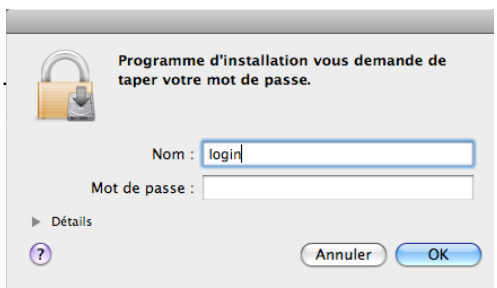


Illustration 2.8: Authentification

rgement de

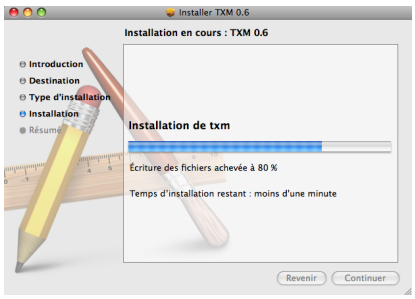


Illustration 2.7: Installation

Installation des librairies statistiques R

Par défaut, TXM installe automatiquement les librairies statistiques R dont il a besoin pour les calculs de spécificités, AFC, cooccurrences et progression.

Pour les utilisateurs avancés : il est possible d'installer soi-même les librairies statistiques R en modifiant les options d'installation lors de l'étape .

http://txm.sourceforge.net/wiki/index.php/Build_the_toolbox_or_the_application#Install_R_yourself

Fin de l'installation

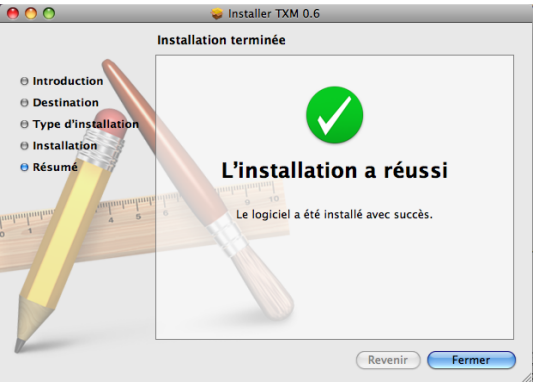


Illustration 2.9: Fin de l'installation

Si l'écran 2.9 s'affiche, tout s'est bien déroulé et l'installation est terminée. Cliquer sur « Fermer ».

Si l'installation n'a pas abouti, veuillez suivre les instructions de retour de problème à la section 2.8 page 32.

Vous pouvez obtenir le journal d'installation complet à l'aide du raccourcis « Pomme + L » avant de quitter l'installateur. Ce raccourcis affiche le journal d'installation que vous pouvez ensuite copier-coller.

2.1.3.3 Premier lancement de TXM

Vous devez lancer TXM une première fois pour finaliser le processus d'installation, voir la section 1 page 34.

Remarque : les mises à jour automatiques et les extensions ne sont disponibles qu'à partir du deuxième lancement de TXM, voir la section 2.3 page 23.

2.1.4 Installation sur Linux Ubuntu

Pour installer TXM sur Linux Ubuntu (version 12.04 et supérieur) :

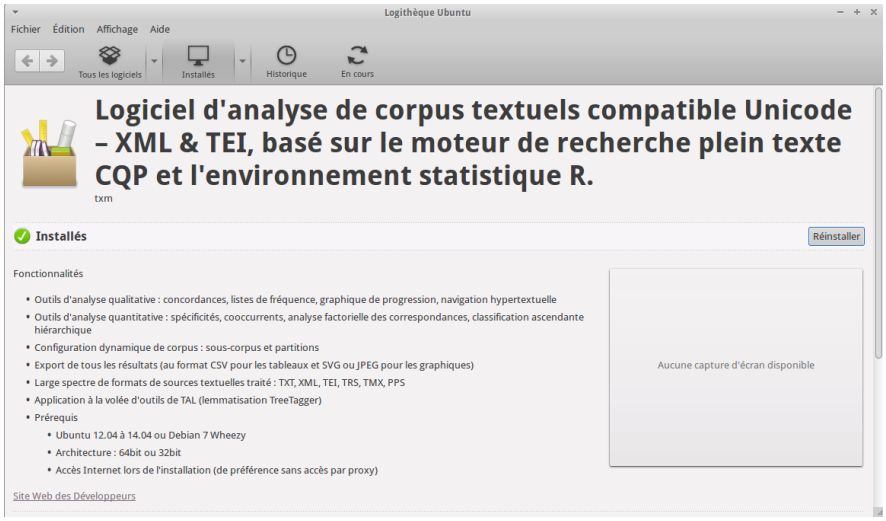
- Télécharger le fichier d'installation « `TXM_0.7.7_LinuxXX.deb` » à l'adresse ci-dessous, XX étant l'architecture de votre machine ((32-bit ou 64-bit)[4] : `https://sourceforge.net/projects/txm/files/software/0.7.7`
- Puis installer avec la logithèque ou Gdebi ou par ligne de commande.

[4] Pour connaître l'architecture 32 ou 64 bits de votre système Linux : Sous Ubuntu (testé avec un Ubuntu 12.04 LTS), dans le menu tout en haut à droite (petite roue dentée), Paramètres système… > Système : Détails > Type d'OS (= 32 ou 64 bits).

2.1.4.1 Installation avec la logithèque Ubuntu

Ouverture de TXM_0.7.7_LinuxXX.deb

Pour démarrer la logithèque avec l'installeur Linux, un double-clic sur le fichier .deb suffit. Sinon, faire un clic droit sur l'icone de l'installeur, puis sélectionner « Ouvrir avec une autre application... », dans la liste des applications affichée, sélectionner « Logithèque Ubuntu ».



Démarrage de l'installation

S'il s'agit d'une première installation de TXM, il faut cliquer sur le bouton « Installer », sinon sur le bouton « Mettre à jour » (ou « Réinstaller » s'il s'agit d'une réinstallation de la même version de TXM).

Étapes suivantes de l'installation

Les étapes suivantes correspondent aux sections et .

2.1.4.2 Installation avec Gdebi

Il s'agit d'une autre voie d'installation de TXM qui peut être utile en cas de panne de la logithèque.

Gdebi est par défaut installé dans les distributions Debian. S'il n'est pas installé pour pouvez l'installer via la logithèque ou via un Terminal avec la ligne de commande suivante :

```
sudo apt-get install gdebi
```

Ouverture de TXM_0.7.7_LinuxXX.deb

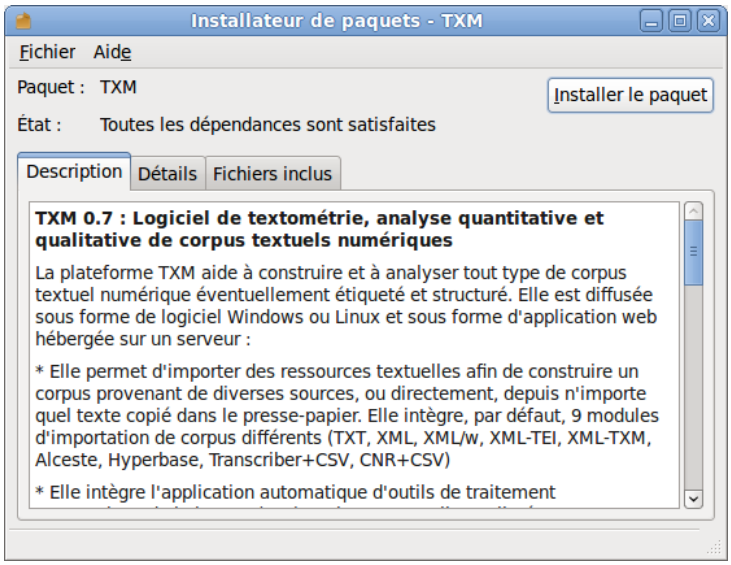


Illustration 2.10: Ouverture avec Gdebi

Une fois l'écran d'accueil ouvert, cliquer sur « Installer le paquet ».

Si la ligne du champ « état » est de couleur rouge, cela signifie que Gdebi n'a pas pu trouver les paquets nécessaires à l'installation de TXM. Dans ce cas, le plus simple est de nous contacter et de vérifier si votre version d'Ubuntu a bien été testée. Dans l'illustration 2.10, toutes les dépendances sont disponibles.

Étapes suivantes de l'installation

Les étapes suivantes correspondent aux sections et .

2.1.4.3 Installation par ligne de commande

Il s'agit d'une autre voie d'installation de TXM qui peut être utile en cas de difficulté avec les précédentes. Dans un Terminal, lancer la ligne de commande suivante :

```
sudo dpkg -i TXM_0.7.7_LinuxXX.deb
```

Acceptation de la licence

- La première étape de l'installation consiste à accepter les termes de l'accord de licence du logiciel :

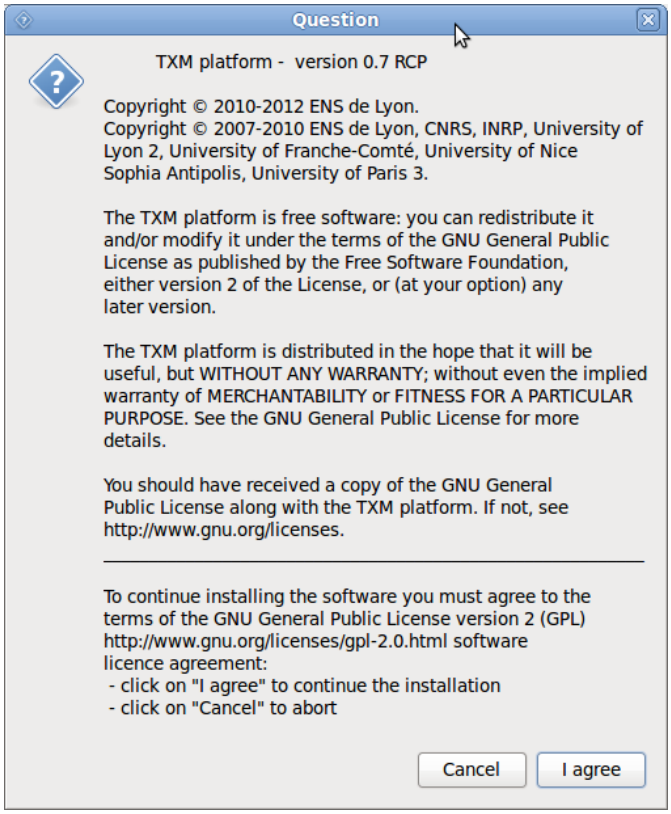


Illustration 2.11: Acceptation de la licence de TXM

- Cliquer sur le bouton « I agree » pour continuer.
- Attention : Cela a valeur d'acceptation des termes d'accord de la Licence publique générale GNU du logiciel.
    - Vous pouvez lire la Licence publique générale GNU complète à l'adresse : http://www.april.org/gnu/gpl_french.html ;
    - ainsi qu'une introduction à la notion de logiciel libre à : http://www.april.org/sites/default/files/documents/html/logiciel-libre.html

Installation des librairies statistiques R

- Après avoir validé la licence, TXM propose d'installer lui-même les librairies statistiques R :

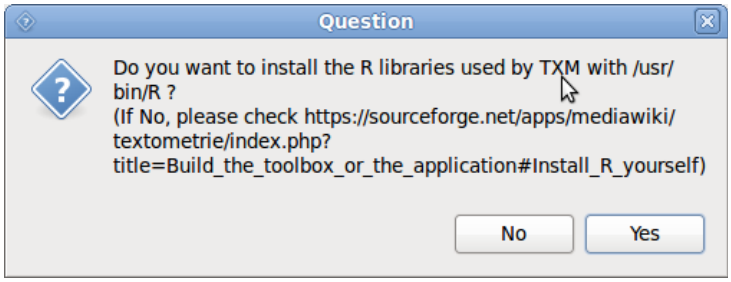


Illustration 2.12: TXM va installer les librairies statistiques R

- Pour une installation standard de TXM, cliquez sur « Yes ».

Pour les utilisateurs avancés, il est possible d'installer soi-même les librairies statistiques R. Pour plus d'informations sur la procédure, veuillez vous référer à la page http://sourceforge.net/apps/mediawiki/txm/index.php?title=Build_the_toolbox_or_the_application#Install_R_yourself

Progression de l'installation

Illustration 2.13: Fenêtre de progression de l'installation

Pendant l'installation, les dépendances suivantes seront installées :

- r-base
- r-recommended
- zenity
- default-jre
- libwebkitgtk-1.0-0
- debconf
- libc6 (>= 2.15)

Pour obtenir des détails sur l'avancement, vous pouvez cliquer sur « Terminal »

Fin de l'installation du package

Pour Gdebi, tout s'est bien déroulé et l'installation est terminée si l'écran 2.14 s'affiche. Vous pouvez alors fermer les fenêtres de Gdebi.

Illustration 2.14: Fin de l'installation

Si l'installation n'a pas abouti, veuillez suivre les instructions de retour de problème à la section 2.8 page 32.

2.1.4.4 Premier lancement de TXM

Vous devez lancer TXM une première fois pour finaliser le processus d'installation, voir la section 1 page 34.

Remarque : les mises à jour automatiques et les extensions ne sont disponibles qu'à partir du deuxième lancement de TXM, voir la section 2.3 page 23.

2.1.4.5 Reconnexion lors de la première installation

La première fois que vous installez TXM sur une machine, il est également nécessaire de quitter votre session de travail (déconnexion) puis de vous reconnecter.

2.1.5 Vérification de l'installation des packages R

TXM peut vérifier la bonne installation des packages R dont il a besoin pour les calculs statistiques. Pour cela, il faut utiliser la commande « Fichier > Vérifier les packages R » dans le menu principal.

La vérification se déroule de la façon suivante :

- TXM teste la présence de chaque package
- Si un package est présent, TXM vérifie la version du package et le met à jour si nécessaire

- Si le package est absent, TXM l'installe

Attention, pour les opérations d'installation ou de mise à jour de package, TXM a besoin d'être connecté à Internet.

2.2 Installer TreeTagger pour ajouter automatiquement des propriétés morphosyntaxiques et des lemmes aux mots

Pour être en mesure d'automatiser la lemmatisation et l'étiquetage morphosyntaxique de votre corpus lors de son importation dans TXM, ce tutoriel va vous guider pour :

1. Récupérer le logiciel TreeTagger et un ou plusieurs de ses modèles linguistiques parce que nous ne pouvons pas le livrer avec TXM[*] ;
2. Indiquer à TXM où se trouve votre TreeTagger et le modèle linguistique choisi.

2.2.1 À l'aide d'un navigateur et de votre explorateur de fichiers

En étant connecté à Internet :

1. Télécharger l'archive du logiciel TreeTagger correspondant à votre système d'exploitation à partir du site de TreeTagger :
   - Windows (32bit et 64bit)
   - Mac OS X
   - Linux 64bit
   - Linux 32bit
2. Extraire le contenu de l'archive compressée (*.zip) dans un dossier nommé '**treetagger**' :

   - Sous Windows **`C:\Programmes\treetagger`**

   - Sous Windows XP **`C:\Program Files\treetagger`**

   - Sous Mac OS X **`/Applications/treetagger`**

   - Sous Linux **`/usr/lib/treetagger`**

   Vérification : Une fois extrait, ce dossier doit contenir les dossiers et fichiers suivants : bin, cmd, doc, FILES, LICENSE et README.
3. Créer le sous-dossier '**models**' dans votre dossier 'treetagger' qui contiendra les modèles de langues de TreeTagger.
4. Télécharger le modèle (fichier compressé '*.gz') de chaque langue dont vous souhaitez une lemmatisation à partir du site de TreeTagger :

- français : french-par-linux-3.2-utf8.bin.gz (fr)
- anglais : english-par-linux-3.2-utf8.bin.gz (en)
- allemand : german-par-linux-3.2-utf8.bin.gz (de)
- italien : italian-par-linux-3.2-utf8.bin.gz (it)
- espagnol : spanish-par-linux-3.2-utf8.bin.gz (es)
- russe : russian-par-linux-3.2-utf8.bin.gz (ru)
- latin classique : latin-par-linux-3.2.bin.gz (la)
- ancien français du projet BFM (sans lemmes) : fro.zip (fro)
- autres langues : voir la liste de tous les modèles de langue TreeTagger disponibles (à la section 'Parameter files')

5. Décompresser chaque fichier compressé de modèle dans votre dossier 'models'.

   Sous Windows, si vous n'avez pas de logiciel extracteur-décompresseur compatible avec les fichiers '*.gz', nous vous recommandons le logiciel libre 7-zip.

6. Renommer chaque fichier de modèle en utilisant les codes de langues ISO 639-1 à deux lettres.
   Par exemple :
   - 'french.par' en 'fr.par' pour le fichier modèle français
   - 'english.par' en 'en.par' pour le fichier modèle anglais
   - etc.

   Sous Windows et Mac OS X : Par défaut, ces systèmes masquent à l'utilisateur les extensions de fichiers dont il gère le type. Dans ce cas, on peut se trouver dans une situation où l'on pense avoir renommé un fichier 'fr.bin' en 'fr.par' alors que le nom complet réel du fichier reste 'fr.par.bin'.
   Dans ce cas il faut accéder à l'affichage complet des noms de fichiers puis les renommer :
   - Sous Windows :
     1. Pour afficher les noms complets des fichiers avec leur extension, vous pouvez suivre ce tutoriel : Afficher-les-extensions-et-les-fichiers-caches-sous-windows
     2. Vous pouvez alors renommer le nom complet.
   - Sous Mac OS X :
     1. Faire un clic droit sur l'icone du fichier (Ctrl-clic avec la souris ou bien cliquer à deux doigts sur le trackpad)
     2. Lancer la commande 'Lire les informations'
     3. Éditer le champ 'nom et extension' : supprimer l'extension '.bin'.
     4. Fermer la fenêtre d'informations.

   Vérification : Le dossier 'models' doit contenir le fichier 'fr.par' qui fait environ 17 Mo, et éventuellement les fichiers d'autres modèles de langues ('en.par', 'de.par', etc.).

2.2.2 Dans TXM

7. Aller dans les préférences de réglage de TreeTagger (voir figure 1) :
    - Menu 'Outils / Préférences'
    - Aller à la page 'TXM / Avancé / TAL / TreeTagger'
    - Renseigner le champ 'Chemin du dossier d'installation de TreeTagger' : cliquer sur 'Parcourir...', puis sélectionner votre dossier 'treetagger' (voir étape 2.) et terminer par 'OK'

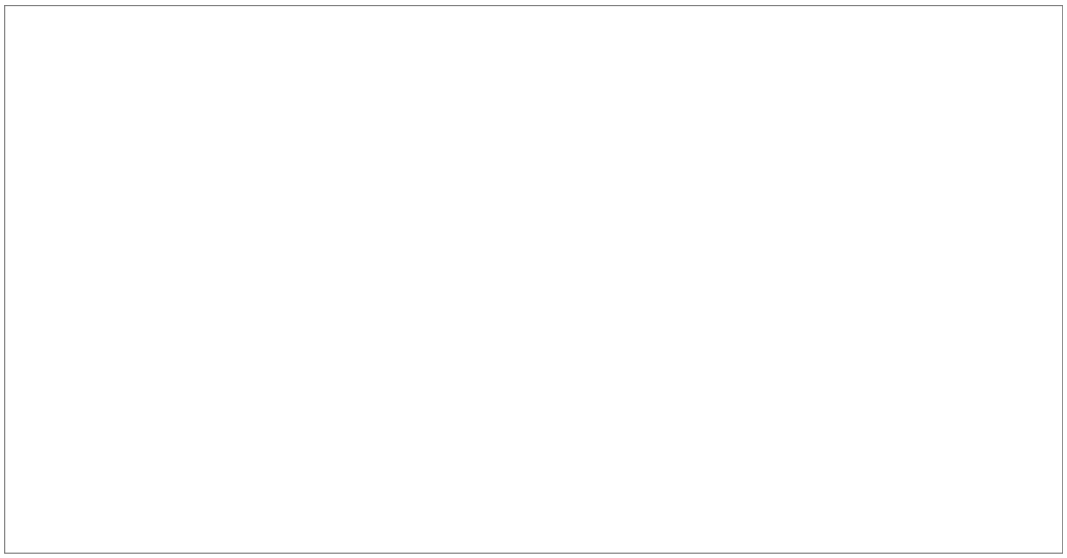


Illustration 2.15: Préférences TreeTagger

    - Renseigner le champ 'Chemin du dossier de modèles linguistiques de TreeTagger' : cliquer sur 'Parcourir...', puis sélectionner votre dossier 'models' et terminer par 'OK'
    - Terminer par 'OK' pour enregistrer ces réglages
8. Vérifier votre installation

En cas de problème, vous trouverez de l'aide supplémentaire dans la FAQ.

Si vous ne parvenez pas à aller jusqu'au bout de cette procédure d'installation, veuillez nous contacter via la liste de diffusion des utilisateurs de TXM (txm-users@cru.fr) après vous être inscrit à la liste de diffusion txm-users.

2.3 Mises à jour automatiques

À partir de sa version 0.7.5, TXM se met à jour automatiquement ce qui permet de ne plus avoir à télécharger de nouvel installateur TXM pour chaque version. Quand une mise à jour est disponible, un petit encadré s'affiche en bas à droite de la fenêtre principale de TXM et propose son téléchargement et son installation (voir section suivante).

2.3.1 Niveaux de mise à jour

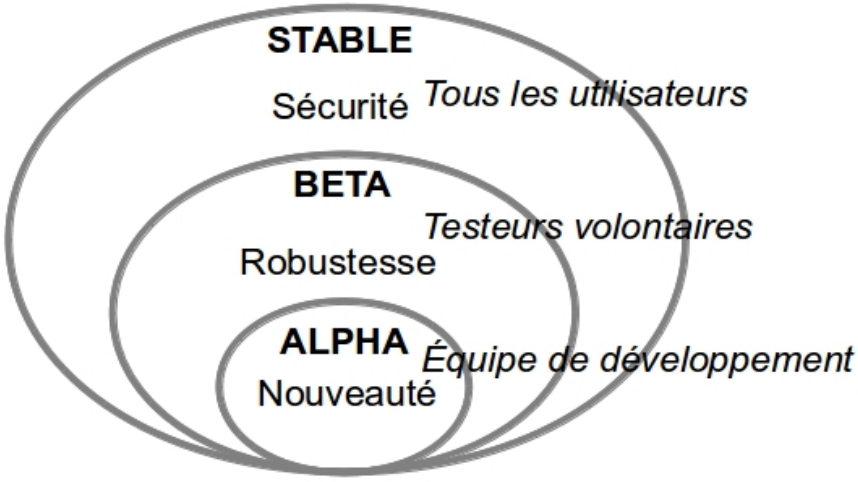


Illustration 2.16: Niveaux de mise à jour

Les mises à jour s'organisent selon quatre niveaux, du simple utilisateur de TXM à celui plus impliqué dans le développement de la plateforme :

- le niveau « STABLE » (par défaut) permet à tous les utilisateurs de bénéficier des améliorations progressives de la plateforme au fur et à mesure de leur disponibilité ;
- le niveau « BETA » permet aux personnes souhaitant tester les améliorations candidates au niveau « STABLE » mais encore non éprouvées dans divers environnements (différents systèmes d'exploitation et versions) et dans tous les cas de figure (compatibilité avec certains types de corpus, certains outils externes comme les étiqueteurs, etc.). Les personnes testant les améliorations de niveau « BETA » sont invitées à faire leurs retours à l'équipe de développement par mail ou dans le wiki txm-users : https://groupes.renater.fr/wiki/txm-users/public/retours_de_bugs_logiciel ;
- le niveau « ALPHA » permet à l'équipe de développement de TXM de tester des améliorations prototypes, candidates au niveau « BETA ». Ces améliorations correspondent à des ébauches. Elles ne font pas encore forcément consensus dans l'équipe de développement, leur interface utilisateur n'est pas encore développée, leur documentation est à peine ébauchée, leur usage peut provoquer des plantages de TXM et laisser la plateforme dans un état incohérent.
- Le niveau "DEV" permet aux développeurs de TXM de débugger rapidement des modifications de TXM. L'usage de ce niveau est fortement déconseillé car les mises à jour correspondantes ne sont pas forcément compatibles avec la version publique de TXM, ce qui peut provoquer des plantages et laisser TXM dans un état incohérent.

Pour choisir le niveau de mise à jour de son TXM, l'utilisateur règle la préférence « TXM / Avancé / Niveau de mise à jour » à la valeur souhaité (STABLE, BETA, ALPHA, DEV).

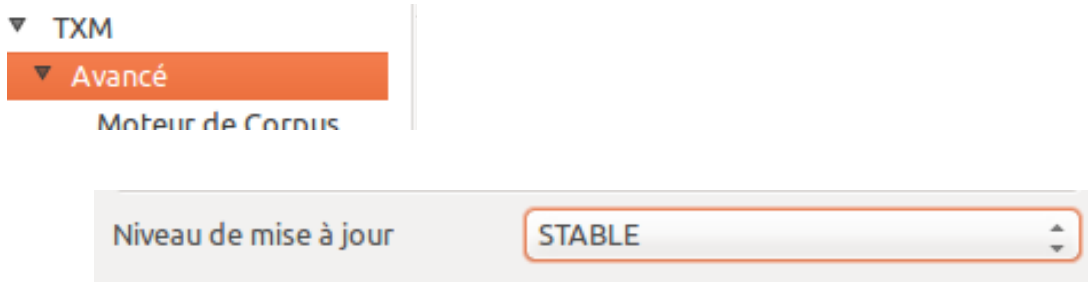


Illustration 2.17: Choix du niveau de mise à jour.

2.3.2 Lancer une mise à jour

Par défaut, les mises à jour ont lieu automatiquement au lancement de TXM. Pour les désactiver, il faut se rendre dans les préférences du logiciel à la page « TXM » et décocher l'option « Rechercher automatiquement les mises à jour et m'avertir ». Vous pouvez alors déclencher vous-même une mise à jour manuellement avec la commande « Aide / Vérifier les mises à jour ».

Au démarrage, si TXM est connecté à Internet et les mises à jour automatiques sont demandées, il vérifie si il y a une mise à jour disponible. Si une mise à jour est disponible , TXM affiche la fenêtre de mise à jour (illustration 2.18).

2.3.3 Effectuer une mise à jour

Le lancement d'une mise à jour affiche la liste des nouveaux composants disponibles.

2.3.3.1 Étape 1

Sélectionner les composants à mettre à jour puis passer à l'étape suivante en cliquant sur « Next ».

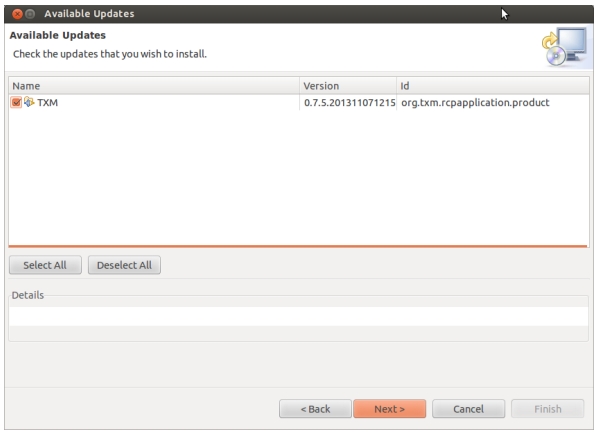


Illustration 2.18: Etape 1 : Mises à jour disponibles

2.3.3.2 Étape 2

L'étape suivante affiche une description plus précise des mises à jour que l'on peut ignorer dans la plus part des cas. Il suffit de faire « Next » pour passer à l'étape suivante.

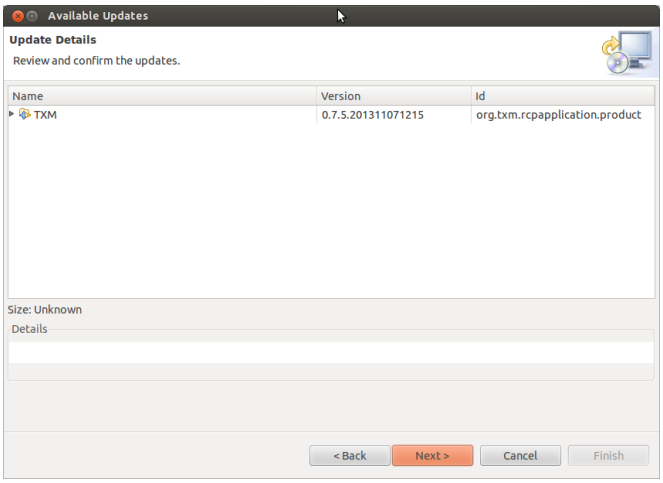


Illustration 2.19: Etape 2 : Détail des mises à jour

2.3.3.3 Étape 3

L'étape suivante consiste à accepter les licences de diffusion de chaque composant. Ces licences sont en général identiques à celle utilisée pour la diffusion de TXM – la licence GNU public licence V3. Cocher l'option « I accept the terms of the license agreements » puis sélectionner «Finish» pour lancer le téléchargement des mises à jour.

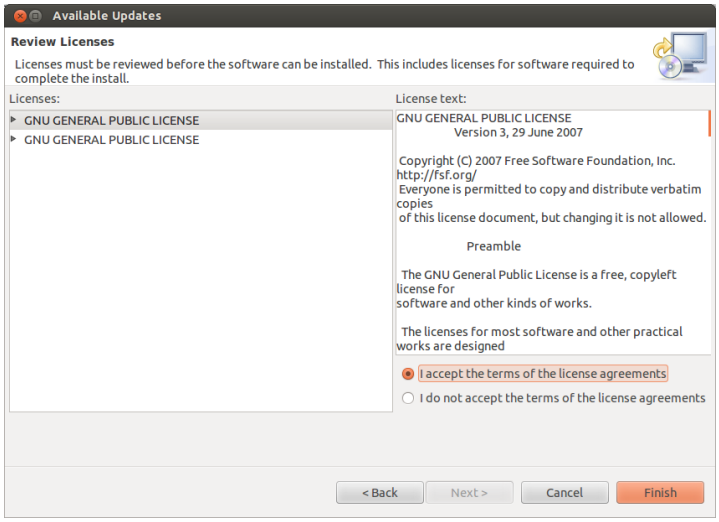


Illustration 2.20: Etape 3 : Acceptation des licences de diffusion

2.3.3.4 Étapes 4 à 6

Une nouvelle fenêtre s'ouvre indiquant la progression du téléchargement.

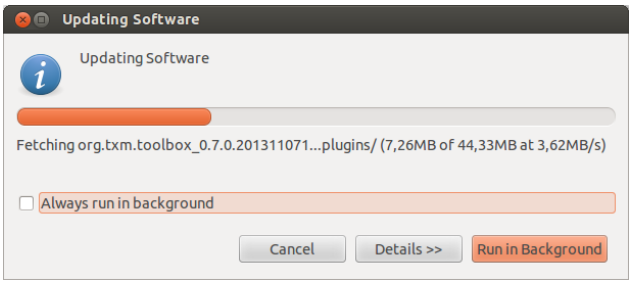


Illustration 2.21: Etape 4 : Téléchargement des mises à jour

Une fois le téléchargement terminé, TXM pourra éventuellement vous demander une dernière confirmation pour les mises à jour non signées avec un certificat d'authenticité garantissant leur origine du projet TXM (cas de toutes les mises à jour pour l'instant).

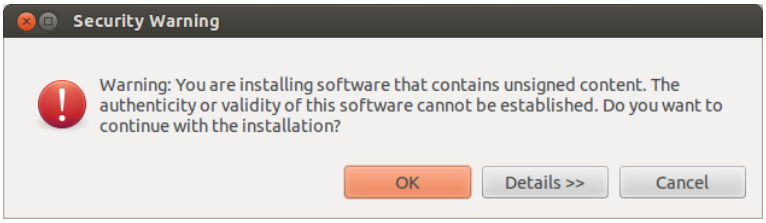


Illustration 2.22: Etape 5 : Dernière confirmation de sécurité avant installation

Enfin, TXM demandera à être relancé pour que la mise à jour soit effective. Cliquer alors sur le bouton « Yes ».

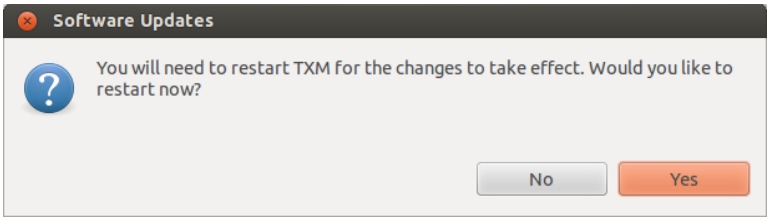


Illustration 2.23: Etape 6 : Relancer TXM pour appliquer les mises à jour

2.4 Installer une extension

À partir de sa version 0.7.5, il est possible d'installer dans TXM des extensions ajoutant de nouvelles fonctionnalités. Ces extensions sont proposées par le projet de développement de TXM comme fonctionnalités optionnelles ainsi que par nos partenaires.

Pour installer une extension dans TXM, utiliser la commande « Aide / Ajouter une extension ». Cela ouvre la fenêtre des extensions disponibles avec leur description. Sélectionner les extensions souhaitées puis cliquer sur « Next ».

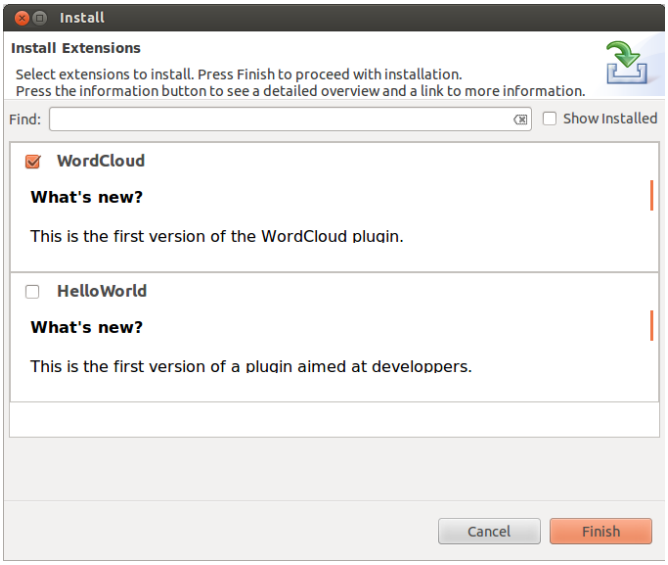


Illustration 2.24: Liste des extensions disponibles

Les étapes suivantes sont similaires aux étapes 2 à 6 des mises à jour de TXM, voir la section 2.3.3.

2.4.1 Documentation des extensions

Chaque extension ajoute une entrée au menu « Aide > Extensions » pour l'accès à sa documentation.

2.4.2 Installer une extension tierce dans TXM

À partir de sa version 0.7.5, il est possible d'installer dans TXM des extensions tierces (ou « plugins ») développés par des projets indépendants compatibles avec l'architecture Eclipse RCP de TXM[5].

Pour découvrir les extensions disponibles, on utilise des portails d'extensions publics comme celui du consortium Eclipse, l'Eclipse marketplace : http://marketplace.eclipse.org.

Quand on souhaite installer une extension, il faut que TXM connaisse l'adresse de son entrepôt de mise à jour. Cette adresse est fournie par les portails d'extensions ou bien par les sites de développement de ces extensions.

[5] Conforme au standard OSGi : http://en.wikipedia.org/wiki/OSGi

2.4.2.1 Étape 1

Pour lancer l'installation, on appelle la commande « Fichier / Ajouter une extension tierce ». Cette commande ouvre une boite de dialogue dans laquelle (voir illustration 5.1) :

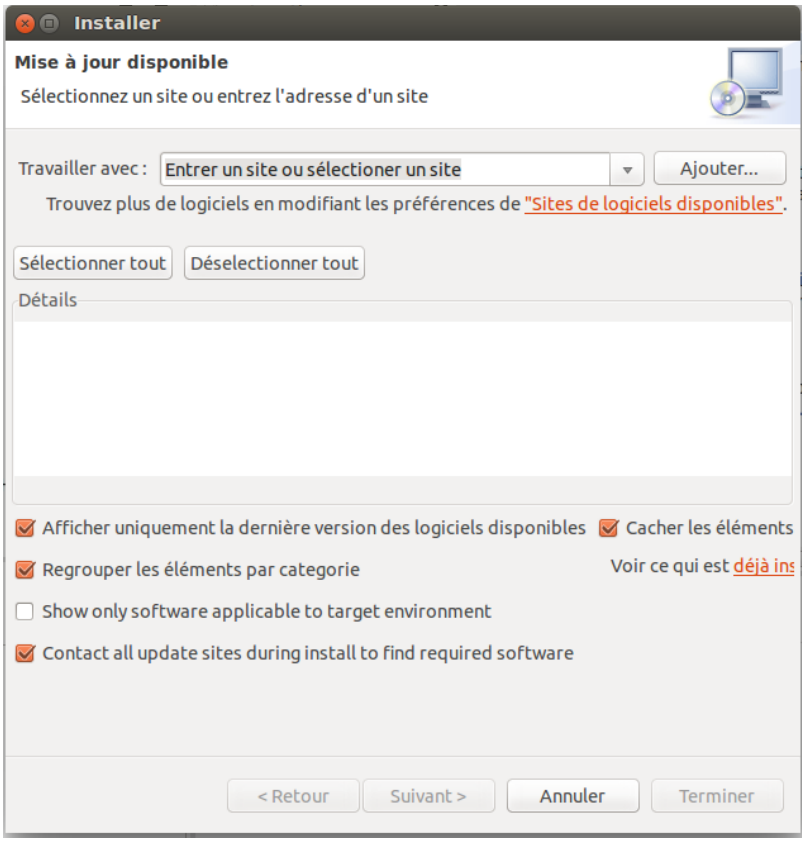


Illustration 2.25: installation d'extensions tierces

- la champ « Work with : » permet de saisir ou de coller l'adresse de l'entrepôt de mises à jour (update site) de l'extension ou bien de choisir une des adresses de la liste des entrepôts connus (en ouvrant le bouton situé sur la droite du champ « flèche vers le bas ») ;
- on peut filtrer les noms d'extensions disponibles dans l'entrepôt choisi en saisissant une partie du nom dans le champ suivant (par exemple « XML ») ;
- la liste des extensions s'affiche alors en dessous ;
- on sélectionne alors les extensions souhaitées et on clique sur « Next » pour passer à l'étape suivante.

2.4.2.2 Étapes suivantes

Les étapes suivantes sont similaires aux étapes 2 à 6 des mises à jour de TXM, voir la section 2.3.3.

Si lors de l'étape 2 on obtient le message « The operation cannot be completed. See the details », l'ajout de l'extension nécessite d'indiquer à TXM des entrepôts de mise à jour supplémentaires pour pouvoir télécharger d'autres extensions dont elle dépend. Ces extensions sont indiquées dans le champ « Details ». Il faut alors :

- abandonner l'installation de l'extension ;
- ajouter les entrepôts de mise à jour nécessaires[6] ;
- relancer l'installation de l'extension depuis le début.

2.5 Désinstaller une mise à jour, une extension ou une extension tierce

On désinstalle les extensions et les extensions tierces de la même façon :

- Ouvrir la fenêtre "À propos" à partir du menu principal "Aide > À propos"
- Sélectionner le bouton "Détails de l'installation" (anglais : "Installation détails")
- Une nouvelle fenêtre s'ouvre et affiche l'historique des installations des état de mise à jour et d'extension de TXM
- Sélectionner une ligne d'état de TXM :
  - Le bouton "Delete/Supprimer" supprime les fichiers d'une installation précédente (qui ne sont donc pas actuellement en cours d'utilisation)
  - Le bouton "Revert/Rétablir" re-télécharge la mise à jour ou l'extension et l'applique (et donc supprime les mises à jour et extensions plus récentes)

Il faut alors redémarrer TXM pour que la désinstallation prenne effet.

[6] Il peut être intéressant d'ajouter d'emblée l'entrepôt « http://download.eclipse.org/releases/indigo » qui contient toutes les extensions de base de la plateforme Eclipse utilisée par TXM. Ces extensions étant susceptibles d'être demandées par des extensions tierces.

2.6 Réglages de l'accès au réseau par proxy

Si vous accédez à votre réseau par le biais d'un proxy, vous pouvez configurer TXM de la façon suivante :

- aller à la page des préférences réseau « General > Network Connection » ;
- positionner le champ « Active Provider » à la valeur « Manual » ;
- double-cliquer sur le champ « HTTP » et régler ses valeurs de « Host » et de « Port » pour qu'ils correspondent à votre proxy ;
- double-cliquer sur le champ « HTTPS » et régler ses valeurs de « Host » et de « Port » pour qu'ils correspondent à votre proxy ;
- terminer avec « OK »

Remarque : en cas d'accès par proxy, au premier démarrage de TXM vous pourrez avoir à attendre jusqu'à 10 min le déblocage de la tentative d'accès au réseau par TXM avant de pouvoir régler le proxy dans les préférences.

2.7 Visualisation de l'espace mémoire utilisé

Il est possible de visualiser l'espace mémoire Java utilisé par TXM. Pour cela :

- aller à la page des préférences générales « General » ;
- activer le champ « Show heap status » ;
- terminer avec « OK ».

Remarque : il s'agit de la visualisation de la consommation mémoire Java seulement, elle ne tient pas compte de celle des moteurs de recherche et statistique utilisés par TXM. Mais cela donne malgré tout une indication proportionnelle à la totalité de la mémoire utilisée par TXM à un instant donné.

2.8 En cas de problème avec le logiciel

Si vous rencontrez un problème dans TXM :

A) Vérifier si le problème persiste après redémarrage

B) Vérifier dans les archives de la liste de diffusion txm-users si le problème n'a pas déjà été évoqué : https://groupes.renater.fr/sympa/arc/txm-users
(ne pas hésiter à utiliser le moteur de recherche)

C) Vérifier si le problème a déjà été rapporté dans le wiki txm-users dans les pages suivantes :

a) la F.A.Q. : https://groupes.renater.fr/wiki/txm-users/public/faq

b) les retours de problème pour votre version de TXM. : https://groupes.renater.fr/wiki/txm-users/public/retours_de_bugs_logiciel/txm_0.7
(les problèmes marqués par **OK** ont déjà été résolus et leur correction sera disponible dans la prochaine version de TXM)

Si le problème a déjà été rapporté mais reste non résolu, vous pouvez ajouter des précisions à sa description dans le wiki ou nous demander par mail où nous en sommes de son traitement.

Si le problème est nouveau, merci de faire votre retour sur le wiki ou dans la liste de diffusion txm-users.

Pour nous aider à faire le meilleur diagnostic du problème, il est important de nous fournir les informations les plus détaillées sur votre configuration (TXM n'a pas les mêmes comportements selon les systèmes d'exploitation en particulier) et sur les journaux d'activité de TXM au moment où le problème s'est posé.

Pour indiquer votre configuration :

- depuis TXM aller dans le menu « Aide / À propos de TXM » ;
- cliquer sur « Installation Details » ;
- ouvrir l'onglet « Configuration » ;
- cliquer dans « Copy to Clipboard » et coller l'information dans un fichier ou dans le corps d'un mail.

Pour fournir le journal de lancement de TXM :

- depuis TXM aller dans le menu « Aide / À propos de TXM » ;
- cliquer sur « Installation Details » ;
- ouvrir l'onglet « Configuration » ;
- cliquer dans « View Error Log » et sauver dans un fichier ou copier dans le corps d'un mail.

Pour fournir le journal d'activités détaillé :

- Régler l'affichage des messages dans la console au maximum : depuis les préférences avancées, régler le niveau de journalisation (ou log) à « ALL » ;
- Faire enregistrer les messages dans un fichier : depuis les préférences avancées, sélectionner « Copier le journal dans un fichier ». Le fichier sera créé dans le dossier de travail de TXM ;
- Provoquer le problème et nous fournir une copie du fichier journal résultant.

En cas de problème d'installation, merci de nous fournir une copie des fichiers « TXMPostInstallLogs.txt » et « TXMPostInstallErrorLogs.txt » se trouvant :

- Sous Windows, dans le dossier (caché) %APPDATA% :
  Le chemin de ce dossier est variable selon les versions de Windows :
  - Windows XP :
    `C:\Documents and Settings\<utilisateur>\Application Data`
  - Windows Seven :
    `C:\Users\<utilsateur>\Application Data\Roaming`).
    Le plus simple pour aller dans ce dossier est d'utiliser le script `BAT` du dossier d'installation de TXM nommé `OpenStartUpLogsDirectory.bat` qui ouvrira directement l'explorateur avec ce dossier.
- sous Mac OS X et Linux : dans le dossier « `$HOME`[7]`/TXM/.txm` »

1 Lancer TXM

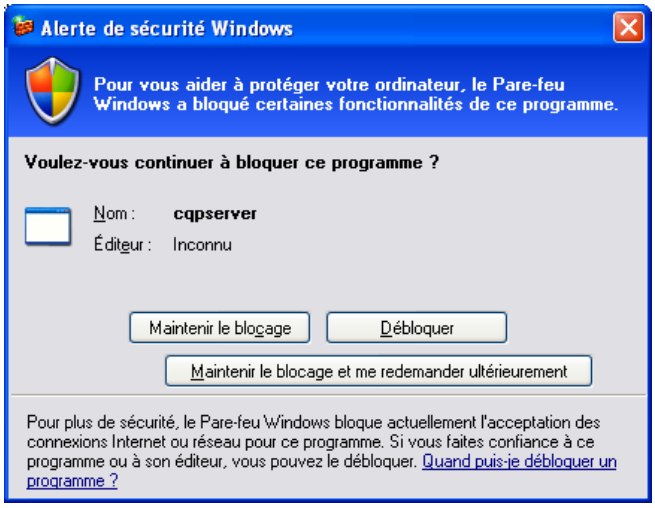
1.1 Sous Windows

1. Menu « Démarrer / TXM / TXM » (faire glisser l'icone de l'application TXM dans la barre de lancement rapide pour rajouter un accès direct)
2. Au premier lancement, en fonction du niveau de sécurité de Windows, vous devrez éventuellement avoir à répondre à une alerte de sécurité de la façon suivante :

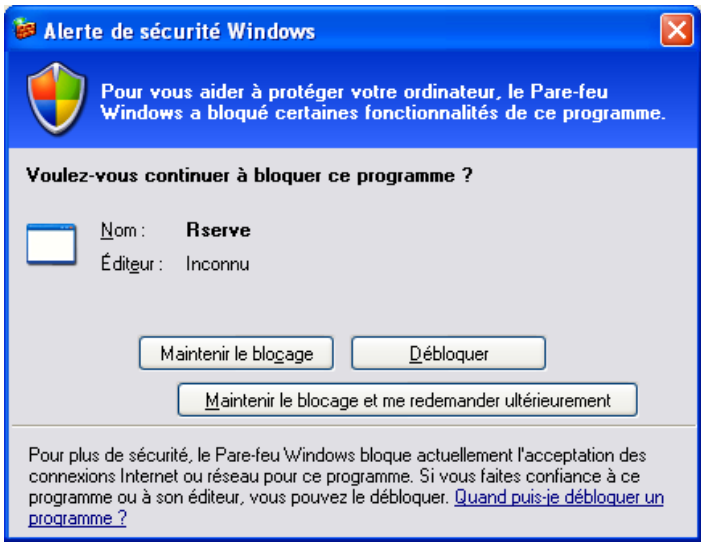
   Attention pour cette étape il faut avoir les droits administrateur pour continuer.

[7] $HOME représente le chemin du dossier personnel de l'utilisateur.

a. Dans cette fenêtre :



Cliquer sur « `Débloquer` »[8]



b. Dans la fenêtre suivante :
Cliquer sur « `Débloquer` »[9]

[8] Le programme « cqpserver » est le moteur de recherche plein texte utilisé par TXM.

[9]Le programme « Rserve » est le moteur statistique utilisé par TXM.

1.2 Sous Mac OS X

Naviguer dans le dossier « Applications / TXM » avec le Finder et double-cliquer sur l'icone de l'application TXM (faire glisser l'icone de l'application TXM dans le dock pour rajouter un accès direct).

1.3 Sous Linux

Naviguer dans la section « Applications installées » du Launchpad Unity et double-cliquer sur l'icone de l'application de TXM (pour rajouter un accès direct, faire un clic droit sur l'icone de TXM dans le Dock et sélectionner 'Garder dans le Dock'). Ou via un terminal en lançant la commande : « TXM& ».

2 Utiliser les fenêtres, les menus, les barres d'outils et les raccourcis clavier

2.1 Vue générale de l'interface graphique

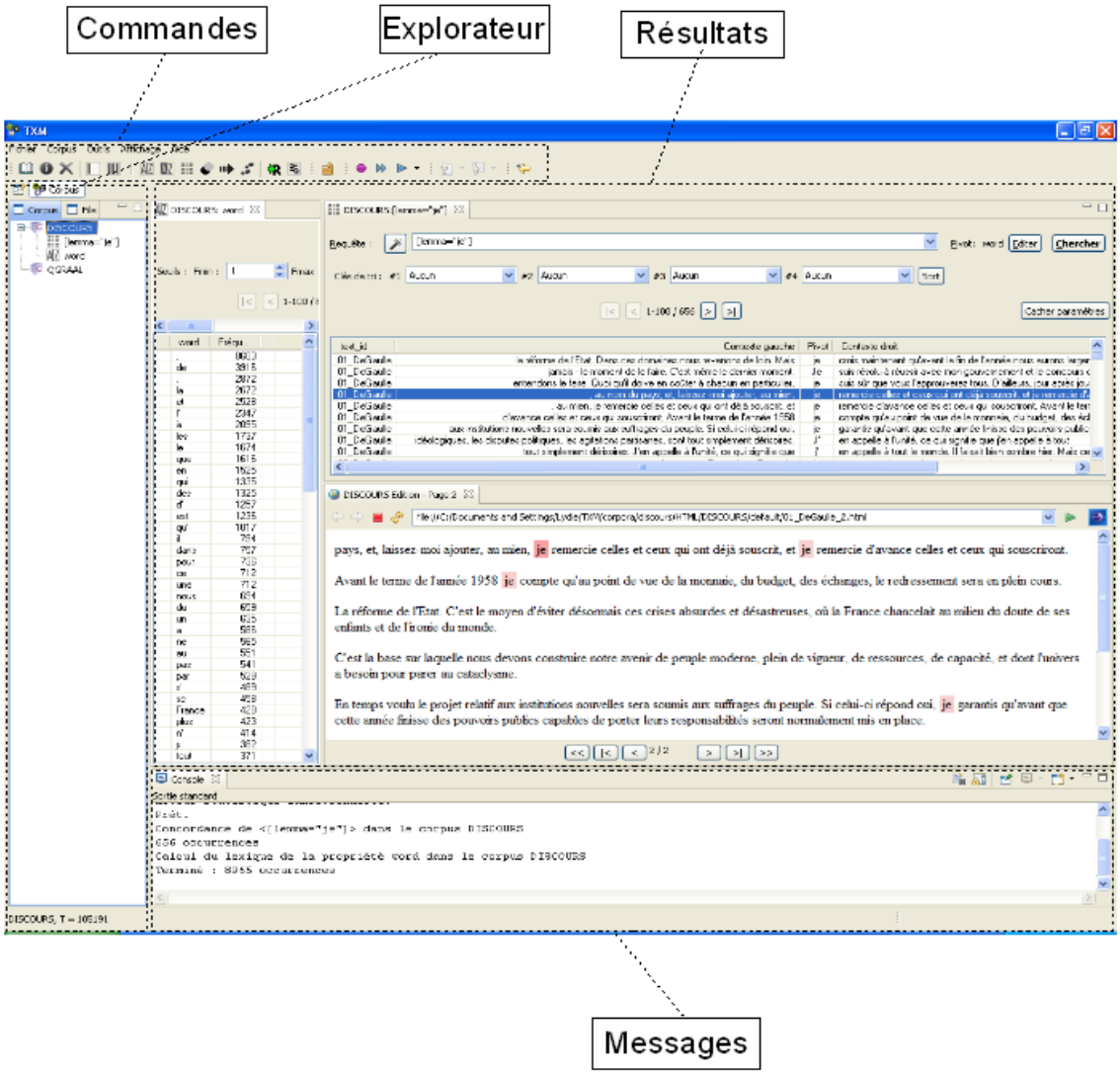


Illustration 2.1 : L'interface générale de TXM

L'interface utilisateur de TXM est divisée en quatre zones, comme indiqué dans l'illustration 2.1 :

- L'explorateur : donnant accès aux corpus et sous-corpus, aux résultats de commandes et aux dossiers et fichiers de scripts. En général, les objets gérés par TXM et sur lesquels s'appliquent les commandes ;
- Les commandes : boutons et menus qui permettent de lancer des actions sur les objets sélectionnés dans l'explorateur ;
- Les résultats : fenêtres de sortie ;

- Les messages : commentaires sur chaque action exécutée.

Toutes les zones sont gérées dans une seule et même fenêtre[10].

Nous allons d'abord présenter les principales zones pour ensuite expliquer comment organiser cette interface dans la fenêtre principale.

2.1.1 L'explorateur

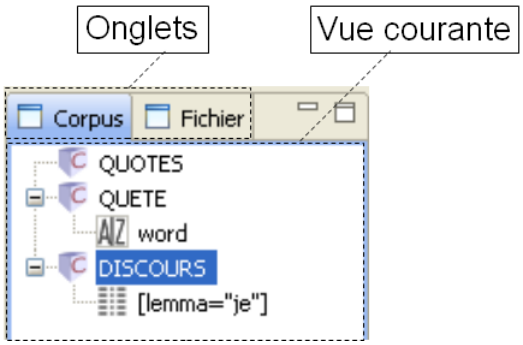


Illustration 2.2 : L'explorateur.

L'explorateur permet à l'utilisateur de sélectionner l'objet sur lequel il souhaite appliquer une commande.

L'explorateur comporte par défaut deux vues :

- La vue « Corpus » : affiche les icones de corpus et de sous-corpus disponibles pour l'analyse ainsi que les différents résultats déjà calculés ;
- La vue « Fichier » : affiche les fichiers situés dans vos dossiers, avec la possibilité de les éditer.

Dans cette même zone, il est possible d'afficher d'autres vues telles que :

- La vue « Variables R » : affiche les objets TXM qui ont été envoyés dans R
- La vue « Requêtes » : affiche toutes les requêtes CQL ayant été utilisées pour chaque corpus.

[10] Nous verrons qu'il est possible d'ouvrir n'importe quelle zone dans une nouvelle fenêtre.

2.1.1.1 La vue « Corpus »

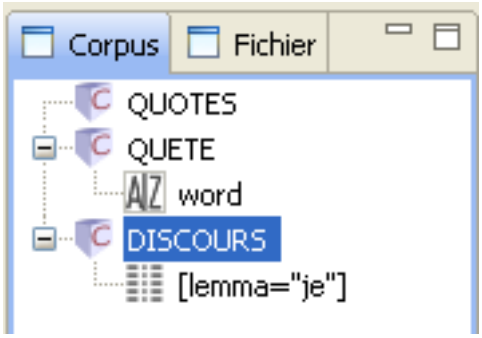


Illustration 2.3 : La vue Corpus.

La vue « Corpus » présente les différents corpus disponibles pour l'analyse dans TXM, ainsi que toutes les icones d'objets construits par TXM durant la session. C'est la vue principale de TXM. Les corpus sont créés soit depuis la commande Import, soit depuis la commande Charger du menu «Fichier».

La vue « Corpus » a une organisation hiérarchique. Chaque racine représente un corpus indépendant. Tous les éléments descendants des racines résultent de l'application de commandes TXM :

- Sous-corpus (icone « C », identique à celle de la 'racine' corpus) depuis 'Créer un sous-corpus' ;
- Partitions (icone « P ») depuis 'Créer une partition' ;
- Lexique ;
- Index ;
- Références ;
- Concordance ;
- Cooccurrences ;
- Spécificités ;
- AFC ;
- Classification ;
- Table lexicale
- Envoyer vers R.

Une branche dans l'arbre des résultats sera créée à chaque nouveau résultat de commande.

Chaque type d'objet peut se voir appliquer un ensemble spécifique de commandes :

- un « Corpus » peut se voir appliquer n'importe quelle commande ;

- un « Sous-Corpus » peut se voir appliquer les mêmes commandes que le corpus, ainsi que la commande « Spécificités ».
- une « Partition » peut se voir appliquer les commandes Spécificités, AFC ou Table lexicale.

Double-cliquer sur un résultat réouvre la fenêtre des résultats si elle a été fermée ou l'affiche si elle était cachée.

2.1.1.2 La vue « Fichier »

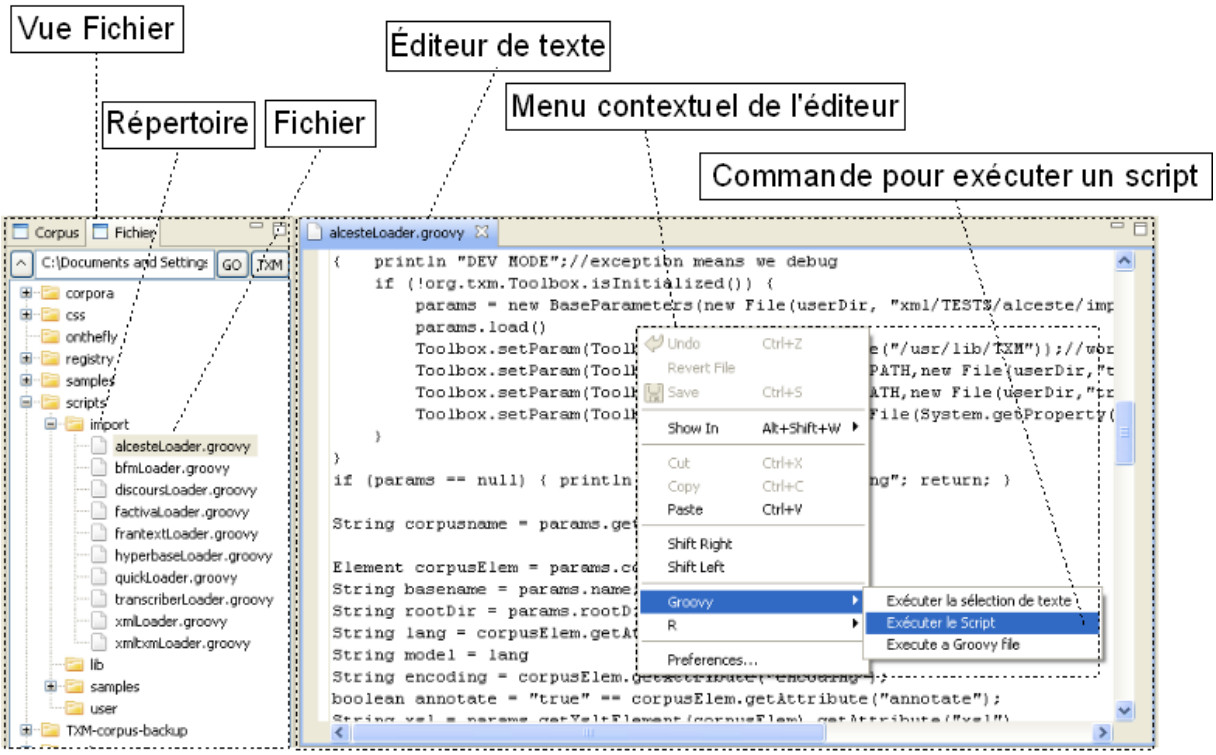


Illustration 2.4 : La vue Fichier.

La vue Fichier présente une arborescence classique des icones de dossiers et de fichiers présents sur votre disque dur.

Depuis cette vue il est possible d'ouvrir un fichier dans l'éditeur de texte de TXM. Ainsi il est possible de modifier les textes sources d'un corpus ou des scripts à partir de TXM.

Navigation

Le bouton « ^ » ouvre le dossier parent du dossier courant.

Dans le champ texte vous pouvez éditer le chemin du dossier courant et cliquer sur le bouton « OK » ou presser la touche « Entrée » pour rafraîchir la vue.

Le bouton « TXM » renvoie directement au dossier TXM de l'utilisateur.

Un double-clic sur un dossier développe son contenu.

Un double-clic sur l'icone d'un fichier ouvre ce dernier dans une nouvelle fenêtre d'éditeur. Un résultat similaire est obtenu via la commande 'Ouvrir un fichier' dans le menu 'Fichier'.

A partir du menu contextuel, on peut :

- créer un nouveau dossier ;
- créer un nouveau fichier ;
- ouvrir un fichier HTML dans le navigateur web ;
- ouvrir un fichier dans l'éditeur de texte ;
- exécuter un fichier en tant que script Groovy
- exécuter un fichier en tant que script R

2.1.1.3 La vue "Console"

La vue Console est l'une des vues les plus importantes. Elle informe sur le bon déroulement ou non de tout calcul de TXM. Si elle est fermée, on peut la ré-ouvrir avec la commande « Affichage / Vues / Console ». Si cette commande n'a pas d'effet, on peut réinitialiser la perspective courante de TXM pour la ré-ouvrir (configuration générale des fenêtres) avec la commande « Affichage / Perspectives / Réinitialiser la perspective ».

2.1.1.4 La fenêtre « Éditeur de texte »

On peut ouvrir autant de fenêtres d'édition de texte que nécessaire pour éditer des fichiers aux formats textuels : TXT, XML, etc. Voir la section 3 « l'éditeur de texte » page 55 pour une présentation de son mode de fonctionnement.

2.1.1.5 La vue « Variables R »

Cette vue fait partie de la perspective R. Elle permet de voir à tout moment les objets TXM de l'environnement R en associant le nom d'un objet de la vue Corpus à son symbole dans l'espace de travail de R (si l'objet à été transmit à R, voir section <à faire>).

Le bouton « Refresh » ré-actualise cette liste. Le bouton « Voir les logs », génère un historique des lignes de commande R de l'utilisateur depuis le lancement de TXM.

Il est possible de copier le symbole d'un objet de l'espace de travail R à l'aide du menu contextuel, ou du raccourci clavier Ctrl+c, pour le coller dans une ligne de commande R par exemple.

2.1.1.6 La vue « R Console »

Elle affiche les sorties de toutes les commandes R exécutées par l'utilisateur. Ces sorties sont exactement les mêmes que celles d'un interpréteur R lancé depuis un terminal.

2.1.1.7 La vue « Requête »

Cette vue recense toutes les requêtes CQL utilisées par chaque corpus (pas pour les sous-corpus).

Il est possible de copier une requête à l'aide du menu contextuel, ou du raccourci clavier Ctrl+c, pour la coller dans un champ de requête par exemple.

Il est aussi possible d'exporter la liste complète des requêtes dans un fichier au format texte brut à l'aide du bouton « Exporter ».

2.1.2 Les commandes

Dans TXM, les commandes principales peuvent être lancées de trois façons différentes :

1) Barre d'outils : quand une icone d'objet est sélectionnée dans la vue Corpus, l'utilisateur peut appliquer une commande à cet objet en cliquant sur l'icone de la commande correspondante dans la **barre d'outils**.

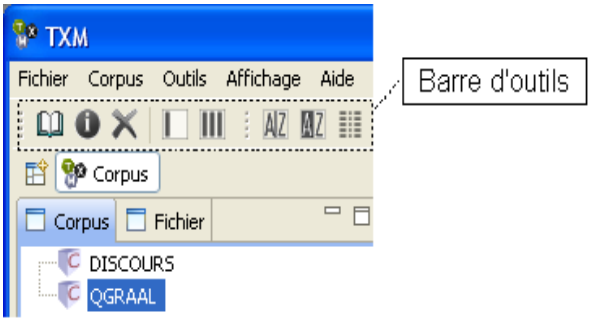


Illustration 2.5 : La barre d'outils.

2) Menu principal : quand une icone d'objet est sélectionnée dans l'explorateur, l'utilisateur peut appliquer une commande à cet objet en sélectionnant l'action correspondante dans les **menus** « Fichier », « Corpus » et « Outils ».

   a. Le menu « Fichier » où l'on retrouve la commande Importer :

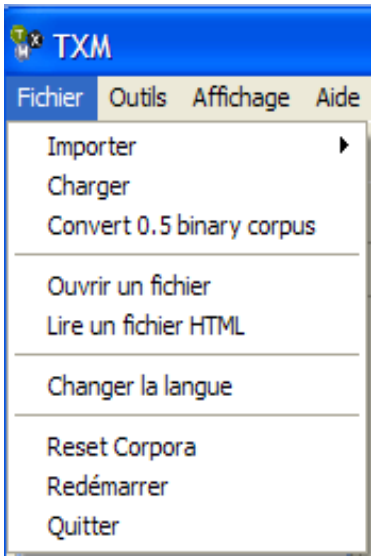


Illustration 2.6 : Le menu Fichier

b. Le menu « Corpus » où l'on retrouve les commandes de description et de manipulation de corpus :
La configuration du menu change en fonction du type d'icone sélectionnée : le premier menu apparaît si un corpus est sélectionné, tandis que le second apparaît quand il s'agit d'une partition.

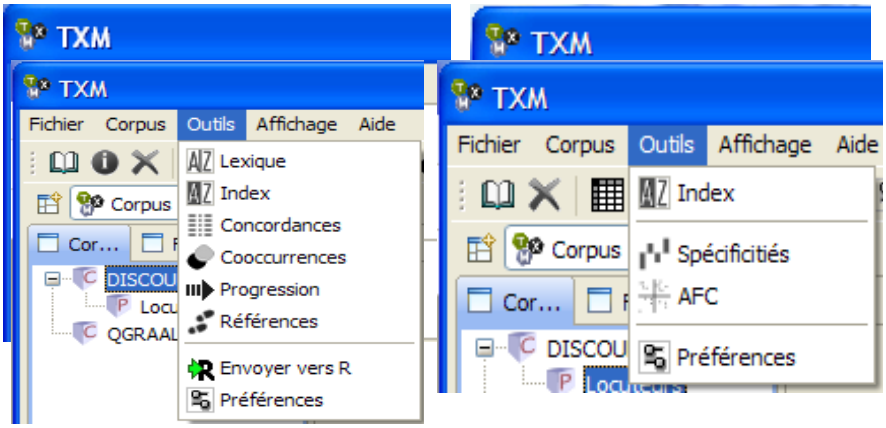
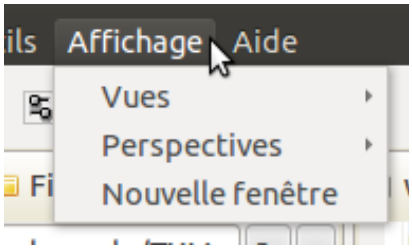


Illustration 2.8 : Le menu Outils, concernant d'une part les corpus et d'autre part les partitions.

c. Le menu « Outils » donne accès aux outils textométriques :

d. Le menu « Affichage » donne accès à la configuration de l'agencement des fenêtres de TXM.



3) L'utilisateur peut ouvrir un menu contextuel en faisant un clic droit sur l'objet qui doit recevoir la commande.

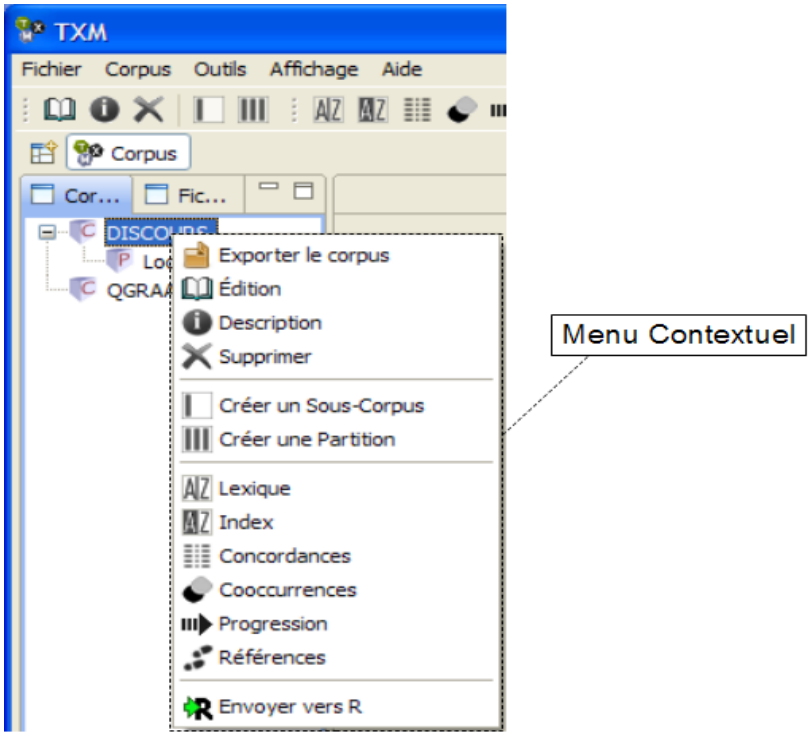


Illustration 2.9 : Menu contextuel du corpus.

Les différentes commandes sont décrites en détail dans la section 4 'Utiliser les commandes de TXM'.

Certaines commandes sont également accessibles par un lien hypertexte (double-clic) depuis les fenêtres de résultats, en fonction des types d'objets contenus dans ces résultats : ligne de tableau, ligne de concordance, etc.

2.1.3 Les icones

Voici la liste des icones de l'interface graphique de TXM, ainsi que leur nom :

2.1.3.1 icones d'objets

	Corpus
	Partition
	Bibliographie
	Édition
	Table lexicale
	Vue interne

2.1.3.2 icones des commandes

	AFC
	Bibliographie
	Concordances
	Cooccurrence
	Description
	Envoi vers R
	Export du corpus
	Export (point d'entrée du menu contextuel)

	Export des données
	Export du graphique
	Index
	Lexique
	Partition
	Préférences
	Progression
	Références
	Requête assistée
	Spécificités
	Sous-Corpus
	Supprimer

2.1.4 Les menus principaux

Tous les menus principaux de TXM qui se trouvent dans le coin supérieur gauche de l'interface sont décrits ci-dessous :

2.1.4.1 Menu « Fichier »

- Exporter : exporte les résultats d'une commande sous divers formats.
- Importer : importe un nouveau corpus à partir de ses sources via les différents modules d'importation disponibles (voir la section « modules d'importation » pour plus de détails sur ces modules) :
  - Presse-papier : importe le texte copié dans le presse-papier ;
  - TXT+CSV: importe les fichiers au format texte brut accompagnés d'un fichier de métadonnées de textes au format CSV '`metadata.csv`' ;

  - XML/w+CSV : importe les fichiers XML ayant leurs mots en texte brut ou balisés avec un élément `<w>` accompagnés d'un fichier '`metadata.csv`' ;
  - XML-TEI BFM : importe les fichiers au format XML-TEI P5, comprenant l'encodage des textes et des métadonnées, selon les recommandations du projet BFM[11] ;
  - XML-TEI Frantext : importe les fichiers au format XML-TEI de l'ATILF[12] ;
  - XML-TEI TXM : importe les fichiers au format interne de TXM[13] ;
  - XML-TRS+CSV : importe les fichiers au format '`.trs`' générés par logiciel Transcriber accompagnés d'un fichier '`metadata.csv`' ;
  - XML-PPS : importe les fichiers au format XML produit par le portail Factiva ;
  - XML-TMX : importe les corpus alignés au format TMX ;
  - Factiva TXT : importe les fichiers au format texte produit par le portail Factiva ;
  - CNR+CSV : importe les fichiers au format CNR (produits par le logiciel Cordial) accompagnés d'un fichier '`metadata.csv`' ;
  - Alceste : importe les fichiers au format du logiciel Alceste ;
  - Hyperbase : importe les fichiers à l'ancien format du logiciel Hyperbase ;
  - CWB : importe les corpus au format source du moteur CQP.
- Charger : charge un nouveau corpus depuis son dossier binaire ;
- Nouveau fichier : ouvre un éditeur de texte sur un nouveau fichier ;
- Ouvrir... : ouvre un fichier dans un nouvel éditeur de texte ;
- Sauvegarder : enregistre le fichier texte en cours d'édition ;
- Tout sauvegarder : enregistre tous les fichiers texte en cours d'édition ;
- Fermer : ferme le fichier texte en cours d'édition ;
- Tout fermer : ferme tous les fichiers texte en cours d'édition ;
- Ouvrir dans un navigateur : affiche un fichier dans un nouveau navigateur web ;
- Changer la langue : choisir la langue de l'interface de TXM ;

[11] `http://bfm.ens-lyon.fr`

[12] `http://www.cnrtl.fr/corpus/frantext`

[13] `http://txm.sourceforge.net/wiki/index.php/Xml-txm-tei`

- Vérifier les mises à jour : vérifier s'il y a des mises à jour ;
- Ajouter une extension : ajoute une extension de TXM ;
- Ajouter une extension tierce : ajoute une extension à TXM ;
- Redémarrer les moteurs : redémarre les moteurs de recherche et statistique ;
- Quitter : ferme l'application TXM.

2.1.4.2 Menu « Corpus »

- Ouvrir une fiche bibliographique : affiche les informations bibliographiques des textes disponibles
- Édition : affiche la première page de l'édition du premier texte du corpus
- Description : affiche les structures et leurs propriétés ainsi que les propriétés des mots du corpus
- Supprimer : supprime l'objet sélectionné.
- Créer un sous-corpus : construit un nouveau sous-corpus
- Créer une partition : construit une nouvelle partition
- Table lexicale : crée une table lexicale à partir d'une partition ou à partir de l'index d'une partition.

2.1.4.3 Menu « Outils »

- Lexique : liste hiérarchique des valeurs d'une propriété de mot sur l'ensemble d'un corpus.
- Index : liste hiérarchique de combinaisons de valeurs de propriétés de mots pour toutes les occurrences d'une requête CQL donnée
- Concordances : cherche les occurrences d'un motif exprimé à l'aide d'une requête CQL et affiche les résultats contextualisés sous forme de concordances kwic
- Références : affiche les références d'un motif CQL
- Progression : affiche l'évolution d'un ou de plusieurs motifs au fil du corpus
- Cooccurrences : calcule les cooccurrents d'une requête CQL
- Spécificités : calcule les valeurs de propriétés les plus spécifiques de chaque partie d'une partition

- AFC : calcule l'analyse factorielle des correspondances d'une partition pour une propriété de mots donnée et affiche le premier plan factoriel.
- Classification : calcul une classification à partir d'une table lexicale ou d'une AFC.
- Envoyer vers R : envoyer dans R les données numériques de l'objet sélectionné.
- Réglages : ouvre la page des préférences des commandes. [Dans cette version, le menu est identique à celui accédé par l'entrée Fichier / Préférences]

2.1.4.4 Menu « Affichage »

Ce menu permet d'ouvrir les vues et perspectives suivantes :

- Perspectives (deux types d'agencement des fenêtres de TXM)
  - Corpus : il s'agit de l'agencement par défaut qui permet de lancer les commandes de Textométrie et de manipuler les résultats.
  - R : il s'agit de l'agencement spécialisé pour travailler avec le moteur de statistiques R.
- Vues
  - Corpus : affiche les icones de corpus et de sous-corpus disponibles pour l'analyse ainsi que les différents résultats déjà calculés ;
  - Fichier : affiche les fichiers situés dans vos dossiers, avec la possibilité de les éditer ;
  - Requête : liste de toutes les requêtes CQL ayant été utilisées pour chaque corpus ;
  - Console : affiche les messages d'erreur ou de résultats de TXM
  - Console R : affiche les sorties de toutes les commandes R exécutées par l'utilisateur.
  - Variables R : liste les objets TXM qui ont été envoyés dans R
  - …
- Réinitialiser la perspective : réorganise la disposition de toutes les fenêtres de la perspective courante à leur position par défaut.
- Nouvelle fenêtre : Duplique l'ensemble de la fenêtre de TXM dans une nouvelle fenêtre.

2.1.4.5 Menu « Aide »

- Raccourcis clavier: affiche tous les raccourcis clavier disponibles ;
- Raccourcis clavier graphiques : affiche les raccourcis pour les manipulations de graphiques ;
- Manuel en ligne : ouvre la version en ligne de ce manuel dans un navigateur ;
- Signaler un bug : ouvre la page web de « rapport de bug » ;
- Demandes de fonctionnalités : ouvre la page web de « demandes de fonctionnalités » ;
- S'inscrire à la liste de diffusion txm-users : ouvre le formulaire d'inscription à la liste de diffusion des utilisateurs de TXM ;
- Installer TreeTagger : ouvre le tutoriel d'installation de TreeTagger dans TXM ;
- Aller au site web de TXM : ouvre la page de téléchargement de TXM ;
- Site web du projet Textométrie ; ouvre la page d'accueil du site du projet Textométrie ;
- Dossier des manuels de TXM : ouvre la page des manuels sur le site du logiciel ;
- Documentation du projet Textométrie : ouvre la page de documentation du projet ;
- À propos : affiche la version de TXM, des informations sur sa licence ainsi que la liste de tous les composants du logiciel et leur licence.

2.1.5 Affichage des résultats

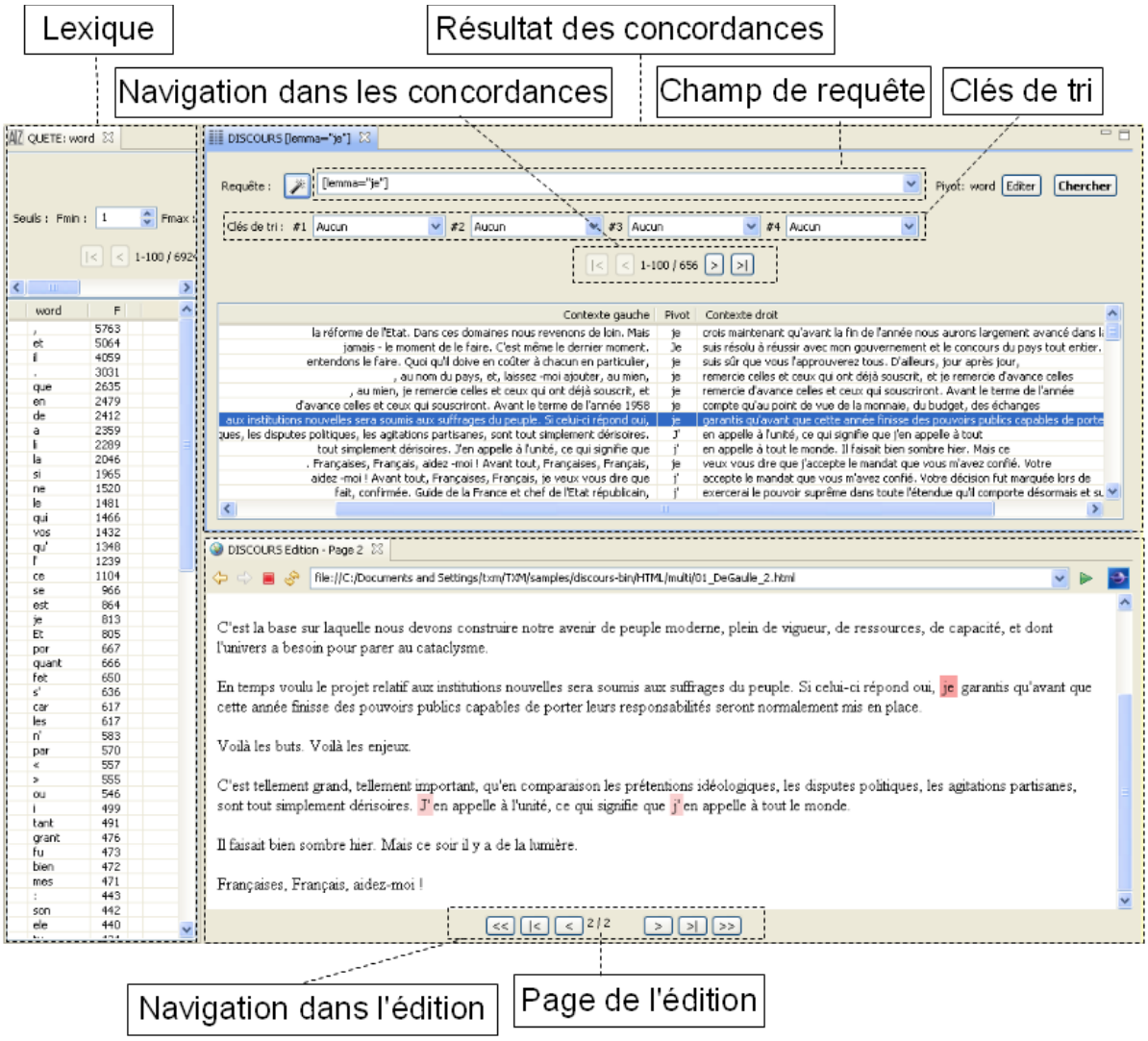


Illustration 2.10 : Exemple de fenêtres de résultats.

Tous les résultats des commandes sont, par défaut, affichés dans la zone des résultats, à droite[14].

Pour chaque nouvelle commande, le résultat est affiché dans une nouvelle fenêtre, dont le nom est en rapport avec la commande en question ainsi que ses paramètres. Une nouvelle icone est également ajoutée, selon le même principe, dans la vue « Corpus ».

[14] Cette zone peut-être déplacée n'importe où, grâce au gestionnaire de fenêtres.

Le nom de la fenêtre est également présent dans l'onglet et dans la légende de l'icone.

Cet onglet permet de gérer l'affichage de la fenêtre comme il sera expliqué dans la partie traitant de la gestion des fenêtres.

Pour augmenter la surface du tableau d'affichage du résultat, on peut cliquer sur le bouton [Cacher paramètres].

Si une fenêtre est fermée par inadvertance pendant la session, elle peut être ré-ouverte en double-cliquant sur l'icone correspondante dans la vue « corpus ».

Recherche dans les résultats

Le raccourci « Ctrl-F » (Find) lance une recherche de chaîne de caractères dans un tableau de résultats.

Copier/Coller de lignes de résultats

Les lignes des résultats sous forme de tableaux peuvent être sélectionnées à la souris et copiées dans le presse-papier avec le raccourci clavier « Ctrl-C » (Copy). Le presse-papier contient alors une version tabulée des lignes du tableau d'origine en texte brut encodé en UTF-8.

2.1.6 Affichage des messages dans la console

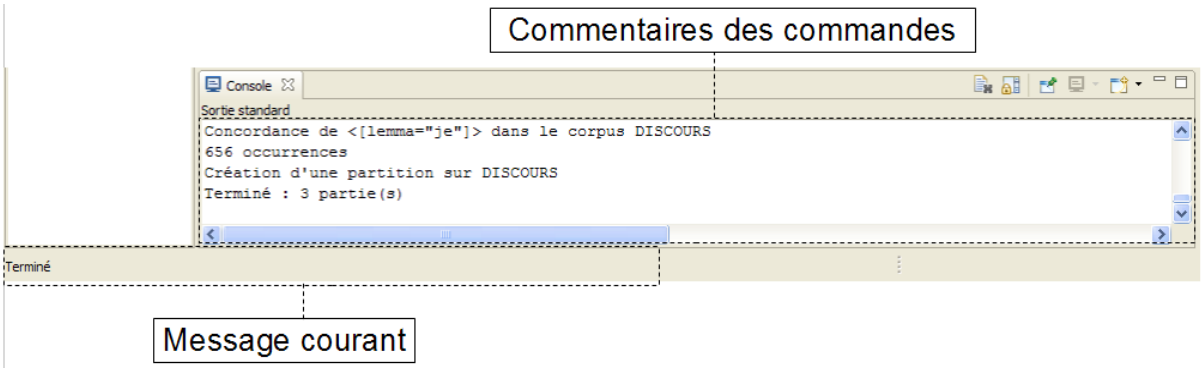


Illustration 2.11 : Les messages.

La zone des messages courants affiche les informations temporaires, comme le nombre de résultats de la dernière commande.

La zone de commentaires des commandes, appelée également « console » ou « journal », située en bas à droite de l'interface par défaut donne plus d'informations relatives à celles-ci. Le texte peut être parcouru, sélectionné, copié et collé. Les messages d'erreurs seront également affichés ici.

2.1.6.1 Réglage du niveau de détails des commentaires de la console

Il est possible de régler le niveau de détails des messages de la console en lançant la commande du menu principal « Outils > Préférences », puis en accédant à la section « TXM > Avancé », et en choisissant le niveau de détails de la préférence « `Niveau de détails du journal` ».

Remarque : les niveaux inférieurs à WARNING provoquent l'affichage de messages de debuggage des modules d'import pendant leur exécution.

2.1.7 Changer d'interface grâce aux perspectives

Les perspectives, accessibles depuis le menu « Affichage > Perspectives », organisent de façon cohérente l'accès aux commandes, les vues et les fenêtres de résultats par type de session de travail :

- la perspective « Corpus » : utilisée par défaut, présente les corpus disponibles et donne accès aux commandes et aux vues de résultats permettant de réaliser une session d'analyse textométrique ;
- la perspective « R » : spécialisée pour les sessions de travail utilisant de façon préférentielle l'environnement statistique R. Elle facilite le transfert d'objets de TXM (table lexicale, tableau de spécificités, corpus, sous-corpus, index, concordance, …) vers l'environnement R, la manipulation de ces objets par le biais de symboles R et l'exécution de scripts R avec une console spécialisée.

2.1.8 Réinitialiser l'interface utilisateur

Les perspectives mémorisent la disposition des fenêtres pour un usage donné de TXM. À tout moment, on peut réinitialiser la disposition des fenêtres de l'interface utilisateur par le biais du menu contextuel de la perspective courante (clic-droit sur le bouton d'appel de la perspective dans la barre d'outils) en appelant la commande « Reset » du menu contextuel du bouton de la perspective.

2.2 Le gestionnaire de fenêtres

Avec le gestionnaire de fenêtres, il est facile d'augmenter ou de minimiser la taille, réduire, rouvrir, déplacer et redimensionner n'importe quelle fenêtre de l'interface avec l'aide de la souris.

Les manipulations qui peuvent être effectuées sont les suivantes :

- afficher la fenêtre en plein écran : double-cliquer sur l'onglet de la fenêtre ;
- remettre la fenêtre à sa taille originale : double-cliquer sur l'onglet ;

- déplacer et redimensionner une fenêtre en « glisser-déposer » : faire glisser l'onglet de la fenêtre à la place choisie. Avant de relâcher le bouton de la souris, une fenêtre fantôme représente la future taille de la fenêtre si on la relâche à cet endroit. Chaque fenêtre d'arrivée possède quatre zones de « dépôt » potentielles en bordure :
  - gauche : permet de séparer verticalement et de déposer la fenêtre sur le côté gauche ;
  - droit : sépare de manière verticale et dépose la fenêtre sur le côté droit ;
  - haut : sépare de manière horizontale et dépose la fenêtre en haut ;
  - bas : sépare de manière horizontale et dépose la fenêtre en bas ;
- réduire la fenêtre : cliquer sur l'icone « réduire » de la fenêtre ;
- fermer toutes les fenêtres de résultats d'un coup : dans le menu contextuel de l'onglet d'une vue, sélectionner « close all »
- détacher une fenêtre : dans le menu contextuel de l'onglet d'une vue, sélectionner « detached ».

Chaque fenêtre des zones des objets et des résultats est gérée de façon logique.

La disposition des fenêtres est automatiquement sauvegardée par TXM et réutilisée au prochain lancement de TXM.

3 Utiliser l'éditeur de texte

L'éditeur de texte intégré à TXM permet de modifier, enregistrer, etc. n'importe quel fichier texte (TXT, XML, etc.) du disque dur. On peut ouvrir autant de fichiers qu'on veut, à partir :

- de la vue « Fichier » (en double-cliquant sur les icones de fichiers à éditer) ;
- du menu « Fichier > Nouveau fichier » (pour éditer un nouveau fichier) ;
- du menu « Fichier > Ouvrir... ».

Les commandes de l'éditeur sont accessibles depuis :

- la barre d'outils (voir illustration 2.5) ;
- le menu contextuel (clic droit de la souris) ;
- des raccourcis clavier (voir section 13 page 212 « Raccourcis clavier »).

3.1.1 Barre d'outils de l'éditeur de texte

Illustration 3.1: Barre d'outils d'un éditeur de texte

Les outils de la barre sont dans l'ordre :

Gestion des préférences

- gestion des préférences de TXM

Gestion de fichiers

- Édition d'un nouveau fichier
- Édition d'un fichier existant
- Enregistrer le fichier courant
- Enregistrer le fichier courant sous un nouveau nom

Édition

- Couper le texte sélectionné
- Copier dans le presse-papier
- Coller le contenu du presse-papier
- Supprimer le texte sélectionné ou le caractère situé à droite du curseur
- Annuler l'édition précédente
- Refaire l'édition annulée
- Sélectionner tout le texte

Chercher-Remplacer

- Chercher-Remplacer une chaîne de caractères ou une expression régulière
  - Précédent (vers le haut)
  - Suivant (vers le bas)
- Recherche incrémentale (la recherche est précisée à chaque caractère saisi) vers le bas
- Recherche incrémentale vers le haut

Général

- Aller à la ligne numéro...
- Afficher les numéros de lignes
- Afficher les caractères non imprimables
- Ouvrir le lien sélectionné dans un navigateur
- Ré-ouvre le fichier avec un encodage de caractères différent
- Revenir à l'état initial du texte

Exécution de scripts

- Exécuter le script (par les interpréteurs Groovy ou R en fonction de l'extension du fichier .groovy ou .R)
- Exécuter les lignes sélectionnées (par les interpréteurs Groovy ou R)
- Exécuter un script Groovy contenu dans un fichier
- Ré-exécuter le dernier script Groovy ou la dernière macro exécuté

3.1.2 Menu contextuel de l'éditeur de texte

- Annuler l'édition
- Revenir à l'état initial du texte
- Enregistrer
- Couper le texte sélectionné
- Copier dans le presse-papier
- Coller le contenu du presse-papier
- Indenter à droite
- Indenter à gauche
- Option de retour à la ligne automatique
- Exécuter par l'interpréteur Groovy
  - la sélection de texte
  - le script
  - un fichier Groovy

- Exécuter par l'interpréteur R
    - la sélection
    - le script
- Préférences de l'éditeur
- Méthodes de saisie au clavier (sous Ubuntu)
    - Cyrillique (translittéré) : pour la saisie de caractères russes avec un clavier occidental
    - etc.

4 Importer un corpus : créer un nouveau corpus dans TXM

4.1 Principes généraux d'import : les trois types de sources textuelles exploitables

La plateforme TXM est conçue pour importer et analyser trois grands types de corpus textuels :

A. les corpus de **textes écrits**, comprenant éventuellement des éditions paginées incluant des images de fac-similés (comme par exemple de manuscrits médiévaux, d'auteur ou encore d'élèves) ;

B. les corpus de **transcriptions d'enregistrements**, éventuellement synchronisées avec la source audio ou vidéo ;

C. les corpus **multilingues alignés** au niveau d'une structure textuelle comme la phrase ou le paragraphe.

L'importation consiste à lire les fichiers sources du corpus pour en construire une représentation interne au sein de TXM qui est ensuite utilisée pour son exploitation. Cette représentation est à la base de tous les calculs qui sont réalisés par TXM. Elle est composée des éléments fondamentaux suivants :

- des unités textuelles : chaque corpus est composé d'un ensemble de **textes** (livre, article, entretien...) pouvant avoir des propriétés appelées « métadonnées » (auteur, titre, date, genre...)
    - des structures textuelles optionnelles : chaque texte peut comprendre des **structures** internes imbriquées (sections, paragraphes, tours de parole...) pouvant avoir des propriétés (titre, numéro...)

- des unités lexicales : chaque texte est composé d'une séquence de **mots** pouvant avoir des propriétés (forme graphique, lemme, catégorie grammaticale...). Les mots peuvent être imbriqués dans des structurelles textuelles et forment la plus petite unité du corpus.

À chaque unité textuelle correspond une édition du texte au format HTML destinée à la lecture cursive et au « retour au texte » depuis les commandes d'analyse. Selon le type de corpus l'édition peut être paginée, disposer de mise en page, de styles ainsi que d'illustrations sous forme d'images.

Pour les corpus de textes écrits, la pagination de l'édition peut être alignée avec les fichiers images de fac-similés (folios de manuscrits, pages d'édition ou de manuscrits, etc.). Ce qui permet la lecture synoptique image de fac-similé et page d'édition en vis-à-vis dans TXM.

L'importation d'un corpus est l'occasion d'équiper automatiquement chaque mot d'un texte avec son lemme et sa catégorie morphosyntaxique à l'aide de logiciels comme TreeTagger.

Les corpus de transcriptions d'enregistrements peuvent être synchronisés avec les fichiers vidéo ou audio d'origine. Ce qui permet de jouer à la demande les passages vidéo ou audio correspondant à la transcription dans TXM.

Les corpus alignés sont alignés au niveau d'une structure interne (phrase, paragraphe, etc.). Ce qui permet de chercher simultanément dans les deux langues, ou dans les deux versions d'un même texte par exemple, l'apparition de mots se trouvant dans des passages alignés.

Enfin, les modules d'import les plus évolués peuvent adapter leurs traitements en fonction de différentes parties de chaque texte pour construire ce qu'on appelle les plans textuels. Ils peuvent par exemple ignorer des parties (le « hors texte »), éditer des parties sans que leurs mots soient indexés par le moteur de recherche, éditer certaines parties sous forme de notes de bas de page, etc.

4.2 Philologie progressive : les trois principaux niveaux de représentation textuelle importables

L'environnement d'importation de sources de TXM est conçu de sorte à pouvoir choisir un niveau de représentation des sources plus ou moins riche, et donc plus ou moins coûteux à préparer, en fonction de ses besoins d'analyse avec l'outil :

1. texte brut (TXT) : une représentation élémentaire comme le texte brut (une séquence de caractères) peut être importée dans un premier temps dans TXM et déjà offrir les services d'analyse de base. Tous les formats textuels non standard peuvent être convertis[15] en TXT pour pouvoir bénéficier de ce premier type d'import (PDF, MS Word, LibreOffice Writer, etc.) ;

[15] Voir les différentes macros TXM utiles pour faire ces conversions <https://groupes.renater.fr/wiki/txm-users/public/macros>.

2. texte encodé en XML: si cela s'avère pertinent, les sources peuvent être augmentées en une représentation plus évoluée comme par exemple avec un balisage XML[16] pour être ré-importées dans TXM et bénéficier d'autres possibilités de manipulation de corpus et d'analyse (disponibilité de structures internes, de pré-codage de mots particuliers, etc.) ;

3. texte encodé en XML selon les recommandations de la TEI : à partir du moment où l'on investit dans l'encodage XML, il devient alors intéressant d'appliquer les principes du consortium TEI pour l'encodage XML pour pouvoir bénéficier de certains services supplémentaires de TXM (réglage de la construction des éditions, construction d'éditions synoptiques, etc.).

Avec TXM on peut donc moduler l'investissement dans la préparation des sources en fonction des besoins d'analyse. Par exemple on peut commencer en texte brut et s'y limiter si les analyses obtenues sont satisfaisantes. Le Tableau 1: Carte des niveaux d'import TXM illustre les différents niveaux de services offerts par la plateforme en fonction du niveau de représentation choisi :

		Niveaux de représentation		
		TXT	**XML/w**	**XML-TEI**
Services	Unités Textuelles	fichiers	fichiers	fichiers
	Métadonnées	CSV	CSV	teiHeader
	Mots	brut	<w>?	<w>?
	Structures	-	toutes	spécifique
	Plans textuels	-	XSL frontale	spécifique

Tableau 1: Carte des niveaux d'import TXM

En colonnes, les niveaux de représentation :

1. « TXT » tel qu'importé par le module d'import « TXT + CSV » ;
2. « XML/w » tel qu'importé par le module d'import « XML/w + CSV »[17] ;
3. « XML-TEI » tel qu'importé par un module comme « XTZ + CSV » ou encore « XML-TEI BFM ».

[16] Voir la macro TXT2XML pour faciliter la conversion par lot de fichiers TXT vers XML.

[17] Le « /w » dans le nom du module exprime le fait que le module interprète spécifiquement les balises XML <w>...</w> dans les sources comme encodant directement des unités lexicales (mots).

Chaque ligne correspond à un élément du modèle de corpus de la plateforme plus ou moins disponible selon le niveau de représentation choisi :

- Unités textuelles : tous les niveaux de représentation associent une unité textuelle à un fichier de sources. Pour chaque fichier source il y aura un texte et une édition dans le corpus ;
- Métadonnées : les deux premiers niveaux importent les métadonnées de textes par le biais d'un tableau de métadonnées appelé « metadata.csv » (voir section ci-dessous). En TEI les métadonnées peuvent provenir d'un encodage dans le teiHeader. Remarque : le fichier « metadata.csv » n'est pas obligatoire pour un import, donc l'utilisateur peut l'encoder et le fournir ou non suivant ses besoins d'analyse. Ceci constitue une deuxième dimension de progression philologique ;
- Mots : en texte « brut », les mots ne peuvent qu'être calculés automatiquement par TXM en fonction du système d'écriture utilisé par le corpus. En XML et en XML-TEI, il est possible de pré-coder de façon précise la délimitation et les propriétés de certains ou de tous les mots du corpus. Remarque : le pré-codage optionnel de mots, à ne réaliser que si cela est utile à l'analyse, est une troisième dimension de progression philologique ;
- Structures : n'ayant pas de mécanisme de délimitation particulier, le texte brut ne peut pas disposer de structures textuelles. En XML tout venant, toutes les délimitations par balises correspondent à des structures textuelles (sauf la balise <w> réservée aux mots). En TEI certaines balises correspondent aux structures textuelles. Remarque : En XML, le balisage progressif, à ne réaliser que si cela est utile à l'analyse, est une quatrième dimension de progression philologique ;
- Plans textuels : le texte brut n'ayant pas de mécanisme de délimitation aucun plan textuel n'est mobilisable. En XML, le module d'import XML/w+CSV propose d'appliquer n'importe quelle feuille de transformation XSLT sur les sources avant de les traiter (la XSL frontale). De cette façon, la XSL va permettre de calculer par exemple le « hors texte » à la volée. Le module d'import XTZ+CSV offre 4 phases successives de transformation XSLT pour offrir plus de possibilités de manipulation de plans textuels. En encodage XML-TEI, certaines balises pourront correspondre à différents plans textuels.

4.3 Panorama des modules d'import et des niveaux de représentation

La figure 4.1 illustre l'imbrication des trois niveaux de représentation. L'imbrication d'un niveau dans un autre signifie qu'il est de même nature et bénéficie des mêmes services que celui dans lequel il est imbriqué. Plus un niveau est imbriqué profondément plus la représentation correspondante est explicite et normalisée :

- le format texte brut (Unicode TXT), correspondant au périmètre le plus externe, est le format le plus élémentaire: une séquence de caractères ;

- le format XML est du texte brut ayant des contraintes supplémentaires comme la convention de la syntaxe des balises pour délimiter et qualifier des parties de texte ;
- le format TEI est du XML ayant des contraintes supplémentaires comme la convention de nommage, de positions respectives et de sémantique de balises.

Le processus d'import consiste à transformer les sources depuis un format de départ plus ou moins riche et précis (les différents formats correspondent à des rectangles blancs dans la figure), en suivant les flèches correspondants aux différents modules d'import disponibles (représentés par des rectangles à la bordure épaisse), jusqu'au format « XML-TEI TXM » qui est un format compatible avec les recommandations standard de la TEI et suffisamment spécialisé pour être traitable par les outils internes de TXM. À partir du moment où les sources d'un corpus sont dans ce format, les outils d'indexation et de construction d'éditions peuvent finaliser directement la représentation interne du corpus (représentée par le rectangle [TXM - CQP / HTML] dans le cercle brun le plus foncé). La représentation XML-TEI TXM est la plus explicite et la plus normalisée de l'environnement d'import de TXM, et tous les corpus importés dans TXM sont représentés sous cette forme normalisée à un moment donné quel que soit leur format source de départ.

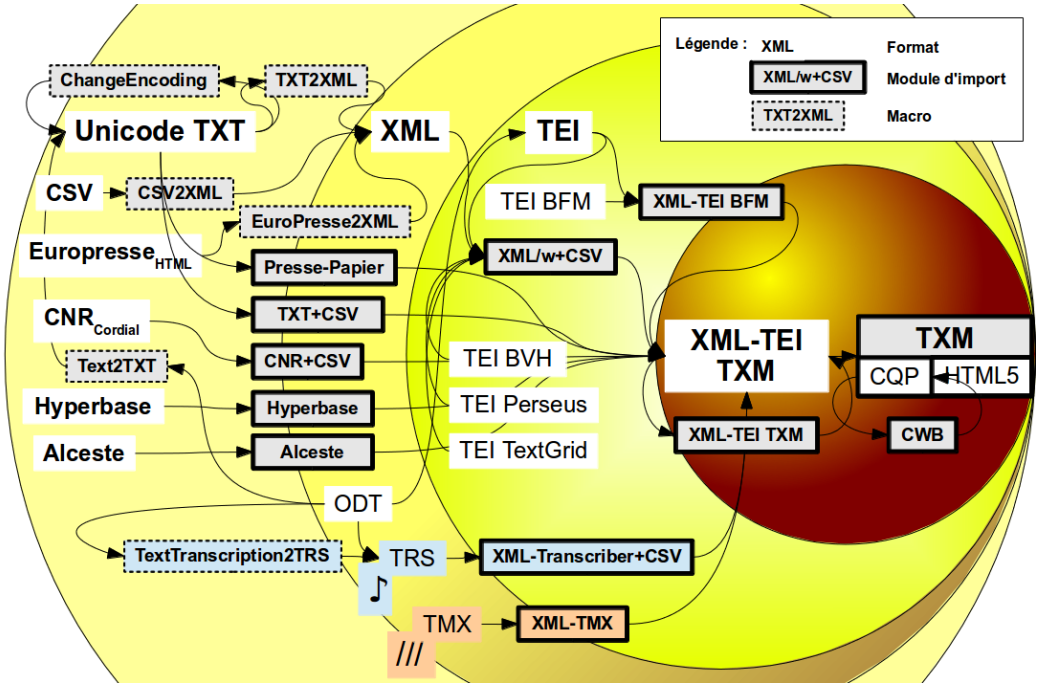


Illustration 4.1: Panorama des modules d'import

4.4 Enchaînement canonique des opérations d'un module d'import

Le module d'import de documents au format texte brut (TXT + CSV) correspond typiquement à la succession des opérations de normalisation suivante :

1. on récupère les fichiers d'extension « .txt » du répertoire source ;
2. on les transforme en XML et on ajoute éventuellement les métadonnées aux textes à partir du fichier « metadata.csv » qui se trouve dans le répertoire des sources ;
3. on crée la version XML-TEI TXM des textes ;
4. on crée une représentation des textes pour leur appliquer le logiciel TreeTagger et on injecte le résultat du logiciel dans les représentations XML-TEI TXM ;
5. on indexe les mots et on crée les éditions de textes ;
6. à ce moment là le corpus est disponible dans TXM pour être partitionné, créer des sous-corpus, manipuler les structures internes des textes et les propriétés de mots.

4.5 Création d'un corpus par appel d'un module d'import

Pour créer un nouveau corpus dans TXM on lance un module d'import à l'aide du menu « Fichier > Importer ». Tous les modules d'import prennent en entrée un chemin vers le répertoire contenant les fichiers sources du corpus sauf le module d'import « Presse-papier » qui utilise directement le contenu du presse-papier du système d'exploitation.

4.5.1 Import à partir du presse-papier

L'import presse-papier est la façon la plus simple et la plus limitée de créer un corpus dans TXM.

Il suffit d'abord de copier du texte dans n'importe quelle application (traitement de texte – MS Word ou OO Writer, lecteur PDF – Adobe Acrobat, navigateur web – FireFox ou Internet Explorer, logiciel de messagerie, etc.), ce qui se fait en général par : sélection du texte puis 'Éditer > Copier' ou raccourcis clavier Control-C (Command-C en Mac), puis ensuite de lancer la commande « Fichier > Importer > Presse-papier ».

Le résultat est un nouveau corpus ayant un nom calculé automatiquement « PRESSEPAPIER* » composé d'un seul texte (factice) contenant les mots trouvés dans le presse-papier considéré comme du texte brut. Si TreeTagger est installé et configuré pour TXM les mots sont également lemmatisés selon la langue choisie dans la préférence « TXM / Utilisateur / Import / Clipboard / Default language ».

4.5.2 Modules d’import à partir de fichiers sources

Chaque format de source (TXT, XML, TEI, etc.) correspond à un module d’import du menu « Fichier > Importer » :

Module	Format
TXT + CSV	fichiers de texte brut (.txt) + tableau de métadonnées (metadata.csv)
ODT/DOC/RTF + CSV	fichiers de traitement de texte (.doc, .odt, etc.)
XML/w + CSV	fichiers XML (.xml)
XTZ + CSV	fichiers XML avec des balises TEI optionnelles
XML-TEI BFM	fichiers XML-TEI de la Base de Français Médiévale (BFM)
XML-TEI Frantext	fichiers XML-TEI libres de droits de Frantext
XML-TEI TXM	fichiers XML-TEI normalisés pour TXM
XML Transcriber + CSV	fichiers XML de transcription selon le schéma XML du logiciel Transcriber
XML Factiva	fichiers exportés au format XML depuis le portail Factiva
XML-TMX	fichiers XML de corpus multilingues alignés (mémoires de traduction)
Factiva TXT	fichiers exportés au format TXT depuis le portail Factiva
CNR + CSV	fichiers résultat du logiciel Cordial
Alceste	fichiers au format étoilé (****) des logiciels IRaMuTeQ et Alceste
Hyperbase	fichiers au format Hyperbase (&&&)

CQP fichiers au format tabulé de CWB-CQP

Le lancement d'un module d'import provoque l'ouverture de son formulaire de paramètres, similaire à celui de l'illustration 4.2) :

Illustration 4.2: Formulaire des paramètres d'import du module TXT + CSV.

Pour remplir le formulaire il faut d'abord commencer par sélectionner le dossier qui contient les fichiers sources, en cliquant sur l'icône « dossier » ou en cliquant sur le lien hypertexte « Sélectionner le répertoire des sources ».

On peut ensuite renseigner les autres paramètres, il faut ouvrir les différentes sections de paramètres en cliquant sur leur intitulé pour y accéder :

- Nom du corpus : c'est l'identifiant dans TXM qui sera notamment affiché dans la vue corpus. Il doit obéir à un format très strict : il ne doit être composé que de majuscules non accentuées ou des chiffres et ne pas commencer par un chiffre. Tant que le nom n'est pas conforme à ce format, l'import ne peut pas commencer ;
- Description : une description optionnelle du corpus en format libre (nom complet, auteur, date de production, numéro de version, licence de diffusion, commentaire, etc.). On peut utiliser des balises HTML pour la mise en forme (mise en gras, italique, intertitres, etc.).

L'affichage exact des sections suivantes dépend du module d'import utilisé :

- Encodage des caractères : à préciser si l'encodage des caractères des textes sources est différent d'Unicode UTF-8[18]. Le système d'encodage des caractères par défaut des textes varie suivant les systèmes d'exploitation :
  - Windows : en général « windows-1252 » ou « cp1252 » ;
  - Mac OS X : « x-MacRoman » ou « MacRoman » ;
  - Linux : « UTF-8 »

    Si l'encodage varie en fonction des textes, ou que vous ne savez pas lequel choisir, vous pouvez sélectionner l'option « deviner » qui essaiera de déterminer l'encodage des textes automatiquement[19] ;
- Langue principale[20] : utilisée pour les tris lexicographiques et le choix du modèle linguistique utilisé par TreeTagger quand l'option « Annoter le corpus » est sélectionnée. Comme pour le paramètre d'encodage des caractères, l'option « deviner » essaiera de déterminer automatiquement la langue des textes[21].
- Segmentation lexicale : vous pouvez régler le comportement du repérage des mots en modifiant quelque-uns de ses paramètres. Pour connaître les noms et valeurs par défaut de ces paramètres, voir le script : https://txm.svn.sourceforge.net/svnroot/txm/trunk/Toolbox/trunk/org.textometrie.toolbox/src/groovy/org/txm/tokenizer/TokenizerClasses.groovy)
- Éditions : pour générer ou non des éditions de chaque texte du corpus (il est pratique de ne pas créer systématiquement des éditions quand on teste l'import d'un corpus de grande taille), le nombre de mots par page pour la pagination automatique ou le nom de la balise XML délimitant les pages dans le cas de modules d'import basés sur le format XML (balise <pb> par défaut).
- Feuille XSL d'entrée : dans le cas de modules d'import basés sur le format XML, avant toute lecture des sources, TXM peut leur appliquer au préalable des feuilles de transformation XSLT[22].

---

[18] La macro « ChangeEncoding » permet si nécessaire de modifier par lots l'encodage des caractères de tous les fichiers sources d'un corpus situés dans un dossier. Elle est à utiliser depuis TXM sur un dossier de sources donné, avant de procéder à l'importation du corpus. Elle est documentée dans la page de documentation des macros de TXM : https://groupes.renater.fr/wiki/txm-users/public/macros#changeencoding. Une fois les sources encodées en Unicode UTF-8, il n'est plus nécessaire de régler le paramètre d'import « Encodage des caractères ».

[19] L'algorithme de recherche de l'encodage est d'abord lancé sur l'ensemble des textes pour trouver une valeur générale. Puis texte par texte. Si un texte est trop petit ce sera la valeur générale qui sera utilisée.

[20] Le code de la langue suit le standard ISO 639-1 : http://fr.wikipedia.org/wiki/Liste_des_codes_ISO_639-1.

[21] Voir 14

[22] Voir la page de documentation des feuilles XSL préparées pour TXM : https://groupes.renater.fr/wiki/txm-users/public/xsl.

- Commandes : permet de régler le comportement de certaines commandes.
  - Noms des structures délimitant les contextes de concordance. Par défaut, ce champ ne contient que la structure « text » (les contextes de concordances ne vont pas au delà des limites de chaque texte). On peut par exemple limiter les contextes de concordances de corpus de transcriptions aux limites des tours de parole en utilisant la structure « sp ». Dans ce cas le paramètre prend la valeur « text,sp » pour combiner les limites de ces deux structures.
- Police d'affichage : choix d'une police de caractères particulière pour l'affichage des résultats et des éditions (utile pour les éditions de textes en langue ancienne).

Tous les paramètres d'importation sont sauvegardés dans un fichier nommé « import.xml » dans le dossier des sources.

Une fois les paramètres renseignés, on lance l'import en cliquant sur le bouton vert avec la flèche ou en cliquant sur le lien hypertexte « Lancer l'import du corpus ».

Le résultat est un nouveau corpus ajouté à la vue « Corpus » auquel on peut appliquer toutes les commandes TXM : Description, Lexique, Concordances, Édition, etc.

4.6 Exporter ou charger un corpus binaire

TXM peut exporter un corpus dans un fichier au format appelé « corpus binaire » facile à copier ou à transmettre et à charger dans un autre TXM.

- **Exporter** : Pour transférer un corpus déjà importé vers une autre installation de TXM (sur un ordinateur différent, par exemple). Sélectionner le corpus dans la vue Corpus et lancer la commande « **Fichier / Exporter** ». Le fichier produit a une extension « .txm ». Il peut être transféré par mail ou copié sur une clé USB par exemple.
- **Charger** : Si vous avez transféré un fichier corpus TXM au format binaire (.txm) sur votre machine, vous pouvez le charger rapidement dans votre TXM à l'aide de la commande « **Fichier / Charger** ». Le résultat de cette commande est un nouveau corpus dans votre vue Corpus. Cette commande est beaucoup plus rapide que celle d'import car elle n'analyse pas les sources du corpus. Il suffit de l'exécuter une seule fois pour que le corpus soit définitivement chargé dans TXM.

4.7 Exporter les sources d'un corpus au format standard XML-TEI P5

Lors du processus d'import, tous les fichiers sources d'un corpus (quel que soit leur format d'origine) ont été encodés dans le format pivot « XML-TEI TXM »[23] compatible TEI dans le dossier de transit du corpus binaire. Vous trouverez ce dossier en suivant le chemin suivant :

- Sous Windows :

« `C:\Utilisateurs\<identifiant de l'utilisateur>\TXM\corpora\<nom du corpus>\txm` »
ou bien
« `C:\Documents and Settings\<identifiant de l'utilisateur>\TXM\corpora\<nom du corpus>\txm` »

- Sous Mac OS X :

« `/Users/<identifiant de l'utilisateur>/TXM/corpora/<nom du corpus>/txm` »

- Sous Linux :

« `/home/<identifiant de l'utilisateur>/TXM/corpora/<nom du corpus>/txm` ».

Ces fichiers, produits intermédiaires du processus d'importation, peuvent être utilisés comme fichiers d'échange de sources entre partenaires, pour l'importation de sources dans d'autres logiciels ou bien pour le stockage pérenne des sources. Ils peuvent également être ré-importés dans TXM avec le module d'import XML-TXM.

5 Modules d'import

5.1 Fichier de métadonnées « metadata.csv »

Les modules nommés « XXX+CSV » sont des modules qui peuvent associer à chaque texte du corpus des métadonnées définies dans un fichier CSV[24]. Ce fichier doit être au format suivant :

- le fichier se nomme « metadata.csv » ;
- le séparateur de colonne est « , » ;
- le séparateur de texte est « " » ;
- l'encodage des caractères doit être UTF-8[25] ;
- la première ligne - d'entête - sert à nommer chaque métadonnée ;

[23] Le format XML-TXM est une extension du format XML-TEI P5 qui représente efficacement le modèle de corpus traité par TXM : textes@métadonnées / structures@propriétés / mots@propriétés.

[24] Les fichiers CSV peuvent être édités et exportés avec les tableurs Calc ou Excel.

[25] Voir The Unicode Consortium : http://www.unicode.org.

- la première cellule de la première ligne - contenant « id » (en minuscule) - **doit obligatoirement être renseignée**. Elle définit la métadonnée « id » qui nommera chaque fichier de texte sans son extension ;
- les cellules suivantes de la première ligne définissant les autres métadonnées sont nommées librement, mais doivent respecter quelques contraintes :
  - le nom est en minuscules
  - sans caractère spécial (par exemple : .,@ç%"#~&) ;
- chacune des lignes suivantes du fichier (en dehors de la première) définit les valeurs des métadonnées d'un seul texte, en commençant dans la première colonne avec le nom du fichier contenant le texte (sans extension : « .txt », « .xml », « .cnr », etc.) et en continuant dans les colonnes suivantes avec les valeurs des métadonnées du texte.

5.1.1 Exemple de fichier « metadata.csv »

Voici les trois premières lignes du fichier « metadata.csv » du corpus exemple DISCOURS.

```
"id","loc","type","date"
01_DeGaulle,de Gaulle,Allocution radiotélévisée,27/06/1958
02_DeGaulle,de Gaulle,Allocution radiotélévisée,28/12/1958
03_DeGaulle,de Gaulle,Allocution radiotélévisée,30/01/1959
```

Pour que la relation entre les métadonnées définies dans ce fichier et les textes - définis dans le dossier source du corpus - puisse s'établir, il faut que les trois premiers textes soient représentés respectivement par des fichiers nommés « 01_DeGaulle.cnr », « 02_DeGaulle.cnr » et « 03_DeGaulle.cnr » (ce corpus est importé par le module « CNR+CSV »).

5.2 Noms des fichiers source

Les noms de fichiers source sont utilisés pour construire l’identifiant unique de chaque texte d’un corpus. La gestion des noms de fichiers est variable selon les systèmes d’exploitation. Il est recommandé :

- de ne pas utiliser de point (.), comme dans 'p.', dans les noms de fichiers ;
- de ne pas utiliser l'espace ( ), comme dans 'p. 9', dans les noms de fichiers ;
- de ne pas utiliser de caractères à diacritiques (accent, cédille), comme dans 'français', dans les noms de fichiers.

5.3 Module Presse-papier

5.3.1 Entrée

Ce module importe le texte brut copié dans le presse-papier du système.
La propriété « lb » est ajoutée aux mots pour encoder le numéro de ligne.

5.3.2 Sortie

En sortie, on obtient une structure de texte (`text`) et des mots segmentés par les caractères séparateurs.

5.3.3 Annotation

Des annotations morphosyntaxiques et le lemme sont ajoutés avec TreeTagger.
Le modèle linguistique utilisé par TreeTagger est précisé par la préférence générale « Default language ».
Pour régler la langue d'annotation par défaut pour l'import presse-papier :
- aller à la page des préférences « Preferences > Import » ;
- saisir le code de langue dans le champ « Default language » : par exemple « fr » pour le français ;
- terminer en cliquant sur « OK ».

5.3.4 Édition

Il y a une édition du texte tenant compte de la gestion des espaces et ponctuations entre mots, et paginée par blocs de n mots.

5.4 Module TXT+CSV

5.4.1 Entrée

Corps de texte

Ce module importe un dossier de fichiers[26] contenant du texte tout venant (format texte brut). L'extension de fichier correspondante est '`.txt`' par défaut.

Les sauts de ligne sont interprétés et chaque mot encode son numéro de ligne dans la propriété « `lb` ».

Métadonnées de texte

[26] Le contenu des sous-dossiers éventuels sera également importé.

Les métadonnées des textes sont encodées dans un fichier au format CSV nommé « `metadata.csv` » situé dans le même dossier que les fichiers sources.
Le séparateur de colonnes est « , ». Le caractère de champ est « " ».

La première colonne doit être nommée « id », les suivantes sont nommées à la discrétion de l'utilisateur mais sans utiliser de caractères accentués ou spéciaux.
La première colonne doit contenir le nom du fichier source (sans extension) qui correspond aux métadonnées de la ligne.

5.4.2 Sortie

En sortie, on obtient des structures de texte (`text`) ayant des propriétés correspondant aux métadonnées, et des mots segmentés par des caractères séparateurs par défaut.

5.4.3 Annotation

Des annotations morphosyntaxiques et le lemme sont ajoutés avec TreeTagger.

5.4.4 Édition

Il y a une édition par texte paginée par blocs de n mots. La première page d'édition de chaque texte reprend la liste des métadonnées.

5.5 Module CWB

5.5.1 Entrée

Ce module importe un fichier WTC ou VRT[27] directement. Si un fichier de déclaration (nommé « registry ») est présent, il sera analysé pour obtenir les propriétés du corpus.

5.5.2 Sortie

Un « id » est ajouté à chaque mot.

5.5.3 Édition

L'édition produite est très simple. Il y a quelques mise en formes :

- un saut de ligne à la fin de chaque élément paragraphe « p »

[27] Format d'entrée des sources d'un corpus pour les outils d'indexation du moteur de recherche CQP. Voir « The IMS Open Corpus Workbench (CWB) Corpus Encoding Tutorial », http://cwb.sourceforge.net/files/CWB_Encoding_Tutorial.pdf.

- un saut de ligne si un élément « lb » ou « br » est rencontré.

5.6 Module XML/w+CSV

5.6.1 Entrée

5.6.1.1 Corps de texte

Ce module importe un dossier de fichiers[28] contenant du texte au format XML. L'extension de fichier correspondante est '`.xml`' par défaut.

Chaque balise XML encode un niveau de structure avec ses propriétés. Le nom de la structure vient du nom de la balise et les propriétés et leur valeur viennent des attributs et de leur valeur. La balise « `text` » est réservée pour ce module.

Si des mots sont délimités par des balises « `<w>` » portant des attributs, ils sont interprétés en tant que tels. Il faut toutefois faire attention à ce que tous les `<w>` aient les mêmes noms d'attributs. Si les balises <w> possèdent un attribut « ref », alors celui-ci sera utilisé pour afficher les références par défaut des concordances.

Avertissements

1. Tous les éléments XML et leurs attributs sont importés dans TXM sous la forme de structures avec propriétés et de mots avec propriétés par le moteur de recherche CQP. Comme la syntaxe du langage de requêtes (CQL) de ce dernier utilise des mots-clés réservés, il n'est pas possible d'utiliser des noms de structures et de propriétés correspondant à ces mots-clés. La liste des mots que l'on ne peut pas utiliser pour nommer une structure ou une propriété, et donc pour nommer un élément ou un attribut XML, est la suivante : `asc`, `ascending`, `by`, `cat`, `cd`, `collocate`, `contains`, `cut`, `def`, `define`, `delete`, `desc`, `descending`, `diff`, `difference`, `discard`, `dump`, `exclusive`, `exit`, `expand`, `farthest`, `foreach`, `group`, `host`, `inclusive`, `info`, `inter`, `intersect`, `intersection`, `join`, `keyword`, `left`, `leftmost`, `macro`, `maximal`, `match`, `matchend`, `matches`, `meet`, `MU`, `nearest`, `no`, `not`, `NULL`, `off`, `on`, `randomize`, `reduce`, `RE`, `reverse`, `right`, `rightmost`, `save`, `set`, `show`, `size`, `sleep`, `sort`, `source`, `subset`, `TAB`, `tabulate`, `target`, `target[0-9]`, `to`, `undump`, `union`, `unlock`, `user`, `where`, `with`, `within`, `without`, `yes` ;
2. La gestion de la récursion des structures par CQP peut interférer avec des suffixes de numérotation de noms d'éléments et d'attributs. Par exemple quand un trois éléments ' ' sont imbriqués, l'indexation CQP les recodera sous les noms (non récursifs)

[28] Le contenu des sous-dossiers éventuels sera également importé.

‘div, div1, div2’. Il vaut donc mieux éviter de nommer les éléments et attributs XML avec des suffixes numériques ;

3. Le caractère underscore (_) étant réservé dans le langage de requêtes CQL, il ne peut pas être utilisé dans les noms de structures et de propriétés donc d’éléments et attributs XML.

5.6.1.2 Métadonnées de texte

Les métadonnées des textes sont encodées dans un fichier au format CSV nommé « `metadata.csv` » situé dans le même dossier que les fichiers sources.
Le séparateur de colonnes est « , ». Le caractère de champ est « " ».

5.6.1.3 Paramètres supplémentaires

Un fichier «import.properties » situé dans le dossier des sources permet de régler les paramètres suivants :

- stopifmalformed : interrompt l'import si un des fichier XML est mal formé.
- ignoredelements : expression régulière des noms de balises que le tokenizeur doit ignorer (le « hors-texte »). Par exemple : "note|teiHeader"
- normalizemetadata : normaliser les valeurs des propriétés de structures = true/false. Les valeurs de métadonnée de texte seront passées en minuscules.
- sortmetadata : nom de la métadonnée utilisée pour définir l'ordre entre les textes

Exemple de fichier « import.properties » :

```
stopifmalformed =false
ignoredelements =note|teiHeader
normalizemetadata=true
sortmetadata=true
```

5.6.1.4 Prétraitements XSL front

Le module d’import XML/w permet d’appliquer une feuille de transformation XSL à l’ensemble des sources avant tout traitement par le module d’import. Ces traitements permettent d’adapter à la volée le format des sources aux besoins du module d’import de TXM.

TXM est livré avec une bibliothèque de feuilles XSL utiles à ces prétraitements.

Feuilles d'adaptation de sources XML-TEI P5

- **txm-filter-teip5-xmlw-preserve.xsl** : rend compatible n'importe quel document au format XML-TEI P5 pour un import dans TXM. Par défaut, elle supprime le contenu

des éléments `<teiHeader>` et `<facsimile>` et laisse tous les autres éléments inchangés.
Il est possible d'appliquer cette feuille de style indépendamment du module d'import XML/w+CSV, par exemple avec l'aide de la macro ExecXSL, avec les paramètres suivants :

- deleteAll : liste des noms de balises à supprimer avec leur contenu, les noms sont séparés pas des barres verticales (« `teiHeader|facsimile` » par défaut)
- deleteTag : liste des noms de balises à supprimer en conservant leur contenu, les noms sont séparés pas des barres verticales (liste vide par défaut)

- **txm-filter-teip5-xmlw-simplify.xsl** : rend compatible n'importe quel document au format XML-TEI P5 pour un import dans TXM en ne gardant que les balises `<ab>`, `<body>`, `<div>`, `<front>`, `<lb>`, `<p>`, `<pb>`, `<s>`, `<TEI>`, `<text>` et `<w>` dans le corps du texte.
  Il est possible d'appliquer cette feuille de style indépendamment du module d'import XML/w+CSV, par exemple avec l'aide de la macro ExecXSL, avec les paramètres suivants :
  - deleteAll : liste des balises à supprimer avec leur contenu, les noms des balises sont séparés pas des barres verticales (« `teiHeader|facsimile` » par défaut)
  - copyAll : liste des balises à conserver, les noms des balises sont séparés pas des barres verticales (« `ab|body|div|front|head|lb|p|pb|s|TEI|text|w` » par défaut) ;

  Toutes les autres balises sont supprimées, leur contenu textuel est en revanche conservé.
- **txm-filter-bnc_oral-xmlw.xsl** : adapte les transcriptions de l'oral du BNC pour un traitement dans TXM.
  - Projette dans des attributs de balise `<div>` le contenu de certaines métadonnées du teiHeader (pour faciliter les contrastes internes entre types d'activités) :
    - `titleStmt/title`
    - `profileDesc/creation`
    - `classCode[@scheme='DLEE']`
    - `setting/placeName`
    - `setting/locale`
    - `setting/activity`

        - `setting/activity/@spont`
        - `recording/@date ou profileDesc/creation`
    - Projette dans des attributs de la balise `<u>` le contenu de certaines métadonnées du `teiHeader` (pour faciliter les contrastes internes entre types de locuteurs) :
        - `profileDesc/particDesc/person[…]/@*`

Feuilles d'adaptation de corpus particuliers

- **txm-filter-corpusakkadien-xmlw_syllabes-cuneiform.xsl** : filtre de choix des unités lexicales au niveau des syllabes ou au niveau des mots du corpus de tablettes cunéiformes en Akkadien ;
- **txm-filter-perseustreebank-xmlw.xsl** : filtre d'adaptation des textes du projet Perseus Treebank ;
- **txm-filter-qgraal_cm-xmlw.xsl** : filtre d'adaptation du format diffracté de la Quête du Graal ;
- **txm-filter-rnc-xmlw.xsl** : filtre d'adaptation des textes du corpus National Russe ;
- **txm-filter-teibrown-xmlw.xsl** : filtre d'adaptation des textes du corpus BROWN du projet NLTK/Brown ;
- **txm-filter-teibvh-xmlw.xsl** : filtre d'adaptation des textes TEI du corpus BVH ;
- **txm-filter-teicorpustextgrid-xmlw.xsl** : filtre d'adaptation des textes TEI du corpus TextGrid de DARIAH-DE ;
- **txm-filter-teifrantext-xmlw.xsl** : filtre d'adaptation des textes du corpus Frantext libre.
- **txm-filter-teiperseus-xmlw.xsl** : filtre d'adaptation des textes TEI du projet Perseus après conversion en TEI P5.

5.6.2 Édition

Il y a une édition par texte, paginée par défaut par blocs de n mots.

5.6.2.1 Interprétation des éléments XML pour construire l'édition

Ce module d'import interprète certains éléments XML pour mettre en forme les éditions (HTML) :

- Élément `text`
    - produit un titre `h3` avec le contenu de l'attribut `@id` au début de la première page de l'édition
    - suivi d'un tableau des métadonnées du texte : nom, valeur
    - force un saut une ligne après le tableau

- Élément `head`
  - produit un titre `h2` avec le contenu de l'élément
- Élément `note`
  - produit un appel de note avec un `span` contenant le texte « `[*]` » en rouge affichant un tooltip composé du contenu des sous-éléments `w/form` et du texte hors `w`
  - par défaut, le contenu des notes (le contenu du tooltip) est segmenté et indexé par le moteur de recherche (mais le retour au texte est impossible à partir de concordances). Pour que le contenu des notes ne soit pas indexé, il faut ajouter l'élément `note` au paramètre d'import ignoredelements (cf. section #16.5.1. Entrée).
- Élément `graphic`
  - produit un élément `div` contenant un élément `img` ayant l'attribut `@src` à la valeur de l'attribut `@url` de l'élément `graphic` et l'attribut `@align` à la valeur `'middle'`. Si l'attribut `@url` n'est pas renseigné, cet élément est ignoré.
- Éléments `lg`, `p` et `q`
  - produit un paragraphe `p` ayant l'attribut CSS `@class` à la valeur de l'attribut `@rend` de l'élément d'origine. Si l'élément d'origine n'a pas d'attribut `@rend` renseigné, le paragraphe n'a pas de style particulier.
- Éléments `lb` et `br`
  - force un saut de ligne `br`
- Élément `pb` (ou la balise de pagination indiqué par le paramètre d'import editionpage)
  - clos la page courante et ouvre une nouvelle page de l'édition. Si cet élément apparaît au sein d'une imbrication d'éléments, ces derniers sont refermés avant de clore la page courante puis ré-ouverts au début de la nouvelle page.
  - numérote la nouvelle page par un paragraphe `p` centré en haut de la page. Le numéro de page de format « `- n -` » est affiché en rouge à partir du contenu de l'attribut `@n` de l'élément `pb`. Si l'élément `pb` n'a pas d'attribut `@n` on affiche un numéro de page automatique incrémenté à partir de 1.

5.6.2.2 Stylage par CSS

Il est possible de personnaliser le stylage des pages HTML par CSS de deux façons différentes.

A) fichier CSS du corpus : on peut déposer un fichier nommé « MONCORPUS.css » dans le répertoire de l'édition par défaut du corpus `$TXMHOME/corpora/MONCORPUS/html/default`.

Cette CSS est déclarée dans chaque page HTML par la déclaration suivante :

```
<link rel="stylesheet" type="text/css" href="MONCORPUS.css"/>
```

Cette feuille de style CSS définit globalement le style de chaque élément HTML ou bien des classes qui seront associées aux différents types de `lg`, `p` et `q`.

B) fichier CSS de TXM : on peut modifier la feuille de style par défaut de TXM `$TXMHOME/css/txm.css`. Cette CSS est déclarée dans chaque page HTML par la déclaration suivante :

```
<link rel="stylesheet" type="text/css" href="txm.css"/>
```

Attention : la modification de ce fichier impactera le style de toutes les éditions produites ultérieurement par les modules d'import de TXM.

5.7 Module XTZ+CSV

Le module XTZ est capable d'interpréter un jeu minimal de balises TEI appelé « TEI Zero » (dans la lignée des jeux de balises TEI minimaux déjà existants TEI lite, TEI tite, TEI bare ou TEI Simple). Les balises interprétées servent à construire les données habituellement exploitées par TXM dans l'indexation des mots, dans la construction des éditions, etc. Cet import est progressif au sens où il n'est pas nécessaire d'encoder toutes les balises du jeu disponible dans un corpus donné pour pouvoir être importé par ce module. L'utilisateur n'encode que les balises qui lui sont nécessaires dans l'exploitation avec TXM.

Ce module remplace le module XML/w+CSV comme module interprétant des balises a priori et de façon progressive, et il s'utilise dans le même esprit.

5.7.1 Balises TEI interprétées

5.7.1.1 Unités textuelles

`text`

- les attributs `text@MM` présents deviennent des métadonnées ;
- ces métadonnées sont fusionnées avec celles du fichier *metadata.csv ;*
- cette balise est la seule balise obligatoire.

5.7.1.2 Unités lexicales

`w`

- `w` pré-encode certains ou tous les mots ;
- les attributs `w@PP` deviennent des propriétés de mots.

5.7.1.3 Autres éléments

Tous les autres éléments XML (les autres balises) sont transférés tels quels comme structures intermédiaires entre l'unité textuelle et les unités lexicales, leurs attributs devenant les propriétés des structures. Ces éléments ne sont pas disponibles dans les éditions par défaut, produites par le module. En revanche, ils peuvent être conservés dans les éditions (sous la

forme d'éléments HTML `span` avec l'attribut `@class`, par exemple) dans les éditions produites par des feuilles de style XSLT (voir plus bas).

Avertissements

1. Tous les éléments XML et leurs attributs sont importés dans TXM sous la forme de structures avec propriétés et de mots avec propriétés par le moteur de recherche CQP. Comme la syntaxe du langage de requêtes (CQL) de ce dernier utilise des mots-clés réservés, il n'est pas possible d'utiliser des noms de structures et de propriétés correspondant à ces mots-clés. La liste des mots que l'on ne peut pas utiliser pour nommer une structure ou une propriété, et donc pour nommer un élément ou un attribut XML, est la suivante : `asc`, `ascending`, `by`, `cat`, `cd`, `collocate`, `contains`, `cut`, `def`, `define`, `delete`, `desc`, `descending`, `diff`, `difference`, `discard`, `dump`, `exclusive`, `exit`, `expand`, `farthest`, `foreach`, `group`, `host`, `inclusive`, `info`, `inter`, `intersect`, `intersection`, `join`, `keyword`, `left`, `leftmost`, `macro`, `maximal`, `match`, `matchend`, `matches`, `meet`, `MU`, `nearest`, `no`, `not`, `NULL`, `off`, `on`, `randomize`, `reduce`, `RE`, `reverse`, `right`, `rightmost`, `save`, `set`, `show`, `size`, `sleep`, `sort`, `source`, `subset`, `TAB`, `tabulate`, `target`, `target[0-9]`, `to`, `undump`, `union`, `unlock`, `user`, `where`, `with`, `within`, `without`, `yes` ;
2. La gestion de la récursion des structures par CQP peut interférer avec des suffixes de numérotation de noms d'éléments et d'attributs. Par exemple quand un trois éléments ' ' sont imbriqués, l'indexation CQP les recodera sous les noms (non récursifs) 'div, div1, div2'. Il vaut donc mieux éviter de nommer les éléments et attributs XML avec des suffixes numériques ;
3. Le caractère underscore (_) étant réservé dans le langage de requêtes CQL, il ne peut pas être utilisé dans les noms de structures et de propriétés donc d'éléments et attributs XML.

5.7.2 Éditions

Toutes les pages d'édition sont encodés en HTML5 + CSS3 + Javascript.

5.7.2.1 Production de l'édition par défaut

Page de garde

L'édition possède une page de garde contenant :

- un titre `h3` contenant la valeur de `text@id` ;
- le tableau des métadonnées (sauf 'id').

Intertitres

- head crée un élément h2 ;
- si head@rend est présent il est transféré dans h2@rend.

Paragraphes

- p crée un paragraphe p ;
- si p@rend est présent il est transféré dans p@rend.

Mises en évidence

- hi rend le texte en gras b ;
- si hi@rend='**i**|**italic**', hi est converti en i ;
- si hi@rend='**b**|**bold**', hi est converti en b ;
- emph rend le texte en italique i ;
- si emph@rend='**i**|**italic**', emph est converti en i ;
- si emph@rend='**b**|**bold**', emph est converti en b.

Sauts de ligne

- lb crée un saut de ligne forcé (élément br) ;

Listes à puces

- list crée une nouvelle liste à puces :
  - si list@type = unordered → liste à puces ul ;
  - si list@type = ordered → liste numérotée ol.
- si list@rend est présent il est transféré dans ul@rend ou ol@rend ;
- item crée une nouvelle entrée li ;
- si **item@rend** est présent il est transféré dans *li@rend*.

Tableaux

- table crée un nouveau tableau table *;*
- si table@rend est présent il est transféré dans table@rend *;*
- row crée une nouvelle ligne tr *;*
- si row@rend est présent il est transféré dans tr@rend *;*
- cell crée une nouvelle cellule td *;*
- si cell@rend est présent il est transféré dans td@rend *;*

Illustrations

- graphic insère une image (élément img avec img@align = “middle” - l'image est centrée -, inséré dans une div) ;
- si graphic@url est présent il est transféré dans img@src.

Liens hypertextes

- `ref` insère un lien hypertexte `a` avec `a@target` = “_blank” (le lien s'ouvre dans un nouvel onglet) ;
- si `ref@target` est présent il est transféré dans `a@href`.

Notes de bas de page

- `note` insère :
  - un appel de note `a` numéroté à partir de 1 dans la page avec `a@id`=noteref_N, `a@title`='contenu de la note' et `a@href`=note_N ;
  - une note de bas de page composée :
    - d'un lien retour vers l'appel `a@id`=note_N et `a@href`=noteref_N ;
    - du contenu de la note dans un *span.*

Pagination

- l'élément de pagination termine la page courante en fermant les listes, paragraphes, sections, etc. ouverts ;
- par défaut l'élément de pagination est `pb` (valeur du paramètre « pageBreakTag ») ;
- si un attribut `pb@facs` est renseigné et la construction de l'édition synoptique est demandée (option “Build 'facs' edition”), alors l'URL est utilisée pour accéder à l'image de la page pour construire l'édition facsimilé ;
- si on n'utilise pas d'éléments de pagination, cette dernière est réalisée en nombre de mots par page (réglable dans le formulaire de paramètres du module) ;
- crée une nouvelle page en ré-ouvrant si nécessaire certains éléments au préalable (listes, etc.) ;
- numérote la page dans l'entête avec un paragraphe centré de contenu ”- `p@n` -” en rouge.

Mots

- `w` génère un mot inséré dans un élément `span` à `span@id` unique ayant un `span@title` contenant la liste de toutes ses valeurs de propriétés.

Stylage par CSS

Il est possible de personnaliser le stylage des pages HTML par CSS de trois façons différentes.

A) répertoire de CSS du corpus : on peut créer un répertoire « css » dans le répertoire des sources et y déposer des feuilles CSS (d’extension « .css »). Le répertoire « css » et les feuilles qu’il contient seront copiés dans le répertoire de l’édition par défaut du corpus `$TXMHOME/corpora/MONCORPUS/html/default`. Il suffit alors de déclarer les CSS dans le HTML produit pour l’édition, par exemple :

```
<link rel="stylesheet" type="text/css" href="MyCSS.css"/>
```

B) fichier CSS du corpus : on peut déposer un fichier nommé « MONCORPUS.css » dans le répertoire de l'édition par défaut du corpus `$TXMHOME/corpora/MONCORPUS/html/default`.

Cette CSS est déclarée dans chaque page HTML par la déclaration suivante :

```
<link rel="stylesheet" type="text/css" href="MONCORPUS.css"/>
```

Cette feuille de style CSS définit globalement le style de chaque élément HTML ou bien des classes qui seront associées aux différents types de `lg`, `p` et `q`.

C) fichier CSS de TXM : on peut modifier la feuille de style par défaut de TXM `$TXMHOME/css/txm.css`. Cette CSS est déclarée dans chaque page HTML par la déclaration suivante :

```
<link rel="stylesheet" type="text/css" href="MONCORPUS.css"/>
```

Attention : la modification de ce fichier impactera le style de toutes les éditions produites ultérieurement par les modules d'import de TXM.

Images et Javascript

Les images et les scripts Javascript utilisés par les pages HTML d'édition peuvent être fournis à l'édition par le biais de répertoires « images », respectivement « js », situés dans le répertoire des sources. Si ces répertoires sont présents dans les sources leur contenu est transféré dans le répertoire de l'édition par défaut `$TXMHOME/corpora/MONCORPUS/html/default`. Le HTML peut alors y accéder par des URL relatives de la forme « images/image1.jpg ».

5.7.2.2 Production de l'édition "fac-similé"

Le module XTZ peut produire des éditions synoptiques affichant côte-à-côte différentes versions de chaque page :

- l'image du facsimilé de la page (son scan ou sa photo) ;
- l'édition critique de la page ;
- une autre édition de la page ;
- une traduction de la page ;

etc.

Par défaut, seule une édition simple, non synoptique, est produite à l'import.

Le module construit une édition incluant les images de pages (de fac-similé) quand on coche l'option « Construire l'édition fac-similé/Build 'facs' edition » du formulaire de paramètres d'import. L'édition des textes est alors implicitement synoptique en combinant au moins l'édition du texte de base et l'édition fac-similé.

Les sources doivent contenir des éléments XML de saut de page, dont on peut choisir le nom avec le paramètre « Balise de saut de page/Page break tag » (valeur « `pb` » par défaut).

Les images des pages peuvent se trouver sur la machine de l'utilisateur (locales) ou bien être accessibles depuis Internet (distantes).

Désignation des images de pages à partir de fichiers locaux

Toutes les images des pages d'un texte donné doivent être regroupées dans un répertoire ayant comme nom l'identifiant du texte.

Tous les répertoires d'images de pages de textes doivent être regroupés dans un répertoire de base des images du corpus.

Quand le chemin vers ce répertoire de base est fournit au paramètre « Répertoire d'images/Images directory », le module d'import ajoute ou modifie les attributs `@facs` de tous les éléments de saut de page du corpus à partir des noms de fichiers images. L'ordre alphabétique des noms de fichiers images sera utilisé pour correspondre à l'ordre des sauts de page au fil du texte. Les répertoires d'images sont recopiés dans le corpus binaire.

Désignation des images par URLs encodées dans les sources

Si le paramètre « Répertoire d'images/Images directory » est laissé vide, le module d'import va interpréter directement les valeurs des attributs `@facs` de chaque élément de saut de page.

Ces valeurs doivent être des URLs absolues ou relatives, distantes (avec le préfixe « http:// », pour désigner une image sur un serveur web) ou locales (avec le préfixe « file:// », c'est à dire désignant des fichiers se trouvant sur la machine de l'utilisateur). Les URLs ne sont pas vérifiées au moment de l'import. Il faut s'assurer de la disponibilité de l'accès aux images au moment de l'exploitation du corpus. Le module XTZ+CSV produit des éditions par défaut en interprétant certains éléments TEI.

5.7.2.3 Production d'éditions supplémentaires par XSL

L'étape « 4-edition » permet de produire des éditions supplémentaires à l'aide de feuilles de transformation XSL s'appliquant aux représentations XML-TEI TXM de chaque texte du corpus (voir la section 5.7.4 page 83).

Une édition est produite par deux XSL appliquées successivement :

- une première nommée « <n° ordre>-<nom de l'édition>-html.xsl » qui produit un fichier HTML à partir d'un fichier XML-TEI TXM ;
- une seconde nommée « <n° ordre+1>-<nom de l'édition>-pager.xsl » qui pagine l'édition en découpant le fichier HTML initial en autant de fichiers HTML que de pages.

L'édition est stockée dans le sous-répertoire « <nom de l'édition> » du répertoire « HTML » du corpus binaire. Le nom de l'édition ne doit pas contenir de tiret « - ».

Le nom de l'édition apparaitra dans le menu des éditions disponibles de l'Édition du corpus.

5.7.3 Plans textuels

Le module XTZ peut ignorer certaines balises ou certains contenus de balises lors de l'indexation pour le moteur de recherche ou lors de la production des pages d'édition.

5.7.3.1 Hors texte

Ces éléments sont supprimés entièrement en amont de l'étape de tokénisation. Ils ne sont pas disponibles pour la production des éditions, ni pour la création de références.

5.7.3.2 Hors texte à éditer

Ces éléments sont conservés, mais le texte qu'ils contiennent n'est pas tokénisé et indexé par le moteur de recherche. En revanche, ce texte est affiché dans les éditions. Exemples d'usage : Introduction à une édition scientifique, titres ajoutés par l'éditeur, entêtes TEI dont on veut utiliser des métadonnées.

5.7.3.3 Notes

Un type particulier de hors texte à éditer qui prend la forme de notes de bas de page dans les éditions par défaut.

5.7.3.4 Milestones

Le moteur de recherche CQP de TXM ne peut pas prendre en compte les éléments milestone XML. Cette option permet de déplacer l'information utile dans des propriétés de mots. Pour chaque balise milestone indiquée, on projette dans les mots :

- -id : l'identifiant du milestone précédent le mot
- -start : la distance en mots au milestone précédent le mot
- -end : la distance en mots au milestone suivant le mot

Par exemple avec la valeur de paramètre : lb, cb, pb

on ajoute à tous les mots les propriétés suivantes :

- lbid, lbstart et lbend
- cbid, cbstart et cbend
- pbid, pbstart et pbend

5.7.4 Traitements XSL intermédiaires à certaines étapes clés du traitement du module

Le module XTZ permet d'appliquer des XSL aux sources en cours de traitement lors de 4 étapes clés :

- **1-split-merge** : traitement initial permettant de changer l'architecture des fichiers, par exemple pour la rendre conforme au modèle de corpus de TXM 1 fichier = 1 texte ;
- **2-front** : prétraitement pour changer le contenu des fichiers, par exemple pour éliminer (eg teiHeader) ou transformer certains éléments ;
- **3-posttok** : traitement final de la version XML-TEI TXM pivot de chaque texte où chaque mot est déjà encodé, par exemple pour recomposer les mots césurés du corpus ;
- **4-edition** : production d'éditions HTML à partir de la version XML-TEI TXM pivot de chaque texte et de la pagination réalisée par le composant de création d'édition par défaut (le Pager).

À chacune de ces étapes clés correspond un répertoire de même nom contenant les XSL des traitements à appliquer à cette étape clé. Les XSL sont appliquées dans l'ordre lexicographique de leur nom. Les répertoires de traitements XSL intermédiaires sont regroupés dans un répertoire « xsl » situé lui-même dans le répertoire des sources. Les traitements sont déclenchés en fonction de la présence des ces répertoires et des XSL.

Chaque XSL peut utiliser les paramètres suivants :

- number-words-per-page : le nombre de mots par page si indiqué dans le formulaire d'import ;
- pagination-element : l'élément XML à utiliser pour les sauts de page si indiqué dans le formulaire d'import ;
- import-xml-path : le chemin du fichier de stockage des paramètres d'import dans le répertoire source.

Cas des XSL de production de l'édition HTML (point '4-edition') :

- elles reçoivent un paramètre supplémentaire 'output-directory' qui est le chemin du répertoire de sortie des résultats ;
- elles doivent produire un élément <meta name="description" content="{id du 1er mot de la page}"/> dans l'entête de chaque page produite. Si la page ne contient pas de mot (eg page de garde), elle doivent produire un élément <meta name="description" content="w_0"/> dans l'entête de la page.

Les DTD ou schémas utilisés par les XSL doivent être fournis dans le répertoire 'dtd' du répertoire des sources.

5.7.4.1 Bibliothèque de feuilles XSL de transformation intermédiaire

TXM est livré avec une librairie de feuilles XSL utiles aux traitements intermédiaires. Les versions les plus à jour de ces XSL se trouvent en ligne à l'adresse https://sourceforge.net/projects/txm/files/library/xsl.

1-split-merge

- t**xm-rename-files-no-dots.xsl** : remplace les points (.) dans les noms de fichiers source par un souligné (_). Utile pour contourner le bug sur les noms de fichiers source ;
- t**xm-split-teicorpus.xsl** : éclate en plusieurs fichiers un fichier source contenant un élément `teiCorpus` composé de plusieurs éléments `TEI`. Utile pour obtenir une unité textuelle et une édition par texte du corpus TEI dans TXM ;

2-front

Traitements génériques

- **filter-keep-only-select.xsl** : cette feuille de transformation générique supprime le contenu de tous les éléments XML à l'exception de ceux désignés par le paramètre « `select` » et de ses descendants (voir la ligne 43). Si le paramètre `select` n'est pas renseigné, aucune modification n'est effectuée. La valeur du paramètre `select` peut également être transmis en paramètre à la feuille XSL ;
- **filter-out-p.xsl** : cette feuille de transformation exemple supprime le contenu de tous les éléments `<p>` ayant un attribut `@type` à la valeur 'ouverture' (voir l'expression XPath de la ligne 42). Elle peut être adaptée et utilisée pour filtrer le contenu de différentes balises XML à la volée ;
- **filter-out-sp.xsl** : cette feuille de transformation exemple supprime le contenu de tous les éléments `<sp>` ayant un attribut `@who` à la valeur 'enqueteur' (voir l'expression XPath de la ligne 42). Elle peut être adaptée et utilisée pour filtrer les prises de tour de différents locuteurs à la volée ;
- **filter-number-act-scene-line.xsl** : cette feuille de transformation exemple numérote tous les actes, scènes et lignes de l'édition XML de la pièce « All's Well That Ends Well » de William Shakespeare publiée en ligne : https://www.ibiblio.org/xml/examples/shakespeare/all_well.xml.
  Elle peut être utilisée pour numéroter à la volée lors de l'import. Le pré-traitement XSL est strict, il est nécessaire de déposer dans le répertoire source (contenant la pièce) le fichier de DTD de la pièce disponible en ligne : http://www.ibiblio.org/xml/examples/play.dtd.
  Elle peut également être appliquée de façon définitive sur la pièce au préalable, avant import, avec l'aide de la macro ExecXSL ;
- **txm-front-teiHeader2textAtt.xsl** : cette feuille de transformation générique extrait des métadonnées de l'élément `teiHeader` pour les projeter sous forme d'attributs de l'élément `text`. Elle peut aider à s'affranchir d'un fichier « metadata.csv ».

`Feuilles d'adaptation de corpus particuliers`

- **p4top5_perseus.xsl** : conversion des fichiers XML du projet Perseus du format TEI P4 au format TEI P5.

3-posttok

- **txm-posttok-addRef.xsl** : cette XSL construit des références qui seront affichées par défaut dans les concordances. Elles seront composées de :
    - la valeur de l'attribut `@id` de l'élément `text` ou, le cas échéant, du nom du fichier sans extension ;
    - éventuellement, le numéro de page : l'attribut `@n` de la première balise `<pb/>` qui précède le mot (attention, ce traitement affecte sérieusement la performance sur de gros fichiers) ;
    - éventuellement, le numéro de paragraphe : l'attribut `@n` du premier ancêtre `<p>` ;
    - éventuellement, le numéro de ligne : : l'attribut `@n` de la première balise `<lb/>` qui précède le mot (attention, ce traitement affecte sérieusement la performance sur de gros fichiers).
- **txm-posttok-unbreakWords.xsl** : cette XSL permet de reconstruire certains mots qui auraient été découpés par la segmentation lexicale initiale (tokenisation), par exemple à cause de sauts de ligne ou de sauts de page situés au milieu des mots.
- **txm-filter-teitextgrid-xmlw-posttok.xsl** : This styleheet should be used to adjust word properties in the tokenized version of DARIAH-DE Textgrid texts.

4-edition

- **1-default-html.xsl** : cette XSL permet de construire une édition alternative à celle construite par le module par défaut. Elle transforme chaque élément TEI par un élément `span` ayant un attribut `@class` spécifique. Cette XSL est à utiliser conjointement avec la XSL 2-default-pager.xsl ;
- **2-default-pager.xsl** : Cette XSL est à utiliser conjointement avec la XSL 1-default-html.xsl pour créer les pages de l'édition ;
- **txm-edition-xtz-corpusakkadien-translit.xsl** : XSL a utiliser pour régler les éditions translittérées de tablettes cunéiformes en Akkadien[29] ;
- **txm-edition-xtz-cuneiform.xsl** : XSL a utiliser pour créer les éditions de tablettes cunéiformes en Akkadien.

5.7.5 Ordre des textes

L'ordre des textes d'un corpus concerne l'ordre d'apparition des occurrences dans les progressions ou dans les concordances, l'ordre des éditions de textes, etc.

Si le répertoire source du corpus contient un fichier metadata.csv et si ce dernier contient une colonne "`textorder`" alors les textes du corpus seront ordonnés selon l'ordre **alphanumérique** des valeurs de cette colonne.

[29] Voir la page de suivi du projet : https://groupes.renater.fr/wiki/txm-users/public/umr_proclac_corpus_akkadien#etape_3_facultative_mise_a_jour_de_l_edition_translitteree_affichage_des_sauts_de_lignes_et_de_traits_d_union_entre_les_syllabes.

Par exemple, pour un répertoire contenant :

- un fichier a.xml :
  ```
  <text id="a">A AA AAA</text>
  ```
- b.xml :
  ```
  <text id="b">B BB BBB</text>
  ```
- z.xml :
  ```
  <text id="z">Z ZZ ZZZ</text>
  ```
- metadata.csv :
  ```
  id      textorder
  a       003
  b       001
  z       002
  ```

L'ordre des textes sera 'b', 'z' et 'a'.

5.7.6 Tokenisation

5.7.6.1 Élément mot

Par défaut, l'élément utilisé pour pré-coder la tokenisation est « w », mais il est possible d'en choisir un autre en renseignant le paramètre “Word tag” avec le nom de l'élément à utiliser.

5.7.7 Options supplémentaires

Décocher le paramètre « suppression des répertoires intermédiaires » permet de vérifier le résultat d’un certain nombre de traitements intermédiaires de l’import en analysant le contenu de répertoires situés dans le répertoire `$TXMHOME/corpora/MONCORPUS` :

- **src** : ce répertoire contient le résultat final des étapes « split-merge » et « front » combinées ;
- **tokenized** : contient le résultat du tokeniseur et de l’étape « posttotk ». Les fichiers qui s’y trouvent sont au même format XML que les fichiers du répertoire « src » avec l’élément « w » en plus encodant chaque mot ;
- **wtc** : contient la représentation du corpus directement indexable par l’outil « cwb-encode » du moteur « CQP ». Elle peut être ré-importée très rapidement avec le module « CQP ».

5.8 Module XML-PPS

5.8.1 Entrée

Ce module prend en entrée les fichiers produit par l'export XML du portail Factiva[30]. Les fichiers sont traités puis importé à l'aide du module XML/w+CSV.

Le traitement des fichiers PPS est celui proposé par Daniel Marin et Florent Bédécarrats.

5.9 Module Transcriber+CSV

5.9.1 Entrée

Corps de texte
Ce module prend en entrée un dossier de transcriptions au format XML-TRS (extension '.trs'). Elles doivent être accompagnées du fichier « `trans-14.dtd` » pour être valides. Chaque transcription sera considérée comme une seule unité documentaire ou texte.

Les transcriptions doivent répondre au cahier des charges défini : http://sourceforge.net/projects/txm/files/documentation/Guide_de_Transcription_d_entretiens Transcriber-TXM 0.2_FR.pdf/download

Métadonnées de texte
Les métadonnées des textes sont encodées dans un fichier au format CSV
nommé « `metadata.csv` » situé dans le même dossier que les fichiers sources.
Le séparateur de colonnes est « , ». Le caractère de champ[31] est « " ».
La première ligne d'entête nomme chaque métadonnée.
La première colonne doit être nommée « id », les suivantes sont nommées à la discrétion de l'utilisateur mais sans utiliser de caractères accentués ou spéciaux.
La première colonne doit contenir le nom du fichier source (sans extension) qui correspond aux métadonnées de la ligne.
Les métadonnées seront injectées au niveau de chaque transcription, si elles sont présentes.

Paramétrage
Ce module utilise un fichier de paramètres appelé « `import.properties` » se trouvant dans le même dossier que les transcriptions.
Il permet de définir trois paramètres :

- `removeInterviewer` : vaut « true » ou « false », indique s'il faut ignorer les paroles des interviewers. Les interviewers **de chaque** texte sont définis dans les colonnes « enqN » (N un nombre) du fichier metadata.csv;

[30] Voir http://www.factiva.com.

[31] le caractère de champ permet d'encadrer des valeurs complexes contenant notamment des espaces ou des caractères délimiteurs de colonnes.

- `metadataList` : la liste des métadonnées. Chaque métadonnée est séparée de la suivante par le caractère « | », ex : `titre|date|lieu`
- `csvHeaderNumber` le nombre de lignes d'entête du fichier CSV (s'il existe) :
  - 1 = il n'y a que les identifiants des métadonnées ;
  - 2 = il y a une ligne d'identifiants et une ligne d'identifiants longs ;
  - 3 = il y a une ligne d'identifiants, une ligne d'identifiants longs puis le type de la métadonnée[32].

5.9.2 Sortie

La structure des fichiers XML de Transcriber est reproduite :

- une section Transcriber correspond à la structure div ;
- un tour de parole correspond à la structure « u » (pour 'utterance', de la TEI) ;
- un segment de parole correspond à la structure sp.

Les deux formes d'événements Transcriber sont gérées :
- ponctuels : commentaires, bruit court ;
- sur empan de mots : prononciation, incertitudes...

Les descriptions associées aux événements ponctuels sont portées par le mot suivant.
Pour les événements à empan, les descriptions sont concaténées dans la propriété lexicale « event » des mots compris entre les événements « begin » et « end ».

Certaines métadonnées sont dupliquées au niveau des mots (`spk`) et des structures (`u@spkattrs`, `textAttr@`<metadata>, `div@topic@endtime@starttime@type`, `sp@speaker@endtime@starttime@overlap`, `event@type@desc`) pour faciliter la construction de sous-corpus.

5.9.3 Annotation

Des annotations morphosyntaxiques et le lemme sont ajoutés avec TreeTagger[33].

5.9.4 Édition

L'édition reproduit celle de Transcriber. On retrouve au début de chaque texte (ou transcription) la liste des métadonnées correspondantes.
Les textes sont paginés par nombre de mots après un tour de parole.
Les événements et commentaires apparaissent entre parenthèses.
Les indications de synchronisation apparaissent entre crochets.

---

[32] Cette dernière information n'est pas utilisée dans cette version du logiciel.

[33] Pour les transcriptions en français, il est conseillé d'utiliser le modèle linguistique TreeTagger développé pour l'écrit et pour l'oral dans le cadre du projet PERCEO <http://www.cnrtl.fr/corpus/perceo>.

5.10 Module XML-TEI BFM

5.10.1 Entrée

Le format d'entrée est défini dans la documentation d'encodage de la Base de Français Médiéval (BFM). Chaque texte est représenté dans un fichier au format XML TEI P5 qui encode à la fois le corps des textes et leurs métadonnées.

En plus des fichiers de textes, un fichier de paramètres appelé « import.properties » contient les expressions XPath[34] servant à extraire les métadonnées des textes dans leur entête TEI.
Voici un exemple de contenu de ce fichier:

```
titre=/tei:TEI/tei:teiHeader/tei:fileDesc/tei:titleStmt/tei:title[@type='reference']/text()

auteur=/tei:TEI/tei:teiHeader/tei:fileDesc/tei:titleStmt/tei:author/text()
```

Glose :
- la valeur de la métadonnée « titre » correspond au contenu d'un élément <title> se trouvant à un endroit précis de l'entête TEI et dont l'attribut « type » vaut « reference » ;
- la valeur de la métadonnée « auteur » correspond au contenu d'un élément <author> se trouvant à un endroit précis de l'entête TEI.

Durant l'import, chaque mot reçoit une nouvelle propriété "ref", qui sera utilisée pour afficher la référence par défaut des concordances. Cette propriété est construite à partir de plusieurs informations :

- le sigle du texte, valeur de l'attribut text@sigle
- le numéro de paragraphe, valeur de l'attribut p@n
- si le texte contient des vers : le numéro de vers, valeur de l'attribut lb@n

La valeur de l'attribut text@sigle est construit lors de l'étape "importer" du module d'import en récupérant l'information, par ordre de préférence, dans :

- le retour de l'XPath associée au paramètre "idbfm" déclaré dans le fichier "import.properties"
- le retour de l'XPath associée au paramètre "sigle" déclaré dans le fichier "import.properties"
- le nom du fichier source XML

Attention, le module d'import ne gère pas les balises TEI `div1`, `div2`, etc. Il faut les remplacer par des balises `div` soit directement dans les sources XML, soit à l'aide d'une feuille XSL d'entrée.

Pour plus d'informations sur l'encodage des textes de la BFM :

- Manuel d'encodage XML-TEI des textes de la BFM : `http://bfm.ens-lyon.fr/article.php3?id_article=158`

[34] XML Path Language (Xpath) 2.0 : http://www.w3.org/TR/xpath20.

- Consortium de la Text Encoding Initiative : http://www.tei-c.org

5.10.2 Annotation

Des annotations morphosyntaxiques sont ajoutés avec TreeTagger au moyen du modèle linguistique « fro.par ». Le jeu d'étiquettes utilisé par ce modèle est CATTEX2009 (voir http://bfm.ens-lyon.fr/article.php3?id_article=176).

5.10.3 Édition

L'édition des textes est assez proche de celle réalisée pour le projet « Queste del Saint Graal » (voir http://portal.textometrie.org/txm). Toutefois cette partie du module sera remplacée à terme par les feuilles de styles XSLT+CSS d'Alexis Lavrentiev pour produire une édition équivalente et pérenne.

5.11 Module XML-TEI Frantext

Ce module est similaire au module XML-TEI BFM. Avant de lancer l'import XML-TEI BFM, une feuille XSL prétraite les fichiers sources :

- l'élément '
' est recodé en '<lb/>'
- les mots étoilés sont réencodés en '<w type="caps">…</w>'
- l'entête TEI est corrigée, en particulier l'élément '<auteur>' devrait être '<author>'
- les éléments '<seg>' sont réencodés en '<w>…</w>'

5.12 Module XML-TMX

5.12.1 Entrée

Ce module prend en entrée un fichier XML au format TMX[35].

5.12.2 Sortie

Le module créé un corpus par langue alignée et leurs relations.

5.12.3 Édition

Il s'agit d'une édition simple (il n'y a aucune information éditoriale dans le format TMX).

[35] Voir TMX 1.4b Specification, http://www.gala-global.org/oscarStandards/tmx/tmx14b.html.

5.13 Module XML-TXM

5.13.1 Entrée

Ce module importe les textes au format pivot XML-TXM[36] directement. Ce module ne réalise pas de tokenisation car le format XML-TXM encode déjà les mots dans des éléments « `<w>` ».

5.13.2 Sortie

Les métadonnées de textes sont issues des propriétés encodées au niveau des entêtes des textes.
Les structures intra-textuelles et leurs propriétés sont issues de toutes les éléments XML hors schéma XML-TXM.
Les mots et leurs propriétés sont issus des éléments '<w>'.

5.13.3 Annotation

Il n'y a pas d'annotation ajoutée dans ce module.

5.13.4 Édition

Il y a une édition par texte paginée par blocs de n mots.

5.14 Module CNR+CSV

5.14.1 Entrée

Corps de texte
Les textes sont des fichiers au format CNR de Cordial, c'est à dire un TSV avec comme caractère séparateur de colonne la tabulation et sans caractère de champ.
Dans l'ordre, les colonnes des fichiers CNR sont :

- para : le numéro de paragraphe
- sent : le numéro de phrase
- form : la forme graphique d'une unité lexicale
- lem : le lemme
- pos : la propriété morphosyntaxique
- func : la fonction syntaxique

Métadonnées de texte
Les métadonnées des textes sont encodées dans un fichier au format CSV nommé « `metadata.csv` » situé dans le même dossier que les fichiers sources.

[36] Voir https://txm.sourceforge.net/wiki/index.php?title=XML-TXM.

Le séparateur de colonnes est « , ». Le caractère de champ[37] est « " ».
La première ligne d'entête nomme chaque métadonnée.
La première colonne doit être nommée « id », les suivantes sont nommées à la discrétion de l'utilisateur mais sans utiliser de caractères accentués ou spéciaux.
La première colonne doit contenir le nom du fichier source (sans extension) qui correspond aux métadonnées de la ligne.

5.14.2 Sortie

On obtient en sortie des structures pour les paragraphes (p), les phrases (s) et les textes (text). Les mots sont équipés de toutes les propriétés correspondant aux colonnes CNR.

5.14.3 Annotation

Il n'y a pas d'annotation ajoutée dans cet import, il ne nécessite pas la présence de TreeTagger.

5.14.4 Édition

Les pages d'édition sont découpées par blocs de n mots. La première page d'édition de chaque texte reprend la liste des métadonnées lues dans le fichier CSV.

5.15 Module Alceste

5.15.1 Entrée

Ce module prend en entrée un fichier au format utilisé par le logiciel Alceste (format également utilisé par IRaMuTeQ - http://www.iramuteq.org). Il s'agit d'un format de texte brut utilisant quelques conventions d'encodage simples.
Pour encoder un début de texte (qui correspond à l'« uci », unité de contexte initiale dans la terminologie d'Alceste) et ses métadonnées, il y a deux façons de faire au choix :

1. une ligne de la forme : `0001 *Meta1_Val1 *Meta2_Val2... *MetaN_ValN`
2. une ligne de la forme : `**** *Meta1_Val1 *Meta2_Val2... *MetaN_ValN`

*Meta_val, déclare une métadonnée de texte dont « Meta » est le nom et « val » la valeur. Si une métadonnée de texte n'a pas de valeur, on peut utiliser la notation suivante : *Meta.
Pour TXM, les noms d'attribut sont composés uniquement de lettres sans distinction de casse tout sera ramené en minuscules) et sans diacritique (sans accent).

Pour pré-encoder un mot composé, on peut remplacer les espaces entre ses constituants par un caractère « _ ». Par exemple, « l'assemblée_nationale » peut être segmenté en deux mots : « l' » et « assemblée_nationale ».

[37] le caractère de champ permet d'encadrer des valeurs complexes contenant notamment des espaces ou des caractères délimiteurs de colonnes.

Le format Alceste propose également un moyen de coder des sections à l'intérieur des uci, sections caractérisées par une variable étoilée (notation : `-*Meta_Val` sur une ligne), mais ce module d'import TXM ne le gère pas encore.

5.15.2 Sortie

En sortie, on obtient des structures de texte (`text`) et des mots segmentés par les caractères séparateurs.

5.15.3 Annotation

Des annotations morphosyntaxiques et le lemme sont ajoutés avec TreeTagger.

5.15.4 Édition

Il y a une édition par texte paginée par blocs de n mots.

5.16 Module Hyperbase

5.16.1 Entrée

Ce module prend en entrée un fichier au format Hyperbase ancienne version. C'est à dire avec des lignes séparatrices de textes de la forme suivante :

…
```
&&& Nom du texte long, NomduTexte, NomCourt &&&
```
…

Les lignes de saut de pages (codées par « `\$` ») sont interprétées. Elles se répercutent par des structures p.

5.16.2 Annotation

Des annotations morphosyntaxiques et le lemme sont ajoutés avec TreeTagger.

5.16.3 Édition

Il y a une édition par texte paginée par blocs de n mots.

5.17 Module Factiva TXT

Ce module convertit les fichiers sources du format export Mail Factiva au format Alceste, puis applique l'import Alceste.

5.17.1 Entrée

Ce module prend en entrée le format d'export Mail de Factiva, en suivant les recommandations d'export Factiva de Pierre Ratinaud et Lucie Loubere :

- faites votre recherche classique ;
- une fois vos articles sélectionnés, au dessus de ces derniers, dans « Options d'affichage » sélectionnez « article complet et indexation » ;
- demandez à voir les articles ;
- copier la totalité du contenu dans un document .txt.

5.18 Module Factiva XML

Ce module prend en charge le format Factiva XML. Ce format n'est plus disponible pour les licences éducative et recherche de Factiva.

5.18.1 Entrée

Le module commence par restructuré les informations du le header pour quelles soient exploitables par CQP.

La suite du module applique un import XML/w+CSV sur les fichiers XML créés.

6 Les corpus exemples

La livraison de TXM comporte quelques corpus exemples, encodés dans des formats représentatifs de ce que peut traiter la plateforme. Ils sont tous diffusés sous une licence Creative Commons BY-NC-SA[38].

D'autres corpus exemples sont disponibles en ligne : https://groupes.renater.fr/wiki/txm-users/public/corpus.

6.1 Le corpus VOEUX

Le corpus «VOEUX» a été édité par Jean-Marc Leblanc du laboratoire Céditec (Centre d'étude des discours, images, textes, écrits, communication) à Créteil Val de Marne. Il est composé de 54 transcriptions de vœux présidentiels aux caractéristiques suivantes :

[38] Obligation de citation, pas d'usage commercial, diffuser selon la même licence.

- septs présidents français : Pompidou (5 discours), de Gaulle (10 voeux), Giscard (7 voeux), Mitterand (14 voeux), Chirac (12 voeux), Sarkozy (5 voeux) et Hollande (1 voeu);
- sur une période allant de 1959 à 2012.

Chaque transcription a été lemmatisée avec le logiciel TreeTagger en utilisant le modèle fr.par. Le jeu d'étiquettes morpho-syntaxiques est décrit sur le site de TreeTagger : http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html.

Le corpus est composé des éléments suivants :

- unités structurelles : text (vœu) / p (paragraphe) / s (phrase)
- chaque « text » comporte les propriétés suivantes :
  - `annee` : au format « yyyy »
  - `loc` : le nom du président
- chaque unité lexicale comporte les propriétés suivantes :
  - `word` : forme graphique du mot ;
  - `frpos` : l'étiquette morphosyntaxique de TreeTagger ;
  - `frlemma` : le lemme de TreeTagger ;
  - `lbn` : le numéro de ligne dans le fichier source ;
  - `sn` : le numéro de la phrase calculé lors de l'import ;
  - `pn` : le numéro du paragraphe calculé lors de l'import.

6.2 Le corpus GRAAL

Le corpus «GRAAL» a été édité par Christiane Marchello-Nizia et Alexei Lavrentiev, du laboratoire ICAR (UMR CNRS), à Lyon. Il est basé sur l'édition du manuscrit K (Lyon, Bibliothèque municipale, Palais des arts 77) du roman de la *Queste del saint Graal* (http://catalog.bfm-corpus.org/qgraal_cm).

Chaque mot du texte est étiqueté avec des étiquettes morphosyntaxiques du jeu CATTEX2009 (étiquettes pour l'ancien français, dont la définition est accessible à l'adresse `http://bfm.ens-lyon.fr/article.php3?id_article=176`).

L'importation de ce corpus dans la plateforme TXM encode les éléments suivants :

- unités structurelles principales (qui concernent tout ou une partie importante des mots du texte) : `p` (paragraphe) / `q` (discours direct) / `s` (phrase)
  - les unités p et s sont numérotées avec l'attribut « n »
- unités structurelles supplémentaires (qui concernent certains fragments du texte) : `add` (mots ou syntagmes ajoutés par le scribe), `corr` (mots ou syntagmes corrigés par l'éditeur), `damage` (passages endommagés dans le manuscrit), `orig` (ponctuations scribales), `subst` (mots remplacés par le scribe), `supplied` (mots ou syntagmes ajoutés par les éditeurs). Ces structures correspondent aux balises TEI utilisées dans l'édition numérique à l'origine du corpus ;
- chaque unité lexicale porte les attributs suivants :
  - `word` : la forme graphique « courante »[39] ;
  - `dipl` : la forme « diplomatique » ;
  - `n` : le numéro d'ordre du mot dans le texte ;
  - `pos` : l'étiquette morphosyntaxique ;
  - `ref` : la référence pour les concordances ;
  - `q` : le niveau d'imbrication de discours direct (de « 0 » pour des passages hors discours direct à « 3 » pour des passages au discours direct imbriqués dans deux autres niveaux de discours direct)

Les propriétés `ana`, `rend` et `type` représentent des annotations obsolètes ou partielles dans l'état actuel du corpus.

7 Outils d'analyse

Les outils sont lancés par le biais de commandes de menu, de barre d'outils ou de lien hypertextuels. En général les commandes de TXM s'appliquant à un corpus ouvrent une nouvelle fenêtre qui permet de paramétrer, lancer et parcourir le résultat du calcul. Un calcul peut être interrompu en appuyant sur le bouton « Cancel » de la fenêtre de progression.

7.1 Description d'un corpus

7.1.1 Appliquée à un corpus

Cette commande calcule une synthèse complète de la structure du corpus sélectionné : les éléments structurels, les unités lexicales et leurs propriétés :

[39] Voir l'Introduction (http://txm.ish-lyon.cnrs.fr/bfm/pdf/qgraal_cm_2013-07-intro.pdf), p. 21-41.

- nombre de mots : le nombre total d'unités lexicales du corpus
  - nombre de propriétés de mot : le nombre d'annotations différentes pour chaque mot
  - pour chaque type d'annotation : on donne le nom de l'annotation et le nombre total de valeurs différentes pour cette annotation, ainsi que quelques exemples de ces valeurs.
- nombre d'unités structurelles : le nombre des différentes unités structurelles du corpus
  - pour chaque type d'unité structurelle : le nom de la structure et la liste de ses attributs avec leurs valeurs
    - pour chaque attribut : les n premiers éléments de la liste des valeurs

L'illustration 7.1 montre un exemple de description du corpus DISCOURS.

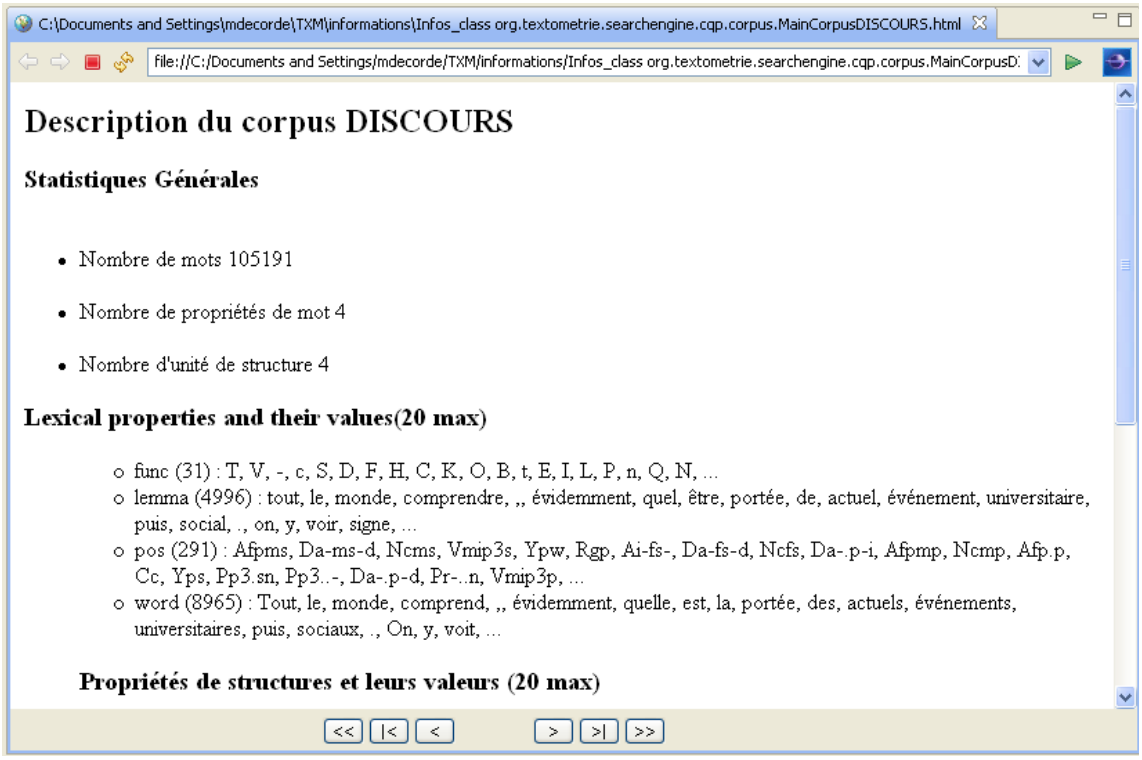


Illustration 7.1 : Description du corpus DISCOURS

7.1.2 Appliquée à une partition

Moins complète, la description d'une partition produit un graphique présentant la taille des parties de la partition. A partir des préférences, on peut choisir de trier l'affichage des parties par taille ou par nom.

7.2 Édition d’un texte

La commande Édition affiche la première page de l'édition du premier texte du corpus sélectionné. Le préambule de l'édition, situé en haut de la première page, affiche toutes les métadonnées du texte.

Dans cette édition, on peut naviguer :

- vers la page suivante '`[>]`' ou la page précédente '`[<]`' ;
- vers la fin du texte '`[>|]`' ou le début du texte '`[|<]`' ;
- vers le texte suivant dans le corpus '`[>>]`' ou le texte précédent '`[<<]`'.

Une autre façon d’accéder à l’édition se fait par retour au texte depuis une concordance. Double-cliquer sur une ligne de concordance (voir ci-dessous) vous mène directement à la page concernée de l'édition, où le pivot de la concordance sera surligné en rouge (s'il y a plusieurs occurrences de la requête dans la même page de concordance, elles seront surlignées en rouge clair).

L'illustration 7.2 présente la première page de l'édition du premier texte du corpus DISCOURS :

- dans cet exemple, les métadonnées sont : `id`, `file`, `loc`, `type`, `date`
  - `loc` : nom du locuteur
  - `type` : type de discours
  - `date`
- chaque mot peut être survolé avec la souris afin d'afficher ses propriétés dans une infobulle : `pos`, `func`, `lemma`
  - dans cette exemple, la souris est placée sur le mot « équilibre », l'infobulle affiche :
    - `pos` = « Ncms » : nom commun nom masculin singulier (étiquetage Multext) ;
    - `func` = « - » : aucune
    - `lemma` = « équilibre »

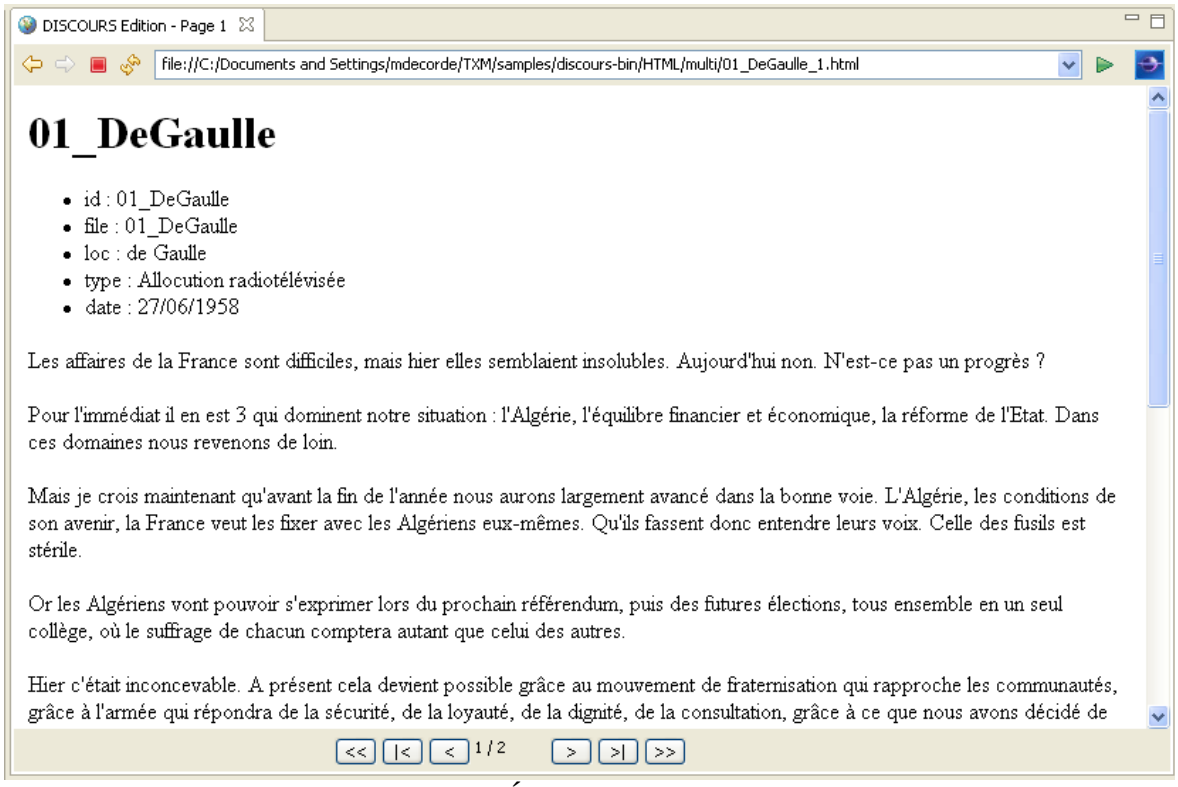


Illustration 7.2 : Édition du corpus DISCOURS

7.3 Lexique et Index

Les listes de mots peuvent être obtenues via deux commandes complémentaires :

- Lexique : calcule la liste hiérarchique de toutes les valeurs d'une propriété de mot donnée d'un corpus ou sous-corpus (la fréquence de chaque forme graphique, de chaque lemme, etc.) ;
- Index : calcule la liste hiérarchique des combinaisons de valeurs de propriétés correspondant aux occurrences d'une requête CQL cherchée dans un corpus ou un sous-corpus (la fréquence de chaque lemme de substantifs, des formes graphiques des occurrences de la séquence « Adj Subst », etc.).

7.3.1 Lexique

La commande Lexique  calcule la liste des fréquences de toutes les valeurs de propriétés lexicales d'un corpus ou d'un sous-corpus (par exemple : des formes de mots, des étiquettes morphosyntaxiques, des lemmes, etc). Par défaut, à l'ouverture, la commande calcule le lexique de la propriété lexicale « word » (celui des formes).

Le résultat se présente sous forme d'un tableau :

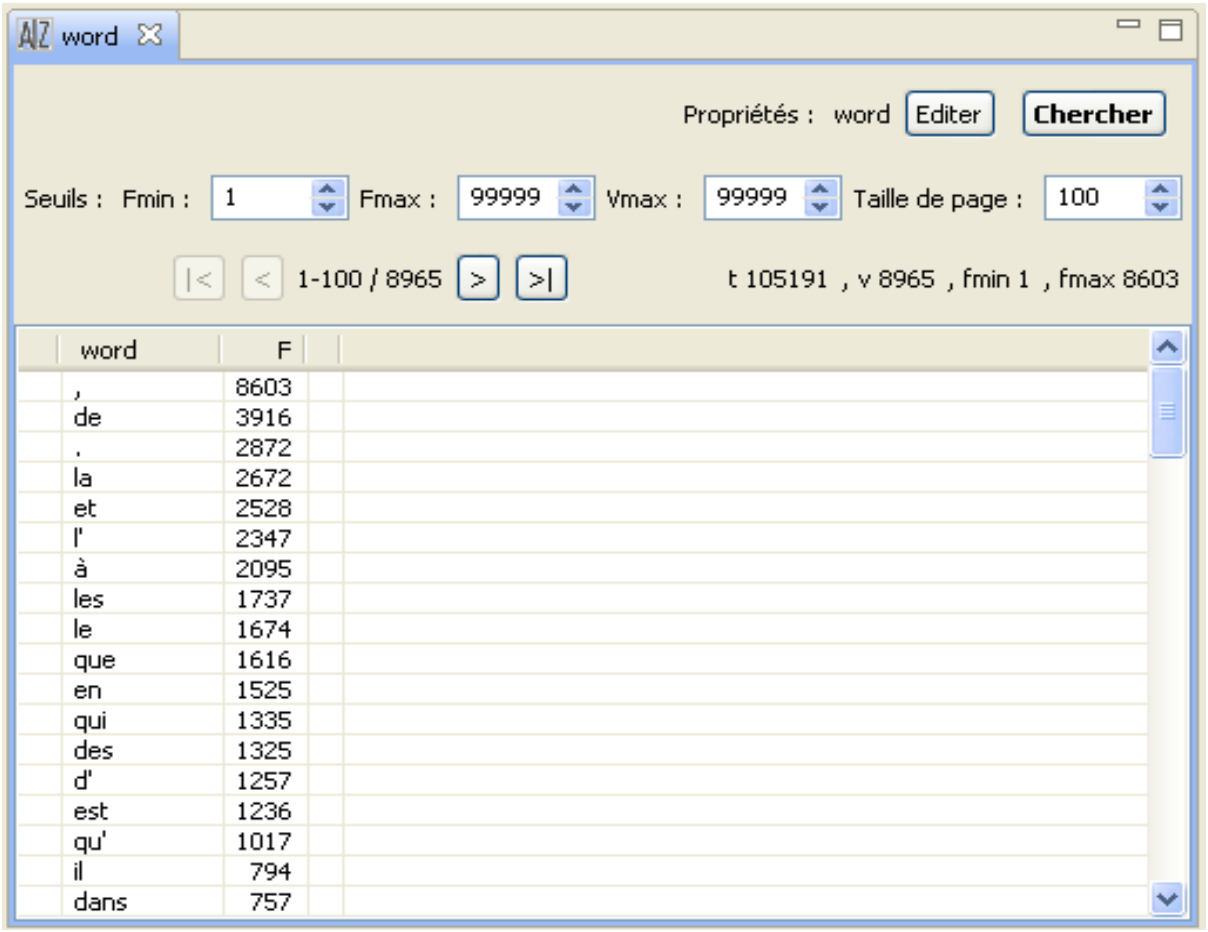


Illustration 7.4 : liste hiérarchique des formes graphiques des mots du corpus DISCOURS.

Vous pouvez trier le tableau par chaque colonne en cliquant sur son entête (exemple tri par les formes ou tri par les fréquences). Un nouveau clic inverse l'ordre de tri.

Vous pouvez exporter ce tableau au format CSV en sélectionnant l'icone du lexique de la vue Corpus.

7.3.2 Index

La commande Index établit la liste de fréquences des propriétés des occurrences d'une requête CQL pour un corpus, sous-corpus ou une partition donnée.

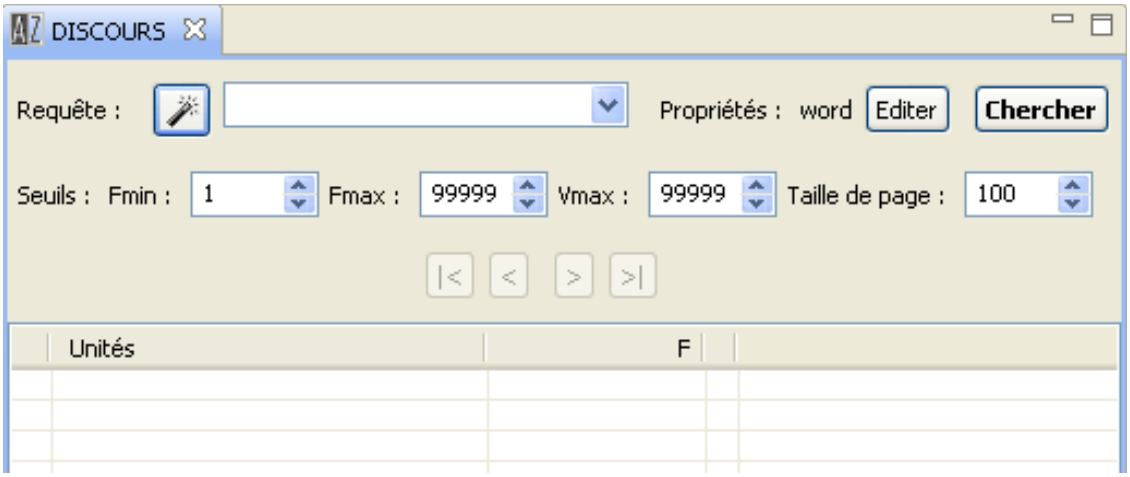


Illustration 7.5 : Fenêtre de la commande Index.

7.3.2.1 Choix du jeu de propriétés de mots à lister

Les occurrences sont décomptées en fonction des propriétés de mots sélectionnées. Par défaut ce sont les formes des mots des occurrences de la requête qui sont listées et décomptées (« word »). TXM permet également de construire la liste à partir des catégories grammaticales des mots, de leur lemme ou de toute propriété de mots encodée dans le corpus et de combinaisons de ces propriétés.

On peut sélectionner le jeu de propriétés à combiner avec le bouton « `Éditer` »[40] :

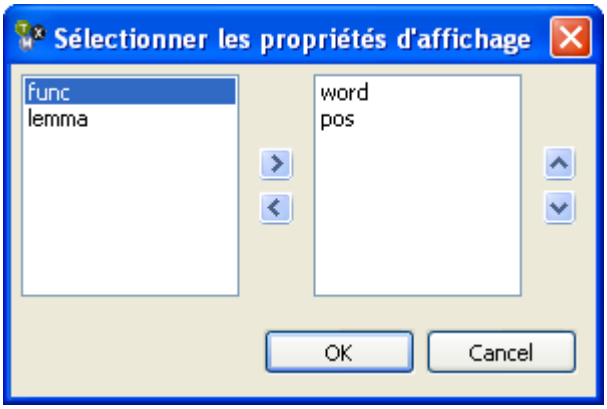


Illustration 7.6 : Fenêtre d'édition des propriétés de mot.

[40] Dans l'exemple, la propriété 'word' désigne la forme graphique du mot.

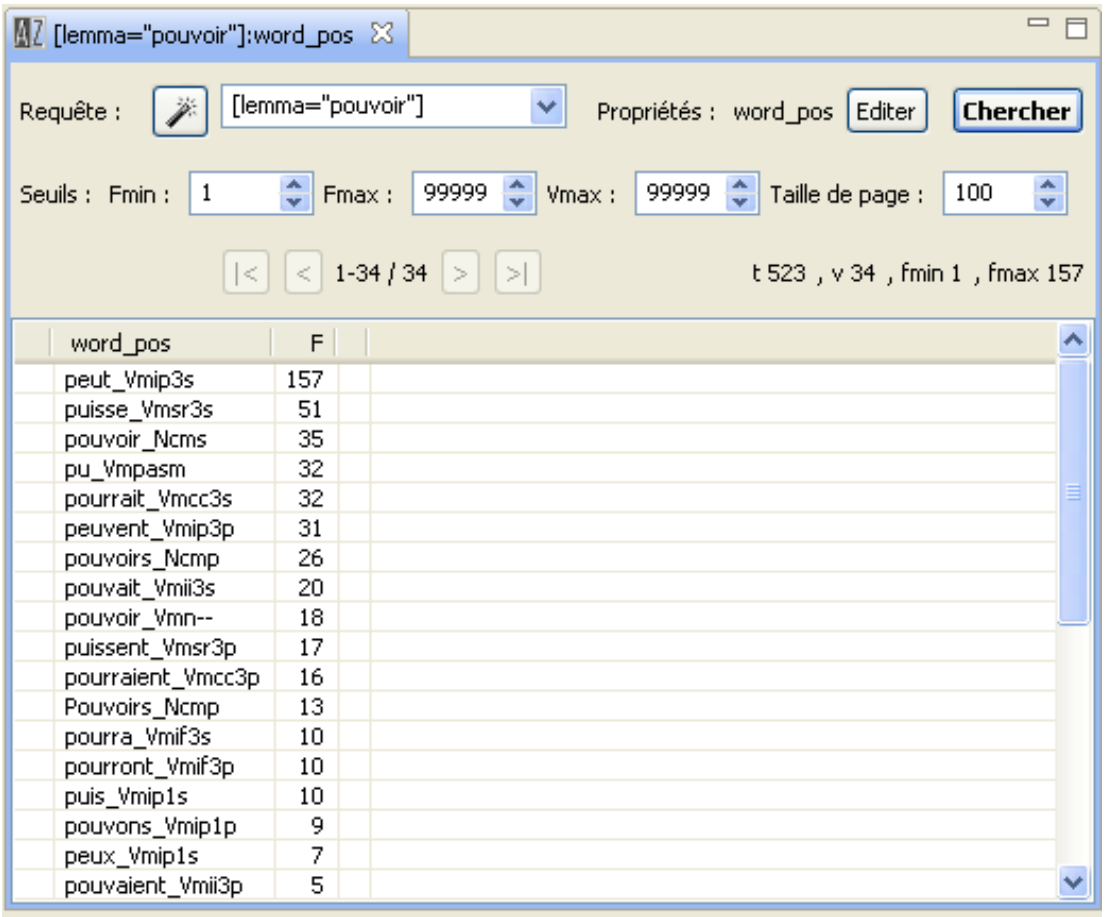
Sélectionner dans la liste de gauche les propriétés que l'on souhaite ajouter[41]. Faîtes les basculer grâce aux flèches qui permettent d'ajouter ou de retirer les propriétés :

- **« > »** : permet d'ajouter une propriété (on peut aussi double-cliquer sur une propriété dans la liste de gauche) ;
- **« < »** : permet de retirer une propriété (on peut également double-cliquer sur une propriété dans la liste de droite) ;
- **« ^ »** : permet de modifier l'ordre d'une propriété vers le haut (la propriété qui se trouve tout en haut sera celle qui s'affichera en premier) ;
- **« v »** : permet de modifier l'ordre d'une propriété vers le bas.

7.3.2.2 Requêtes

Vous pouvez utiliser les mêmes requêtes CQL que pour les concordances (ainsi que l'assistant de requêtes).

[41] Un double-clic sur un mot le fait basculer à droite directement.

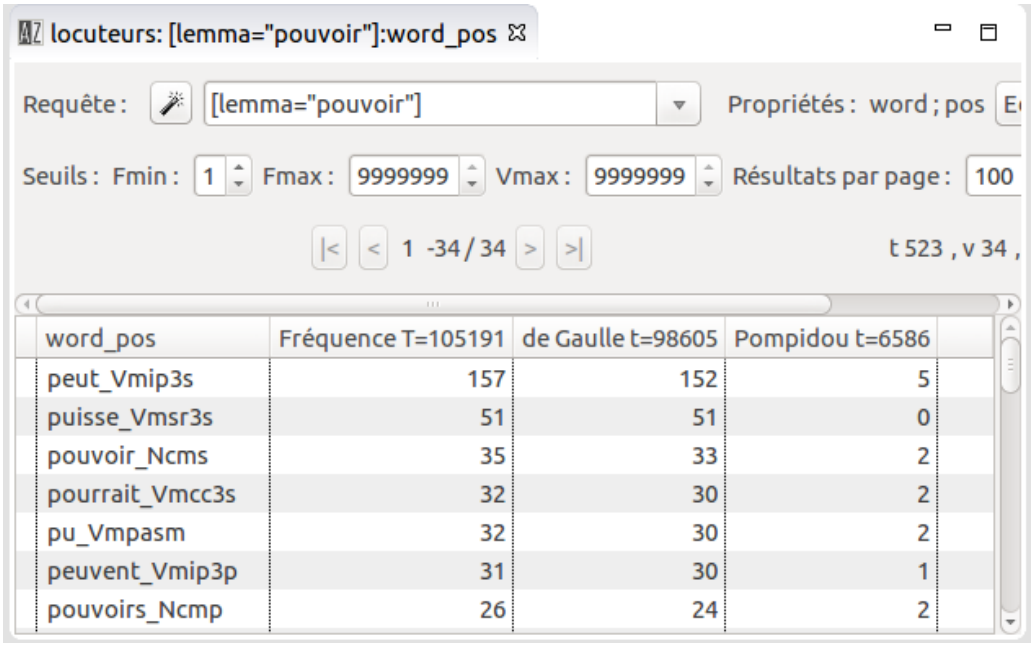


word_pos	F
peut_Vmip3s	157
puisse_Vmsr3s	51
pouvoir_Ncms	35
pu_Vmpasm	32
pourrait_Vmcc3s	32
peuvent_Vmip3p	31
pouvoirs_Ncmp	26
pouvait_Vmii3s	20
pouvoir_Vmn--	18
puissent_Vmsr3p	17
pourraient_Vmcc3p	16
Pouvoirs_Ncmp	13
pourra_Vmif3s	10
pourront_Vmif3p	10
puis_Vmip1s	10
pouvons_Vmip1p	9
peux_Vmip1s	7
pouvaient_Vmii3p	5

Illustration 7.7 : Index formé sur les propriétés 'word' et 'pos' pour le lemme « pouvoir », dans le corpus DISCOURS.

7.3.2.3 Index d'une partition

L'Index appliqué à une partition calcule le tableau des fréquences ventilées par parties. Ce tableau peut alors être transformé en une table lexicale pour être soumis au calcul des spécificités ou à une AFC.



word_pos	Fréquence T=105191	de Gaulle t=98605	Pompidou t=6586
peut_Vmip3s	157	152	5
puisse_Vmsr3s	51	51	0
pouvoir_Ncms	35	33	2
pourrait_Vmcc3s	32	30	2
pu_Vmpasm	32	30	2
peuvent_Vmip3p	31	30	1
pouvoirs_Ncmp	26	24	2

Illustration 7.8: Index de partition

7.3.2.4 Filtrage des résultats

Vous pouvez élaguer les résultats avec :

- Fmin : fréquence minimum à partir de laquelle on ajoute un résultat à la liste ;
- Fmax : fréquence maximum ;
- Vmax : nombre maximum de résultats à afficher. Par exemple si Vmax = 100, on obtiendra les 100 premières valeurs triées par la fréquence ;
- page size : nombre de résultats par page.

7.3.2.5 Navigation dans les résultats

L'index affiche d'abord la première page de résultats.

Vous pouvez naviguer dans l'ensemble des résultats avec les boutons suivants :

- « [|<] » : retour à la première page des résultats ;
- « [<] » : retour à la page précédente ;
- « [>] » : aller à la page suivante ;
- « [>|] » : aller à la dernière page.

7.3.2.6 Appel de commandes à partir des résultats

La commande index est liée aux commandes Concordance et Progression.

Vous pouvez sélectionner certaines lignes de l'index avec la souris[42], puis par l'intermédiaire du menu contextuel, choisir la commande à exécuter :

- « Envoyer vers les concordances » : une requête CQL correspondante sera créée afin de construire la concordance.
- « Envoyer vers progression » : autant de requêtes CQL que de lignes sélectionnées seront créées pour construire une progression.

7.4 Concordances

Cette commande construit une concordance kwic à partir des résultats de recherche correspondant à une requête CQL sur un corpus ou un sous-corpus sélectionné.

La boîte de dialogue de recherche est organisée de la façon suivante :

- un champ pour saisir la requête CQL ;
- un bouton pour accéder à l'historique des requêtes ;
- un bouton pour accéder à l'éditeur des propriétés affichées des unités lexicales afin de sélectionner quelles propriétés seront affichées dans la colonne des pivots ;
- le bouton « `chercher` » pour lancer le calcul.
- le bouton « Cacher paramètres »: cache les paramètres de la concordance pour améliorer le confort de lecture.

[42] Shift-clic gauche permet de sélectionner des lignes contiguës. Ctrl-clic gauche permet de sélectionner plusieurs lignes non contiguës.

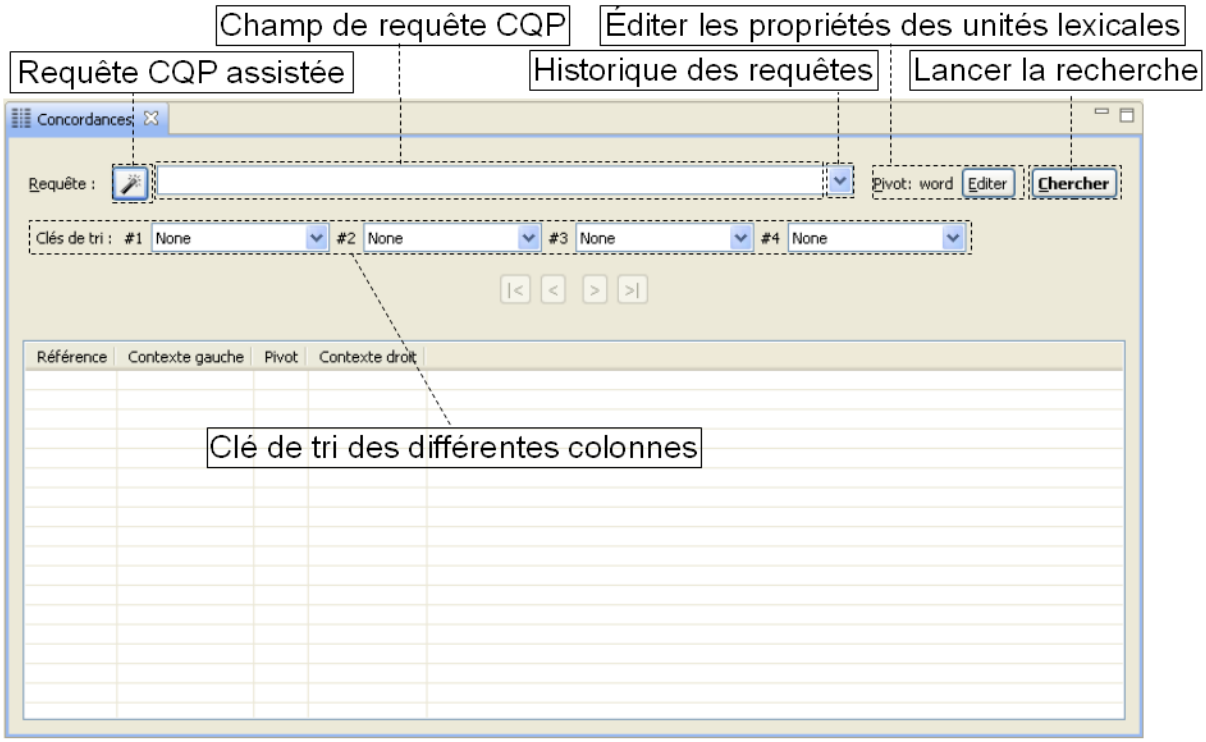


Illustration 7.9 : La fenêtre de concordance

7.4.1 Requêtes

Le moteur de recherche vous permet d'exprimer dans le langage formel CQL (voir ci-dessous la section 5 « la syntaxe du moteur de recherche »).

TXM utilise une syntaxe simplifiée basée sur le langage CQL, afin d'écrire facilement des requêtes. Par exemple, pour rechercher « je », vous n'avez qu'à écrire « `je` » dans le champ « `Requête` ».

Pour des recherches plus complexes, vous pouvez utiliser toute la variété du langage CQL. Par exemple, pour chercher :

le mot « je » suivi d'un verbe

dans le corpus DISCOURS, vous pouvez saisir la requête suivante :

```
"je" [pos="V.*"]
```

Cette requête peut être décomposée ainsi :

- `"je"` désigne le mot « je » ;
- `[pos="V.*"]` indique que le verbe sera sur la droite du mot « je » :
  - les crochets `[...]` indiquent qu'il ne doit y avoir qu'une seule unité lexicale à la droite du mot « je » ;
  - `pos="V.*"` indique que l'occurrence doit porter l'étiquette morphosyntaxique « `V.*` ». Dans le corpus DISCOURS, étiqueté par Cordial et le jeu d'étiquettes Multext, cette requête sélectionne tous les verbes (dans ce corpus, tous les verbes ont une étiquette qui commence par « V »).

Il est également possible de lancer une requête assistée. En cliquant sur l'icone « Assistant de Requête »  une fenêtre permet de construire plus facilement des requêtes CQL :

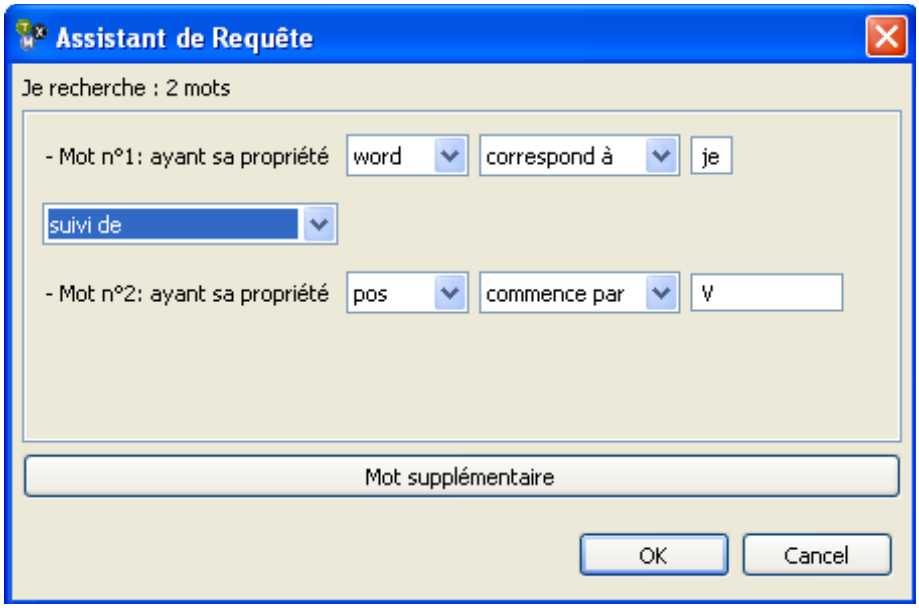


Illustration 7.10 : Construction d'une requête sur le mot "je" suivi d'un verbe.

- Le bouton « `mot supplémentaire` » permet d'ajouter un mot à la requête.
- Le premier menu déroulant permet de sélectionner une propriété de mot
- Le second menu déroulant permet de sélectionner un champ de recherche plus ou moins restreint
- Le dernier champ permet de saisir un mot ou quelques lettres.

- Le menu déroulant situé entre les expressions de mots permet de préciser si les mots sont consécutifs ou non.

Si vous validez votre requête avec « `OK` », elle apparaîtra sous sa forme CQL dans le champ « requête ».

Pour lancer la recherche, cliquer sur le bouton « `chercher` ».

Avant d'afficher les résultats de la concordance, la zone de commentaire ainsi que la ligne de statut vous donneront le nombre total de résultats.

L'illustration 7.11 montre les résultats :

- il y a 206 occurrences ;
- les résultats affichés vont de 22 à 41 (il s'agit de la deuxième page) ;
- la colonne « pivot » recense les deux mots ciblés par la requête « `je` » suivi d'un verbe ;
- les concordances sont triées par défaut alphabétiquement dans la colonne « pivot » ;
- la référence prend la forme ici du nom du locuteur ;
- on peut ouvrir le menu contextuel en cliquant à droite sur une concordance :
    - Définir le patron des références : régler les informations de la colonne référence ;
    - Définir la propriété de tri : définir la propriété de mot qui déterminera l'ordre de tri initial ;
    - Tri multiple : définir plusieurs clés de tri ;
    - Définir la taille des contextes : choisir le nombre maximum de mots dans les contextes de gauche et de droite ;
    - Lignes par page : définir le nombre de résultats dans la page
    - Sélectionner les propriétés : choisir les propriétés de mots qui seront affichées dans chaque colonne.

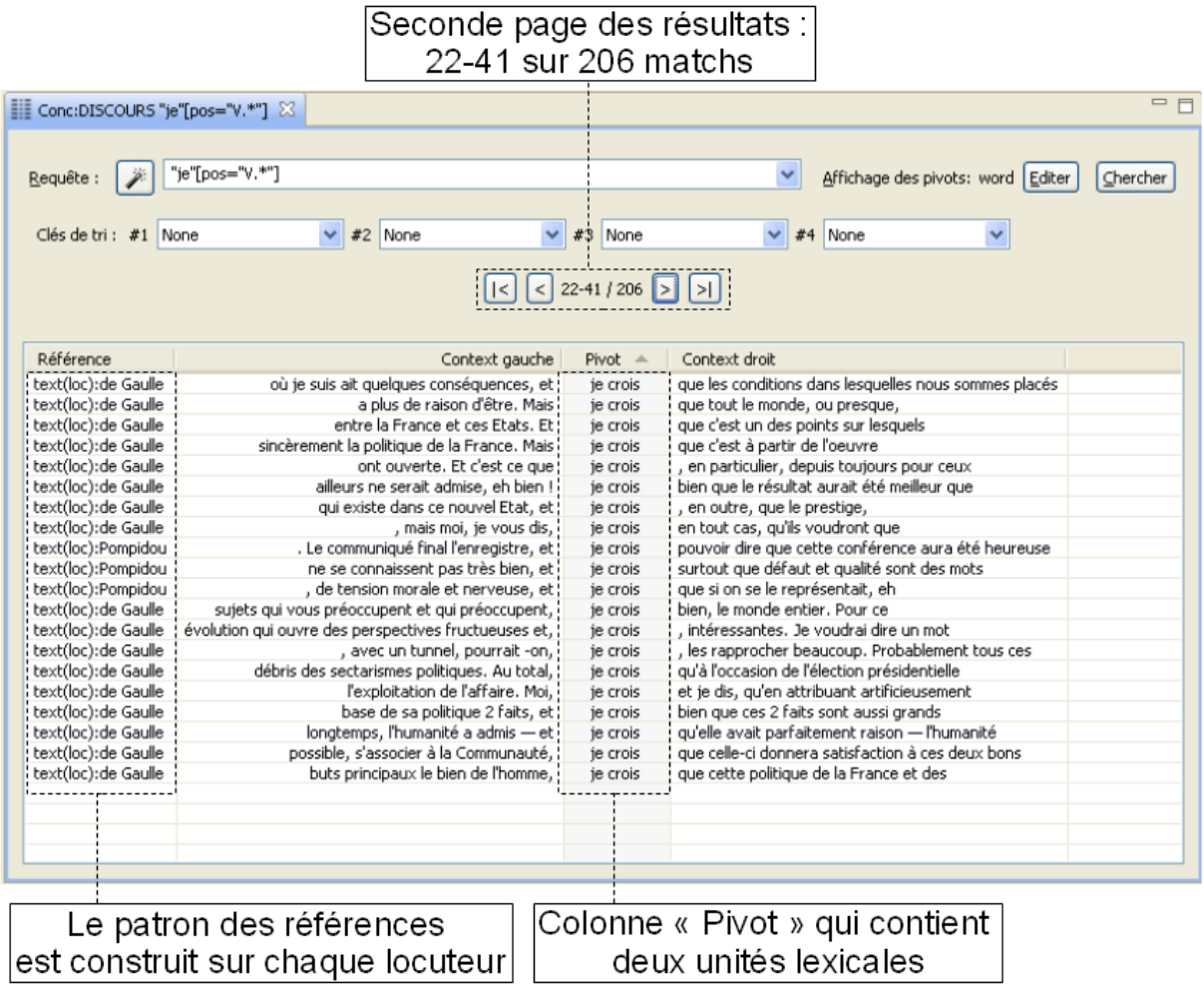


Illustration 7.11 : Concordance du mot « je » suivi d'un verbe dans le corpus DISCOURS.

7.4.2 Navigation

Une concordance commence par afficher la première page des résultats.

Les boutons de navigation permettent de visionner tous les résultats :

- « [|<] » : retour à la première page ;
- « [<] » : retour à la page précédente ;
- « [>] » : aller à la page suivante ;
- « [>|] » : aller à la dernière page.
- « [Cacher/Montrer les paramètres] » : cache ou affiche les paramètres de la concordance pour plus de confort de lecture.

Le nombre de lignes par page par défaut peut être réglé via le menu « Fichier / Préférences », puis la fenêtre « TXM>Utilisateur>concordances ». Pour un réglage uniquement dans la fenêtre courante de la concordance, il faut passer par le menu contextuel du tableau de la concordance.

7.4.3 Retour au texte

En double-cliquant sur une ligne de la concordance, on retourne à la page de l'édition qui contient le pivot. L'édition est ouverte dans un nouvel éditeur.

Au sein de la page, le pivot est surligné en rouge, tandis que les autres pivots de la concordance se trouvant dans la même page sont surlignés en rouge clair.

Si on re-double-clique sur une ligne de la concordance, le même éditeur est utilisée. Pour une navigation dans l'édition, vous pouvez placer l'éditeur de l'édition à côté de l'éditeur de la concordance.

7.4.4 Tri

Vous pouvez trier les concordances selon chaque colonne : « Références », « Contexte gauche », « Pivot » et « Contexte droit » en cliquant sur leurs entêtes. Vous pouvez changer l'ordre de classement en cliquant une nouvelle fois sur l'entête. Vous remarquerez qu'alors les clés de tri changent en fonction de l'entête sélectionnée. Le tri par défaut se fait selon le pivot. Toutefois vous avez la possibilité de changer les propriétés de tri en cliquant sur « Définir la propriété de tri » dans le menu contextuel. Enfin vous pouvez effectuer un tri multiple en changeant chaque clé de tri.

7.4.5 Propriétés de mot

Chaque colonne contenant une propriété de mot peut être personnalisée de deux façons différentes :

- les propriétés affichées pour le pivot peuvent être réglées en appuyant sur le bouton « `éditer` », situé à côté du champ requête ;
- sur une concordance, sélectionner dans le menu contextuel « Définir les propriétés ».

7.4.6 Références

Vous pouvez choisir quelles informations seront affichées dans la colonne « référence » (sur la gauche dans chaque ligne de concordance).

Dans le menu contextuel, sélectionner « Définir le patron des références ». Une fenêtre s'ouvre, comme vous pouvez le constater dans l'illustration 7.12 :

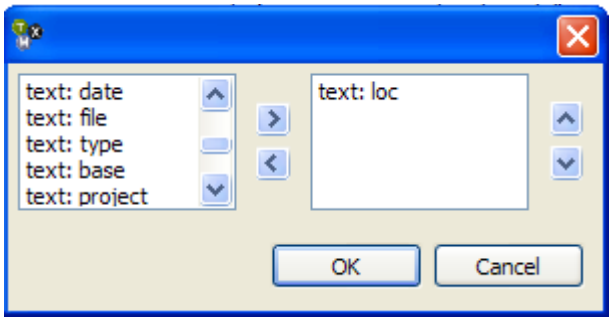


Illustration 7.12 : Boîte de dialogue « patron des références »

Toutes les propriétés d'unités de structure et d'unités lexicales se trouvent dans la liste de gauche.

Par exemple, `text:loc` représente la propriété « `loc` » de la structure « `text` ».

Pour choisir une propriété, sélectionnez-là puis cliquez sur le bouton « > » pour la faire glisser dans le champ de droite. La liste qui se formera sur la droite correspondra à l'affichage dans la colonne référence.

Afin de retirer une propriété, sélectionnez-là dans la liste de droite et appuyez sur le bouton « < » afin de la faire basculer dans la liste de gauche.

Afin de changer l'ordre des propriétés dans la liste de droite, utiliser les boutons monter « ^ » et descendre « v ».

7.4.7 Export

Les concordances peuvent être exportées au format CSV : sélectionnez l'icone de la concordance dans la vue « corpus » et cliquez sur l'icone dans la barre d'outils ou sur la commande Export dans le menu contextuel.

7.5 Cooccurrences

La commande Cooccurrences calcule le tableau des différents cooccurrents des occurrences d'une requête CQL, trié par défaut par l'indice de cooccurrence[43] (un indicateur de probabilité de rencontre). Elle permet donc de calculer les cooccurrents d'une forme, d'un lemme, d'une combinaison d'un lemme et d'une catégorie, etc.

L'appel de cette commande ouvre une fenêtre composée d'une zone de paramètres et d'une zone affichant les cooccurrents (Illustration 7.13).

[43] P. Lafon, “Sur la variabilité de la fréquence des formes dans un corpus,” *Mots*, no. 1 (1980): 127-165.

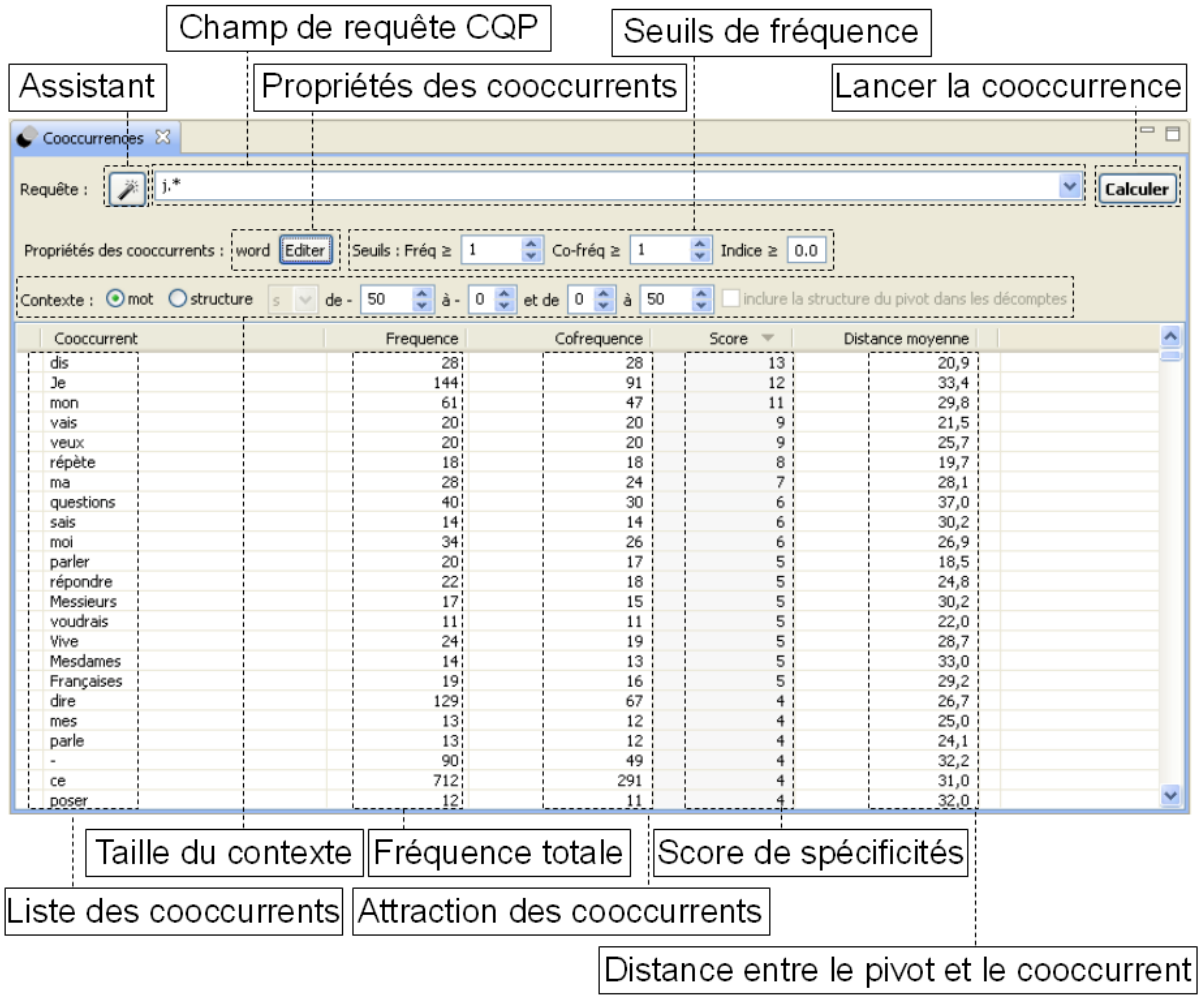


Illustration 7.13 : Cooccurrents des mots commençant par "j".

La zone de paramètres permet de :

- Saisir une expression CQL dans le champ de requête (on peut aussi utiliser l'assistant de requête).
- Choisir les propriétés lexicales utilisées pour construire les cooccurrents (forme, lemme, etc.)
- Régler les seuils de fréquence, de co-fréquence et d'indice pour élaguer les résultats. La co-fréquence est le nombre de rencontres entre les cooccurrents et les occurrences de la requête
- Choisir le type et la taille du contexte de rencontre :
  - Contexte en structure, si on coche « structure »

  - Contexte en fenêtre de mots, si on coche « forme »
  - On peut définir la taille du contexte à gauche et à droite du pivot (en nombre de structures ou en nombre de mots suivant le type de contexte).
  - On peut ignorer des contextes en décochant « Contexte gauche actif » ou « Contexte droit actif ».

- Trier la liste des cooccurrents en cliquant sur l'entête d'une colonne.

Pour lancer le calcul, cliquer sur « `Calculer` ».

7.6 Progression

Voir aussi la documentation commune à toutes les visualisations dans la section « 7.14 Visualisation graphique des résultats » page 141.

Une progression affiche l'évolution d'un ou de plusieurs motifs au fil du corpus. Cette commande est lancée sur un corpus. Elle produit au choix un graphique cumulatif ou de densité et superpose la position des structures du corpus à la demande. À son lancement cette commande ouvre une boîte de dialogue de paramètres, telle qu'à l'illustration 7.14 :

- On doit d'abord y choisir le type de progression: cumulatif ou densité
- On peut choisir une unité structurelle et une de ses propriétés : chaque limite d'unité pour chaque valeur de la propriété sera représentée sur le graphique sous la forme d'une barre verticale. Ceci est une option d'affichage, c'est à dire qu'une barre sera affichée par unité de structure.
- On peut filtrer les valeurs de la propriété au moyen d'une expression régulière (pour limiter le nombre de barres par exemple)
- Ensuite, on peut ajouter une ou plusieurs requêtes de motif CQL à afficher (éventuellement avec l'aide de l'assistant) au moyen du bouton « `ajouter` ». On peut supprimer une requête avec le bouton « `supprimer` »

Si le mode « densité » est sélectionné, on peut faire varier la fenêtre de densité par un facteur multiplicatif. Par défaut, la taille de la fenêtre, est la taille minimum entre chaque unité de structure (entre chaque texte si la structure sélectionnée est « text »).

Des options d'affichage sont disponibles :

- Afficher le graphique en noir & blanc
- Répéter ou pas les valeurs de propriétés de structure.
- Utiliser des styles de ligne différents

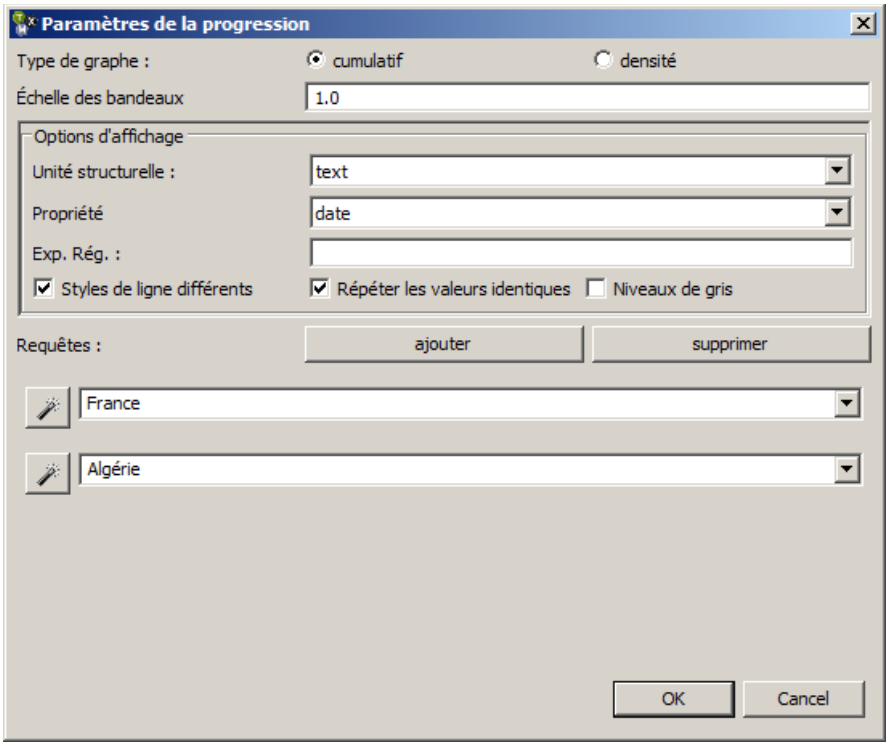


Illustration 7.14 : Calcul de la progression des mots « France » et « Algérie » dans les discours de Pompidou et De Gaulle.

En cliquant sur « OK » on obtient le graphique de progression tel que dans l'illustration 7.15. Dans ce graphique, le nom des locuteurs sont affichés en début de discours. Les courbes représentent les progressions respectives des mots « France » et « Algérie ».

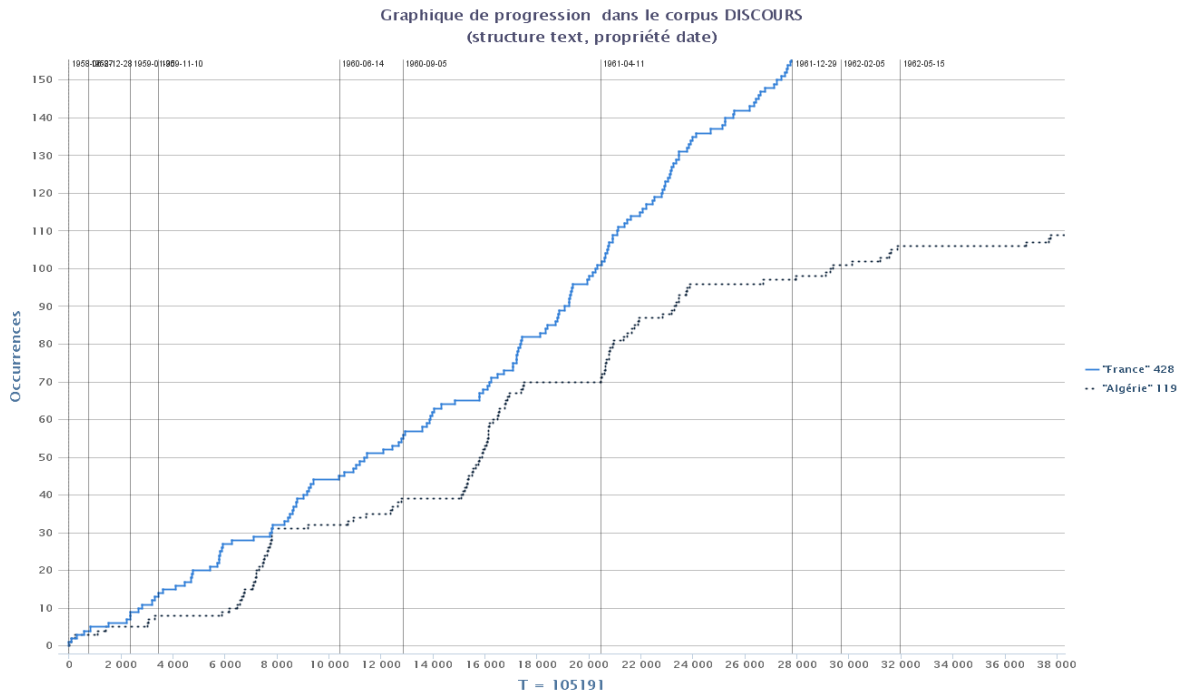


Illustration 7.15 : Graphique de la progression cumulatif du mot France et Algérie dans les discours de De Gaulle et Pompidou.

Le graphique est exportable sous forme d'image via le bouton « Export » de la barre d'outils.

7.7 Références

La commande Références affiche la liste toutes les références des valeurs retournées par une requête CQL à partir des informations des unités structurelles les contenant.

A côté de chaque référence, on trouve, entre parenthèses, la fréquence de la référence. C'est à dire le nombre de fois qu'un pivot à cette référence. Les références peuvent être triées par fréquence ou alphabétiquement.

Si la requête CQL correspond à une succession d'unités lexicales, c'est alors la première unité qui est prise en compte.

Utilisation :

- On doit saisir une requête CQL dans le champ requête
- On choisit la propriété d'affichage des occurrences, et ainsi la façon de les regrouper
- On peut choisir les propriétés de structures à utiliser. Tout comme la commande concordance, il s'agit d'un patron.

– Enfin, on lance le calcul à l'aide du bouton « Chercher »

7.8 Sous-corpus

Cette commande construit un sous-corpus du corpus sélectionné. Le sous-corpus est représenté comme un descendant du corpus dans la vue « corpus ».

Cette commande ouvre une boîte de dialogue de nom « Créer un sous-corpus ». Elle est composée de trois onglets : ils permettent de construire des sous-corpus en mode simple, en mode assisté ou en mode avancé.

7.8.1 Construire un sous-corpus : mode « simple »

L'illustration 7.16 affiche la boîte de dialogue du mode simple de la commande « construire un sous-corpus ».

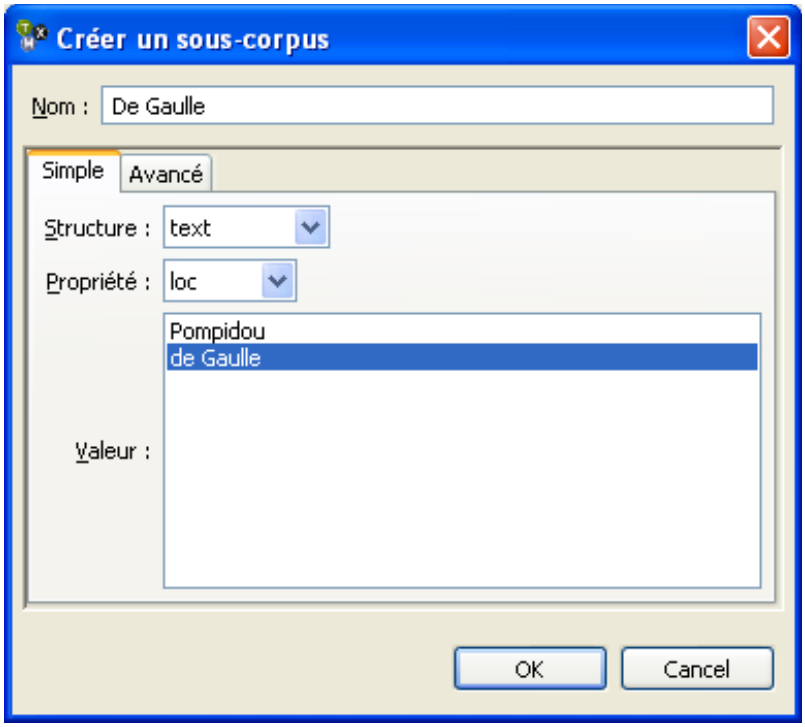


Illustration 7.16 : Mode « simple » : construction d'un sous-corpus de tous les discours de De Gaulle.

Ici, on doit :

- OPTIONNEL : entrer le nom du nouveau corpus : il sera affiché dans la vue « corpus »
- sélectionner une unité structurelle
- sélectionner la propriété de cette unité.
- sélectionner une ou plusieurs valeurs

Le nouveau corpus contiendra toutes les unités lexicales se trouvant dans les unités structurelles ainsi désignées.

7.8.2 Construire un sous-corpus : mode « assisté »

L'illustration 7.17 présente le formulaire de création de sous-corpus en mode « assisté ». qui permet de formuler la requête de création de sous-corpus à partir de différentes propriétés d'une structure

Dans cette fenêtre, on doit :

- OPTIONNEL : Saisir le nom du sous-corpus
- Cocher « tous les critères » pour considérer tous les critères de recherche saisis ou cocher « certains critères » pour ne considérer que certains d'entre eux.
- Sélectionner la structure du sous-corpus qui sera utilisée
- Saisir des critères de sélection :
  - ajouter un critère avec le bouton « + »
  - supprimer un critère avec le bouton « - »
  - choisir la propriété utilisée par le critère :
    - qui contient ou qui ne contient pas l'attribut sélectionné
- Rafraîchir la requête de création du sous-corpus
- **Modifier si besoin la requête**
- Cliquer sur « OK » pour créer le sous-corpus

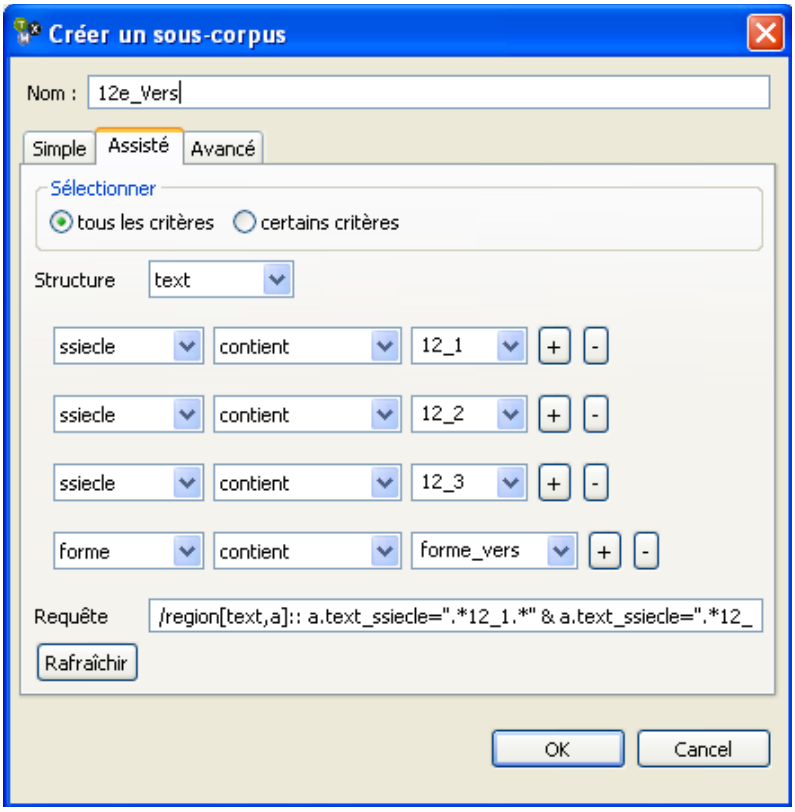


Illustration 7.17: Mode « assisté » : création d'un sous-corpus des entretiens radiotélévisés de Pompidou.

Attention, rajouter un critère de sélection rajoute une contrainte logique de type « ET ». Vous pouvez remplacer les « & » de la requête par des « | » si vous voulez « ajouter ».

7.8.3 Construire un sous-corpus : mode « avancé »

L'illustration 7.18 présente la boîte de dialogue du mode avancé[44] qui permet à une utilisateur expert de construire des sous-corpus à l'aide du langage de requête CQL.

Ici on doit :

- OPTIONNEL : entrer le nom du nouveau corpus qui apparaîtra dans la vue corpus
- écrire une requête CQL qui sélectionnera les unités lexicales du nouveau sous-corpus

Le sous-corpus contiendra toutes les unités lexicales sélectionnées par la requête.

[44] L'expression régulière complète est : `/region[text,a]:: a.text_loc="Pompidou"& a.text_date=".*1970"`

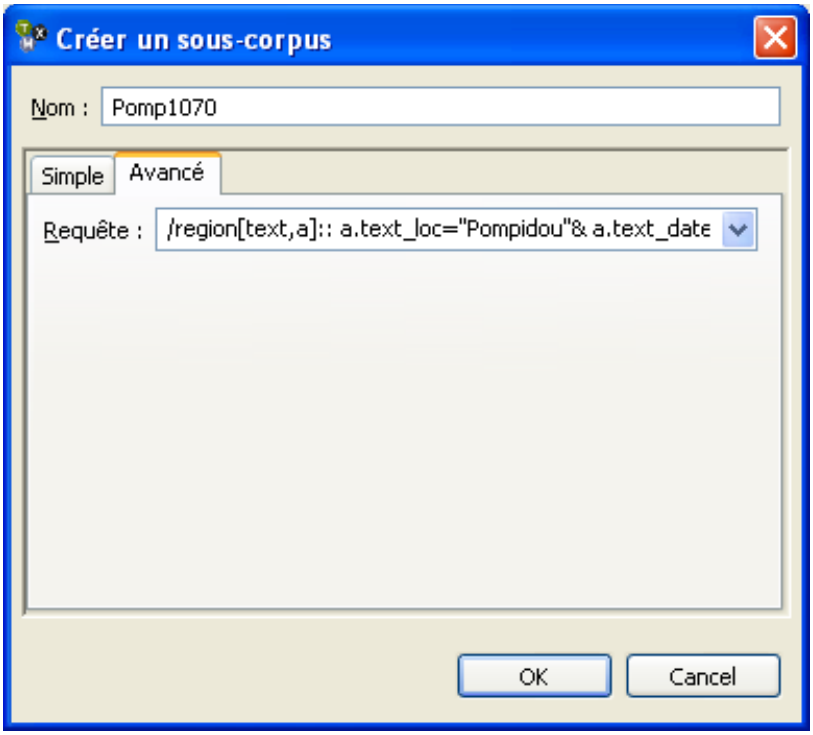


Illustration 7.18 : Mode « avancé » : construire un sous-corpus des discours de Pompidou datant de 1970.

7.9 Partition

Cette commande construit une partition à partir du corpus sélectionné. La nouvelle partition apparaît comme un descendant dans la vue « corpus ».

Cette commande ouvre une boîte de dialogue intitulée « Créer une partition ». Elle est composée de trois onglets : mode simple, assisté et avancé.

7.9.1 Construire une partition : mode « simple »

L'illustration 7.19 montre la fenêtre du mode simple.

Ici on doit :

- OPTIONNEL : entrer le nom de la nouvelle partition qui apparaîtra dans la vue « corpus »
- sélectionner une unité structurelle
- sélectionner la propriété de l'unité structurelle sélectionnée.

Les parties de la nouvelle partition seront construites, en tant que sous-corpus, en fonction des différentes valeurs de l'unité structurelle sélectionnée. On ne peut pas accéder aux parties individuellement mais elles sont accessibles via l'objet partition et les commandes qui permettent de mettre ces parties en contraste : Spécificités et AFC.

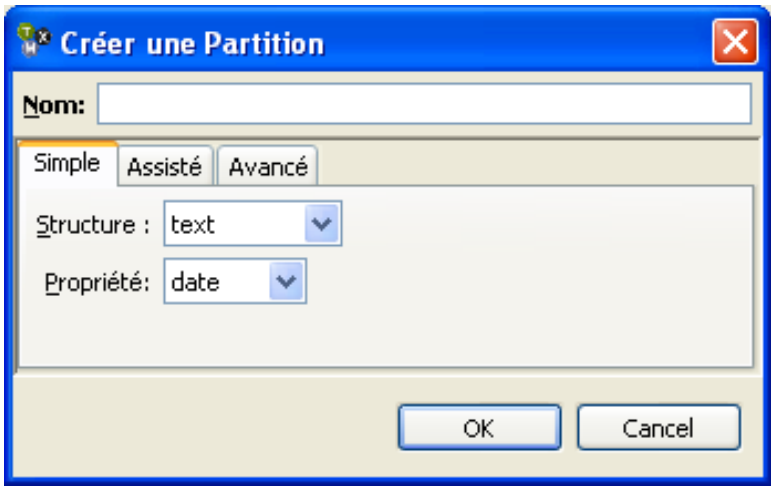


Illustration 7.19 : Mode simple : construire une partition sur chaque date d'un discours.

7.9.2 Construire une partition : mode « assisté »

Le mode assisté permet de définir plus finement les parties de la partition en offrant la possibilité de sélectionner les différentes valeurs de la propriété de structure à utiliser pour composer chaque partie.

L'illustration 7.20 présente la fenêtre de création de partition en mode assisté.

Ici, il faut :

- OPTIONNEL : entrer le nom de la partition qui apparaîtra dans la vue « corpus »
- sélectionner une unité de structure, ainsi qu'une de ses propriétés
- sélectionner les valeurs qui constitueront une partie de la partition
- cliquer sur « `nouvelle partie` » pour créer une partie supplémentaire
  - entrer le titre de la partie dans le champ correspondant
  - cliquer sur « `affecter` » afin de basculer les valeurs précédemment sélectionnées dans cette partie
  - on peut cliquer sur « `supprimer` » afin d'enlever certaines valeurs à cette partie

    - on peut cliquer sur la croix pour supprimer la partie
    - on peut cliquer sur « `Supp. toutes les parties` » afin de supprimer en une seule fois toutes les parties d'un coup
- cliquer sur « `OK` » crée la partition ainsi paramétrée.

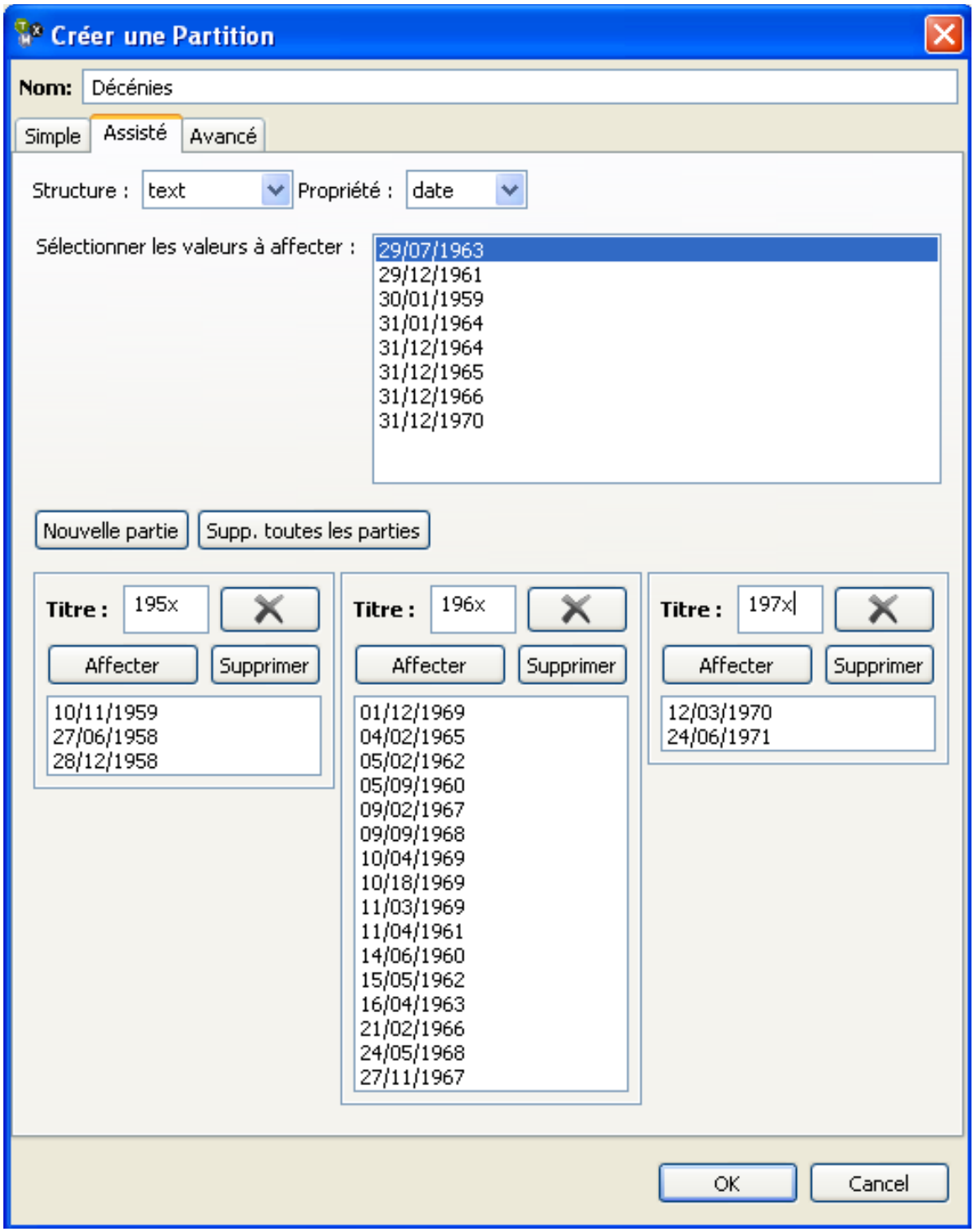


Illustration 7.20 : Mode assisté : construire une partition sur les dates du corpus DISCOURS.

7.9.3 Construire une partition : mode « avancé »

L'illustration 7.21 présente la fenêtre de création de partition en mode avancé[45].

Ici on doit :

- OPTIONNEL : entrer le nom du nouveau corpus qui apparaîtra dans la vue « corpus »
- écrire autant de requêtes CQL qui sélectionnent chacune les unités lexicales qui composent chaque partie
  - utiliser le bouton '+' pour ajouter une nouvelle partie et saisir la requête correspondante
  - utiliser le bouton '-' pour supprimer une partie

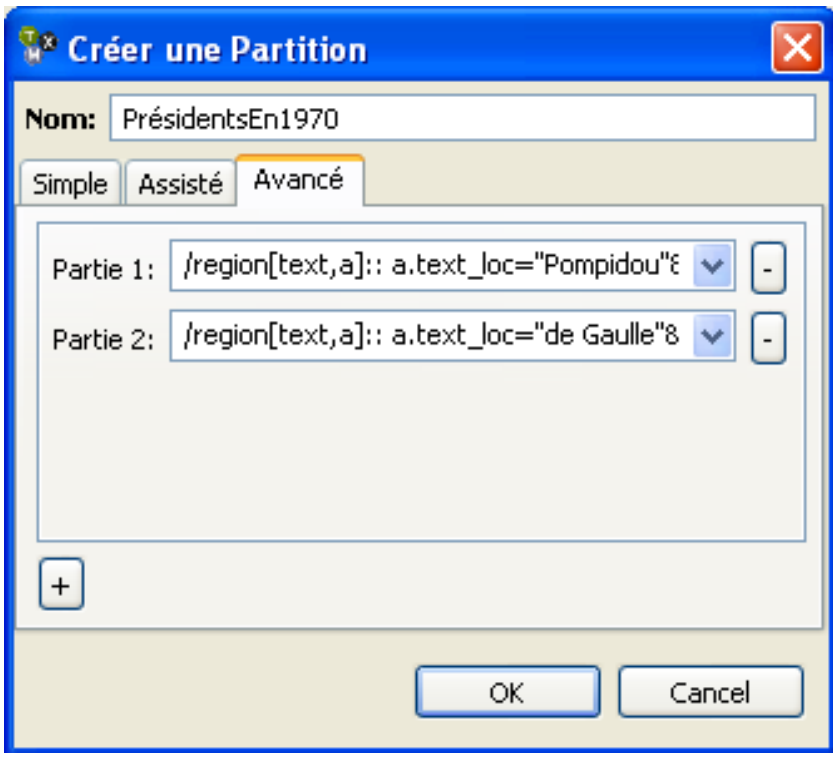


Illustration 7.21 : Construire une partition sur chaque président pour l'année 1970.

La nouvelle partition sera composée de toutes les parties définies, chacune contenant les unités lexicales sélectionnées par la requête correspondante.

Attention, la bonne couverture du corpus total par l'union des différentes parties est de la responsabilité de l'utilisateur.

Les parties de partitions avancées sont nommable en cliquant sur leur nom.

[45] Les requêtes complètes sont :

```
- [_.text_loc="Pompidou" & _.text_date=".*1970"]
- [_.text_loc="de Gaulle" & _.text_date=".*1970"]
```

7.10 Table lexicale

Une table lexicale réunit dans un tableau les différentes unités lexicales d'une partition.

Ce tableau peut être généré à partir d'une partition ou depuis l'index d'une partition. Une fois la partition sélectionnée, il faut choisir la propriété de mot sur laquelle se construira la table lexicale, comme ce qui apparaît dans l'illustration 7.22 :

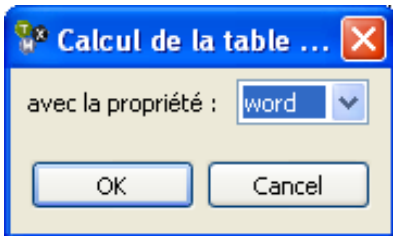
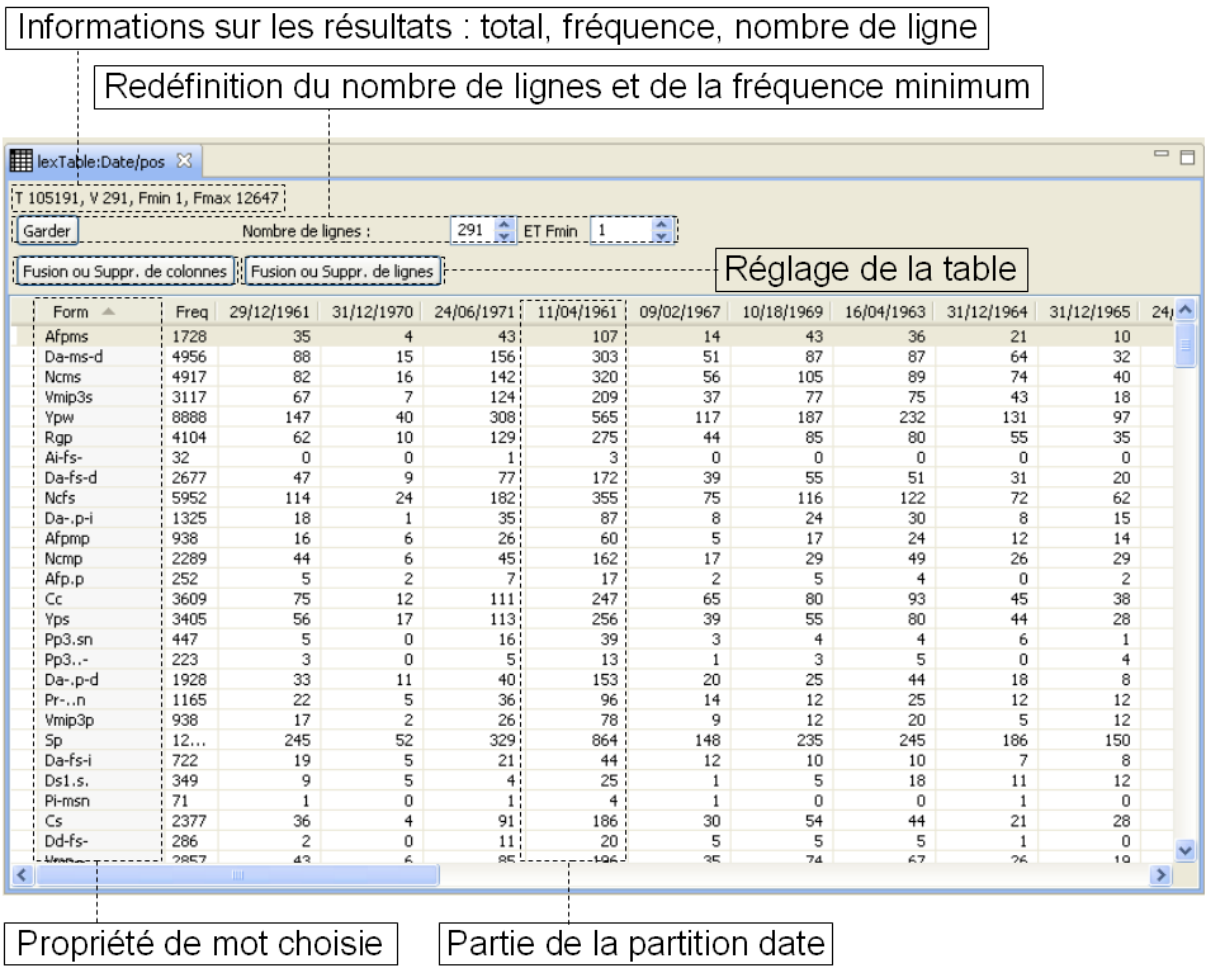


Illustration 7.22 : Propriété de la table lexicale.

Le tableau se présente de la façon suivante : une entrée par ligne, une partie par colonne. C'est un tableau éditable où les lignes et les colonnes peuvent être fusionnées ou supprimées. Il est également possible de ne retenir que certaines lignes en fonction de leur fréquence, la taille du tableau pouvant être limitée par un nombre de lignes maximum.

Enfin, une table lexicale est créée automatiquement dès qu'une commande AFC ou Spécificités est appliquée à un corpus et apparaîtra comme descendante de ce corpus.



Form	Freq	29/12/1961	31/12/1970	24/06/1971	11/04/1961	09/02/1967	10/18/1969	16/04/1963	31/12/1964	31/12/1965
Afpms	1728	35	4	43	107	14	43	36	21	10
Da-ms-d	4956	88	15	156	303	51	87	87	64	32
Ncms	4917	82	16	142	320	56	105	89	74	40
Vmip3s	3117	67	7	124	209	37	77	75	43	18
Ypw	8888	147	40	308	565	117	187	232	131	97
Rgp	4104	62	10	129	275	44	85	80	55	35
Ai-fs-	32	0	0	1	3	0	0	0	0	0
Da-fs-d	2677	47	9	77	172	39	55	51	31	20
Ncfs	5952	114	24	182	355	75	116	122	72	62
Da-.p-i	1325	18	1	35	87	8	24	30	8	15
Afpmp	938	16	6	26	60	5	17	24	12	14
Ncmp	2289	44	6	45	162	17	29	49	26	29
Afp.p	252	5	2	7	17	2	5	4	0	2
Cc	3609	75	12	111	247	65	80	93	45	38
Yps	3405	56	17	113	256	39	55	80	44	28
Pp3.sn	447	5	0	16	39	3	4	4	6	1
Pp3..-	223	3	0	5	13	1	3	5	0	4
Da-.p-d	1928	33	11	40	153	20	25	44	18	8
Pr-..n	1165	22	5	36	96	14	12	25	12	12
Vmip3p	938	17	2	26	78	9	12	20	5	12
Sp	12...	245	52	329	864	148	235	245	186	150
Da-fs-i	722	19	5	21	44	12	10	10	7	8
Ds1.s.	349	9	5	4	25	1	5	18	11	12
Pi-msn	71	1	0	1	4	1	0	0	1	0
Cs	2377	36	4	91	186	30	54	44	21	28
Dd-fs-	286	2	0	11	20	5	5	5	1	0

Illustration 7.23 : Table lexicale de la partition date du corpus DISCOURS.

Dans l'illustration ci-dessus on peut voir la table lexicale formée à partir de la partition Date du corpus DISCOURS. Il est possible de :

- Régler le nombre de ligne ainsi que la fréquence minimum. Il faut valider le choix en cliquant sur le bouton « `Garder` »
- Fusionner ou supprimer des colonnes : en cliquant sur le bouton « `Fusion ou Suppr. de colonnes` ». Ceci ouvre une boîte de dialogue (voir illustration 7.24) :

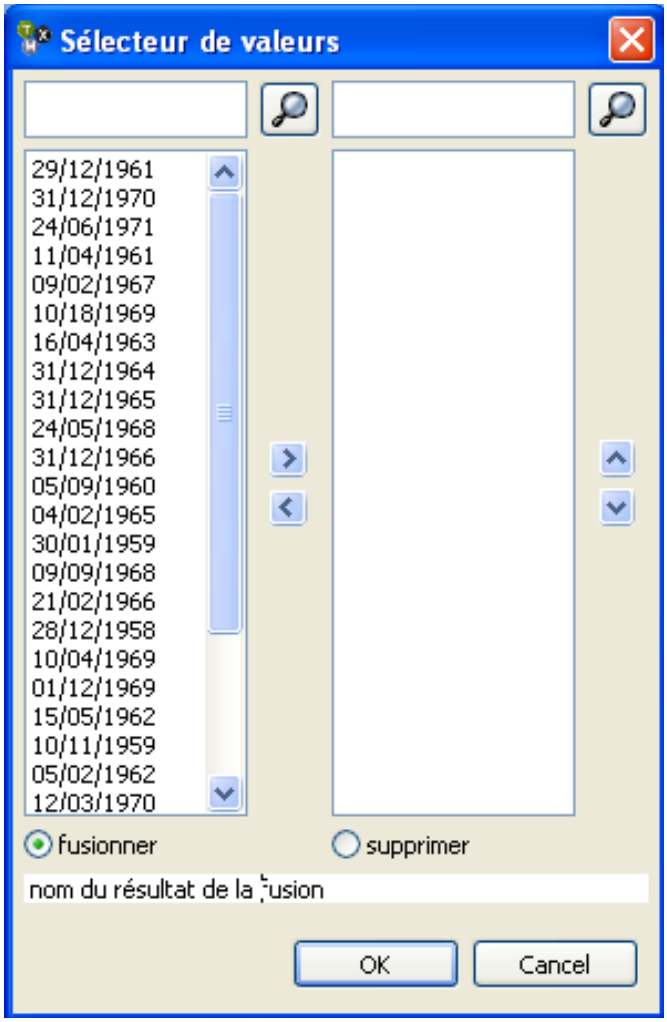


Illustration 7.24 : Fenêtre d'édition de colonnes

- Cette fenêtre offre la possibilité de sélectionner certaines colonnes. Soit via le champ de recherche (qui filtre par mots ou expressions régulières) en haut de la fenêtre, soit en sélectionnant directement une partie.
- « > » sert à ajouter une colonne en particulier
- « < » sert à supprimer une colonne
- Il faut ensuite cocher « `fusionner` » ou « `supprimer` » en fonction du résultat souhaité. Dans le cas de la fusion, il faut préciser le nom du nouveau champ.

– Fusionner ou supprimer des lignes :

  – en cliquant sur le bouton « `Fusion ou Suppr. de lignes` » : une fenêtre similaire à celle des colonnes vous permet d'éditer le nombre de lignes du tableau.

  - en sélectionnant directement les lignes que vous souhaitez supprimer ou fusionner, puis en accédant au menu contextuel via un clic-droit.
  - en cliquant sur « OK », on retourne à la table lexicale mise à jour
- Il est possible d'exporter la table obtenue grâce au menu contextuel.
- Enfin, vous pouvez trier les colonnes en cliquant sur leurs en-têtes.

7.10.1 Sauvegarde d'une table lexicale

Les tables lexicales, comme les autres résultats de calcul, sont perdus quand vous quittez TXM (seuls les sous-corpus et les partitions sont conservés). Si vous souhaitez conserver une table lexicale pour la réutiliser au-delà de la session de travail courante, le principe est de la conserver dans un fichier en l'exportant, puis de la récupérer dans une nouvelle session de TXM en l'important.

7.10.1.1 Exporter une table lexicale

Attention, ce n'est pas la commande « export » habituelle qu'il faut utiliser (celle qui sert à récupérer la table pour l'afficher dans un tableur comme Calc et Excel). C'est un autre export (au format R), qui est accessible en faisant un clic droit sur une des cellules de la table elle-même : commande « exporter la table ». Vous enregistrez alors votre fichier avec comme extension « .csv » (l'encodage des caractères utilisé est UTF-8),

7.10.1.2 Importer une table lexicale

Il vous faut préalablement créer une table lexicale quelconque sous la même partition. Ensuite, par un clic droit sur une cellule de cette table, vous retrouvez dans le menu contextuel qui s'ouvre la commande « importer une table », vous indiquez alors le fichier contenant la table que vous aviez exportée : celle-ci va remplacer le contenu de la table courante.
Remarque : les modifications que vous effectuez sur la table dans TXM ne sont pas enregistrées dans le fichier depuis lequel vous avez importé la table. Donc si vous voulez enregistrer les modifications faites lors d'une session il vous faut faire un nouvel export de la table.

7.11 Spécificités

La commande Spécificités  calcule une statistique indiquant si les occurrences d'un mot ou d'une requête CQL quelconque paraissent en surnombre (ou en sous-effectif) dans chaque partie d'une partition, ou dans un sous-corpus donné (par rapport à son corpus parent).

7.11.1 Indice de spécificité

Afin d'analyser la spécificité d'apparition d'un événement textuel dans une partie d'un corpus plutôt qu'une autre, un événement étant défini comme l'apparition d'un mot ou d'une expression CQL quelconque, on peut progressivement estimer le nombre d'apparitions le plus vraisemblable de la manière suivante :

- Le décompte des occurrences de l'expression CQL (ou d'une forme graphique simple) dans chaque partie, soit la fréquence, permet de se faire une première idée contrastive entre les parties.
- Diviser cette fréquence par le nombre total d'occurrences se trouvant dans la partie considérée (ou dira aussi la taille de la partie) permet d'utiliser les « fréquences relatives » (comme dans le moteur Stella de la base de textes FRANTEXT par exemple). On a alors « normalisé » la fréquence ou encore on l'a pondérée indépendamment de la taille de chaque partie. Ce qui permet de comparer plus sereinement les fréquences entre elles.
- On peut faire plus précis que cela encore : c'est l'objet du calcul de la mesure de spécificité d'une apparition dans une partie mise en œuvre dans TXM. En effet, normaliser en divisant par la taille de la partie nous fait considérer implicitement (ou non) que les fréquences relatives sont représentatives des fréquences d'origine (avant la division par la taille). Pour ce faire, en se trompant le moins possible en dehors de toute information complémentaire, on peut considérer la fréquence relative comme étant le maximum de vraisemblance du nombre d'apparition dans une partie de taille quelconque selon une loi d'apparition normale. On considère en quelque sorte que la fréquence relative se comporte comme le mode d'une distribution de probabilité normale (le milieu de la cloche de Gauss, là où c'est le plus élevé et donc le plus probable), soit la moyenne (cf. propriétés de la loi normale : moyenne, écart-type...). Or, il se trouve que la probabilité d'apparition d'une forme graphique - ou de façon plus générale d'une expression CQL - dans une partie n'a aucune raison de se comporter selon une loi normale. C'est-à-dire dont la distribution ressemble à une belle cloche de Gauss, avec une moyenne, un écart-type, etc. C'est

$$mode\left(card\left\{A \in V \mid A \in p\right\} = f\right) = \frac{\left(F+1\right)\times\left(t+1\right)}{T+2}$$

Équation 7.25: Maximum de vraisemblance d'apparition dans une partie.

- *A : l'événement recensé ;*
- *V : l'ensemble des événements possibles (le vocabulaire pour les mots) ;*
- *p : la partie considérée ;*
- *f : la fréquence de l'événement dans la partie ;*
- *F : la fréquence totale de l'événement dans le corpus ;*
- *t : le nombre total d'événements ayant lieu dans la partie ;*
- *T : le nombre total d'événements ayant lieu dans l'ensemble des parties.*

ce qu'a fait remarquer Pierre Lafon dans sa thèse [?], en insistant sur la déformation de la distribution pour les petites fréquences (≪20 par exemple) qui ne ressemble pas du tout à une cloche de Gauss. Il a formalisé cette apparition et constaté qu'elle était plutôt du type hypergéométrique. Cette loi de probabilité est très générale et apparaît sous diverses formes. Mais le plus souvent dans le cas qui nous préoccupe, elle ressemble à une cloche de Gauss dissymétrique vers la droite avec une queue s'affaissant petit à petit vers les hautes fréquences. Et le mode de cette distribution, c'est à dire le maximum de vraisemblance d'apparition que nous cherchons à estimer ne s'obtient pas par une moyenne arithmétique mais plutôt par l'équation 7.25.

Dans TXM, le calcul de la probabilité qu'une forme *A* apparaisse *f* fois dans une partie *p* de longueur *t*, la forme apparaissant *F* fois en tout dans l'ensemble du corpus dont la longueur totale est de *T* occurrences, a été modélisé par Pierre Lafon [Lafon80] et peut s'exprimer formellement par l'équation 7.26 [46].

$$Prob_{spécif}(card\{A \in V | A \in p\} = f) = \frac{C_F^f \times C_{T-F}^{t-f}}{C_T^t}$$

Équation 7.26: Probabilité d'apparition dans une partie.

$C_n^k = \frac{n!}{k!(n-k)!}$ *est le nombre d'échantillons de k éléments parmi n éléments, ou le nombre de parties de k éléments dans un ensemble de n éléments.*

$n! = 1 \times 2 \times 3 \times ... \times (n-1) \times n$

Le calcul exact de l'indice de spécificité utilisée dans TXM est celui du calcul de la probabilité du fait que l'événement apparaisse autant de fois qu'on l'observe effectivement dans la partie (soit f_{obs}) **ou plus fréquemment encore** à concurrence de la taille de la partie (en suivant la loi hypergéométrique décrite par l'équation 7.26 qui dépend de *f*, *t*, *F* et *T*). Concrètement, on obtient cette mesure en sommant les valeurs de la probabilité $Prob_{spécif}$ pour chaque fréquence d'apparition possible comme le montre l'équation 7.27 .

$$Prob_{spécif}(card\{A \in V | A \in p\} \geqslant f_{obs}) = \sum_{f=f_{obs}}^{card\{A \in V | A \in p\}} Prob_{spécif}(card\{A \in V | A \in p\} = f)$$

Équation 7.27: Indice de spécificité

[46] On peut obtenir cette équation en procédant grossièrement de la manière suivante. Si il y a C_F^f manières d'obtenir f éléments parmi F et C_{T-F}^{t-f} manières de combiner les formes restantes du corpus alors il y a $C_F^f \times C_{T-F}^{t-f}$ manières d'obtenir f fois la forme A dans un échantillon de t occurrences. Le quotient de ce nombre par le nombre de manières d'obtenir des échantillons différents de t occurrences parmi T (c'est-à-dire C_T^t) nous donne la probabilité recherchée.

7.11.2 Calcul direct de l’indice de spécificité

La macro livrée avec TXM « ExecR » permet de calculer l'indice de spécificité pour différentes valeurs de ses paramètres[47]. En effet, par défaut, le script R exemple qu'elle exécute affiche la courbe de la distribution de probabilité de la spécificité.

Pour utiliser cette macro :

- ouvrir la vue « Vues / Macro » ;
- double-cliquer sur la macro « ExecR » :

[47] https://groupes.renater.fr/wiki/txm-users/public/macros#execr

Illustration 7.28: Paramètres macro de la ExecR exemple

- la fenêtre des paramètres s'ouvre (ill. 7.28). Les paramètres par défaut sont ceux de l'exemple du mot « peuple » prononcé dans le discours D9 de Robespierre illustré dans [Lafon80] (voir la Figure 1, pp 140-141) :
  - f la fréquence de la forme dans la partie ;
  - F la fréquence totale de la forme dans le corpus ;
  - t le nombre total d'occurrences de la partie ;
  - T le nombre total d'occurrences du corpus.

- cliquer ensuite sur « Exécution » pour afficher la courbe de la densité de probabilité avec ces paramètres (ill. 7.29) :
  - le nombre d'apparitions le plus probable (le mode) est de 5 ;
  - la probabilité d'apparaître exactement 11 fois dans le discours D9 est de 0,01013 % ;
  - la probabilité d'apparaître 11 fois et plus dans le discours D9 (l'indice de spécificité) est de 0,01699 %

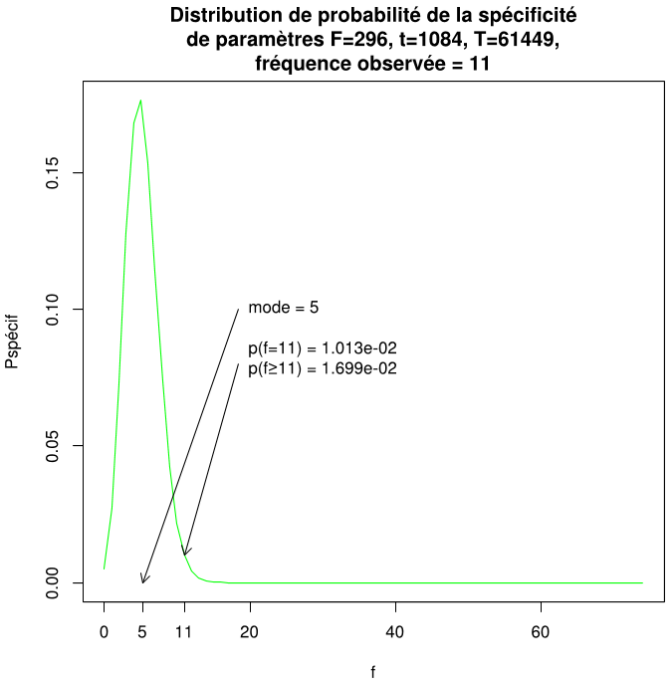


Illustration 7.29: Distribution de probabilité de la spécificité de paramètres 296, 1084 et 61449.

Pour reproduire la forme de la distribution correspondant à un cas précis se trouvant dans un tableau de résultats de spécificités, il suffit donc de lancer la macro ExecR avec les paramètres f, F, t et T correspondants à la cellule du tableau.

7.11.3 Présentation des résultats

Dans TXM, la spécificité est représentée par la partie entière des logarithmes en base 10 (log_{10}) des estimations de probabilité de spécificité car, comme le nom hypergéométrique le suggère, les probabilités obtenues par les calculs varient dans un domaine exponentiel et l'ordre de grandeur de la probabilité suffit souvent à la comparer aux autres. On compare donc des ordres de grandeur plutôt que les probabilités elles-mêmes.

Par convention, la représentation de la sous spécificité (ou sous-représentation) se distingue de celle de la sur spécificité (ou sur-représentation) par un signe moins (-) situé devant l'indice. On s'intéressera alors aux faibles probabilités (donc aux valeurs de log_{10} importantes) qui rendent compte :

- soit d'un nombre d'apparitions plus faible que prévu si l'observation est inférieure au mode de la distribution théorique (c'est-à-dire si le nombre d'apparitions de l'événement dans la partie est inférieur au maximum de vraisemblance estimé par notre modélisation hypergéométrique de la distribution (cf. l'équation 7.25)). On parlera alors de sous-spécificité ou spécificité négative ;
- soit d'un nombre d'apparition plus important que prévu si l'observation est supérieure au mode de la distribution théorique. On parlera alors de sur-spécificité ou spécificité positive.

À ne pas confondre avec les fortes probabilités (par exemple supérieures à 5% de chance), donc aux valeurs de log_{10} faibles, qui indiqueront plutôt la banalité de l'apparition dans la partie (car prévisibles d'après le modèle des spécificités).

Pour les personnes intéressées par la valeur exacte de la probabilité calculée plutôt qu'au classement des événements entre eux par le biais de l'ordre de grandeur de cette probabilité (qui est, notre usage principal des estimations de probabilité), une macro TXM permet non seulement de réaliser directement le calcul de l'indice de spécificité en fonction des paramètres du modèle mais surtout de situer cette valeur dans la courbe de densité de probabilité (voir la section 7.11.2 page 131 « Calcul direct de l'indice de spécificité »).

7.11.4 Spécificités d'une partition

La commande Spécificités appliquée à une partition ouvre la fenêtre de paramètres suivante :

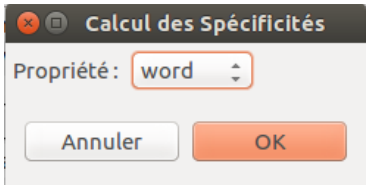


Illustration 7.30: Paramètres des spécificités d'une partition.

– propriété de mot : propriété qui fera l'objet du calcul.

Les résultats sont présentés sous forme de tableau (voir l'exemple figure 7.31) :

- lignes : les différentes valeurs de la propriété de mot considérée (par exemple les différentes formes de mots) ;
- colonnes :
  - la première colonne affiche la valeur de la propriété correspondant à la ligne (par exemple la forme « nous ») ;
  - la deuxième colonne affiche la fréquence totale 'F' de cette valeur dans tout le corpus (par exemple 694 « nous » dans le corpus). Dans le titre de la colonne, 'T' représente le nombre total d'occurrences du corpus (par exemple une taille totale de 100 810 mots) ;
  - les autres colonnes fonctionnent par paire :
    - une première colonne affiche la fréquence de la valeur dans la partie (par exemple 6 occurrences de « nous » dans la partie « Allocution radiotélévisée »). Dans le titre de cette colonne, 't' représente la taille de la partie ;
    - la seconde affiche l'indice de spécificité de la valeur pour la partie (par exemple 21,3 de spécificité pour « nous » dans la partie).

L'illustration 7.31 présente les résultats de la commande Spécificités portant sur la forme graphique de tous les mots de la partition sur le type de discours du corpus DISCOURS. La tableau est trié dans l'ordre décroissant de la colonne d'indice de spécificité de la partie « Allocution radiotélévisée ». On peut y lire que les formes les plus spécifiques du discours de type « Allocution radiotélévisée » sont :

- « nous » ayant un indice de spécificité de 21,3 pour 241 apparitions dans ce genre sur un total de 694 apparitions dans le corpus ;
- « notre » ayant un indice de spécificité de 13,6 pour 124 apparitions dans ce genre sur un total de 335 apparitions ;
- etc.

type: word:[]

Unités	Fréquence T 100810	Allocution radiotélévisée t=19396	score	Conférence de presse t=73501	score	Entretien radiotélévisé t=7913	score
nous	694	241	21,3	411	-14,4	42	-1,4
notre	335	124	13,6	196	-8,1	15	-2,0
Vive	24	22	13,5	2	-10,3	0	-0,9
!	177	76	12,3	82	-13,1	19	1,0
Françaises	19	18	11,7	0	-10,8	1	-0,3
année	64	35	9,5	25	-7,8	4	-0,4

Illustration 7.31 : Spécificités des mots de la partition sur la propriété de texte (ou variable) appelée « type » du corpus DISCOURS .

7.11.4.1 Tri des résultats

On peut trier le tableau en cliquant sur les entêtes de colonnes. Cliquer une seconde fois inverse l'ordre de tri.

Trier une colonne d'indice de façon décroissante, permet d'accéder rapidement aux mots considérés comme étant les plus sur-utilisés par rapport à l'ensemble du corpus. Les derniers mots de la liste sont considérés comme sous-utilisés et les mots intermédiaires – autour de l'indice 0 – sont considérés comme banals (ni sur- ni sous-représentés).

7.11.4.2 Visualisation graphique des indices de spécificité

Voir aussi la documentation commune à toutes les visualisations dans la section « 7.14 Visualisation graphique des résultats » page 141.

Les indices de spécificité peuvent être visualisés sous forme graphique. On sélectionne dans le tableau de résultats au moyen de la souris[48] les lignes pour lesquelles on souhaite une visualisation puis on lance la commande « Calculer le graphique des lignes sélectionnées » via le menu contextuel. Cela produit un graphique comme illustré figure 7.32 :

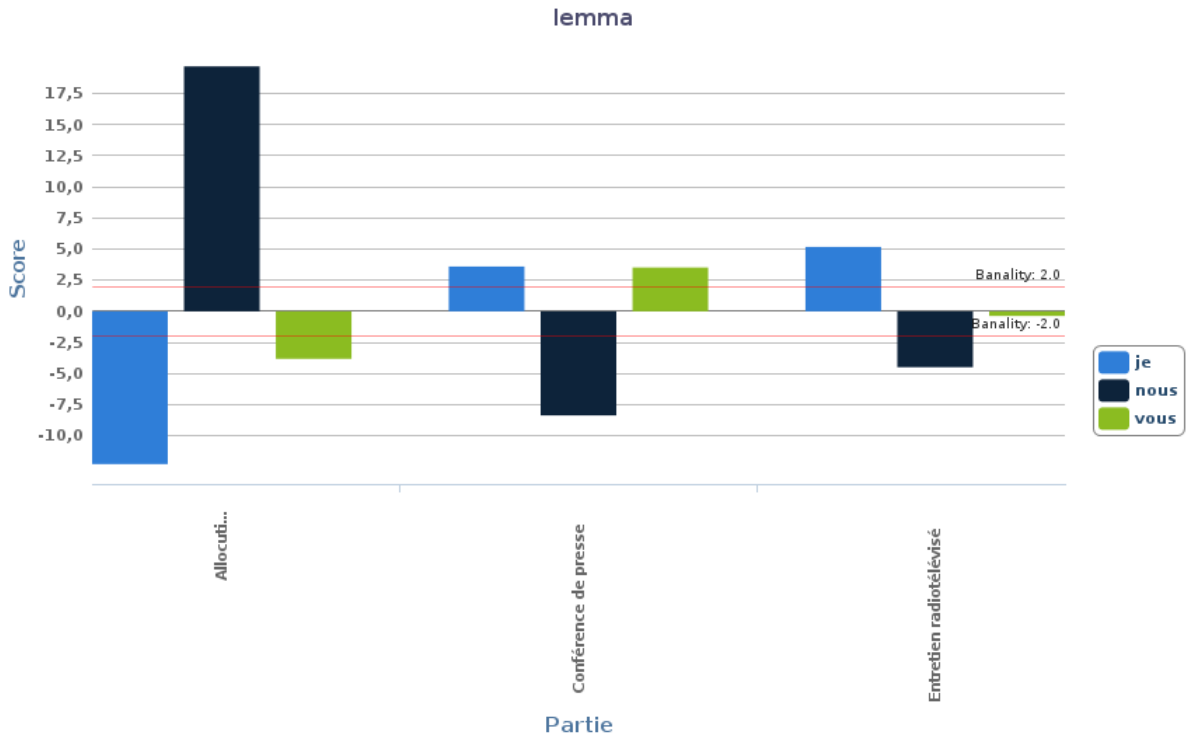


Illustration 7.32 : Graphique de spécificité des lemmes « je », « nous » et « vous » des trois types de discours dans le corpus DISCOURS.

[48] Shift-clic gauche permet de sélectionner plusieurs lignes contiguës. Ctrl-clic gauche permet de sélectionner plusieurs lignes non contiguës.

Dans le graphique :

- chaque partie est représentée par un ensemble de barres contiguës, classées dans le même ordre que dans le tableau ;
- chaque propriété de mot (forme graphique du mot dans cet exemple) sera représentée par une barre de la même couleur dans chaque partie ;
- les couleurs sont légendées dans le coin inférieur droit du graphique ;
- deux lignes rouges délimitent la bande de banalité autour de l'axe d'indice 0 (les barres qui n'en sortent pas sont à considérer comme banales).

Le graphique est exportable sous forme d'image via le bouton « Export » de la barre d'outils.

7.11.5 Spécificités d'une table lexicale

On peut appliquer le calcul de spécificités sur une table lexicale (issue d'une partition). Dans ce contexte, la propriété de mot à considérer a déjà été choisie et le calcul se lance directement.

7.11.6 Spécificités d'un sous-corpus

La commande Spécificités sur un sous-corpus permet de choisir la propriété de mot sur laquelle seront appliqués les calculs, par le biais d'une fenêtre de paramètres similaire à celle de la commande Lexique, comme on peut le voir sur l'illustration Erreur : source de la référence non trouvée page Erreur : source de la référence non trouvée.

Alloc_Radio: word

Unités	FréquenceT105191	Alloc_Radio t=20290	score	DISCOURS \ Alloc_Radio t=84901	score
-	90	16	-0,4	74	0,4
—	114	4	-6,2	110	6,2
,	8603	1724	1,5	6879	-1,5
;	151	31	0,4	120	-0,4
:	140	21	-0,9	119	0,9
!	177	76	12,3	101	-12,3

Illustration 7.33 : Spécificités des formes graphiques de la partie « Allocution radiotélévisée » du corpus DISCOURS.

Les résultats sont présentés sous forme de tableau (voir l'exemple figure 7.33) :

- lignes : les différentes valeurs de la propriété de mot considérée (par exemple les différentes formes de mots) ;
- colonnes :
  - la première colonne affiche la valeur de la propriété correspondant à la ligne (par exemple la forme « - ») ;

- la deuxième colonne affiche la fréquence totale 'F' de cette valeur dans tout le corpus (par exemple 90 « - » dans le corpus). Dans le titre de la colonne, 'T' représente le nombre total d'occurrences du corpus (par exemple une taille totale de 105 191 mots) ;
- la troisième colonne affiche la fréquence de la valeur dans le sous-corpus (par exemple 16 occurrences de « - »). Dans le titre de cette colonne qui mentionne le nom du sous-corpus, 't' représente la taille de la partie ;
- la quatrième colonne affiche l'indice de spécificité de la valeur pour la partie (par exemple spécificité de -0,4 pour « - » dans le sous-corpus) ;
- la cinquième colonne affiche la fréquence de la valeur dans le complémentaire du sous-corpus (par exemple 74 occurrences de « - »). Dans le titre de cette colonne qui mentionne le « nom du corpus \ le nom du sous-corpus », 't' représente la taille du complémentaire ;
- la sixième colonne affiche l'indice de spécificité de la valeur pour le complémentaire (par exemple spécificité de 0,4 pour « - » dans le complémentaire).

7.12 Analyse Factorielle des Correspondances (AFC)

Voir aussi la documentation commune à toutes les visualisations dans la section « 7.14 Visualisation graphique des résultats » page 141.

La commande AFC  calcule l'analyse factorielle des correspondances[49] d'une partition, où chaque partie est représentée par un vecteur de fréquence d'une propriété de mot (`word form`, `lemma`, `pos`...).

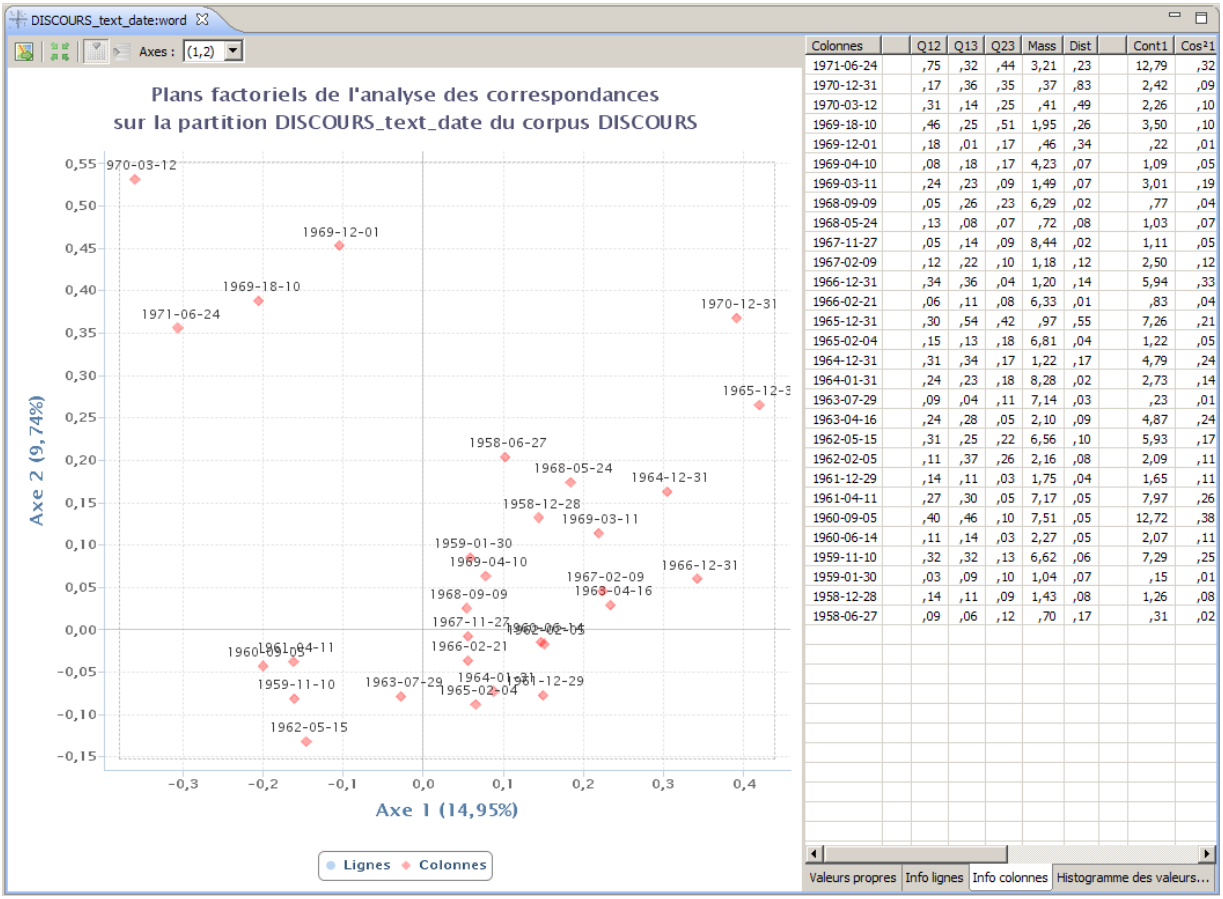
Cette commande doit être appliquée à une partition constituée d'au moins quatre parties. Elle peut également être appelée depuis une table lexicale. Il faut tout d'abord choisir une propriété de mot. Ensuite, une fenêtre similaire à celle de la commande `Lexique` s'ouvre, comme le montre l'illustration Erreur : source de la référence non trouvée.

Les résultats sont affichés dans deux fenêtres :

- la première fenêtre affiche le graphique du premier plan factoriel
- la seconde fenêtre réunit des tableaux de données qui permettent d'interpréter le graphique. Elle se subdivise en quatre onglets :
  - les valeurs propres

[49] Jean-Paul Benzécri et al., *L'analyse des correspondances* (Paris: Dunod, 1973).

- les informations sur les lignes
- les informations sur les colonnes
- l'histogramme des valeurs propres



Colonnes		Q12	Q13	Q23	Mass	Dist		Cont1	Cos²1
1971-06-24		,75	,32	,44	3,21	,23		12,79	,32
1970-12-31		,17	,36	,35	,37	,83		2,42	,09
1970-03-12		,31	,14	,25	,41	,49		2,26	,10
1969-18-10		,46	,25	,51	1,95	,26		3,50	,10
1969-12-01		,18	,01	,17	,46	,34		,22	,01
1969-04-10		,08	,18	,17	4,23	,07		1,09	,05
1969-03-11		,24	,23	,09	1,49	,07		3,01	,19
1968-09-09		,05	,26	,23	6,29	,02		,77	,04
1968-05-24		,13	,08	,07	,72	,08		1,03	,07
1967-11-27		,05	,14	,09	8,44	,02		1,11	,05
1967-02-09		,12	,22	,10	1,18	,12		2,50	,12
1966-12-31		,34	,36	,04	1,20	,14		5,94	,33
1966-02-21		,06	,11	,08	6,33	,01		,83	,04
1965-12-31		,30	,54	,42	,97	,55		7,26	,21
1965-02-04		,15	,13	,18	6,81	,04		1,22	,05
1964-12-31		,31	,34	,17	1,22	,17		4,79	,24
1964-01-31		,24	,23	,18	8,28	,02		2,73	,14
1963-07-29		,09	,04	,11	7,14	,03		,23	,01
1963-04-16		,24	,28	,05	2,10	,09		4,87	,24
1962-05-15		,31	,25	,22	6,56	,10		5,93	,17
1962-02-05		,11	,37	,26	2,16	,08		2,09	,11
1961-12-29		,14	,11	,03	1,75	,04		1,65	,11
1961-04-11		,27	,30	,05	7,17	,05		7,97	,26
1960-09-05		,40	,46	,10	7,51	,05		12,72	,38
1960-06-14		,11	,14	,03	2,27	,05		2,07	,11
1959-11-10		,32	,32	,13	6,62	,06		7,29	,25
1959-01-30		,03	,09	,10	1,04	,07		,15	,01
1958-12-28		,14	,11	,09	1,43	,08		1,26	,08
1958-06-27		,09	,06	,12	,70	,17		,31	,02

Illustration 7.34 : AFC obtenue à partir d'une table lexicale sur les « Dates » du corpus DISCOURS.

La fenêtre de visualisation des plans factoriels permet de choisir quels éléments sont affichés dans le graphique : pour cela, cliquez sur les boutons « Afficher les colonnes » ou « Afficher les lignes » de la barre d'outils des graphiques.

L'échelle du graphique peut être modifiée avec la molette de la souris et sa position avec le bouton droit de la souris.

L'échelle et la position du graphique peuvent être réinitialisées en cliquant sur le bouton « Rétablir la vue initiale ».

La vue courante du graphique peut être exportée avec la commande « Exporter la vue » dans différents formats sélectionnables dans la boîte de dialogue de l'export.

Voir également les raccourcis graphique de zoom, déplacement etc. dans la section 7.14.

Par défaut, l'AFC affiche seulement les parties (colonnes) dans le plan factoriel.

Ce paramètre peut être modifié dans les préférences de l'AFC, dans la section « Rendu des graphiques » :

- « Afficher les lignes » : affiche les propriétés de mot ;
- « Afficher les colonnes » : affiche les parties.

Dans le volet de droite, diverses informations sont disponibles afin d'aider l'utilisateur à interpréter les coordonnées des colonnes (variables) ou des lignes (individus).

Le tableau des valeurs propres indique leur rang, leur valeur, leur pourcentage d'inertie ainsi que le cumul des pourcentages.

Le graphe en barres des valeurs propres en donne un aperçu analogique.

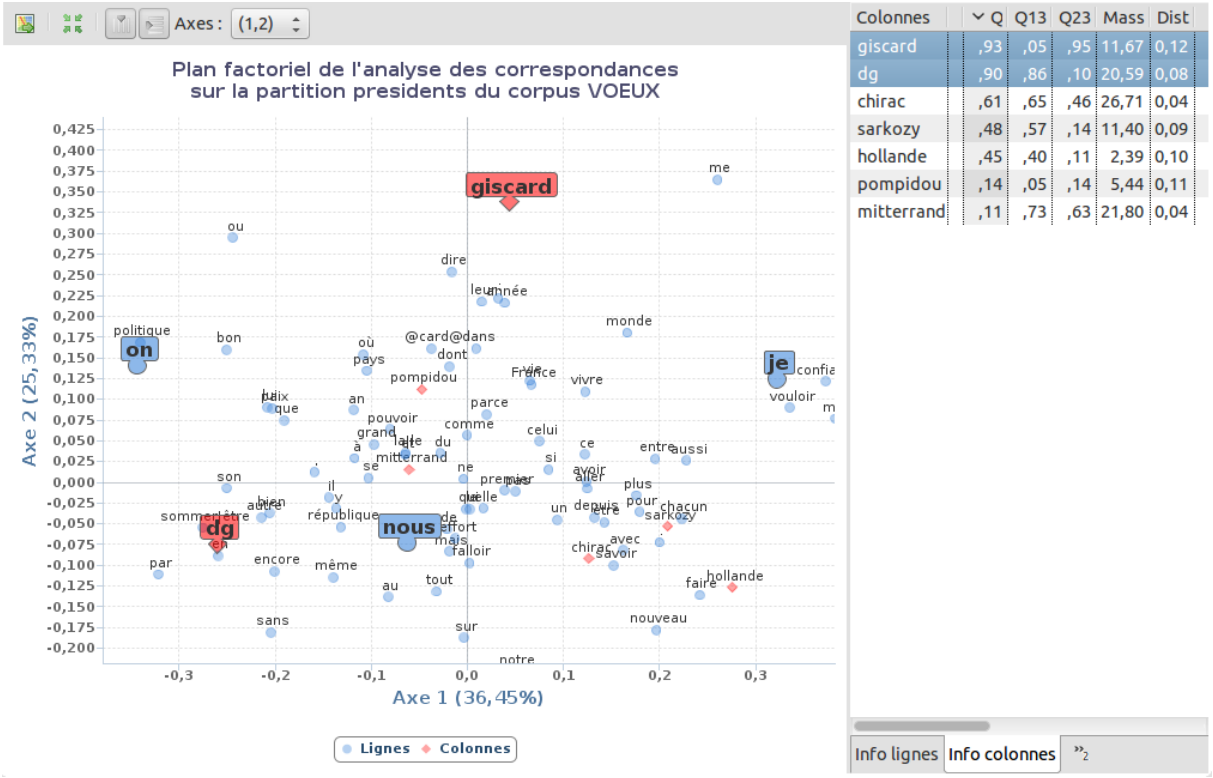
Les tableaux d'information sur les colonnes et les lignes indiquent :

- la qualité des plans « Q- »: la représentation du point dans chaque plan, calculée comme la somme des cos² du point sur les deux axes concernés : plus la qualité est proche de 1, moins la position du point est déformée par la projection dans le plan.
- le poids relatif « Mass »: la fréquence est rapportée à la somme des fréquences des autres mots (lignes).
- le carré de la distance du point à l'origine « Dist » (l'origine est le centre de gravité du nuage de points : plus la distance est grande, plus le point s'écarte du profil moyen, autrement dit plus il est original par rapport au reste du corpus)
- la participation du point à la construction de l'axe « Cont- ». La somme des contributions vaut 100 et les points qui présentent les plus fortes contributions pour un axe donné servent à interpréter l'axe.
- le cos² du point sur chaque axe « Cos² »: la mesure de l'angle entre le vecteur représentatif du point et l'axe. Un cos² proche de 1 indique que le point est bien représenté sur l'axe alors qu'un cos² proche de 0 indique que la projection déforme fortement le point par rapport à cet axe et qu'il vaut mieux donc éviter d'interpréter la position du point par rapport aux autres selon la coordonnée sur cet axe. En particulier, un point qui a un cos² faible sur les deux axes de la représentation choisie a une position trompeuse ; sa proximité apparente avec d'autres points ne doit pas être interprétée dans ce plan.
- les coordonnées des points dans l'espace des trois premiers axes « c- ».

Les fenêtres de résultats offrent un mécanisme de sélection multiple et de mise en évidence des points combiné entre les points des graphiques et les lignes des tableaux de données. Cliquer sur un point dans le graphique ou dans l'un des tableaux a pour effet de le mettre en

surbrillance. La sélection multiple se fait par le mécanisme habituel du système d'exploitation : Ctrl-Clic gauche (Windows et Linux) et Cmd-Clic gauche (Mac OS X) permute entre l'ajout et le retrait d'un point dans la sélection en cours.

La commande de recherche par expression régulière dans un tableau de données (raccourcis Ctrl-F) peut être utilisée conjointement avec la mise en évidence par sélection multiple.



Colonnes	Q	Q13	Q23	Mass	Dist
giscard	,93	,05	,95	11,67	0,12
dg	,90	,86	,10	20,59	0,08
chirac	,61	,65	,46	26,71	0,04
sarkozy	,48	,57	,14	11,40	0,09
hollande	,45	,40	,11	2,39	0,10
pompidou	,14	,05	,14	5,44	0,11
mitterrand	,11	,73	,63	21,80	0,04

Illustration 7.35 : Exemple de mise en évidence de points par sélection multiple dans une AFC créée depuis une partition sur les présidents dans le corpus VOEUX.

La commande « Exporter la vue » conserve ces sélections multiples dans les fichiers générés.

L'algorithme de la commande AFC est implémenté par le package FactoMineR.

Pour de plus amples informations, notamment d'un point de vue R, merci de consulter la documentation de ce package :

- documentation R officielle : http://cran.r-project.org/web/packages/FactoMineR/index.html
- manuel PDF : http://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf

- site web de référence : http://factominer.free.fr
- documentation de référence (dont monographies) : http://factominer.free.fr/docs/index.html

7.13 Classification (CAH)

Voir aussi la documentation commune à toutes les visualisations dans la section « 7.14 Visualisation graphique des résultats » page 141.

Après avoir calculé une AFC, on peut lancer une classification sur l'AFC elle même.

Les options principales sont :

- le nombre de classes
- le choix de la dimension à classer : colonnes ou lignes

On trouvera des paramètres complémentaires dans les préférences de la classification :

- format de l'arbre des classes (dendrogramme) : 2D ou 3D
- dimension de classement par défaut : lignes ou colonnes
- nombre de classes par défaut
- distance à utiliser pour le calcul

L'algorithme de cette commande est implémenté par le package FactoMineR. Merci de consulter la documentation de ce package pour de plus amples informations.

7.14 Visualisation graphique des résultats

Certaines commandes de TXM produisent des représentations graphiques dans des onglets dédiés contenant une barre d'outils spécifique à la visualisation et à la manipulation des graphiques.

7.14.1.1 Manipulation interactive

Vous pouvez interagir avec les graphiques :

- changement d'échelle : molette de la souris ou Ctrl + et Ctrl - (Cmd + et Cmd - sous Mac OS X)
- translation de la vue : clic droit et déplacement de la souris ou flèches du clavier

- revenir à la vue initiale : bouton de la barre d'outils de l'onglet des graphiques

7.14.1.2 Affichages complémentaires

Différentes informations sont affichées en info-bulle lorsque le curseur de la souris se trouve au dessus d'un élément du graphique (ex. barre, point, ligne). Ces données complémentaires dépendent du type de graphique (ex. AFC, dimensions de partition, etc.).

7.14.1.3 Export des graphiques

Les représentations graphiques peuvent être exportées vers d'autres logiciels avec :

- exporter la vue courante : bouton de la barre d'outils de l'onglet des graphiques. Cette commande exporte la graphique tel qu'il apparaît dans l'onglet de visualisation, c'est-à-dire en tenant compte des changements d'échelle et des déplacements latéraux effectués.
- l'export de graphique peut également se faire à partir d'un nœud de résultat de type « graphique » dans la vue « Corpus » (ex. AFC, CAH, etc.) par le biais de l'entrée « Exporter → Graphique... » du menu contextuel. Dans ce cas, l'export n'est lié à aucun onglet de graphique particulier et ne tient donc pas compte de changement d'échelle ou de déplacement latéral.

7.15 Exploitation des résultats

7.15.1 Sauvegarde et Exportation des résultats

Chaque résultat d'une commande TXM (liste, tableau, graphique) peut être exporté dans un fichier pour pouvoir être traité dans un autre logiciel (traitement de texte pour publication, tableur pour analyses complémentaires, etc.). Ce fichier est disponible au moins au format CSV pour les listes et tableaux et au format SVG pour les graphiques. La commande d'export est accessible depuis le menu contextuel, en cliquant sur l'icone de résultat dans la vue « Corpus » ou avec le bouton « Export » dans la barre d'outils si l'objet est sélectionné.

7.15.2 Traitement des résultats avec R

Pour les utilisateurs de R, il est possible de manipuler les résultats dans l'espace de travail de R. Certains résultats sont par défaut déjà disponibles dans l'environnement R : Spécificités, AFC, Classification, Progression, Table lexicale. Les autres peuvent être transférés à la demande avec la commande « Envoyer vers R » : lexique, index, concordance, corpus.

La façon d'accéder à ces résultats depuis R est documentée à la section Utilisation des résultats et objets TXM depuis R page 202.

7.15.3 Exploiter les graphiques de résultats dans d'autres logiciels

TXM produit les graphiques de résultats aux formats suivants :

Vectoriels

- SVG - Scalable Vector Graphics
  Format ouvert d'image vectoriel standardisé par le W3C
  http://www.w3.org/Graphics/SVG,
  http://fr.wikipedia.org/wiki/Scalable_Vector_Graphics ;
- PS - PostScript
  Format propriétaire d'image vectoriel de la société Adobe Systems
  http://fr.wikipedia.org/wiki/PostScript ;
- PDF - Portable Document Format
  Format propriétaire de document vectoriel de la société Adobe Systems
  http://fr.wikipedia.org/wiki/Portable_Document_Format ;

Bitmaps

- PNG - Portable Network Graphics
  Format ouvert d'image bitmap compressé sans perte normalisé par l'ISO
  http://www.w3.org/TR/PNG,
  http://fr.wikipedia.org/wiki/Portable_Network_Graphics ;
- JPEG - Joint Photographic Experts Group
  Format ouvert d'image bitmap compressé avec perte normalisé par l'ISO/CEI 10918-1 | UIT-T Recommendation T.81 http://fr.wikipedia.org/wiki/JPEG.

Le choix du format d'export se règle dans les Préférences : Préférences / Utilisateur / Export / Format des graphiques R par défaut.

Les formats vectoriels présentent l'avantage de pouvoir varier de taille sans perte de détails (zoom = agrandissement ou réduction général de l'image) et de pouvoir être édités par des logiciels spécialisés (par exemple pour ajuster la typographie en fonction de consignes éditoriales, pour améliorer la lisibilité en agrandissant ou réduisant les caractères sans changer l'échelle globale du graphique, pour déplacer la légende, etc.).

Nous recommandons :

- le logiciel gratuit et open-source « InkScape » pour éditer le format SVG
  http://www.inkscape.org/fr ;

- le logiciel commercial « Adobe Illustrator » pour éditer le format PS http://www.adobe.com/fr/products/illustrator.html.

TXM privilégie l'export des graphiques de résultats au format SVG.

Les formats bitmaps sont pris en charge par un plus grand nombre d'outils de travail et surtout sont plus faciles à manipuler dans les traitements de texte. Le format JPEG est un peu mieux pris en charge que PNG sous Windows. C'est donc le format le plus facile à manipuler en dehors de TXM, même s'il n'est pas encore très pratique à manipuler au sein de TXM lui-même (nous devons homogénéiser la façon avec laquelle il est exporté depuis TXM).

7.15.3.1 Import direct d'une image vectorielle au format SVG dans le traitement de texte LibreOffice Writer

- produire et exporter un graphique dans TXM au format SVG. Par exemple :
  - dans TXM créer une partition dans le corpus DISCOURS appelée « loc » à partir de la structure « text » et son attribut « loc » ;
  - régler le champ « Préférences / Utilisateur / Export / Format des graphiques R par défaut » à la valeur « SVG » ;
  - lancer la commande «Dimensions» sur la partition « loc » ;
  - dans la barre d'outils de l'onglet des graphiques, cliquer sur le bouton  , sélectionner le type « *.svg » dans la boîte de dialogue et sauver le graphique dans un fichier ;
- importer l'image dans Writer :
  - dans Writer lancer la commande « Insertion / Image / À partir d'un fichier » ;
  - désigner le fichier SVG en navigant jusqu'à son dossier ;
  - l'image est alors insérée à l'endroit du curseur. Vous pouvez si nécessaire changer la taille de l'image avec la souris :

- cliquer sur l'image → des poignées vertes de manipulation s'allument (voir illustration 14.1) ;

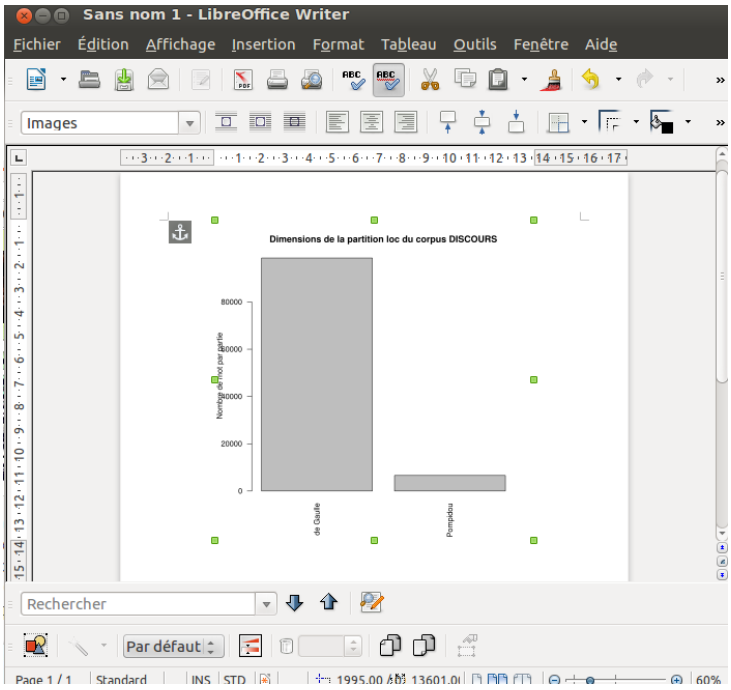


Illustration 7.36: Image SVG importée dans Writer

- « shift-clic » sur une des poignées et déplacer la souris fait varier la taille de l'image de façon homothétique (l'image n'est pas déformée) ;
- cliquer sur l'image et déplacer la souris déplace l'image dans la page.

7.15.3.2 Import direct d'une image bitmap au format JPEG dans le traitement de texte LibreOffice Writer

- produire et exporter un graphique dans TXM au format JPEG. Par exemple :
  - dans TXM créer une partition dans le corpus DISCOURS appelée « loc » à partir de la structure « text » et son attribut « loc » ;
  - lancer la commande «Dimensions» sur la partition « loc » ;
  - dans la barre d'outils de l'onglet des graphiques, cliquer sur le bouton , sélectionner le type « *.jpeg » dans la boîte de dialogue et sauver le graphique dans un fichier ;
- importer l'image dans Writer :
  - dans Writer lancer la commande « Insertion / Image / À partir d'un fichier » ;

- désigner le fichier JPEG en navigant jusqu'à son dossier ;
- l'image est alors insérée à l'endroit du curseur. Vous pouvez si nécessaire changer la taille de l'image avec la souris :
  - cliquer sur l'image → des poignées vertes de manipulation s'allument ;
  - « shift-clic » sur une des poignées fait varier la taille de l'image de façon homothétique (l'image n'est pas déformée).

7.15.3.3 Édition préalable d'un graphique au format SVG avec InkScape

- produire et exporter un graphique dans TXM au format SVG. Par exemple :
  - dans TXM créer une partition dans le corpus DISCOURS appelée « loc » à partir de la structure « text » et son attribut « loc » ;
  - régler le champ « Préférences / Utilisateur / Export / Format des graphiques R par défaut » à la valeur « SVG » ;
  - lancer la commande «Dimensions» sur la partition « loc » ;
  - dans la barre d'outils de l'onglet des graphiques, cliquer sur le bouton  , sélectionner le type « *.svg » dans la boîte de dialogue et sauver le graphique dans un fichier ;
- éditer l'image dans InkScape:
  - ouvrir le fichier SVG depuis InkScape ;
  - avec la souris tracer un rectangle autour de la légende des ordonnées « Nombre de mots par partie » pour la sélectionner (voir illustration 14.2) :

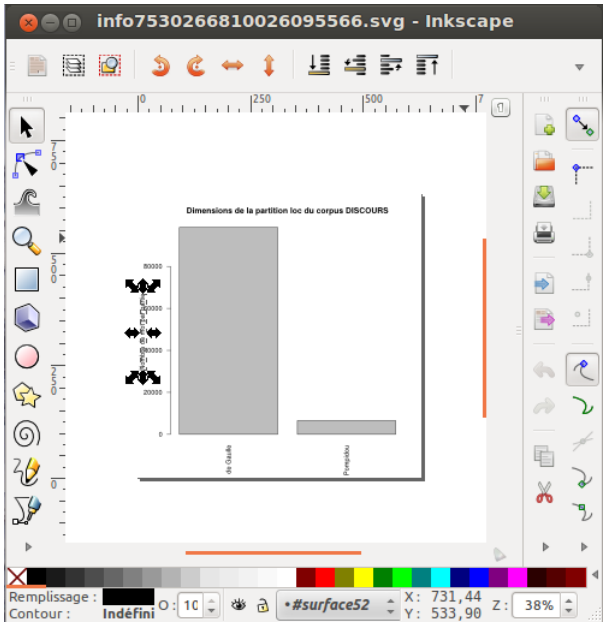


Illustration 7.37: Sélection de la légende des ordonnées

– Utiliser la touche « flèche gauche » du clavier pour translater la légende vers la gauche (ou « cliquer-glisser » avec la souris sur la sélection) :

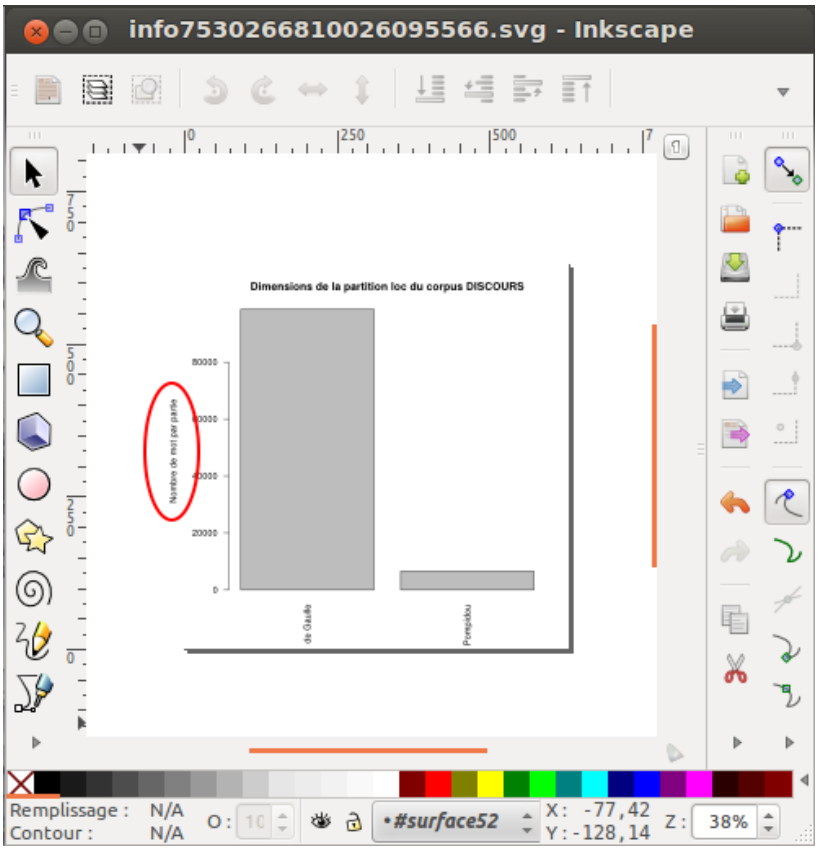


Illustration 7.38: La légende des ordonnées a été déplacée à gauche (indiquée par une ellipse rouge)

– Vous pouvez alors sauvegarder votre travail pour un import ultérieur dans un traitement de texte.

7.16 Récapitulatif des relations entre commandes et résultats dans TXM

Ces relations sont accessibles en général à partir du menu contextuel d'un résultat ou d'une icone.

COMMANDES	DEPUIS	VERS	UTILISÉ PAR
AFC	Partition Table lexicale		
Classification	AFC		AFC
Concordances	Corpus Index Lexique Cooccurrences	Édition	Cooccurrences
Cooccurrences	Corpus	Concordances	
Corpus	Import		Cooccurrences Concordances Corpus Description Édition Index Lexique Partition Progression
Description	Corpus		
Index	Corpus Partition	Concordances Progression	Table lexicale d'une partition
Lexique	Corpus	Concordances Progression	
Partition	Corpus		AFC Édition Index Spécificités Table lexicale
Progression	Corpus		
Références	Corpus	Concordances	
Sous-Corpus	Corpus		Idem que Corpus + Spécificités

COMMANDES	DEPUIS	VERS	UTILISÉ PAR
Spécificités	Partition Table lexicale Sous-corpus		
Table lexicale	Partition Index d'une partition		AFC Spécificités
Édition	Concordances Corpus Sous-Corpus Partition		
Notice	Corpus		

8 Annoter un corpus

L'annotation de corpus consiste à associer à une séquence de mots[50] d'un texte un certain nombre d'informations comme une catégorie, un mot clé, un type, une chaîne de caractères etc. puis à exploiter ces informations avec les outils de TXM.

Plusieurs outils d'annotation commencent à être développés au sein de TXM :

- l'annotation simple ou avancée de séquences de mots par pivots de concordances ;
- l'annotation Analec/Glozz de séquences de mots au sein d'éditions de texte.

Le premier outil est directement utilisable dans TXM, le second par le biais de l'installation de l'extension Analec.

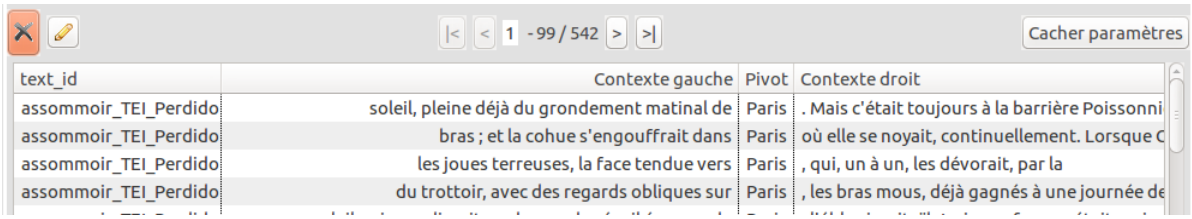
8.1 Annotation simple par concordances

Pour les corpus importés avec le module XTZ+CSV[51], un nouveau bouton d' « Annotation » (bouton crayon) situé en haut à gauche des concordances permet d'annoter des pivots directement.

Par exemple, dans une concordance de « `Paris` » dans l'Assommoir d'Émile Zola:

[50] éventuellement limitée à un seul mot.

[51] Pour pouvoir utiliser ce module d'import, il suffit d'avoir des sources au format XML. Pour passer facilement du format texte brut (TXT) au format XML, voir la macro TXT2XML.



on annote 3 lignes particulières avec l'annotation « ville » :

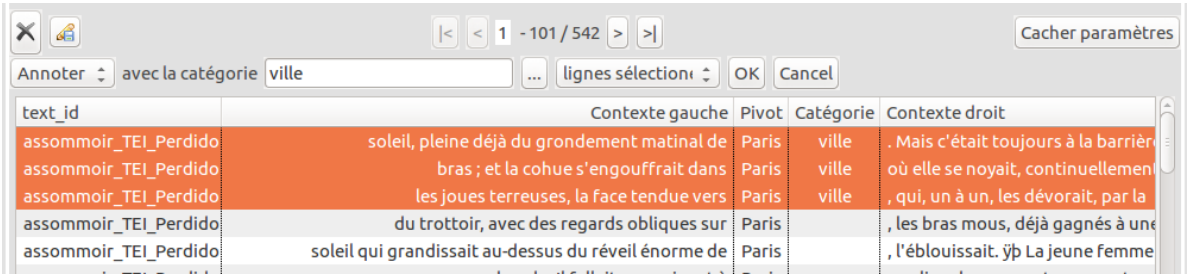


Illustration 8.1: Annotation d'occurrences du mot "Paris" par la catégorie "ville"

8.1.1 Sauvegarde des annotations et exploitation avec TXM

Après avoir cliqué dans une concordance sur le bouton "Annoter", qui a une icone de crayon, pour lancer une session d'annotation, il bascule en bouton "Enregistrer les annotations" avec l'icone crayon+disquette. Cliquer sur le bouton crayon+disquette sauvegarde les annotations de la session courante, ce qui permet ensuite d'exploiter les annotations. On peut utiliser les annotations directement dans des requêtes CQL à l'aide de l'expression de structure « <span_ref="..."> ». C'est à dire que l'annotation crée une nouvelle structure « span » dans le corpus (autour des pivots annotés), et l'annotation devient la valeur de son attribut « @ref ».

Par exemple, pour refaire directement la concordance des 3 pivots annotés précédemment :

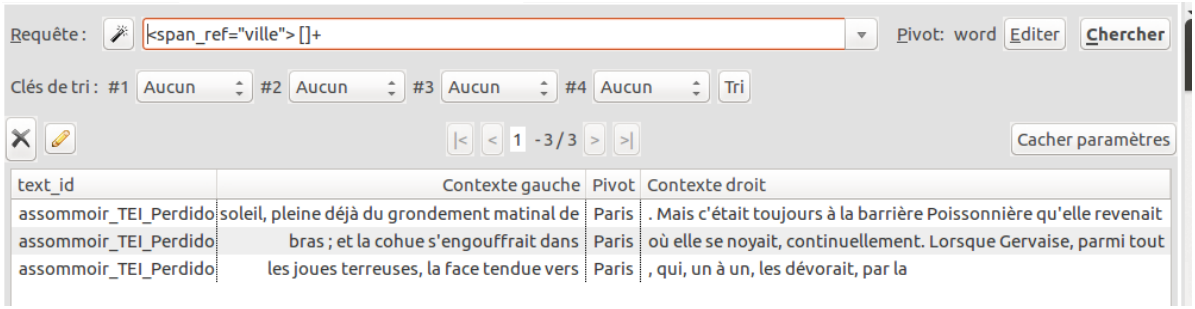


Illustration 8.2: Recherche de l'annotation "ville"

8.1.2 Encodage de plusieurs informations dans l'annotation

L'annotation est une chaîne de caractères quelconque et de n'importe quelle longueur, donc on peut la structurer librement, par exemple avec des séparateurs (« `type=lieu,sous-`

`type=ville,valeur=Paris...` » ou « `lieu,ville,Paris...` » par exemple), puis la déstructurer au moment des extractions par des expressions régulières dans les requêtes CQL. Par exemple : tout ce qui est étiqueté 'lieu-qlqchose' : `<span_ref="type=lieu,.*"> []+ </span>` pour l'encodage « `type=lieu,sous-type=ville,valeur=Paris...` » ou encore `<span_ref="lieu,.*"> []+ </span>` pour l'encodage « `lieu,ville,Paris...` ».

Autres exemples de requêtes d'extraction pour l'encodage « `lieu,ville,Paris...` » :

- le premier mot des séquences annotées par « lieu » : `<span_ref="lieu,.*"> []`
- tous les mots des séquences annotées par « lieu » : `<span_ref="lieu,.*"> []+ expand to span`
- chaque mot des séquences annotées par « lieu » : `[_.span_ref="lieu,.*"]`
- tout ce qui est étiqueté 'ville-qlqchose' en seconde position : `<span_ref=".*,ville,.*"> []+ </span>`
- tout ce qui est étiqueté 'Paris-qlqchose' en troisième position : `<span_ref=".*,.*,Paris"> []+ </span>`
- tout ce qui est étiqueté 'lieu-qlqchose' en première position et étiqueté 'ville-qlqchose' en deuxième position : `<span_ref="lieu,ville,.*"> []+ </span>` ou bien `[_.span_ref="lieu,.*" & _.span_ref=".*,ville,.*"]`
- etc.

8.1.3 Combinaison de recherche d'annotations et de propriétés de mots

L'expression de recherche d'annotations peut se combiner avec la recherche de propriétés de mots au sein d'une expression de recherche d'occurrences.

Par exemple :

- tous les adjectifs composant des annotations de villes : `[_.span_ref=".*,ville,.*" & frpos="ADJ"]`

  Glose : je cherche un mot (`[...]`), dominé par une structure span (`_.span`) dont la propriété ref (`_.span_ref`) vaut `'.*,ville,.*'` (`_.span_ref=".*,ville,.*"`), etc.
- tous les adjectifs non annotés « ville » : `[_.span_ref !=".*,ville,.*" & frpos="ADJ"]`
- tous les adjectifs non annotés : `[!_.span_ref & frpos="ADJ"]`

8.1.4 Visualisation des annotations dans une concordance

Il est possible de visualiser les annotations dans une concordance, il suffit d'y lancer une session d'annotation qui affichera les annotations de pivots existantes dans une colonne supplémentaire. Il n'est alors pas nécessaire de faire une annotation.

Une autre façon est de faire afficher les propriétés « ref » des structures « span » dans les références.

8.1.5 Transmission des annotations entre différents TXM

Après avoir sauvegardé les annotations, ces dernières font partie intégrante du corpus binaire. Il suffit donc d'exporter le corpus binaire (avec la commande 'Fichier > Exporter > Corpus au format binaire) dans un fichier .txm et de le transmettre à un correspondant. Le correspondant charge le corpus binaire (avec la commande 'Fichier > Charger) puis exploite le corpus et ses annotations comme pour n'importe quel corpus binaire.

8.2 Annotation avancée par concordances

Le système d'annotation simple avec une chaîne d'annotation unique décrit dans la section précédente est le comportement par défaut de TXM. Si on positionne la préférence « TXM > Utilisateur > Annotation > Mode » à la valeur « avancé (avec types+valeurs) » :

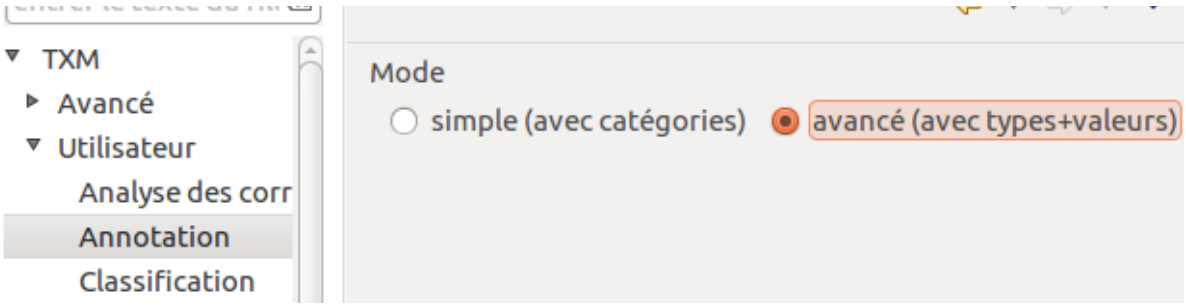


Illustration 8.3: Réglage annotation simple / annotation avancée

On dispose alors d'un système d'annotation équivalent mais combinant deux informations (au lieu d'une seule) :

- une catégorie (ou type) : de préférence un mot en minuscules simple ;
- une valeur (quelconque).

Par exemple, pour revenir au cas de la concordance de « Paris » on peut combiner la catégorie « lieu » et la valeur « Paris » :

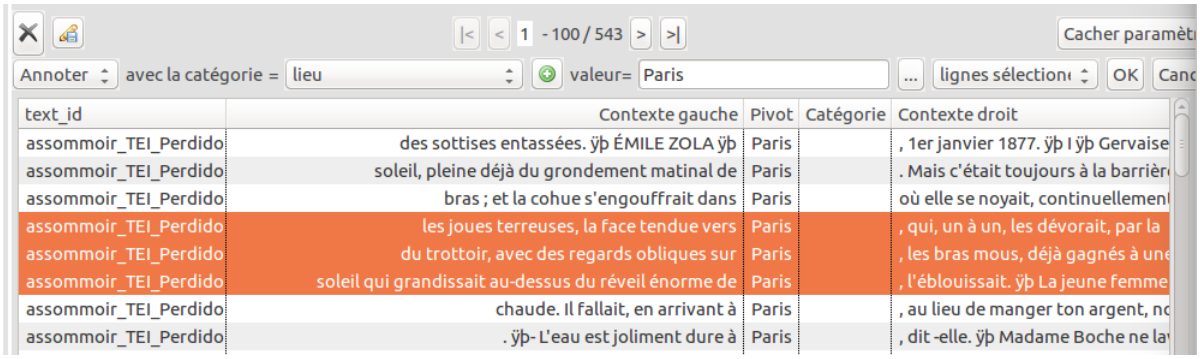


Illustration 8.4: annotation d'occurrences du mot "Paris" avec la catégorie "lieu" et la valeur "Paris"

Avec ce type d'annotation, l'exploitation (après sauvegarde des annotations) se fait à l'aide d'une expression de structure de la forme « <{catégorie}_ref="..."> ». C'est à dire que la catégorie crée une nouvelle structure ayant son nom dans le corpus (autour des pivots annotés), et la valeur devient la valeur de son attribut « @ref ».

Par exemple pour le cas précédent :

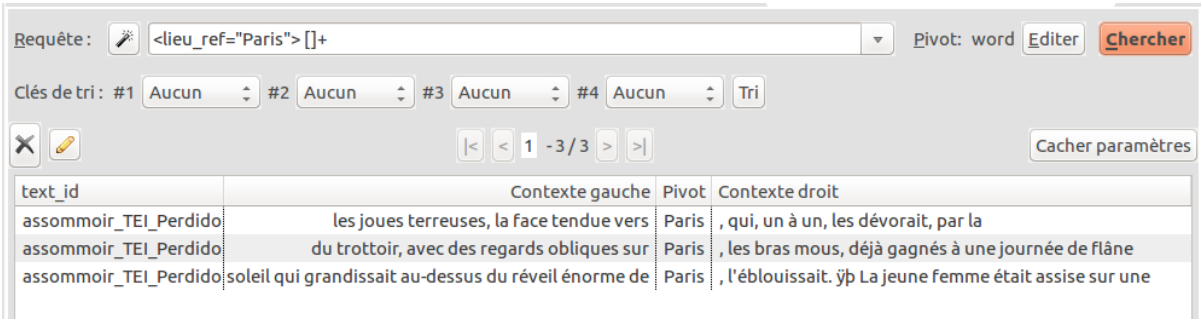


Illustration 8.5: Recherche de la catégorie "lieu" à la valeur "Paris"

8.2.1 Limites de l'annotation simple et avancée

L'encodage de l'annotation dans le corpus consiste en l'ajout d'une structure autour de la séquence de mots. Cela a l'avantage de pouvoir être exploitable immédiatement par le moteur CQP et accessible aux commandes habituelles de TXM. Par contre, cette structure ne pouvant chevaucher une structure pré-existante du corpus, la séquence de mots annotée ne peut pas être à cheval sur deux structures pré-existantes du corpus. Par exemple une séquence de mots annotée ne peut pas commencer à la fin d'un paragraphe et continuer au début du paragraphe suivant. C'est une limite du modèle que n'a pas l'annotation de séquences de mots avec l'extension Analec.

8.3 Annotation Analec/Glozz au sein d'éditions de texte

L'annotation Analec au sein de TXM se fait par le biais d'une extension permettant d'annoter les textes d'un corpus avec un modèle d'annotation de type Glozz et de réaliser diverses exploitations de ces annotations.

Le modèle d'annotation Glozz est documenté dans le manuel « Glozz User's Manual » <http://glozz.free.fr/glozzManual_1_0.pdf>.

La première version de l'extension permet d'annoter interactivement les unités au sein des éditions de texte de TXM, ainsi que d'enrichir l'annotation, de vérifier sa cohérence, de procéder à quelques extractions pour affichage ou décomptes à l'aide de macros TXM.

L'interface d'annotation des unités reproduit celle du logiciel Analec[52], dont nous vous invitons à consulter le manuel <http://www.lattice.cnrs.fr/IMG/pdf/ManuelAnalec_1501.pdf> pour apprendre à utiliser la barre d'outils d'annotation des unités (en particulier la section "Annoter des unités" page 15).

8.3.1 Installation de l'extension Analec

Appeler la commande « Fichier > Ajouter une extension » :

1. choisir l'extension Analec
2. accepter la licence
3. lancer l'installation
4. redémarrer TXM

8.3.1.1 Compatibilité et Prérequis

En mars 2017 sous Windows 7 : pour que l'extension Analec puisse mettre en évidence les unités annotées dans l'édition vous devez disposer d'une version d'Internet Explorer plus récente que celle livrée avec le système d'exploitation (dans ce cas il faut mettre à jour votre Internet Explorer vers une version plus récente) [voir le ticket #2017 <http://forge.cbp.ens-lyon.fr/redmine/issues/2017>].

[52] http://lattice.cnrs.fr/Telecharger-Analec

8.3.2 Préparation d'un corpus pour l'annotation

8.3.2.1 Corpus prêts à l'annotation dans TXM

Pour vous aider à tester l'annotation Analec dans TXM rapidement, voici un exemple de corpus diffusé par le projet ANR DEMOCRAT[53] à charger directement dans TXM :

1. télécharger le corpus « PRINCESSE-modèle-annotation-democrat-sans-annotations.txm »[54]
2. depuis TXM charger le corpus PRINCESSE-modèle-annotation-democrat-sans-annotations.txm avec la commande du menu « Fichier > Charger » → un nouveau corpus PRINCESSEBRUTSIMPLIFIETEXTE apparaît dans la vue Corpus ;
3. voir la section « Créer des annotations… » suivante pour procéder à l'annotation.

8.3.2.2 Corpus déjà annotés dans Analec

Il y a deux façons d'importer dans TXM un corpus déjà annoté dans Analec :

- Import XML-TEI Analec : à partir d'un fichier XML-TEI Analec exporté depuis Analec ;
- Import Glozz : à partir des 3 fichiers .aa, .aam et .ac exportés depuis Analec.

Import XML-TEI Analec

- lancer la commande « Analec > Import XML-TEI Analec Corpus… » en lui fournissant l'argument suivant :
  - xmlFile : le fichier XML-TEI Analec exporté depuis Analec (eg Le_Capitaine_Fracasse_or.xml)
    → un nouveau corpus LECAPITAINEFRACASSEOR est ajouté à la vue Corpus (il a été importé par une version interne du module TXT+CSV de TXM et contient les annotations Analec et leur modèle d'annotation)

Exemple de fichier XML-TEI Analec du projet ANR DEMOCRAT : Le_Capitaine_Fracasse_or.xml

Import Glozz : à partir des trois fichiers .aa, .aam et .ac

- lancer la commande « Analec > Import a Glozz corpus… » en lui fournissant les arguments suivants :
  - aafile : le fichier d'annotations (eg Cleves-brut-simplifie-annotations.aa)
  - aamfile : le fichier modèle d'annotations (eg Cleves-brut-simplifie-structure-annotation.aam)

[53] http://www.lattice.cnrs.fr/democrat

[54] La « Princesse de Clèves » équipé du modèle d'annotation DEMOCRAT mais sans annotations.

- acfile : le fichier texte (eg Cleves-brut-simplifie-texte.ac) → un nouveau corpus CLEVESBRUTSIMPLIFIETEXTE est ajouté à la vue Corpus (il a été importé par une version interne du module TXT+CSV de TXM et contient les annotations Analec et leur modèle d'annotation)

Exemple de fichiers .aa, .aam et .ac exportés depuis Analec pour le projet ANR DEMOCRAT : Cleves-brut-simplifie.zip.

Le module d'import vous demandera de désigner un répertoire : celui qui contient votre corpus (.ac) et vos annotations (.aa). Le fichier modèle d'annotation (.aam) peut se trouver dans un autre répertoire.

8.3.2.3 Corpus TXM quelconque

Tout corpus importé dans TXM peut être annoté selon un modèle Analec[55].

Il faut pour cela lui associer au préalable un modèle d'annotation Glozz :

- soit en important la description d'un modèle d'annotation depuis un fichier d'extension « .aam » à l'aide de la commande « Analec > Import Glozz model… », après avoir sélectionné le corpus qui va recevoir le modèle d'annotation dans la vue Corpus
  - on trouvera le fichier modèle d'annotation Glozz de référence du projet ANR DEMOCRAT à l'adresse suivante democrat.aam.
- soit en l'éditant directement au moyen de la commande « Analec > Edit Annotation Structure »

Dès qu'un modèle d'annotation Glozz est associé à un corpus, son Édition dispose d'un bouton « Annoter » actif (bouton « crayon » situé en bas à gauche) qui permet de lancer une session d'annotation interactive.

8.3.3 Annoter les unités interactivement depuis une édition de texte

8.3.3.1 Lancer une session d'annotation

- ouvrir l'édition du corpus, par exemple PRINCESSEBRUTSIMPLIFIETEXTE, (clic droit sur l'icone du corpus et menu contextuel « Edition ») ;
- cliquer sur le bouton « Annoter » (bouton crayon situé en bas de l'édition à gauche) ;
- la barre d'outils des unités s'ouvre en haut de l'édition ;
- ainsi que la fenêtre d'édition des unités (vue « Unit ») située en bas de l'édition ;
- remarque : dans TXM les annotations Analec sont posées sur les mots et non sur les caractères comme dans Analec. Les mots sont les mots simples définis par TXM (pas composés, souvent étiquetés et lemmatisés, etc.) ou par l'utilisateur selon le module

[55] On trouve sur le site de diffusion de TXM des corpus exemples TXM prêts à l'emploi.

d'import de textes sources qui a été utilisé pour créer le corpus. Le corpus PRINCESSE a été importé avec le module TXT+CSV en appliquant le modèle TreeTagger français (le s long - ſ - ruine les performances). Il s'agit donc de mots standards (par défaut) de TXM. La ponctuation est assimilée aux mots (eg la virgule est un mot que l'on peut sélectionner).

8.3.3.2 Visualiser les unités présentes

- dans la barre d'outils des unités sélectionner un type d'unité → toutes les unités de ce type sont mises en évidence dans la page.

8.3.3.3 Créer des unités

- sélectionner quelques caractères ou la totalité d'un mot dans l'édition
    - créer l'unité correspondant au mot avec le bouton « Créer » de la barre d'outils des unités ou la touche « Entrée » du clavier
- sélectionner plusieurs mots (ou portions de mots) dans l'édition
    - créer l'unité avec le bouton « Créer »
- double-cliquer sur un mot
    - créer l'unité avec la touche « Entrée » du clavier
- les unités sont mises en évidence avec la couleur vert clair et l'unité courante avec du vert foncé

8.3.3.4 Éditer les propriétés d'une unité

- cliquer directement sur l'unité ou bien la sélectionner par son identifiant dans le menu des identifiants d'unités (voir la section « sélection des unités » ci-dessous) ;
- les propriétés de l'unité s'affichent dans la vue « Unit » et sont éditables ;
- la valeur d'une propriété peut être choisie dans la liste des valeurs déjà connues (bouton [▼]) ;
- la valeur peut également être saisie directement :
    - la saisie active automatiquement un mécanisme d'auto-complétion
    - l'auto-complétion affiche la liste des valeurs déjà connues commençant par ce qui a déjà été saisi
        - cette liste sert à choisir directement la valeur souhaitée sans avoir à la re-saisir entièrement
            - navigation dans la liste
                - sélectionner la valeur suivante ou précédente (+1 ou -1)
                    - Flèche_vers_le_bas ou Flèche_vers_le_haut
                - sélectionner la valeur +10 ou -10
                    - Page_vers_le_bas ou Page_vers_le_haut
                - sélectionner la première ou la dernière valeur

- Début ou Fin
- continuer la saisie réduit la liste des valeurs proposées
- la touche Échap ou Esc (en haut à gauche du clavier) permet de quitter le mode d'auto-complétion : la saisie continue là où elle en était

8.3.3.5 Sélectionner des unités

- depuis l'édition :
  - cliquer sur l'unité → l'unité est mise en évidence et ses propriétés s'affichent dans la vue Unit ;
  - aller à l'unité suivante avec le raccourci clavier Ctrl-Flèche_vers_le_bas (tout en maintenant la touche 'Ctrl' enfoncée, appuyer sur la touche 'Flèche_vers_le_bas') ;
  - aller à l'unité précédente avec le raccourci clavier Ctrl-Flèche_vers_le_haut.
- depuis la barre d'outils des unités (située en haut de l'édition)
  - utiliser les boutons de flèches droite [▶] et gauche [◀] pour aller à l'unité suivante ou précédente
  - cliquer sur l'identifiant courant :
    - aller à l'unité suivante avec la touche Flèche_vers_le_bas ou avec la molette de la souris vers le bas ou deux doigts glissés vers le haut sur le trackpad (le sens peut être inversé selon les systèmes d'exploitation) ;
    - aller à l'unité précédente avec la touche Flèche_vers_le_haut ou avec la molette de la souris vers le haut ou deux doigts glissés vers le bas sur le trackpad ;
    - aller 10 unités plus loin avec la touche Page_vers_le_bas ;
    - aller 10 unités en arrière avec la touche Page_vers_le_haut ;
    - la navigation dépassant la fin de la liste cycle au début de la liste (même chose pour le dépassement du début de la liste) ;
    - sélectionner l'unité par la saisie de son identifiant puis validation avec la touche Entrée :
      - pendant la saisie, on peut activer l'auto-complétion avec le raccourcis Ctrl-Espace ;
      - quand l'auto-complétion est activée la liste des identifiants correspondants à ce qui a déjà été saisi s'affiche
        - cette liste sert à choisir directement l'identifiant souhaité sans avoir à saisir la totalité de l'identifiant
          - navigation dans la liste
            - sélectionner l'identifiant suivant ou précédent (+1 ou -1)
              - molette de la souris
              - Flèche_vers_le_bas ou Flèche_vers_le_haut

- sélectionner l'identifiant +10 ou -10
  - Page_vers_le_bas ou Page_vers_le_haut
- continuer la saisie réduit la liste des identifiants proposés

8.3.3.6 Rechercher des unités par la valeur de leurs propriétés

On peut rechercher des unités par leurs propriétés en cliquant sur le bouton “Chercher” (icone de loupe). Cette commande ouvre un formulaire de recherche dans une nouvelle vue, qui s'ouvre par défaut en bas de l'interface de TXM.

Le formulaire de recherche comporte :

- sur la première ligne :
  - à gauche
    - un bouton “Chercher” qui lance la recherche en utilisant les critères courants du formulaire ;
    - un bouton de remise à zéro des critères de recherche.
  - à droite
    - des boutons de navigation dans les résultats : aller au premier résultat, précédent, numéro de résultat courant, suivant, dernier ;
    - un bouton “Concordance” pour afficher la concordance des mots des unités correspondants à la recherche.
- sur les lignes suivantes : un champ de recherche par propriété
  - on saisit la valeur recherchée dans le champ de la propriété concernée ;
  - chaque champ dispose d'un menu déroulant des valeurs possibles de la propriété ;
  - un champ peut contenir soit une valeur exacte soit une expression régulière ;
  - on peut faire la recherche dans plusieurs propriétés à la fois pour une recherche combinée ;
  - si un champ est laissé vide alors la propriété ne participe pas à la recherche.

8.3.3.7 Rectifier les bornes d'une unité

Il y a 3 modes de rectification des bornes d'une unité :

- corriger la borne gauche :
  - 1) sélectionner l'unité
  - 2) cliquer sur le bouton 'corriger la borne gauche' ”[↔”
  - 3) cliquer sur le mot qui sera la nouvelle borne gauche de l'unité
- corriger la borne droite :
  - 1) sélectionner l'unité
  - 2) cliquer sur le bouton 'corriger la borne droite' “↔]”
  - 3) cliquer sur le mot qui sera la nouvelle borne droite de l'unité

- corriger simultanément la borne gauche et la borne droite :
    - 1) sélectionner l'unité
    - 2) cliquer sur le bouton 'corriger les bornes gauche et droite' ”[↔”
    - 3) sélectionner les mots qui formeront les nouvelles bornes de l'unité (comme lors de la création d'une unité)

8.3.3.8 Créer des unités à cheval sur deux pages d'édition

- créer l'unité à partir de ses premiers mots dans la première page (derniers mots de la page) ;
- lancer la rectification des bornes ;
- passer à la page suivante ;
- étendre l'unité à ses derniers mots
  → actuellement l'affichage bugue : les unités s'effacent, on peut changer de page et revenir pour ré-afficher correctement ;
- on procède de la même manière pour créer une unité à partir de ses derniers mots situés dans la deuxième page (premiers mots de la page) ;
- si l'unité s'étend sur plus de deux pages, suivre la même procédure en se déplaçant du nombre de pages nécessaire.

8.3.3.9 Supprimer une annotation

- sélectionner une unité en vert clair par un clic, elle devient vert foncé ;
- la supprimer avec le bouton “Supprimer” ou la touche “Suppr” du clavier.

8.3.3.10 Sauvegarder les annotations

- avec le bouton d'enregistrement (bouton crayon+disquette situé en bas à gauche de l'édition)

8.3.4 Exporter les annotations

8.3.4.1 Au format Glozz

- macro **ExportToGlozz** : exporte les annotations des unités d'un certain type dans un fichier au format Glozz.

8.3.4.2 Dans un corpus binaire TXM

- Sélectionner le corpus annoté ;

- S'assurer que les dernières modifications sont bien enregistrées avec la commande « Analec > Sauvegarder les annotations » ;
- Exporter le corpus dans un fichier .txm avec la commande « Fichier > Exporter ».

8.3.5 Enrichir des annotations Analec avec des macros

8.3.5.1 Utilisation de macros

Pour exécuter les macros il faut :

- ouvrir la vue Macro avec la commande « Affichage > Vues > Macro » ;
- éventuellement glisser-déposer cette vue par son onglet à un endroit pratique, par exemple dans la moitié inférieure de la vue Corpus ;
- ouvrir le répertoire de macros « analec » ;
- double-cliquer sur le nom d'une macro pour la lancer.

En général, il faut sélectionner le corpus sur lequel on veut travailler dans la vue Corpus avant de lancer la macro.

Pour lire ou modifier le code Groovy d'une macro :

- clic droit sur le nom de la macro dans la vue Macro ;
- lancer la commande « Éditer » du menu contextuel :
  → un éditeur de texte s'ouvre avec le code de la macro
- quand on clique dans la vue de l'éditeur de texte (pour y placer le curseur de saisie) la barre d'outils d'édition remplace celle des corpus ;
- ne pas oublier de sauver les modifications (avec le bouton « disquette »/sauver de la barre d'outils) avant de relancer la macro.

8.3.5.2 Macros d'ajouts d'annotations

- **PremierMaillon** : ajoute une propriété NEW à la valeur ‘YES’ aux unités d’un certain type si la valeur de sa propriété REF est rencontrée pour la première fois et ‘NO’ sinon[56].
- **AjoutDefinitude** : ajoute une propriété DEFINITUDE à une des valeurs ‘DEFINI’, ‘INDEFINI’, ‘DEMONSTRATIF’, ‘AMBIGU’ ou ‘NONE’ aux unités d’un certain type ;
- **CreationChaines** : crée des schémas de type 'CHAINE' composés d'unités ayant la propriété 'REF' de valeur identique ;
- **ResetAnnotations** : supprime toutes les annotations du corpus ;

[56] ATTENTION : Le champ « NEW » est ajouté dans la structure, mais il n'est pas affiché dans la vue. On ne peut donc pas le corriger directement. Pour le faire, il faut d’abord enregistrer les annotations et relancer TXM. La vue sera alors rafraîchie.

- **CompUnitProperties** : en travaux ;
- **Frpos2Categorie** : remontée de propriétés morphosyntaxiques de mots en français moderne dans des propriétés d'unités qui les contiennent (en travaux) ;
- **Fropos2Categorie** : remontée de propriétés morphosyntaxiques de mots en ancien français dans des propriétés d'unités qui les contiennent (en travaux) ;
- **CreationRelations** : en travaux.

8.3.6 Exploiter des annotations Analec avec des macros

Les macros qui suivent permettent à l'utilisateur de prototyper des calculs basés sur des annotations Analec et toutes autres informations disponibles dans la plateforme TXM.

8.3.6.1 Macros de vérification de Cohérence

Ces macros sont accompagnées de macros préliminaires de contrôle de la cohérence des annotations (proto-validation des annotations d'un texte par rapport à la structure d'annotation utilisée). Elles ont été développées pour vérifier la cohérence des annotations entre plusieurs annotateurs (les mesures supposent une cohérence parfaite).

- **UnitTypes** : index des types d'unités, des schémas d'un certain type ;
- **CategorieGrammaticale** : index des valeurs de propriétés, des unités d'un certain type, des schémas d'un certain type ;
- **SchemaTypes** : index des types de schémas ;
- **UnitTypesInSchema** : ibid. ;
- **UnitTypesNotInSchema** : index des types d'unités non associées aux schémas d'un certain type ;
- **CheckAnnotationStructureValues** : recense et supprime (si demandé au lancement de la macro) toutes les valeurs d'une propriété d'un type donné d'unité de la structure d'annotation non utilisées par les unités du corpus.

8.3.6.2 Macros de Mesures

Un premier jeu de macros a été réalisé pour calculer différentes mesures à l'occasion de l'écriture d'un article pour Langue française n° 195 (3/2017) (CG, JG, VO).

- **NombreDeChaines** : nombre de chaînes de référence du corpus ;
- **LongueurMoyenne** : longueur moyenne des chaînes de référence et index hiérarchique des longueurs de chaînes du corpus ;

- **NatureDuPremierMaillon** : index des valeurs d'une propriété donnée de la première unité de chaque chaîne du corpus ;
- **CoefficientStabilite** : rapport entre le nombre d'unités ayant la propriété 'Catégorie' à la valeur 'GN Défini' ou 'GN Démonstratif' ou 'Nom Propre' et le nombre de formes différentes représentant ces unités du corpus ;
- **DensiteReferentielle** : rapport entre le nombre d'unités d'un certain type et le nombre de mots du corpus (en %) ;
- **DistanceInterMaillonnaire** : histogramme des distances, en mots ou en caractères, entre le dernier mot d'une unité et le premier mot de la suivante du corpus.

- **AllMesures** : calcul simultané de toutes les mesures précédentes.

8.3.6.3 Macro d'affichage des annotations

- **Units** : affichage de toutes les unités d'un certain type, sous la forme : n° identifiant - adresse premier mot → adresse dernier mot, propriétés ;
- **Chaines** : affichage des valeurs d'une propriété des unités d'une chaîne. Son option buildCQL en particulier permet d'exploiter les autres outils de TXM comme les concordanciers, index, etc.

Cette macro liste pour chaque chaîne d'un corpus la valeur de sa propriété « Nom du référent » (par défaut) suivie des formes (par défaut) de ses unités successives.

Par exemple pour le corpus DESPERIERS pour la propriété 'word' (forme graphique) des mots on obtient :

```
Caillette: à Caillette, le povre Caillette, disoit, il, qu'il, sa, le, le, l'ha, l'ha, l'ha, Caillette, son, tu, Caillette, je, de ce sage homme Caillette, Caillette, qui, moy, luy, Caillette, Caillette, le, luy, luy, son, il, luy, il, Je, va
auteur: LES pages, les pages, les pages, tous ses gens de bien de pages, leur, à tous l'un apres l'autre, vous, qu'ilz, tous, qu'ilz, les, n'en, tant d'honnestes jeunes gens, qui, tous, les autres, avec les pages, je
un Seigneur de cour: un des Seigneurs de court, qui, qui, du Seigneur, ses
lecteur: vous, vous, vous, vous
un page (2): moy, qu'un, lequel, il, je
auteur: mon, je, Moy, m'escouter, je, Mon
Triboulet: Triboulet, Qui, ses, sa, Il, il, il, Il, luy, le, le, luy, vous, vous, vous, Triboulet, qui, son, luy, son, Triboulet, il, il, vous, son, Triboulet, il, je, je
maitre de Triboulet: un maistre, povre maistre, tu, Ce maistre, je, son maistre, son maistre
le cheval: son cheval, le cheval, qu'il, le, luy, luy, cheval, ce meschant cheval, le, il
les hommes: d'hommes, lesquelz, ilz, savent, qu'ilz
Polite: un autre fol nommé Polite, qui, Polite, mit, il, il, qu'il, il, tenoit, il, Polite, tu, Po lite, il
Abbé de Bourgueil: à un abbé de Bourgueil, Monsieur l'abbé, luy, le, l'abbé, Moyne, moy, l'abbé, moy, moy, l'abbé, l'abbé, moy, le moyne, il, il
```

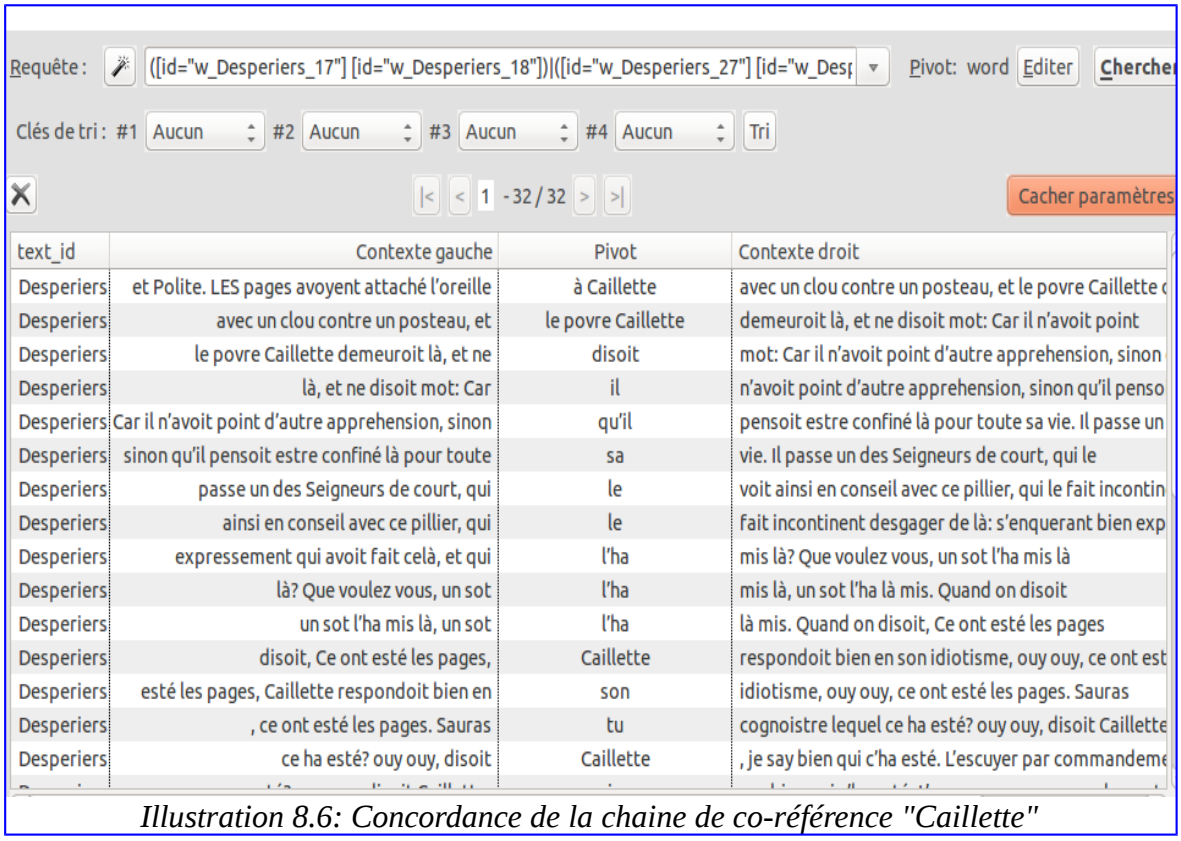
```
le chantre: un chantre, qui, lequel, qu’il, qu’il, l’appelloyent, luy, luy,
son, il, je, je, Je, je, Je, Sa, luy, te, luy, t’en, il, luy, t’oubliera,
tu, luy, qu’il, ce Bassecontre, servoit, il, sa, qu’il, le povre chantre,
luy, le Bassecontre, qu’il, qu’il, son, son, sa, il, il, qu’il, il, qu’il,
il, il, pria, luy, sa, qu’il, il, qu’il, luy, luy, ce Bassecontre, sa, sa,
le chantre, lequel, me, il, il, qu’il, Le chantre, Mon chantre, qui,
venoit, luy, il, il, tu, tes, ce fol, le Bassecontre, il, Je, me, je, t’ha,
tu, il, il, je, qu’il, tu, tu, il, apporta, qu’il, qu’il, son, luy, ses
les chanoines: les chanoines, qui, qu’ilz, d’eux, tes chanoines, nous
partie des chanoines: Monsieur, vous, vous, vous autres, d’eux, ilz,
messieurs, ilz, des messieurs, qu’ilz, leur, ilz
messieurs tel et tel: messieurs tel et tel, ceux, qu’il, leur, ilz, leur,
d’eux, messieurs, eux
principaux chanoines: aux principaux d’entre eux, les, l’un apres l’autre,
qu’ilz, leur, leur, leur, les, leur, Ilz, ilz, qu’ilz, leur, messieurs,
qui, tous, leurs, leur, ilz, ilz, ilz, ilz, chacun, leur, qu’ilz, leur,
ilz, nous, nous, les nostres, les vostres, Les nostres, ilz, chacun, leurs,
ilz, avons, les, Messieurs, voz, ilz, vous, vous autres messieurs, vous,
chacun, soy, vous, vous, vous, vous, vostre, vous, voz, ilz, ilz, nous,
nous, vous, leur, leur, ilz, allerent, conclurent, qu’ilz
un chanoine: monsieur vostre maistre, il, il, il
un chanoine2: je, l’un, j’avois, je
un autre chanoine: l’autre, j’avois, je, moy
valets chanoines: aux valetz, qui, qu’ilz
on: on, qu’on, qu’on, qu’on
```

Si on coche l'option buildCQL de cette macro, on génère à la place une requête par chaîne qui cherche tous les mots de ses unités.

Par exemple pour la chaine Caillette on obtient :

```
Caillette:          ([id="w_Desperiers_17"]          [id="w_Desperiers_18"])|
([id="w_Desperiers_27"]   [id="w_Desperiers_28"]   [id="w_Desperiers_29"])|
([id="w_Desperiers_35"])|([id="w_Desperiers_39"])|([id="w_Desperiers_46"])|
([id="w_Desperiers_53"])|([id="w_Desperiers_65"])|([id="w_Desperiers_75"])|
([id="w_Desperiers_92"])|([id="w_Desperiers_102"])|
([id="w_Desperiers_108"])|([id="w_Desperiers_122"])|
([id="w_Desperiers_126"])|([id="w_Desperiers_139"])|
([id="w_Desperiers_150"])|([id="w_Desperiers_152"])|
([id="w_Desperiers_177"]   [id="w_Desperiers_178"]   [id="w_Desperiers_179"]
[id="w_Desperiers_180"]  [id="w_Desperiers_181"])|([id="w_Desperiers_250"])|
([id="w_Desperiers_255"])|([id="w_Desperiers_264"])|
([id="w_Desperiers_276"])|([id="w_Desperiers_288"])|
([id="w_Desperiers_349"])|([id="w_Desperiers_356"])|
([id="w_Desperiers_363"])|([id="w_Desperiers_368"])|
([id="w_Desperiers_374"])|([id="w_Desperiers_381"])|
([id="w_Desperiers_388"])|([id="w_Desperiers_390"])|
([id="w_Desperiers_394"])|([id="w_Desperiers_402"])
```

On peut alors copier/coller la requête de Caillette dans le champ “Requête” d'une concordance et on obtient :



text_id	Contexte gauche	Pivot	Contexte droit
Desperiers	et Polite. LES pages avoyent attaché l'oreille	à Caillette	avec un clou contre un posteau, et le povre Caillette
Desperiers	avec un clou contre un posteau, et	le povre Caillette	demeuroit là, et ne disoit mot: Car il n'avoit point
Desperiers	le povre Caillette demeuroit là, et ne	disoit	mot: Car il n'avoit point d'autre apprehension, sinon
Desperiers	là, et ne disoit mot: Car	il	n'avoit point d'autre apprehension, sinon qu'il penso
Desperiers	Car il n'avoit point d'autre apprehension, sinon	qu'il	pensoit estre confiné là pour toute sa vie. Il passe un
Desperiers	sinon qu'il pensoit estre confiné là pour toute	sa	vie. Il passe un des Seigneurs de court, qui le
Desperiers	passe un des Seigneurs de court, qui	le	voit ainsi en conseil avec ce pillier, qui le fait incontin
Desperiers	ainsi en conseil avec ce pillier, qui	le	fait incontinent desgager de là: s'enquerant bien exp
Desperiers	expressement qui avoit fait celà, et qui	l'ha	mis là? Que voulez vous, un sot l'ha mis là
Desperiers	là? Que voulez vous, un sot	l'ha	mis là, un sot l'ha là mis. Quand on disoit
Desperiers	un sot l'ha mis là, un sot	l'ha	là mis. Quand on disoit, Ce ont esté les pages
Desperiers	disoit, Ce ont esté les pages,	Caillette	respondoit bien en son idiotisme, ouy ouy, ce ont est
Desperiers	esté les pages, Caillette respondoit bien en	son	idiotisme, ouy ouy, ce ont esté les pages. Sauras
Desperiers	, ce ont esté les pages. Sauras	tu	cognoistre lequel ce ha esté? ouy ouy, disoit Caillette
Desperiers	ce ha esté? ouy ouy, disoit	Caillette	, je say bien qui c'ha esté. L'escuyer par commandeme

Illustration 8.6: Concordance de la chaine de co-référence "Caillette"

La concordance permet alors également de naviguer dans la chaîne par retour au texte :

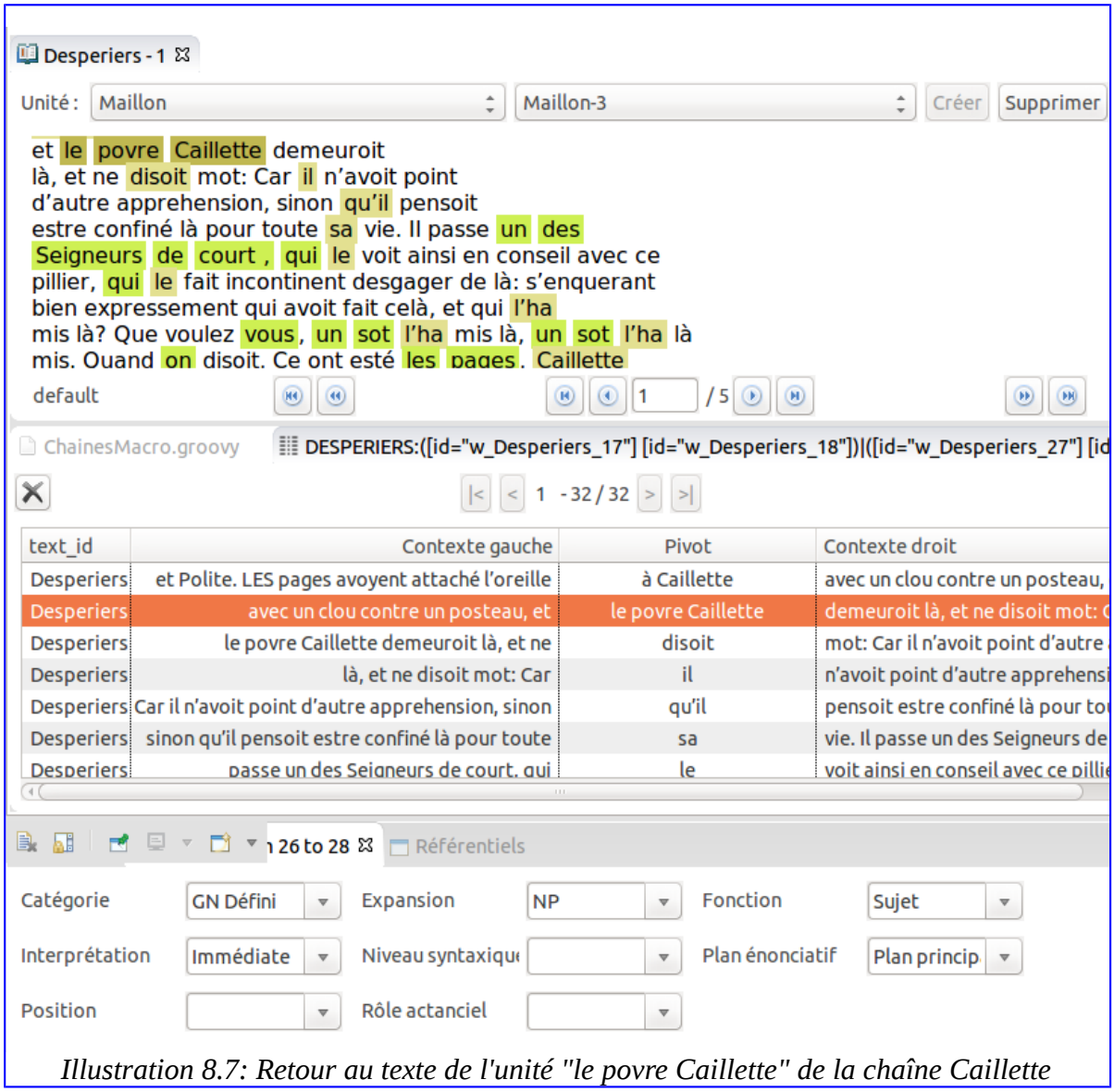


Illustration 8.7: Retour au texte de l'unité "le povre Caillette" de la chaîne Caillette

Pour obtenir cet affichage :

- on double-clique sur la ligne de « le povre Caillette »
- dans l'édition qui s'ouvre on clique sur le bouton « Annoter » pour faire apparaître les unités dans l'édition ainsi que la fenêtre inférieure 'Maillon' qui donne le détail des valeurs des propriétés du maillon 'le povre Caillette'
- en double-cliquant sur chaque ligne de concordance, on visualise l'unité correspondante dans le contexte de l'édition ainsi que la valeur de ses propriétés.

On peut par ailleurs tester par exemple la Progression de plusieurs chaînes, l'index des mots des unités d'une chaîne, etc.

- **Relations** : affichage des valeurs d'une propriété des unités d'une relation d'un certain type ;

- **Schemas2Progression** : affiche le graphique de progression des N chaines les plus longues.

9 Préférences

La fenêtre des Préférences de TXM donne accès au réglage de toutes les valeurs de paramètres par défaut des commandes ainsi qu'au réglage de l'interface utilisateur. On y accède par la commande « Outils > Préférences » :

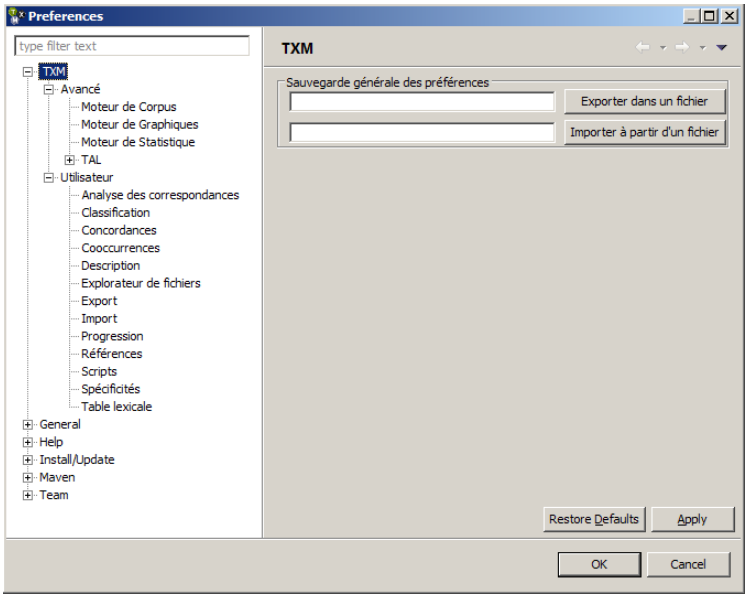


Illustration 9.1: Fenêtre des préférences de TXM

La fenêtre est organisée en sections.

9.1 Section TXM

Section d'utilitaires et de réglages d'ordre général.

- `Exporter dans un fichier` : pour sauvegarder les préférences de TXM dans un fichier (pratique pour importer ses réglages habituels dans un nouveau TXM) ;
- `Importer à partir d'un fichier` : pour importer dans TXM des préférences précédemment exportées ;
- `Vérifier les mises à jour au démarrage` : pour vérifier si une mise à jour est disponible à chaque lancement de TXM. Il peut être utile de désactiver cette préférence quand on travaille hors réseau.

9.2 Section TXM / Avancé

Section de configuration de TXM dont certaines préférences sont à manipuler avec précaution.

- `Dossier d'installation de TXM` : le dossier qui contient le logiciel TXM ;
- `dossier utilisateur` : le dossier de travail de TXM qui contiendra notamment vos corpus et vos macros ;
- `Niveau de détails du journal` : Niveau de détail des messages dans la console. Les niveaux inférieurs à WARNING provoquent l'affichage de messages de debuggage des modules d'import dans la console ;
- `Afficher les temps d'exécution dans la console` : affichage du temps d'exécution de chaque commande (Concordances, Index...) ;
- `Afficher la stacktrace` : affiche le détail de la pile Java dans la console en cas de plantage d'une commande ;
- `Ajouter des commentaires techniques` : affiche les messages internes de TXM dans la console ;
- `Ouvrir une boite de dialogue en cas d'erreur délicate` : avertit directement l'utilisateur par l'ouverture d'une boite de dialogue en cas d'erreur pouvant gêner la bonne poursuite de la session de travail TXM ;
- `Copier le journal dans un fichier` : enregistre le contenu de la console dans un fichier nommé « TXM (date) » dans le dossier utilisateur de TXM ;
- `Niveau de mise à jour` : voir la section Niveaux de mise à jour page 24;
- `Mode avancé` : donne accès aux réglages avancés des mises à jours et des installations d'extensions.

9.2.1 Moteur de Corpus

- `Utiliser le protocole réseau` : pour forcer une communication par socket entre TXM et CQP ;
- `Chemin vers CQPlib` : ne pas modifier (réglage expert) ;
- `Serveur CWB distant` : pour connecter TXM à un serveur CQP distant (voir les paramètres suivants) ;
- `Nom de la machine` : adresse réseau du serveur distant ;
- `Numéro de port du serveur CWB` : port de communication à utiliser avec le serveur distant (« 4877 » par défaut) ;

- `Login au serveur CWB` : nom d'utilisateur sur le serveur distant (« anonymous » par défaut) ;
- `Mot de passe au serveur CWB` : mot de passe sur le serveur distant ;
- `Chemin vers le fichier exécutable « cqpserver »` : exécutable cqpserver se trouvant dans les dossiers d'installation de TXM (pour une connexion réseau en local) ;
- `Chemin vers le dossier de Registre` : dossier « registry » de référence dans le dossier utilisateur de TXM ;
- `Chemin vers le fichier d'initialisation de cqpserver` : fichier de configuration de cqpserver, utilisé au tout début du lancement de cqpserver. Pratique pour changer les variables d'environnement de cqpserver ;
- `Options additionnelles` : options à ajouter à la ligne de commande de cqpserver. Par exemple, l'option « -b 1000000 » permet de définir la taille maximale d'une expression CQL non bornée.

9.2.2 Moteur de Graphiques

- `Moteur de graphiques courant` : détermine le moteur à utiliser pour la production des graphiques (R/SVG ou Java)

9.2.3 Moteur de Statistique

- `Connexion distante` : pour connecter TXM à un serveur R distant ;
- `Chemin vers l'exécutable de R` : chemin vers le logiciel R (pour une connexion locale) ;
- `Utiliser la transmission de données R par fichier` : à utiliser en cas de difficultés de communication avec R (pour une connexion locale) ;
- `Adresse du serveur` : pour la connexion distante ;
- `Utilisateur` : utilisateur du serveur R distant ;
- `Mot de passe` : mot du passe de l'utilisateur.

9.2.4 TAL / TreeTagger

- `Chemin du dossier d'installation de TreeTagger` : dossier de TreeTagger décompressé ;

- `Chemin du dossier de modèles linguistiques de TreeTagger` : dossier contenant les modèles linguistiques de TreeTagger décompressés et renommés d'après la norme ISO 639 ('fr.par', 'en.par', etc.) ;
- `Options [séparées par '  ' (2 blancs)]` : options à ajouter à la ligne de commande de TreeTagger. Les options sont séparées par 2 blancs pour permettre à TXM d'en faire la liste. Exemple : "-arg value -arg2 value2".

9.3 Section TXM / Utilisateur

Section de réglage des commandes de TXM.

Paramètres généraux du comportement de l'interface :

- `Pas de confirmation pour supprimer un objet` : pour ne plus faire de demande de confirmation lors de la suppression d'un objet de la vue Corpus ;
- `Recalculer automatiquement` : pour recalculer certains résultats directement après la modification de certains paramètres.

9.3.1 Analyse factorielle des correspondances

- `Fréquence minimale` : filtrage des lignes par fréquence minimale lors de l'appel sur une partition ;
- `Nombre de lignes` : nombre de lignes maximum lors de l'appel sur une partition ;
- `Format de la colonne Qualité` : format d'affichage des valeurs de qualité de représentation dans le plan. Voir la section 9.3.16.1 ;
- `Format de la colonne Contribution` : format d'affichage des valeurs de contribution aux axes ;
- `Format de la colonne Contribution` : format d'affichage des valeurs de contribution aux axes ;
- `Format de la colonne Masse` : format d'affichage des valeurs de masses ;
- `Format de la colonne Distance` : format d'affichage des valeurs de distance ;
- `Format de la colonne Cos²` : format d'affichage des valeurs de cos^2 ;
- `Format de la colonne Coord` : format d'affichage des valeurs de coordonnées ;
- `Afficher les colonnes` : afficher les points colonnes (ou variables) ;
- `Afficher les lignes` : afficher les points lignes (ou individus).

9.3.2 Annotations

- Mode :
  - simple (avec catégories) : on annote par mots-clés ou catégories ;
  - avancé (avec types+valeurs) : on annote avec des paires catégories (ou type)-valeur.

9.3.3 Classification

- Nombre de clusters : nombre de classes à calculer ;
- Méthode : algorithme de construction des classes ;
- Métrique : distance utilisée par l'algorithme ;
- Afficher les graphiques en 2D ou en 3D : affiche ou non les classes dans le même graphique que les plans factoriels de l'AFC.

9.3.4 Concordances

- Lignes par page : nombre de lignes à afficher par page de concordances ;
- Contexte Gauche (en mots) : nombre de mots du contexte gauche ;
- Contexte Droit (en mots) : nombre de mots du contexte droit.

9.3.5 Cooccurrences

- Format de l'indice : format d'affichage de l'indice de spécificités ;
- Seuil minimum de fréquence du cooccurrent : fréquence minimale pour qu'un mot puisse participer au calcul des cooccurrents ;
- Seuil minimum de fréquence de cooccurences : fréquence minimale des rencontres pour qu'un mot puisse participer au calcul des cooccurrents ;
- Seuil minimum de l'indice de cooccurences : seuil en deçà duquel le cooccurrent ne fait pas partie de la liste des résultats ;
- Minimum à gauche : distance la plus proche des cooccurrents de gauche ;
- Maximum à gauche : distance la plus éloignée des cooccurrents de gauche ;
- Minimum à droite : distance la plus proche des cooccurrents de droite ;
- Maximum à droite : distance la plus éloignée des cooccurrents de droite ;

- `Utiliser le total des fréquences de cooccurrents plutôt que de celles de tous les mots du corpus` : limite les fréquences marginales aux mots cooccurrents plutôt qu'à l'ensemble du vocabulaire du corpus dans le modèle des spécificités.

9.3.6 Description

- `Nombre de valeurs de propriétés affiché` : nombre maximal de valeurs différentes à afficher par propriété ;
- `Ordonner les parties par taille` : ordonner les parties par nombre d'occurrences décroissant pour la description des partitions ;
- `Afficher le nombre de parties dans le titre du graphique` : afficher le nombre de parties dans le titre des diagrammes à bâton de dimensions de partition.

9.3.7 Édition

- `Mise en évidence robuste des mots (calcul plus lent)` : En cas de défauts d'affichage dans la mise en évidence lors du retour au texte.

9.3.8 Explorateur de fichiers

- `Exp. Rég. des fichiers cachés` : expression régulière pour filtrer par leur nom certains fichiers dans l'affichage de l'explorateur de fichiers (surtout utilisé pour masquer les fichiers commençant par '.' en Linux) ;
- `Afficher les fichiers cachés` : afficher les fichiers considérés comme cachés par le système.

9.3.9 Export

Paramètres pour tous les exports de TXM :

- `Encodage des fichiers d'export` : table d'encodage des caractères à utiliser pour l'export. La valeur « UTF-8 » est conseillée car c'est la plus universelle ;
- `Colonnes séparées par` : caractère à utiliser comme séparateur de colonnes pour les exports au format CSV (« ; » par défaut). Ce caractère est très variable selon les logiciels tableurs et les systèmes d'exploitation. Bien que « CSV » soit l'acronyme de « Comma Separated Values », la virgule (« , ») est souvent remplacée par point-virgule (« ; ») voire le caractère de tabulation (« →| ») dans ce rôle ;
- `Séparateur de texte` : caractère à utiliser comme délimiteur de valeurs de colonnes pour les exports au format CSV (« " » par défaut).

- `Afficher le résultat de l'export dans un éditeur de texte` : Si l'export s'est déroulé correctement, le résultat est affiché dans une nouvelle fenêtre de l'éditeur de texte de TXM.
- `Format de fichier d'export des graphiques par défaut` : format à utiliser pour l'export des graphiques.

9.3.10 Import

Paramètres de lecture du fichier de métadonnées « `metadata.csv` » utilisé par certains modules d'import (TXT+CSV, XML/w+CSV, etc.) :

- `Encodage des caractères` : table d'encodage des caractères à utiliser. « UTF-8 » par défaut ;
- `Colonnes séparées par` : caractère à utiliser comme séparateur de colonnes. « , » par défaut ;
- `Séparateur de texte` : caractère à utiliser comme délimiteur de valeurs de colonnes. « " » par défaut.
- `Langue du presse-papier` : langue a utiliser pour TreeTagger dans l'import presse-papier ;
- `Code pour les propriétés sans valeur` : chaîne de caractères à utiliser pour les valeurs de propriétés non renseignées ;

9.3.11 Partition

- `Ordonner les parties par taille` : plutôt que par leur ordre par défaut ;
- `Afficher le nombre de parties dans le titre du graphique` : ajoute l'information directement dans le titre ;

9.3.12 Progression

Valeurs par défaut de la fenêtre de paramètres de la progression :

- `Graphe de progression cumulatif` : produire un graphe « cumulatif » plutôt que par densité ;
- `Niveau de gris` : produire un graphique en niveaux de gris plutôt qu'en couleur (aide à la publication) ;
- `Style de ligne unique` : pouvoir utiliser des styles de trait différents (continu, tirets, petits points, etc.) pour chaque courbe (utile pour les graphiques en niveaux de gris) ;

- `Répéter les valeurs de propriétés de structures` : par défaut les limites des structures partageant la même valeur de propriété ne sont pas affichées ;
- `Échelle des limites de structures` : nombre permettant d'affiner la pente des courbes aux limites de structures dans le graphique en densité.

9.3.13 Références

- `Ordonner les références par fréquence` : plutôt que par ordre alphabétique des valeurs.

9.3.14 Scripts

- `Sauvegarder le script avant exécution` : Enregistrer le script dans son fichier avant de l'exécuter ;
- `Prochain numéro de session` : Les scripts R sont sauvegardés dans des fichiers dont le nom est construit automatiquement. Leur nom est construit avec le n° de session en suffixe. Ce n° est incrémenté à chaque appel de script ;
- `Dossier racine` : dossier à partir duquel TXM est autorisé à exécuter les scripts.

9.3.15 Spécificités

- `Format des indices` : format d'affichage des indices (à documenter) ;
- `Indice maximum` : valeur conventionnelle maximale limite des indices de spécificité ;
- `Banalité` : seuil de banalité pour les graphiques ;
- `Regrouper les barres par les lignes de la table` : transposer la table lexicale avant de produire le graphique ;
- `Niveaux de gris` : produire un graphique en niveaux de gris (pour l'aide à la publication) ;
- `Afficher les lignes` : afficher les valeurs d'indices sous forme de lignes brisées ;
- `Afficher les barres` : afficher les valeurs d'indices sous forme de diagramme à bâtons.

9.3.16 Table lexicale

- `Fréquence minimale` : filtrage des lignes par fréquence minimale lors d'un calcul intermédiaire de table lexicale.

9.3.16.1 Définition du format d'affichage des nombres réels ou entiers[57]

Dans les tableaux de résultats, les nombres peuvent être formatés selon un patron défini à l'aide des caractères élémentaires suivants :

0	représente un chiffre qui devra obligatoirement être présent, même s'il s'agit d'un zéro inutile
#	représente un chiffre en ignorant les zéros inutiles
.	représente le séparateur de la partie décimale
,	représente le séparateur des groupes (milliers, millions, etc.)

Tableau 2: Caractères de formatage des nombres

Le nombre de « 0 » ou de « # » dans le patron détermine la taille des parties entière et décimale de la valeur numérique, sachant que 0 représentera un chiffre obligatoirement présent (et éventuellement remplacé par un zéro inutile) et # un chiffre optionnel (qui ignorera donc les zéros inutiles). Exemples de formats :

Format	**0**	**0,02**	**0,8**	**12,9**
###	0	0	1	13
#.##	0	0,02	0,8	12,9
0.##	0	0,02	0,8	12,9
0.00	0,00	0,02	0,80	12,90
#.00	,00	,02	,80	12,90
#,##0.00	0,00	0,02	0,80	12,90

Tableau 3: Exemples de formats de nombres

[57] D'après http://java.developpez.com/faq/java/?page=langage_chaine#LANGAGE_STRING_nombre_en_chaine_formatee

10 Syntaxe des requêtes CQL

[d'après Bénédicte Pincemin, « mémo CQL », 4 octobre 2012, Ateliers TXM]

10.1 Introduction

10.1.1 CQL, CQP

CQL est l'acronyme de Corpus Query Language, c'est un langage d'expression de requêtes. Une expression (ou équation) CQL est une chaîne de caractères exprimant un motif linguistique (un mot, ou une suite de mots) à partir des valeurs de leurs propriétés (comme la catégorie grammaticale, le lemme, la forme graphique).

CQP est l'acronyme de Corpus Query Processor, c'est un composant logiciel qui traite des requêtes : c'est un moteur de recherche qui permet de trouver toutes les occurrences correspondant à une équation CQL dans un corpus donné.

Le moteur CQP a été développé à l'origine à l'université de Stuttgart <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench> et est désormais un logiciel libre <http://cwb.sourceforge.net>. Il est intégré à TXM où il assure les recherches d'occurrences et d'une façon générale toutes les opérations de sélection à l'intérieur du corpus. Il a été choisi pour l'excellent rapport de ses performances à la complexité des requêtes traitées.

10.1.2 Les requêtes dans TXM : requêtes simples, requêtes assistées, requêtes avancées

CQL est donc un langage formel, avec un lexique et une syntaxe d'opérateurs, qui forment un métalangage permettant de combiner des éléments pour la recherche de motifs structurés.

L'apprentissage du langage CQL n'est pas un passage obligé pour utiliser TXM, mais c'est en langage CQL qu'on a le mode d'expression de motifs le plus riche.

Si l'on saisit un mot dans la zone de requête, c'est interprété comme la recherche des mots présentant exactement cette graphie dans le corpus. Cela permet déjà un certain nombre de recherches simples. Mais on perçoit assez vite deux limites : d'une part, on reste à la « surface » du texte, on ne tire aucun parti des autres informations linguistiques encodées dans le corpus (lemme, catégorie grammaticale, etc). D'autre part, on est rivé à l'empan exact d'un mot : la formulation de la recherche ne peut se faire ni sur une partie du mot (son début par exemple), ni sur des expressions en plusieurs mots - alors que cela devient possible en utilisant CQL.

Le logiciel TXM comporte un assistant à l'écriture de requêtes, accessible via une icone « baguette magique » à gauche du champ de saisie de la requête. Cet assistant permet

d'exprimer une recherche à l'aide de menus déroulants plus intuitifs si l'on est peu familier des langages de requête. En revanche, il ne permet pas d'exprimer autant de choses que le langage CQL, qui reste beaucoup plus souple et plus complet. La connaissance de CQL est donc utile pour avoir les possibilités d'expression les plus larges et les plus précises.

En pratique, on peut apprécier de combiner l'utilisation de l'assistant avec la connaissance du langage CQL. L'assistant peut faciliter l'écriture d'une première version de la requête. La connaissance de CQL permet ensuite de bien comprendre l'équation et de l'ajuster ou de l'affiner si nécessaire.

10.1.3 Dynamique de la construction d'une requête

Une requête se met au point : entre ce qu'on veut repérer (que l'on pense avoir exprimé dans la requête), et ce qu'on trouve effectivement dans le corpus, il y a souvent un écart qui demande à être corrigé. Il est de toutes façons toujours sage de vérifier la portée effective, dans le corpus choisi, de la requête utilisée, avant de l'utiliser pour un calcul statistique.

L'apprentissage et l'utilisation de CQL font donc un usage central de la fonctionnalité Index de TXM. La fonctionnalité Index permet de lister toutes les formes correspondant au motif dans le corpus. On peut les parcourir soit par importance quantitative décroissante (tri par fréquence décroissante, qui est la manière dont se présente le résultat par défaut), soit par ordre alphabétique, ce qui peut faciliter la lecture en regroupant les réalisations de forme proche.

Le parcours de cette liste des configurations trouvées met en évidence les formes indésirables ; en revanche il ne dit rien des formes qui seraient pertinentes mais qui, ne correspondant pas formellement à la requête, n'ont pas été repérées. Méthodiquement, on recommande donc toujours, quand on a un motif linguistique à recherche, de commencer par l'exprimer de façon très ouverte, de veiller à minimiser les a priori qui pourraient être réducteurs. L'examen des occurrences correspondantes trouvées guide alors sur la manière d'ajouter alors peu à peu des contraintes permettant de cibler les formes pertinentes et d'écarter les formes non voulues.

10.1.4 Utilisation pédagogique des exemples

Les exemples ci-après ont été choisis pour illustrer les possibilités de CQL qui nous paraissent les plus utiles : il faut les soumettre à la fonctionnalité Index pour bien voir leur effet. Ils ont été conçus pour être lancés sur le corpus Voeux (http://sourceforge.net/projects/txm/files/corpora/voeux/voeux-bin-0.6.zip/download). Le corpus Discours est quelquefois utilisé en complément si nécessaire. Les exemples sur fond gris sont plus complexes et peuvent être ignorés dans un premier temps.

10.2 Recherche simple [niveau 1 (infralexical) : les valeurs]

10.2.1 Recherche d'un mot

`bonheur`	*Pour chercher un mot donné il suffit de saisir sa graphie.*

`l'amitié`
`l'`
`amitié`
`aujourd'hui`
`parce que`
	L'expression CQL doit correspondre exactement à une unité telle que découpée par la segmentation lexicale, une unité lexicale n'est pas forcément une chaîne de caractères entre deux blancs. Voir par exemple aussi les différences entre Voeux et Discours pour les unités ci-contre.

`ami` `amiti`	*Une partie d'un mot ne rapporte aucun résultat, l'expression doit correspondre à un mot entier attesté dans le corpus.*
	Trois façons équivalentes d'exprimer une recherche sur une graphie :
`bonheur`	*- la graphie telle quelle*
`"bonheur"`	*- la graphie entre guillemets doubles droits*
`[word="bonheur"]`	*- l'usage des crochets et du mot réservé « word ».* *Les moyens les plus verbeux montreront leur utilité dans des cas plus complexes.*
`[word="parce que"]` `[word=" bonheur "]`	*Un blanc à l'intérieur des guillemets est significatif (partie intégrante de la graphie). Le guillemet doit être collé à la graphie cherchée (sans espace supplémentaire).*
`[ word = "bonheur" ]`	*Les blancs à l'extérieur des guillemets sont non significatifs et peuvent être utilisés pour faciliter la lecture.*

10.2.2 Variantes d'écriture

`"gouvernement"%c`	*Neutralisation de la casse (majuscules/minuscules). Les guillemets sont obligatoires.*
`"Etat"%d`	*Neutralisation des signes diacritiques (accents, cédille, etc.).*
`"franc.*"%cd`	*Les deux neutralisations peuvent être cumulées.*

10.2.3 Troncature et joker

`libertés?`	*Le point d'interrogation porte sur le caractère qui précède et signifie qu'il est facultatif (0 ou 1 fois). Il peut se placer n'importe où. C'est utile notamment quand le corpus n'est pas lemmatisé, ou que la qualité de la lemmatisation est insuffisante.*
`âgé?e?s?`	
`"premiere?s?"%d`	
`nation.*`	*Point étoile à la fin = « mot qui commence par … » . Point = « un caractère, n'importe lequel ».*
`.*patri.*`	*Etoile = « 0 à n fois, n aussi grand qu'on veut ». Utile pour chercher un radical.*
`.+patri.*`	*Signe plus = « 1 à n fois ». Ici on impose qu'il y ait un préfixe.*
`.*ables?`	*Ces opérateurs se plaçant n'importe où, on peut chercher des mots partageant les mêmes affixes, le radical variant librement.*
`in.*ables?`	
`"i[mn].*ables?"`	*Les crochets sont pratiques pour indiquer l'ensemble des lettres possibles, une seule devant être choisie.*
`.*`	*Zéro à n caractères, n'importe lesquels. Cette expression attrape tous les mots.*
`.* .*`	(dans Discours) *Graphies incluant un blanc (au moins).*
`.`	*Mots formés d'un seul caractère.*
`…`	*Mots de longueur trois.*

10.2.4 Ponctuations

`\.`	*Les caractères spéciaux (opérateurs), doivent être « endormis » en les précédant d'une barre oblique*
`\?`	
	descendante, si on veut pouvoir les considérer eux-mêmes comme des caractères que l'on recherche.
`.*'`	*Ce n'est pas le cas de toutes les ponctuations : ex. ici mots terminés par une apostrophe.*

10.2.5 Classes de caractères

`.+\p{P}`	*Mot terminé par une ponctuation : permet d'attraper aussi les apostrophes obliques (souvent originaires de Word et qu'on ne peut pas saisir facilement au clavier dans TXM).*
`\p{Lu}+`	*Mot composé de majuscules (y compris diacritiques). Voir FAQ pour autres*

classes.

10.2.6 Alternative

`paix|guerre`	*OU, alternative non exclusive. Élargit la recherche à des variantes de formulation.*
`(inter|supra)nation.*`	*Peut s'utiliser à l'intérieur du mot, avec des parenthèses pour délimiter sa portée.*
`(inter|supra)?nation.*`	*Des opérateurs de facultativité ou répétition peuvent porter sur la parenthèse.*

10.3 Recherche sur les propriétés [niveau 2 (lexical) : les propriétés]

10.3.1 Introduction

Jusqu'alors, les recherches effectuées portaient sur la forme graphique des mots, qui est enregistrée dans la propriété *word* : [word="bonheur"] signifie qu'on recherche la valeur *bonheur* de la propriété *word*, correspondant à la forme graphique. Mais, lorsque le corpus est enrichi, les mots portent d'autres informations que leur seule graphie, sous la forme d'autre propriétés. Les requêtes peuvent alors porter sur d'autres propriétés des mots (et les combiner).

La graphie étant une propriété (presque) comme les autres, tout ce qu'on a vu dans la section précédente s'applique aux valeurs de propriété quelle que soit la propriété, sauf l'écriture simplifiée.

Pour interroger sur les propriétés il faut connaître leur nom et leurs valeurs. En effet, le nom des propriétés dépend de l'import du corpus : dans tel corpus la propriété qui enregistre le lemme est *lemma*, dans tel autre *frlemme*, dans tel autre encore *ttlemme*, etc. De même, les valeurs des catégories grammaticales dépendent du jeu d'étiquettes utilisé. Dans TXM en version locale, la fonction Description montre quelles propriétés sont disponibles et donne pour chacune d'elle un aperçu de quelques valeurs attestées (sur les premières occurrences du corpus). La fonction Lexique permet de lister exhaustivement les valeurs d'une propriété attestées dans le corpus. Dans la version locale, un double-clic sur une de ces valeurs permet de voir son usage en contexte (dans une concordance). Ceci étant il est utile d'avoir les tables descriptives des jeux de catégories utilisés pour le corpus sur lequel on travaille.

10.3.2 Recherche sur une propriété

`[frlemma="beau"]`	*Rechercher un lemme permet de désigner un mot sous ses*

`[frlemma="faire"]`	*formes (très) variables. Il faut expliciter sur quelle propriété on travaille, la formulation à crochets devient nécessaire.*
`[frlemma="je"]`	*Le lemme « je » recouvre ici ses formes élidées ou avec majuscule initiale.*
`[frpos="ADV"]`	*De même, on peut chercher sur d'autres propriétés, comme la catégorie grammaticale.*
`[frpos="VER.*"]` `[frpos="NOM|NAM|VER.*|ADJ"]`	*La valeur que prend la propriété peut utiliser les mêmes opérateurs que précédemment, par ex. pour reconstruire des catégories en regroupant des étiquettes.*
`[frlemma=".*|.*"]`	*Ici la barre verticale fait partie intégrante de l'étiquette (ambiguïtés non résolues par* TT*).*

10.3.3 Alternative (2)

`[frpos="NAM|NOM"]`	*Il y a plusieurs manières d'exprimer l'alternative, plus ou moins factorisées.*
`[frpos="N(A|O)M"]`	*La barre verticale est l'opérateur le plus général, sa portée peut être ciblée par des parenthèses.*
`[frpos="N[AO]M"]`	*Les crochets ne sont utilisables que pour une alternance sur un seul caractère,*
`"[aeiouy]+"`	*mais facilitent l'expression d'un large choix*
`[pos=".*[1-3].*"]`	(dans Discours) *ou d'une gamme.*
`[pos="[^12]*"]`	(dans Discours) *Le chapeau est une négation : ensemble des caractères interdits sur la position.*
`[frpos="VER:(futu|cond|subi)"]`	*Alternance sur des séquences de caractères (de longueurs identiques ou non) : seule la barre verticale est utilisable.*

10.3.4 Combinaison d'informations

`[frlemma="pouvoir" & frpos="NOM"]`	*Désambiguïsation catégorielle d'un lemme.*
`[frpos="ADV" & word=".*ment"]`	*Croisement d'une catégorie et d'un trait morphologique.*
`[frlemma="liber.*"%d & frlemma!="libéral"]`	*Exclusion de cas non souhaités.*
`[frpos="NOM" & word!=".*\p{P}"]`	*Post-taitement des erreurs de segmentation.*

`[pos!="NA|pon" & pos!=fropos]`	(dans la BFM) *Comparaison directe à une autre propriété.*

10.4 Recherche d'un motif de plusieurs mots [niveau 3 (supralexical) : séquences d'unités lexicales]

10.4.1 Succession de mots

`[word="réduction"] [word="du"] [word="temps"] [word="de"] [word="travail"]`	*Paire de crochets = mot.*
`"réduction" "du" "temps" "de" "travail"`	*Notation allégée possible si l'on ne travaille que sur des graphies. Mélange possible.*
`[frlemma="réduction"] "du" "temps" "de" "travail"`	
`[frpos="NOM"][frlemma="de"][frpos="NOM"]`	*Usage avec des catégories (patron).*
`[frpos="NOM"][frlemma="de"][frlemma="le"]?[frpos="NOM"]`	
`[frpos="NOM"]([frlemma="de"][frlemma="le"]|[frlemma="du"]) [frpos="NOM"]`	
`[frpos="DET.*"][frpos="ADV"]?[frpos="ADJ"]+ [frlemma="année"]`	

On retrouve à ce niveau 3 les opérateurs vus au niveau 1, pour gérer les variations.

10.4.2 Traitement des insertions

`[frlemma="il"][][frlemma="y"] [frlemma="avoir"]`	*Une unité lexicale quelconque (joker de mot).*
`[frlemma="il"][]?[frlemma="y"] [frlemma="avoir"]`	*Insertion facultative.*
`[frlemma="il"][][][][frlemma="y"] [frlemma="avoir"]`	*Distance de trois unités lexicales.*
`[frlemma="il"][]{0,3}[frlemma="y"] [frlemma="avoir"]`	*Distance de zéro à trois.*
`[frlemma="paix"][]{0,10}[frlemma="monde"]`	*Distance de 0 à 10, deux formulations équivalentes.*
`[frlemma="paix"][]*[frlemma="monde"] within 10`	

Si l'on utilise []* il faut absolument borner l'expansion.

```
[frlemma="je"][frpos!="V.*"]*[frlemma="souhaiter"][frpos!="V.*"]*[frlemma="année"]
within 25
```

Distances avec mots exclus, contrôle davantage syntaxique.

```
[lemma="je"][pos!="V.*"]*[lemma="souhaiter"][pos!
="V.*"]*[lemma="année"] within s
```

(dans Discours) *Empan sur structure (si disponible)*

```
[lemma="République"][]*[lemma="France"]
within 2s
```

(dans Discours*) Structure multipliée.*

10.4.3 Étude distributionnelle

```
[frlemma="très"][]
```

On prend un motif (contexte), et on rend variable une place, soit complètement librement,

```
[frpos="NOM"][frlemma="français"]
```

soit avec une indication de catégorie.

```
[frlemma="ne"][frpos="VER.*"]
```

Recherche des verbes avec négation.

```
[frlemma="ne"]([frpos!="VER.*|NOM|ADJ"]|[frlemma="être|avoir"])*[frpos="VER.*"  &
frlemma!="être|avoir"] within 10
```

Idem, plus affinée.

10.4.4 Alternatives

```
([word="président"%c][][][word="république"%c]|[word="chef"%c][][]
[word="état"%cd])
```

Expressions.

```
([frlemma="paix"][]*[frlemma="monde"]|[frlemma="monde"][]*[frlemma="paix"])
within 10
```

```
([frlemma="travail.*"][]*[frlemma="famil.*"]|[frlemma="famil.*"]
[]*[frlemma="travail.*"]) within 20
```

Cooccurrences.

10.4.5 Lien entre deux mots

```
a:[frpos="NAM|NOM|ADJ|VER.*" & word!=".*\p{P}"][]*[word=a.word]
within 10
```

Répétition, accord,...

10.5 Informations contextuelles

10.5.1 Utilisation des structures

`<s> [pos="V.*"]`	(dans Discours) *Verbes qui commencent une phrase.*
`<s> [pos="V.*"] expand to s`	(dans Discours) *Phrases qui commencent par un verbe.*
`<s> []{1,5}</s>`	(dans Discours) *Phrases d'au plus cinq mots.*
`[pos="Vmsm.*"] expand to s`	(dans Discours) *Phrases contenant un motif donné (ici subjonctif imparfait).*

10.5.2 Utilisation d'une propriété de structure

`[word="Algérie" & _.text_loc!="dg"]`	*« Algérie » dans un texte dont le locuteur n'est pas De Gaulle*.
`<sp_speaker="P"> [frpos!="PRO:PER"]* [frpos="PROPER"]`	*Le premier pronom personnel de chaque tour de parole du professeur d'une transcription de cours*.

10.6 Lien d'alignement entre corpus parallèles

On dispose d'un corpus latin CorpusLAT aligné avec un corpus d'ancien français CorpusFRO (textes existant dans les deux langues, en relation de traduction). Les requêtes suivantes sont effectuées sur CorpusLAT.

`[lemme="HIC"] :CorpusFRO [lemme="CIST"]`	*Occurrences du lemme HIC pour lesquelles on trouve le lemme CIST dans le passage aligné en ancien français.*
`[lemme="HIC"] :CorpusFRO ! [lemme="CIST"]`	*Occurrences du lemme HIC pour lesquelles on ne trouve pas le lemme CIST dans le passage aligné en ancien français.*
`[lemme="HIC"] expand to seg :CorpusFRO [lemme="CIST"]`	*Segments contenant le lemme HIC et pour lesquels on trouve le lemme CIST dans le segment aligné en ancien français.*
`[] expand to seg :CorpusFRO [lemme="CIST"]`	*Segments latins alignés avec ceux contenant le lemme CIST en ancien français (construction d'un sous-corpus pour calcul de résonance).*
`<seg>[lemme!="HIC"]*<seg> :CorpusFRO [lemme="CIST"]`	*Segments ne contenant pas le lemme HIC et pour lesquels on trouve le lemme CIST dans le passage aligné en ancien français.*

10.7 Stratégies de résolution des opérateurs itérateurs

Le nombre d'occurrences « attrapées » par les opérateurs ?, *, + dépend de la stratégie de résolution courante du moteur de recherche CQP. Par exemple, pour la requête suivante[58] :

```
[enpos="DET"]? [enpos="ADJ"]* [enpos="NN"] ([enpos="PREP"] [enpos="DET"]?
[enpos="ADJ"]* [enpos="NN"])*
```

Avec le texte suivant à interroger :

```
the old book on the table in the room
```

On obtient les résultats suivants pour chaque stratégie :

1. `shortest` : 3 matches
   ```
   r1 =         book
   r2 =                     table
   r3 =                                 room
   ```
2. `longest` : 1 match
   ```
   r1 = the old book on the table in the room
   ```
3. `standard` : 3 matches
   ```
   r1 = the old book
   r2 =                 the table
   r3 =                               the room
   ```
4. `traditional` : 7 matches recouvrants
   ```
   r1 = the old book
   r2 =     old book
   r3 =         book
   r4 =                 the table
   r5 =                     table
   r6 =                               the room
   r7 =                                   room
   ```

La stratégie de résolution par défaut est '`standard`'. Actuellement il faut utiliser la macro SetMatchingStrategy de la catégorie 'cqp' pour choisir une autre stratégie de résolution du moteur CQP pour la session de travail courante dans TXM.

10.8 Documentation complémentaire

Pour une description plus complète du langage de requêtes CQL, vous pouvez consulter (en Anglais) :

[58] d'après The CQP Query Language Tutorial, (CWB version 2.2.b90), Stefan Evert, 10 July 2005.

– Oliver Christ, Bruno M. Schulze, Anja Hofmann, and Esther König, « The IMS Corpus Workbench : Corpus Query Processor (CQP), User's Manual », August 16, 1999 (CQP V2.2), University of Stuttgart , <http://www.ling.ohio-state.edu/~cbrew/2004/684.02/assignments/cqpman.pdf>.

11 Syntaxe des expressions régulières

L'expression des chaines de caractères dans les requêtes CQL suit la syntaxe des expressions régulières PCRE (Perl-Compatible Regular Expressions). La syntaxe complète est décrite à la section « Specification of the regular expressions supported by PCRE » du manuel de PCRE : http://regexkit.sourceforge.net/Documentation/pcre/pcrepattern.html.

Voici un résumé en français des opérateurs les plus courants.

.	matche n'importe quel caractère
\	neutralise l'opérateur situé à droite
|	alternance
( )	regroupement
[…]	classe de caractères entre crochets (eg « [aeiouy] » pour une voyelle, ou « [a-z] » pour n'importe quelle minuscule)
[^…]	ensemble complémentaire de la classe de caractères entre crochets, le caractère ^ joue le rôle de négation (eg « [^aeiouy] » pour un caractère qui n'est pas une voyelle)

Tableau 4: Méta-caractères (ou Opérateurs)

?	matche 0 ou 1 fois l'expression située à gauche
*	matche 0 fois ou plus
+	matche 1 fois ou plus
{n}	matche n fois
{n,}	matche au moins n fois
{n,m}	matche entre n et m fois

Tableau 5: Quantifieurs

\x{CC}	caractère de valeur CC (exprimée en hexadécimal) (eg « \x{E9} »pour « é »)
\xCC	caractère de valeur CC (exprimée en hexadécimal)

Tableau 6: Codes de caractères

\d	un chiffre
\D	pas un chiffre
\w	un caractère de « mot »
\W	pas un caractère de « mot »
\s	un caractère d'espace
\S	pas un caractère d'espace
\p{Classe}	un caractère de la classe Unicode « Classe » (eg « \p{Lu} » pour un caractère majuscule)
\P{Classe}	pas un caractère de la classe Unicode « Classe »
[[:ClassePOSIX:]]	un caractère de la classe « ClassePOSIX » (eg « [[:upper:]] » pour un caractère majuscule)

Tableau 7: Classes de caractères

\p{L}	lettre
\p{Ll}	caractère minuscule
\p{Lu}	caractère majuscule
\p{N}	caractère numérique
\p{Xan}	caractère alphanumérique
\p{Pd}	caractère de tiret (eg « - », « — »…)
\p{P}	caractère de ponctuation (eg « , », « . »…)
\p{Ps}	caractère de ponctuation ouvrante (eg « (»)
\p{Pe}	caractère de ponctuation fermante (eg «) »)
\p{Sm}	caractère de symbole mathématique (eg « ~ »)
\p{Cyrillic}	caractère en alphabet russe
\p{Arabic}	caractère en alphabet arabe
\p{Greek}	caractère en alphabet grec

Tableau 8: Classes Unicode courantes[59]

alpha	caractère alphabétique (usage : "`[[:alpha:]]`")
alnum	caractère alphanumérique
ascii	caractère du code ASCII
digit	chiffre décimal
graph	caractère imprimable, sans l'espace
lower	caractère minuscule
print	caractère imprimable, incluant l'espace
punct	caractère de ponctuation
space	caractère d'espace
upper	caractère majuscule
word	caractère de mot
xdigit	chiffre hexadécimal

Tableau 9: Classes POSIX courantes (système plus ancien et plus grossier que les classes Unicode)

Références de caractères mémorisés

[59] Toutes les classes Unicode sont décrites à la section « Unicode character properties » du manuel de PCRE.

- **\2** contenu du premier groupe de parenthèses mémorisé (suppose la présence de parenthèses « ..(..).. » auparavant dans l'expression)
- **\3** contenu du deuxième groupe de parenthèses mémorisé
- …
- **\g{nom}** contenu du groupe de parenthèses mémorisé nommé « nom » (suppose la présence de « ..**(?<nom>**..**)**.. » auparavant dans l'expression)
- ...

Exemples :

- **([[:lower:]])([[:lower:]]).*m.*\3\2** : deux minuscules suivies de « m » suivi des deux premières minuscules en ordre inverse ;
- **(.*)\2** : une même chaine deux fois de suite
- **(?<groupe1>.*)\g{groupe1}** : une même chaine deux fois de suite

12 Macros et scripts Groovy ou R

12.1 Utiliser des macros

Les macros sont le moyen le plus simple pour piloter la plateforme à partir de scripts. Il s'agit de scripts Groovy (voir la section suivante) pouvant prendre leurs paramètres depuis une boite de dialogue ou à partir d'objets sélectionnés dans l'interface de TXM. C'est une façon pratique de faire exécuter un script par l'utilisateur sans avoir besoin de connaissances du langage de programmation Groovy (sans avoir à lire le code du script).

Différentes macros sont disponibles :

- des macros livrées avec TXM et accessibles par la vue « Macro » ;
- des macros complémentaires téléchargeables depuis le site du logiciel TXM sur Sourceforge : http://sourceforge.net/projects/txm/files/software/TXM%20macros (ce dossier contient les versions les plus à jour des macros de TXM)
- des macros de la communauté des utilisateurs de TXM recensées dans le wiki https://groupes.renater.fr/wiki/txm-users/public/macros.

12.1.1 Exécuter une macro

Les macros se lancent à partir de la vue « Macro », ouverte à partir du menu « Affichage > Vues > Macro », en double-cliquant sur leur icône ou bien avec le raccourcis clavier « F12 » (exécuter le dernier script ou la dernière macro exécutée) en cas d'appels successifs.

Il y a deux façons de fournir des paramètres à une macro :

- avant de la lancer : en sélectionnant au préalable un ou plusieurs objets dans l'interface de TXM (corpus, sous-corpus, lexique, index, caractères dans un éditeur de texte, lignes de concordance, etc.) ;
- au lancement : en renseignant les valeurs de paramètres dans la boite de dialogue de saisie des paramètres qui s'ouvre.

Les paramètres et la façon de les renseigner dépendent de chaque macro, par exemple un corpus, une requête CQL, un dossier contenant des fichiers à traiter, un fichier de sortie pour des résultats, etc.).

Les valeurs des paramètres saisies dans la boite de dialogue sont mémorisées entre chaque appel d'une macro. Cette mémorisation s'appuie sur des fichiers d'extension « .properties » situés dans le même répertoire que celui du fichier de la macro (visible dans la vue « Fichier » de TXM ou depuis le navigateur de fichier du système d'exploitation) situé à partir du dossier $HOME/TXM/scripts/macro. La suppression du fichier « .properties » d'une macro supprime

les valeurs de paramètres mémorisées, la macro utilisera les valeurs par défaut définies dans le script. L'édition du fichier « .properties » permet d'éditer les valeurs de certains paramètres.

L'exécution de la macro est terminée quand le message « Terminé : » s'affiche dans la console suivi de son temps d'exécution en millisecondes.

Les résultats des macros sont très variables, comme pour les commandes TXM (nouveau sous-corpus ou concordance, nouveaux fichiers, messages dans la console, etc.).

12.1.2 Installer une macro

De nouvelles macros sont disponibles depuis plusieurs sources, voir la section 12.1.

Une macro correspond à un fichier script Groovy dont le nom se termine par « ...Macro.groovy ». Certaines macros peuvent être accompagnées par des scripts Groovy secondaires (sans suffixe « ...Macro »). Dans ce cas la macro se télécharge sous forme d'archive ZIP.

Pour installer une nouvelle macro dans TXM :

- copier le ou les fichiers .groovy de la macro dans le dossier « $HOME/TXM/scripts/macro » ;
- ouvrir la vue macro (menu « Affichage > Vues > Macro ») ou la rafraîchir (bouton ↺ de la vue Macro) pour accéder à la macro.

12.1.3 Modifier une macro

Dans la vue « Macro », clic-droit sur l'icone de la macro puis lancer la commande « Éditer » du menu contextuel. La modification du script requière des connaissances en programmation dans le langage Groovy (voir section suivante).

Une macro peut être associée à plusieurs scripts Groovy secondaires qui ne sont pas visibles dans la vue Macro. Pour les éditer, il faut utiliser la vue Fichier ouverte sur le dossier contenant la macro et les scripts, faisant partie de la hiérarchie du dossier « $HOME/TXM/scripts/macro ».

12.1.4 Créer une macro

Pour ajouter une macro, il suffit de déposer un script Groovy dont le nom se termine par « ...Macro.groovy » à partir du dossier « $HOME/TXM/scripts/macro » et de rafraîchir la vue « Macro ». La hiérarchie des packages Groovy pour les macros commence à partir de ce répertoire.

Un façon assez simple de commencer est de partir d'un squelette de macro type créé par le bouton «Nouvelle macro» de la vue Macro. Ce bouton provoque la saisie du nom de la macro

puis ouvre un éditeur de texte sur un nouveau fichier de script contenant des exemples de lignes de code Groovy :

- le script commence par des exemples de déclarations de paramètres de différents types (booléen, date, fichier... voir la section suivante pour la liste de tous les types disponibles). Chaque ligne de déclaration commence par le préfixe « `@Field @Option ».` Il faut décommenter une ligne exemple de déclaration pour l'utiliser (supprimer les '//' au début de la ligne). Exemple :

  ```
  @Field @Option(name="query", usage="an example query",
  widget="Query", required=true, def='[pos="V.*"]')

  def query
  ```

  Glose : déclaration du paramètre obligatoire « query » de type « requête CQL » avec la valeur par défaut « `[pos="V.*"]` ».
  Détail :

  - « @Field » commence la déclaration
  - « @Option(...) » contient la déclaration
  - « name="query" » définit le nom du paramètre comme « query »
  - « usage="an example query" » définit la documentation du paramètre
  - « widget="Query" » déclare le type du paramètre
  - « required=true » déclare le paramètre comme obligatoire à renseigner
  - « def='[pos="V.*"]' » déclare la valeur par défaut du paramètre (qui sert à pré-remplir le champ de saisie du paramètre dans la boite de dialogue)
  - « def query » déclare la variable Groovy du paramètre

- vient ensuite une ligne exemple pour l'ouverture de la boite de dialogue de saisie des paramètres. Exemple :

  ```
  if (!ParametersDialog.open(this)) return;
  ```

  Glose : ouvrir la boite de dialogue de saisie des paramètres puis, exécuter la suite de la macro si appui sur le bouton « Exécution » ou abandonner l'exécution de la macro si appui sur le bouton « Annuler ».
  Remarque : si cette ligne est commentée ou omise la macro ne demandera pas de saisie de valeurs de paramètres.

- viennent ensuite des exemples de lignes de manipulation de différents objets couramment sélectionnés dans l'interface utilisateur (éléments de la vue Corpus : corpus, sous-corpus, partition, résultat - d'autres vues, éditeurs de texte, etc.).

Exemple :

```
println "Corpora selection: "+corpusViewSelection
```

Glose : Affiche dans la console le nom de l'élément couramment sélectionné dans la vue Corpus.

12.1.4.1 Types de paramètres d'une macro

Les paramètres d'une macro peuvent être de types suivants :

- Boolean : booléen oui/non (comme une option de traitement) ;
- Date : date au format mois/jour/année ;
- FileOpen ou File : chemin de fichier à ouvrir dans le disque dur (pour la lecture d'un fichier TSV par exemple) ;
- FileSave : chemin de fichier à créer dans le disque dur (pour la sauvegarde d'un résultat par exemple) ;
- Folder : chemin de dossier à sélectionner dans le disque dur (pour désigner un dossier d'entrée par exemple) ;
- Float : nombre réel (comme un seuil d'indice) ;
- Integer : nombre entier (comme un seuil en fréquence) ;
- Query : requête CQL ;
- String : chaîne de caractères (comme un nom de sous-corpus) ;
- Text : chaîne de caractères sur plusieurs lignes (comme un commentaire).

12.1.4.2 Variables prédéfinies

Les macros peuvent accéder à divers objets de TXM par le biais de variables prédéfinies :

Nom de la variable	*Description*	*Type Groovy*
corpusViewSelection	l'élément couramment sélectionné dans la vue Corpus (corpus, sous-corpus, partition, résultat)	Object
corpusViewSelections	liste d'éléments couramment sélectionnés dans la vue Corpus	List<Object>

selection	l'élément couramment sélectionné (sélection de texte dans un éditeur, icône de requête dans la vue Requête...)	Object
selections	une liste d'éléments couramment sélectionnés	List<Object>
monitor	objet interne permettant d'accéder à l'interface utilisateur pendant l'exécution de la macro	org.txm.rcpapplication.utils .JobHandler
gse	objet interne permettant l'appel mutuel entre scripts Groovy et donc entre macros	org.txm.rcpapplication.utils .GSERunner qui étend la classe groovy.util.GroovyScriptE ngine
editor	fenêtre de résultats ou éditeur de texte couramment sélectionné	org.eclipse.ui.IWorkbenchP art

12.1.4.3 Appel d'une macro depuis une autre macro

Il est possible d'appeler une macro depuis une autre macro à l'aide de la variable « gse ».

Pour appeler une macro « B » à partir d'une macro « A », on peut par exemple utiliser la ligne de code suivante dans la macro « A » :

```
gse.runMacro BMacro, ["param1":value1, "param2":value2]
```

Remarques :

- le nom du script de la macro B est « BMacro.groovy », d'où la désignation par le nom « BMacro » (la macro est affichée sous le nom «B » dans la vue Macro) ;
- « param1 » et « param2 » sont des paramètres de la macro « B ».

Si des paramètres obligatoires de la macro appelée (B dans l'exemple) ne sont pas renseignés lors de l'appel, la macro ouvrira la boite de dialogue pour saisir les valeurs des paramètres manquants obligatoires.

12.1.5 Macros TXM prédéfinies

TXM est livré avec plusieurs macros utilitaires prédéfinies, accessibles depuis la vue Macro. Ces macros sont organisées par répertoires thématiques :

- texte : pour les utilitaires de traitements de fichiers aux formats de traitements de texte (DOCX, ODT, RTF, etc.) ;
- txt : pour les utilitaires de traitements de fichiers au format texte brut (TXT) ;
- xml : pour les utilitaires de traitements de fichiers au format XML ;
- xsl : pour l'utilitaire donnant accès aux traitements XSL ;
- csv : pour les utilitaires de traitements de fichiers au format CSV ;
- cqp : pour les utilitaires utilisant les services du moteur CQP ;
- stats : pour les utilitaires offrant divers services statistiques ;
- etc.

12.1.5.1 Assistance à la préparation des fichiers sources d'un corpus

- **text/Text2TXT** : Conversion par lot du format de tous les fichiers textes (MS Word, LibreOffice Writer, etc.) d'un dossier (.doc, .docx, .odt, .rtf, .html…) vers le format TXT (texte brut) ;
- **txt/ChangeEncoding** : Conversion par lot de l'encodage des caractères de tous les fichiers d'un dossier ;
- **txt/CharList** : Calcule la liste de tous les caractères utilisés par un fichier source encodé en Unicode UTF-8 et indique leur fréquence ;
- **txt/SearchInDirectory :** Recherche par lot d'une expression régulière dans tous les fichiers TXT d'un dossier;
- **txt/SearchReplaceInDirectory** : Chercher/Remplacer par lot d'une expression régulière dans tous les fichiers TXT d'un dossier ;
- **txt/TXT2XML** : Conversion par lot de tous les fichiers TXT (texte brut) d'un dossier en fichiers XML de base ;
- **xml/XMLStatistics** : Calcul de la table des fréquences des balises et attributs XML utilisés dans les fichiers d'un dossier source (utile pour une vue d'ensemble quantitative de l'usage des balises dans des documents dont on ne connaît pas les principes d'encodage) ;

- **xsl/ExecXSL** : Application par lot d'une feuille de transformation XSLT sur tous les fichiers XML d'un dossier ;
- **transcription/TextTranscription2TRS** : Conversion par lot de transcriptions d'enregistrements saisies au moyen d'un traitement de texte (.doc, .odt) ou du logiciel Transana (.rtf) vers le format XML-TRS du logiciel Transcriber pour l'import dans TXM avec le module d'import Transcriber+CSV ;
- **misc/EuroPresse2XML** : Assistance à la récupération et à la transformation des exports HTML du portail EuroPresse pour l'import dans TXM avec le module d'import XML/w+CSV ;
- **csv/CSV2XML** : Transforme un fichier tableau au format CSV dont certaines colonnes contiennent du texte (typiquement les réponses aux questions ouvertes d'un sondage) et d'autres des données (typiquement des caractéristiques de répondants ou des réponses à des questions fermées) en un fichier XML importable dans TXM avec le module d'import XML/w+CSV (typiquement pour réaliser des analyses sur les réponses aux questions ouvertes d'un sondage).

12.1.5.2 Assistance à la correction de l'annotation des mots d'un corpus

- **annotation/BuildWordPropTable** : Exportation d'une concordance TXM dans un tableau TSV (colonnes séparées par une TABULATION) contenant les propriétés de mots pour corrections ou ajouts dans un tableur (comme Calc) ;
- **annotation/InjectWordPropTable** : Mise à jour des sources pivot XML-TXM d'un corpus TXM à partir du tableau édité pour ré-importation avec le module d'import XML-TXM.

12.1.5.3 Appel de script R

- **r/ExecR** : Exemple de macro appelant un script R générant un graphique affiché dans TXM (calcul de l'histogramme d'un index) ;
- **r/PlotSpecif** : Appel de la fonction “specificities.distribution.plot” du package R 'textometry' pour afficher la courbe de densité du modèle statistique des spécificités en choisissant les valeurs de ses paramètres T, t, F et f.

12.1.5.4 Assistance à l'appel de commandes CQP

- **cqp/CreateCQPList** : Définit une liste de mots (ou plus généralement de valeurs de propriétés : liste de lemmes, liste de catégories...) utilisable dans les requêtes CQL ;
- **cqp/ExecCQP** : Fait exécuter une ligne de commande quelconque au moteur CQP ;

- **cqp/SetMatchingStrategy** : Change l'option “matchingstrategy” de CQP (influençant la façon de calculer les séquences de mots de longueur variable) pour la session de travail courante.

12.1.5.5 Assistance pour l'appel répétitif de commandes de TXM

- **commands/CrossedPartitionBuilder** : Construit une partition en croisant les valeurs de plusieurs propriétés d'une même structure. L'usage de cette macro est plus souple que la saisie de requêtes CQL dans le mode avancé de la commande Partition.

12.1.6 Partager vos macros avec la communauté des utilisateurs de TXM

Vous pouvez éditer la page https://groupes.renater.fr/wiki/txm-users/public/macros du wiki des utilisateurs de TXM pour documenter et mettre un lien vers vos macros.

12.2 Utiliser des scripts Groovy

Les scripts servent à piloter la plateforme TXM pour :

- appeler n'importe quelle commande TXM : lancer une recherche à partir d'une requête CQL, appliquer un modèle statistique, exporter et sauvegarder des résultats dans un fichier, etc.
- utiliser des paramètres personnalisés pour chacune de ces commandes ;
- enregistrer et lancer une séquence de commandes pour des recherches usuelles.

Leur usage permet à l'utilisateur d'étendre les fonctionnalités de la plateforme à l'aide de scripts[60].

Les macros sont des scripts spécialisés offrant des services de saisie et de mémorisation de paramètres.

Les scripts et macros sont écrits en langage Groovy (`http://groovy.codehaus.org`).

Vous trouverez une courte introduction à l'utilisation de ce langage à l'adresse :
`http://onjava.com/pub/a/onjava/2004/09/29/groovy.html`

Trois livres de référence pourront vous donner plus d'informations :

- *Groovy in action*[61]

---

[60] À l'image de ce qui se fait dans MS Word au moyen des macros Visual Basic.

[61] Dierk König et al., *Groovy in action* (Greenwich: Manning, 2007).

- *Groovy programming: an introduction for Java developers*[62]
- *Programming Groovy*[63]

Le texte des scripts à exécuter peut se trouver dans un fichier ou être simplement sélectionné dans une fenêtre (voir la section « **Éditeur de texte** »).

La meilleure manière de commencer à écrire vos propres script Groovy est de copier un des scripts exemples inclus dans la plateforme, dossier « `C:\Documents and Settings\<identifiant de l'utilisateur>\TXM\scripts\groovy\samples` »[64][65]. Par exemple, le script « `conc.groovy` »[66] calcule automatiquement une concordance du mot « je » dans le corpus DISCOURS puis exporte les résultats au format CSV dans un fichier nommé « `conc.txt` ».

Pour que vous puissiez exécuter vos propres scripts Groovy depuis TXM, ces derniers doivent se trouver dans le dossier « `C:\Documents and Settings\<identifiant de l'utilisateur>\TXM\scripts\user` ». La hiérarchie des packages Groovy commence à partir de ce répertoire.

Pour créer un script, déplacer un script dans ce dossier ou aller dans la vue « Fichier » (voir la section 3.2.1.1.2 La vue « Fichier » et « l'éditeur de texte »), clic-droit sur le dossier « user », puis sélectionner « créer un fichier », donner un nom. Par exemple « test.groovy ». Pour modifier le script, il faut double-cliquer sur son icone.

12.2.1 Exécuter un script

On peut exécuter un script Groovy par sept moyens différents :

- depuis un éditeur de texte[67]:

---

[62]Kenneth A. Barclay et W. J. Savage, *Groovy programming: an introduction for Java developers* (Morgan Kaufmann Publishers, 2007).

[63]Subramaniam Venkat, *Programming Groovy: dynamic productivity for the Java developer*, Pragmatic Bookshelf. (Raleigh: Daniel H. Steinberg ed., 2008).

[64] Aucune protection de sécurité lors de l'exécution de scripts n'a été intégrée à la plateforme TXM pour le moment, il faut donc être vigilant sur la provenance des scripts utilisés.

[65] En Linux : `/home/<identifiant de l'utilisateur>/TXM/scripts/user`

[66] Vous avez également accès à ce script en ligne, à l'adresse :

`http://txm.svn.sourceforge.net/viewvc/txm/trunk/Toolbox/trunk/org.textometrie.toolbox/src/groovy/org/txm/test/conc.groovy?revision=1080&view=markup`

[67]On ouvre un éditeur de texte en ouvrant un fichier de la vue Fichier ou par le menu principal “Fichier / Nouveau fichier” ou “Fichier / Ouvrir...”.

  - menu contextuel « Groovy / Exécuter la sélection de texte » exécute le texte sélectionné dans l'éditeur ;
  - menu contextuel « Groovy / Exécuter le Script » exécute le script se trouvant dans l'éditeur ;
  - menu contextuel « Groovy / Exécuter un fichier Groovy » exécute un script se trouvant dans un fichier ;
  - raccourcis clavier « F11 » exécute le texte sélectionné dans l'éditeur ;
  - raccourcis clavier « Ctrl-F11 » exécute le script se trouvant dans l'éditeur ;
- depuis la vue Fichier :
  - menu contextuel sur l'icone d'un fichier « / Exécuter le Script » exécute le script contenu dans le fichier ;
- raccourcis clavier « F12 » exécute le dernier script exécuté.

12.2.2 Utilisation de librairies tierces (fichiers .jar ou .so)

Pour qu'une librairie tierce (fichier « .jar »[68] ou « .so »[69]) soit accessible depuis un script Groovy, il suffit de la déposer dans le dossier « $HOME\TXM\scripts\lib ». Après avoir relancé TXM, les packages et fonctions correspondants sont importables depuis les scripts Groovy.

12.2.3 Comment utiliser les objets de TXM depuis Groovy

Groovy étant basé sur le langage Java[70], il donne accès à tous les modules Java de TXM.

La façon d'appeler les commandes de TXM depuis Java et donc depuis Groovy est documentée dans la Javadoc de TXM située à l'adresse : http://txm.sourceforge.net/javadoc.

Par exemple, les paramètres de la commande Concordance sont décrits dans le package « org.txm.rcpapplication.editors.concordances », documenté à l'adresse http://txm.sourceforge.net/javadoc/TXM/RCP/org/txm/rcpapplication/editors/concordances/ConcordancesEditor.html.

Toutes les commande décrites dans cette documentation peuvent être exécutées dans un script Groovy.

---

[68] Les fichiers .jar contiennent des classes Java pré-compilées.

[69] Les fichiers .so contiennent des fonctions natives du système d'exploitation hôte, par exemple les fonctions d'une librairie C.

[70] Tout script Groovy est compilé en byte-code Java afin de pouvoir être exécuté.

12.3 Utiliser des scripts R

TXM permet de faire exécuter un script R au moteur statistique R intégré.

12.3.1 Exécuter un script

Un script R peut être exécuté par cinq moyens différents :

- depuis un éditeur de texte[71]:
  - menu contextuel « R / Exécuter le Script » exécute le script se trouvant dans l'éditeur ;
  - menu contextuel « R / Exécuter la sélection comme script R » exécute le texte sélectionné dans l'éditeur ;
  - raccourcis clavier « F11 » exécute le texte sélectionné dans l'éditeur ;
  - raccourcis clavier « Ctrl-F11 » exécute le script se trouvant dans l'éditeur ;
- depuis la vue Fichier :
  - menu contextuel sur l'icone d'un fichier « R / Exécuter le Script » exécute le script contenu dans le fichier.

Session exemple

Exécution du script de démonstration « **HelloWorld.R** » :

- accéder au dossier des scripts R d'exemple livrés avec TXM à partir de la vue Fichier : ouvrir le dossier « $HOME/TXM/scripts/samples/R »
- ouvrir le script « HelloWorld.R » dans un éditeur en double-cliquant sur son icone
- ouvrir la “Console R” pour pouvoir lire les sorties du script : ouvrir “Affichage / Vues / 'Console R' ” à partir du menu principal
- dans le menu contextuel de l'éditeur contenant le script “HelloWorld.R”, lancer “R / Exécuter le Script”

12.3.2 Utilisation des résultats et objets TXM depuis R

Tous les résultats de calculs TXM dont l'icone est affichée dans la vue Corpus peuvent être transférés dans R sous la forme d'une structure de données : pour cela utiliser la commande « Envoyer vers R » du menu contextuel du corpus ou de l'icone de résultat depuis la vue Corpus.

[71]On ouvre un éditeur de texte en ouvrant un fichier de la vue Fichier ou par le menu principal “Fichier / Nouveau fichier” ou “Fichier / Ouvrir...”.

La console affiche alors le nom de l'objet R créé pour vous permettre de l'utiliser dans des scripts R. Un petit « R » rouge en exposant à gauche est ajouté à l'icone de l'objet TXM pour confirmer le transfert dans R. Certains transferts sont réalisés automatiquement par TXM pour certains calculs (tables lexicales, graphique de progression, spécificités...).

La vue « Variables R » affiche la liste des données déjà transférées dans R. Pour ouvrir la vue « Variables R », ouvrir le menu « Affichage / Vues / Variables R ». Vous pouvez copier dans le presse-papier le nom de la donnée transférée avec l'entrée « Copier » du menu contextuel de son icone ou en pressant Contrôle-C au clavier quand l'icone est sélectionnée. Vous pouvez alors coller le nom du symbole dans un script R.

12.3.3 Utiliser la perspective R pour organiser son accès à R

La perspective R vous permet de configurer rapidement l'interface de TXM avec les vues utiles au travail avec R :

- un bouton « Nouvelle session » de création de script de session R est ajouté à la barre d'outils : pour ouvrir un éditeur de texte sur un fichier de script nommé « sessionX.R » créé automatiquement dans le dossier « $USER_HOME/TXM/scripts/R ».
- un onglet « Variables R » est ajouté dans le panneau de gauche, donnant un accès rapide à la vue « Variables R »;
- la partie droite de l'interface empile trois fenêtres :
  - la zone d'édition des scripts R ;
  - la console R ;
  - la console TXM.

Le basculement dans la perspective « Corpus », pour reprendre votre session d'analyse de corpus, fermera la vue « Variables R » et la console R.

Vous pouvez basculer d'une perspective à l'autre à n'importe quel moment.

Par défaut, l'interface de TXM est configurée selon la perspective « Corpus ». Pour changer de perspective, vous pouvez utiliser la barre d'outils des perspectives située en haut à gauche de l'interface, sous la barre d'outils principale, contenant les boutons « R » et « Corpus ». Vous pouvez également utiliser le menu principal (en haut à gauche) « Affichage / Perspectives / R ou Corpus ».

12.3.4 L'environnement R de TXM

Au démarrage, TXM effectue quelques réglages du moteur statistiques R (Rserve) :

- L'encodage des caractères par défaut est initialisé à Unicode UTF-8

- Le chemin du workspace R est initialisé à :
    - Sous Windows :
      « `C:\Utilisateurs\<identifiant de l'utilisateur>\TXM\R` »
      ou bien
      « `C:\Documents and Settings\<identifiant de l'utilisateur>\TXM\R` »
    - Sous Mac OS X :
      « `/Users/<identifiant de l'utilisateur>/TXM/R` »
    - Sous Linux :
      « `/home/<identifiant de l'utilisateur>/TXM/R` ».
- Le dossier d'installation des packages R est initialisé à :
    - Sous Windows :
      « `C:\Utilisateurs\<identifiant de l'utilisateur>\TXM\R\libraries` »
      ou bien
      « `C:\Documents and Settings\<identifiant de l'utilisateur>\TXM\R\libraries` »
    - Sous Mac OS X :
      « `/Users/<identifiant de l'utilisateur>/TXM/R\libraries` »
    - Sous Linux :
      « `/home/<identifiant de l'utilisateur>/TXM/R\libraries` ».

La configuration des options de lancement de R (et Rserve) est accessible depuis la page de préférences avancées de R (TXM > Avancé > Moteur statistique). Pour plus d'informations sur R, voir les pages suivantes :

- Options de R : http://www.r-project.org
- Options de Rserve : http://rforge.net/Rserve/doc.html#start

12.3.5 Exemple de session de travail utilisant R

12.3.5.1 Affichage de l'histogramme des fréquences d'un index de lemmes calculé avec R

L'objectif de cette session exemple est d'afficher l'histogramme des fréquences des 10 lemmes de noms les plus fréquents du corpus Discours :

- construire l'Index des lemmes de noms :
    - commande Index avec les paramètres :

        - corpus : DISCOURS
        - requête : `[pos="N.*"]`
        - propriété : lemma
        - Vmax: 10
- envoyer l'Index dans R :
    - dans le menu contextuel de l'icone de résultat d'Index, descendante de l'icone du corpus DISCOURS dans la vue Corpus et nommée « [pos=”N.*”]:lemma », lancer la commande « Envoyer vers R »
- l'icone de résultat d'Index doit recevoir une lettre R rouge pour confirmer le transfert
- repérer et copier le nom de l'objet R créé dans la console : par exemple « Index1 »
- ouvrir la perspective R en cliquant sur le bouton « R » de la barre d'outils des perspectives
- démarrer une nouvelle session en cliquant sur le bouton « Nouvelle session » de la barre d'outils
- copier le script R suivant dans l'éditeur de texte nommé « sessionX.R » (où « X » est le numéro de la session):

```
svg("/tmp/test.svg")
barplot(t(IndexN$data), space=c(1,35), horiz=F, las=2, beside=T)
dev.off()
```

- éditer la chaine « IndexN » en collant le nom de l'objet R créé lors du transfert vers R (par exemple « Index1 »)
- exécuter le script en cliquant sur le bouton « Exécuter » de la barre d'outils (flèche verte)
- vérifier le résultat en ouvrant le fichier « /tmp/test.svg » dans votre navigateur à partir du gestionnaire de fichiers.

12.3.5.2 Affichage de l'histogramme directement dans TXM avec un script Groovy

Vous pouvez afficher le fichier SVG généré directement dans TXM en double-cliquant sur son icone depuis la vue « Fichier » (dont l'onglet est par défaut situé à côté de celui de la vue « Corpus »).

Vous pouvez également afficher directement le fichier SVG en exécutant le script Groovy suivant :

```
import org.txm.rcpapplication.commands.*
monitor.syncExec(new Runnable() { @Override

     public void run() { OpenSVGGraph.OpenSVGFile("/tmp/test.svg",
"histogram plot") }

});
```

- d'abord copier le script dans un éditeur de texte (menu « Fichier / Nouveau fichier »);
- ensuite sélectionner le texte du script ;
- enfin, lancer « Groovy / Exécuter la sélection de texte » depuis le menu contextuel de l'éditeur.

Si l'éditeur de texte est ouvert sur un fichier du dossier « $HOME/scripts/user » ayant une extension « .groovy », alors vous pouvez exécuter le script directement par le menu contextuel de l'éditeur « Groovy / Exécuter le Script ».

12.3.5.3 Exécution du script R et de l'affichage de l'histogramme depuis un seul script Groovy

Il est possible d'assembler le script R et le script d'affichage du SVG dans un seul script Groovy :

- dans un éditeur de texte ouvert sur un fichier du dossier « $HOME/scripts/user » ayant une extension « .groovy », copier le script Groovy suivant :

```
import org.txm.rcpapplication.commands.*
import org.txm.stat.engine.r.RWorkspace


def r = RWorkspace.getRWorkspaceInstance()


r.eval("""
##### début du script R #####


svg("/tmp/test.svg")
barplot(t(IndexN\$data), space=c(1,35), horiz=F, las=2, beside=T)
dev.off()
```

```
##### fin du script R #####
""")


import org.txm.rcpapplication.commands.*
monitor.syncExec(new Runnable() { @Override public void run() {
     OpenSVGGraph.OpenSVGFile("/tmp/test.svg", "histogram plot") }
});
```

- remplacer le nom du symbole « IndexN » par celui de l'Index transféré à R (par exemple « Index1 ») ;
- enfin, lancer “Groovy / Exécuter le Script” dans le menu contextuel de l'éditeur.

12.3.5.4 Saisie du nom de l'index depuis une boite de dialogue à l'aide d'une macro TXM

Pour faciliter la mise à jour du nom du symbole dans le script, on peut le faire saisir dans une boite de dialogue en transformant le script en macro TXM :

- dans un éditeur de texte ouvert sur le fichier « testMacro.groovy » du dossier « $HOME/scripts/macro », copier le script Groovy suivant :

```
import org.kohsuke.args4j.*
import groovy.transform.Field
import org.txm.rcpapplication.swt.widget.parameters.ParametersDialog


import org.txm.rcpapplication.commands.*
import org.txm.stat.engine.r.RWorkspace


@Field @Option(name="symbol_name",usage="symbol name of the Index to
use", widget="String", required=true, def="Index1")
def symbol_name
if (!ParametersDialog.open(this)) return;


def r = RWorkspace.getRWorkspaceInstance()
```

```
r.eval("""
##### début du script R #####

svg("/tmp/test.svg")
barplot(t(${symbol_name}\$data), space=c(1,35), horiz=F, las=2, beside=T)
dev.off()

##### fin du script R #####
""")

monitor.syncExec(new Runnable() {
     @Override
     public void run() {     OpenSVGGraph.OpenSVGFile("/tmp/test.svg",
"histogram plot") }
});
```

- ouvrir la vue « Macro » (menu principal « Affichage / Vues / Macro »)
- double-cliquer sur l'icone du fichier « test » dans la vue Macro pour l'exécuter :
  - dans la boite de dialogue, saisir le nom de symbole voulu dans le champ « symbol_name » et presser « Run ».

12.3.5.5 Récupération du nom de l'index directement depuis la macro

Pour aller plus loin encore dans la simplification de la désignation de l'index concerné, il est possible d'obtenir le nom de l'index transféré depuis la macro en désignant l'index dans la vue Corpus :

- dans un éditeur de texte ouvert sur le fichier « testMacro.groovy » du dossier « $HOME/scripts/macro », copier le script Groovy suivant :

```
import org.txm.rcpapplication.commands.*
import org.txm.stat.engine.r.RWorkspace
```

```
import org.txm.functions.vocabulary.*


if (!corpusViewSelection || !(corpusViewSelection instanceof Vocabulary))
{
     println "Error: this macro should be run with an Index selected"
     return
}
def symbol_name = corpusViewSelection.getSymbol()


def r = RWorkspace.getRWorkspaceInstance()


r.eval("""
##### début du script R #####


svg("/tmp/test.svg")
barplot(t(${symbol_name}\$data), space=c(1,35), horiz=F, las=2, beside=T)
dev.off()


##### fin du script R #####
""")


monitor.syncExec(new Runnable() {
     @Override
     public void run() {     OpenSVGGraph.OpenSVGFile("/tmp/test.svg",
"histogram plot") }
});
```

- sélectionner l'icone d'Index résultat dans la vue Corpus pour lequel vous souhaitez afficher l'histogramme (ne pas oublier de transférer l'Index vers R au préalable) ;
- ouvrir la vue « Macro » (menu principal « Affichage / Vues / Macro ») ;
- double-cliquer sur le fichier macro pour l'exécuter.

Pour lancer successivement la même macro sur plusieurs résultats d'Index, une configuration pratique consiste à afficher la vue Macros dans le panneau de gauche avec la vue Corpus. Par exemple, le moitié supérieure pour la vue Corpus et le moitié inférieure pour la vue Macros (pour déplacer la vue Macros, faire glisser son onglet dans le panneau de gauche).

12.3.6 Description des principaux objets TXM transférés à R

a) un objet `Lexique` est transféré sous forme d'une matrice (*matrix*) d'un seul vecteur colonne (*vector*) de fréquences (*numeric*) dont chaque ligne est nommée par la valeur de la propriété décomptée.

b) un objet `Index` issu d'un corpus ou d'un sous-corpus est transféré sous forme d'une liste (*data.frame*) contenant un objet nommé `data` . L'objet `data` est une matrice d'un seul vecteur colonne de fréquences dont chaque ligne est nommée par la valeur de la propriété décomptée ou par la concaténation des valeurs des propriétés décomptées quand il y en a plusieurs (dans ce cas les valeurs sont séparées par le caractère souligné : « _ »).

c) un objet `Index` issu d'une partition est transféré sous forme d'une liste contenant un objet nommé `data` . L'objet `data` est une matrice d'autant de vecteurs colonnes de fréquences que de parties. Chaque ligne est nommée par la valeur de la propriété décomptée ou par la concaténation des valeurs des propriétés décomptées quand il y en a plusieurs (dans ce cas les valeurs sont séparées par le caractère souligné : « _ »). La première colonne contient les marges de lignes, elle est nommée `F.`

d) un objet `Tablelexicale`, issu d'une partition ou d'un index de partition est transféré sous forme d'une matrice d'autant de vecteurs colonnes de fréquences que de parties. Chaque ligne est nommée par la valeur de la propriété décomptée ou par la concaténation des valeurs des propriétés décomptées quand il y en a plusieurs (dans ce cas les valeurs sont séparées par le caractère souligné : « _ »).

e) un objet `Concordance` est transféré sous forme d'une liste contenant des objets nommés `data`, `leftcontext` et `rightcontext`. L'objet `data` est une matrice de 4 vecteurs colonnes de chaînes nommés `refs`, `leftcontext`, `keywords` et `rightcontext`. Ils contiennent respectivement, les références, les contextes gauches, les pivots et les contextes droits. Les objets `leftcontext` et `rightcontext` contiennent la taille en mots des contextes gauche et droit.

f) un objet `Corpus` est transféré sous forme d'une liste contenant des objets de base stockant les codes des valeurs de propriétés de mots pour chaque occurrence dans l'ordre des textes du corpus (`data`), ainsi que les dictionnaires de ces valeurs (`wordlex` pour le dictionnaire des formes et `idlex` pour le dictionnaire des identifiants de mots).

Si les mots du corpus ont été annotés, les valeurs de ces annotations sont stockées dans autant de dictionnaires supplémentaires que nécessaire nommés `frposlex`, `frlemmalex` (c'est-à-dire « code de langue|nom de propriété|`lex` » pour des annotations TreeTagger) ou bien de

n'importe quel nom d'annotation supplémentaire. L'objet data est une matrice de vecteurs colonnes pour chaque propriété de mot. Les colonnes sont nommées par les noms de propriétés. Les positions de début et de fin de textes sont stockées dans les vecteurs `$struct$text$start` et `$struct$text$end` de l'objet R créé.

13 Raccourcis clavier

Description de tous les raccourcis clavier de TXM.

13.1 Tableaux de résultats

Commandes	Raccourcis
Chercher une chaîne de caractères	Ctrl-F
Copier les lignes sélectionnées dans le presse-papier	Ctrl-C

13.2 Graphiques

Commandes	Raccourcis
Déplacer la vue	Clic droit et déplacement souris / flèches du clavier
Changement d'échelle	Molette souris / Ctrl + et Ctrl -

13.1 Éditeur de texte

Commandes	Raccourcis
Aide	
Afficher les raccourcis disponibles	Ctrl-Shift-L
Sélection	
Sélectionner tout	Ctrl-A
Sélectionner la première ligne	Shift-Home
Sélectionner la dernière ligne	Shift-End
Sélectionner le mot suivant	Ctrl-Shift-Right
Sélectionner le mot précédent	Ctrl-Shift-Left
Éditer	
Copier	Ctrl-C, Ctrl-Insert
Coller	Ctrl-V, Shift-Insert

Couper	Ctrl-X,Shift-Suppr
Supprimer	Suppr
Annuler	Ctrl-Z
Rétablir	Ctrl-Y
Mettre en majuscules	Ctrl-Shift-X
Mettre en minuscules	Ctrl-Shift-Y
Chercher	
Chercher / Remplacer	Ctrl-F
Chercher suivant	Ctrl-K
Chercher précédent	Ctrl-Shift-K
Recherche « incrémentale »	Ctrl-J
Retour à la recherche simple	Ctrl-Shift-J
Se déplacer	
Début du texte	Ctrl-Home
Fin du texte	Ctrl-End
Début de ligne	Home
Fin de ligne	End
Mot suivant	Ctrl-Right
Mot précédent	Ctrl-Left
Aller à la ligne	Ctrl-L
lieu d'édition précédent	Ctrl-Q
Suppression	
Supprimer une ligne	Ctrl-D
Supprimer jusqu'à la fin de la ligne	Ctrl-Shift-Suppr
Supprimer le mot suivant	Ctrl-Suppr
Supprimer le mot précédent	Ctrl-Backspace
Déplacer les lignes	
Monter la ligne courante	Alt-Up
Descendre la ligne courante	Alt-Down
Insérer une ligne	
Insérer une ligne au dessus de la ligne courante	Ctrl-Shift-Enter
Insérer une ligne en dessous de la ligne	Shift-Enter

courante	
Autre	
Concaténer les lignes	Ctrl-Alt-J
Faire défiler les lignes vers le haut	Ctrl-Up
Faire défiler les lignes vers le bas	Ctrl-Down
Dupliquer les lignes	Ctrl-Alt-Up
Copier les lignes	Ctrl-Alt-Down
Mode	
Basculer en mode passage à la ligne automatique	Ctrl-PavéNum_Diviser
Basculer en mode insertion	Ctrl-Shift-Insert
Basculer en mode remplacement	Insert
Basculer en mode sélection	Alt-Shift-A
Basculer en mode comparaison rapide	Ctrl-Shift-Q
Montrer le menu contextuel de comparaison	Ctrl-F10
Fichier	
Nouveau	Ctrl-N
Enregistrer	Ctrl-S
Fermer	Ctrl-W, Ctrl-F4
Tout fermer	Ctrl-Shift-W
Imprimer	Ctrl-P
Propriétés	Alt-Enter
Rafraîchir	F5
Divers	
Complétion de mot	Alt-/
Scripts	
Exécuter le texte sélectionné	F11
Exécuter le Script	Ctrl-F11
Exécuter le dernier script ou la dernière macro exécutée	F12

13.2 Fenêtres

Gérer les fenêtres

Fenêtre suivante	Ctrl-F6
Fenêtre précédente	Ctrl-Shift-F6
Menu déroulant des fenêtres	Ctrl-E
Retour au menu fenêtre	Ctrl-Shift-E
Afficher le menu de la fenêtre	Alt -

Vues

Agrandir la fenêtre actuelle	Ctrl-M
Vue suivante	Ctrl-F7
Vue précédente	Ctrl-Shift-F7
Afficher le menu Vue	Ctrl-F10
Afficher les raccourcis	Ctrl-Shift-L
Afficher la console	Alt-Shift-Q, C

13.3 Raccourcis généraux

Exécuter le dernier script exécuté	F12
Quitter TXM	Ctrl-Q

14 Jeu d'étiquettes morphosyntaxiques du français

Valeurs de la propriété « frpos » pour le modèle de français contemporain de TreeTagger[72].

Étiquette	Glose
ABR	abréviation
ADJ	adjectif
ADV	adverbe
DET:ART	article
DET:POS	pronom possessif (ma, ta...)
INT	interjection
KON	conjonction
NAM	nom propre
NOM	nom
NUM	numéral
PRO	pronom
PRO:DEM	pronom démonstratif
PRO:IND	pronom indéfini
PRO:PER	pronom personnel
PRO:POS	pronom possessif (mien, tien...)
PRO:REL	pronom relatif

Étiquette	Glose
PRP	préposition
PRP:det	préposition plus article (au, du, aux, des)
PUN	ponctuation
PUN:cit	ponctuation de citation
SENT	balise de phrase
SYM	symbole
VER:cond	verbe au conditionnel
VER:futu	verbe au futur
VER:impe	verbe à l'impératif
VER:impf	verbe à l'imparfait
VER:infi	verbe à l'infinitif
VER:pper	verbe au participe passé
VER:ppre	verbe au participe présent
VER:pres	verbe au présent
VER:simp	verbe au passé simple
VER:subi	verbe au subjonctif imparfait
VER:subp	verbe au subjonctif présent

[72] Source : French TreeTagger Part-of-Speech Tags Achim Stein, April 2003 <http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html>

15 Glossaire

Définition des notions essentielles de la textométrie et de TXM.

Catégories :

- com : Commande
- don : Modèle de données
- for : Format de fichier
- int : Interface utilisateur
- tal : Traitement Automatique de la Langue (TAL)
- req : requête CQL
- log : Composant logiciel
- mét : Méthodologie Textométrique

Entrée	**Cat**	**Description**
AFC	com	action de réduire le nombre de dimensions d'une matrice (de type « parties x mots ») avec l'algorithme d'analyse factorielle des correspondances. Les nouvelles dimensions sont représentées par des vecteurs propres appelés facteurs. Les parties et les mots de la matrice originelle peuvent être affichés simultanément dans les plans factoriels résultants.
AFR	tal	code standard pour l'ancien français.
Alceste	log	logiciel commercial textométrie.
annotation	don	propriété d'une unité (lexicale ou structurelle) d'un point de vue logique.
balise	don	représentation bornée d'un élément, qui contient ses propriétés, en langage XML
caractère	don	unité élémentaire constituant la forme d'un mot.
CATTEX2009	tal	jeu d'étiquettes morphosyntaxiques pour l'ancien français.
module d'importation	com	composant logiciel qui importe des éléments dans la plateforme TXM, depuis une source.

ClipN	int	corpus créés à partir du presse-papier sont nommés 'Clip'+<un numéro>.
CNR	for	format de données de Cordial.
commande	com	action disponible dans TXM.
concordance	com	manière de présenter les résultats d'une recherche, où chaque occurrence apparaît centrée sur sa propre ligne, au milieu de son contexte.
console	int	TXM affiche divers messages lors de son exécution, dans une fenêtre appelée « console ».
Cordial	tal	étiqueteur morphosyntaxique et lemmatiseur commercial.
corpus	don	ensemble de mots. Ces ensembles viennent de textes, entiers ou lacunaires. Les corpus « racines » sont construits à partir de bases.
CQL	req	pour <Corpus Query Language>, langage de requêtes géré par CQP, appliqué aux corpus.
CQP	log	pour <Corpus Query Processor>, module logiciel gérant les requêtes pour construire des index, concordances, etc.
CSV	for	signifie « Comma Separated Values ». C'est un fichier texte où chaque ligne de résultat est séparée par saut de ligne et où les valeurs sont séparées par un caractère séparateur (comme la virgule).
Ctrl	int	touche « Ctrl » ou « Control » sur le clavier.
document	don	texte logique.
éditeur	com	fenêtre où un texte (comme un fichier source ou un script) peut être modifié.
encodage	don	façon dont une information est représentée dans le corpus source.
espace de travail	int	ensemble de tous les objets disponibles dans TXM (corpus, sous-corpus...).
étiqueteur	log	logiciel indépendant, capable de segmenter les mots, de leur

		associer une étiquette morphosyntaxique ou un lemme, à partir de sources textuelles.
étiquette	tal	propriété morphosyntaxique d'un mot
export	com	action d'enregistrer dans un fichier les résultats d'une commande TXM.
fichier	don	élément du système d'exploitation contenant des informations sur le disque dur de l'utilisateur : comme un texte ou un corpus source. Un fichier peut être désigné par un chemin d'accès.
focus	int	focaliser une commande sur un événement lexical particulier, par exemple à travers une requête.
forme graphique	don	forme graphique d'un mot, généralement calculée par les tokeniseurs.
fréquence	mét	nombre total d'occurrences d'un événement (une occurrence de mot, une occurrence de séquence de mots, etc.) dans un corpus.
Groovy	log	langage informatique dans lequel les scripts de TXM sont écrits.
HTML	for	format de représentation des données des pages web.
Hyperbase	log	logiciel académique de textométrie.
import	don	fait d'intégrer un corpus à la plateforme, à partir de fichiers source.
index	com	lister toutes les combinaisons de propriétés de mots, avec leur fréquence, pour toutes les occurrences d'une requête.
index	log	fichier créé par TXM afin d'accélérer les réponses aux requêtes.
indice	tal	valeur numérique fournie pour un modèle statistique.
infobulle	int	fenêtre temporaire qui s'affiche lorsqu'on survole un objet avec le curseur de la souris, par exemple, un mot dans une édition.
Java	log	langage dans lequel est programmé TXM.
jeu d'étiquettes	don	ensemble des valeurs morphosyntaxiques possibles de mots.

langage	don	langage dans lequel est écrit un texte ou un corpus.
lem	don	voir lemme.
lemme	don	entrée d'un mot dans le dictionnaire courant.
lemmatiseur	log	module logiciel qui fait correspondre une entrée de dictionnaire à chaque mot du texte
lexique	com	lister toutes les formes possibles de mots, ou de fréquences de propriétés de mot, dans un corpus.
ligne de statut	log	TXM affiche des commentaires temporaires sur les opérations qu'il effectue, dans un espace situé en bas à gauche de l'interface.
littéral	req	caractère considéré pour lui-même dans une requête.
localisation	int	l'interface de TXM peut s'afficher dans différentes langues, qui peuvent être paramétrées dans le menu « localisation » des préférences.
matcher	tal	correspondance structurelle dans l'algèbre des caractères de propriétés ou des occurrences
metadonnées	don	propriétés d'un texte ou d'un document entiers. Chaque métadonnée a un nom, un type et une valeur.
modifieur	req	caractère spécial utilisé pour exprimer certaines variantes dans une requête.
mot	don	unité lexicale identifiée grâce à sa forme graphique et à sa position dans la séquence des mots. Elle est généralement construite par les tokeniseurs.
Multext	tal	jeu d'étiquettes standard européen.
occurrence	mét	apparition d'un événement textuel dans un corpus, comme une occurrence de mot.
opérateur	req	caractère spécial ayant une signification particulière dans une requête.
page	don	segment de texte affiché sur un support, correspondant généralement à une page d'une édition papier.

partie	don	élément d'une partition d'un corpus.
partition	don	découpage d'un corpus en différentes parties. La somme de toutes ces parties correspond au corpus dans son ensemble. On utilise les partitions pour analyser les contrastes entre les parties (comme entre les dates de discours, des auteurs, des sections d'un même texte, etc.)
phrase	tal	séquence de mots, syntaxiquement homogène, construite par les tokeniseurs.
pivot	com	colonne centrale d'une concordance, affichant toutes les occurrences d'une requête donnée.
pos	don	pour « part of speech », les informations morphosyntaxiques d'un mot.
préférence	int	chaque commande de TXM possède des paramètres. Certains de ces paramètres peuvent être réglés dans la fenêtre « Préférences ».
presse-papier	don	fonction du système d'exploitation permettant de stocker une sélection de texte, grâce à la commande « copier ».
propriété	don	information sur une unité lexicale ou structurelle.
référence	int	information affichée au début d'une ligne de concordance, qui provient des propriétés des unités lexicales et structurelles.
dossier	don	dossier contenant des fichiers ou d'autres dossiers, sur le disque dur de l'utilisateur. Un dossier peut être désigné par un chemin.
requête	com	chaine de caractères exprimant une combinaison de mots et de propriétés de mots.
script	log	fichier contenant une description d'actions précises qui peut être exécutée par TXM.
sélection	mét	liste de séquences de mots. Le résultat d'une recherche pour une requête est une sélection.
source	don	représentation initiale d'un corpus, dans un format propre, contenue dans plusieurs fichiers ou dossiers. Par exemple, le format peut être du TXT (texte brut), du XML ou de la TEI.

spécificité	com	action de lister des formes de mots spécifiques, ou des propriétés de mot, à chaque partie d'une partition, conformément au modèle quantitatif des spécificités.
T	met	le nombre total d'occurrences dans un corpus
TAL	log	pour « Traitement Automatique de la Langue ».
TEI	for	pour « Text Encoding Initiative », la façon standard d'encoder les textes. Consortium international de standardisation de l'encodage des sources de corpus. Voir http://www.tei-c.org. Le format TEI est exprimé en XML.
texte	don	séquence de mots de structure homogène, décrite par des propriétés appelées métadonnées.
textométrie	mét	méthodologie qu'applique TXM. La textométrie vous aide à analyser les corpus de textes, au moyen d'outils quantitatif et qualitatifs. Voir http://textometrie.ens-lyon.fr.
tokeniseur	log	composant logiciel capable de séparer les mots et de les caractériser par des propriétés, dans les fichiers source.
TreeTagger	log	logiciel étiqueteur indépendant académique
TXT	for	format de données d'un fichier en texte brut (sans aucune annotation).
unité	don	unité lexicale ou structurelle d'un texte.
unité structurelle	don	élément marquant la structure logique d'un texte. Dans TXM, toutes les unités structurelles sont organisées de manière hiérarchique : chaque unité est imbriquée dans une autre unité – jusqu'à l'unité 'text'. La plus petite unité structurelle se trouve juste au-dessus de l'unité lexicale.
V	mét	nombre total de formes graphiques différentes d'un corpus.
vocabulaire	com	générer un lexique ou un index.
Weblex	log	logiciel de textométrie académique.
window manager	int	logiciel qui permet d'organiser son interface de travail.

XML	for	format de données principal des sources des corpus.

16 Bibliographie (incomplète)

[en chantier]

Barclay, Kenneth A., et W. J. Savage. *Groovy programming: an introduction for Java developers*. Morgan Kaufmann Publishers, 2007.

Benzécri, Jean-Paul, et al. *L'analyse des correspondances*. Paris: Dunod, 1973.

Oliver Christ, Bruno M. Schulze, Anja Hofmann, and Esther König, « The IMS Corpus Workbench : Corpus Query Processor (CQP), User's Manual », August 16, 1999 (CQP V2.2), University of Stuttgart , <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML> ou <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/PDF/cqpman.pdf>.

König, Dierk, Andrew Glover, Paul King, Guillaume Laforge, et al. *Groovy in action*. Greenwich: Manning, 2007.

Lafon80 Lafon, P. “Sur la variabilité de la fréquence des formes dans un corpus.” *Mots*, no. 1 (1980): 127-165.

<http://www.persee.fr/web/revues/home/prescript/article/mots_0243-6450_1980_num_1_1_1008>.
[d'abord présenté à la conférence « Association for Literary and Linguistic Computing », à Oxford les 4 et 5 Avril 1976]

Venkat, Subramaniam. *Programming Groovy: dynamic productivity for the Java developer*. Pragmatic Bookshelf. Raleigh: Daniel H. Steinberg ed., 2008.

17 Index

Index des illustrations

Index des tables

Table des mises à jour (suite)

13/03/10	Serge Heiden (SH)	Création
02/07/10	Matthieu Decorde (MD)	Mise à jour pour la version 0.4.7
15-29/07/10	SH	Réécriture pour la version 0.4.7
27/08/10	SH	Numérotation des titres, réorganisation du plan
08/10/10	Lauranne Bertrand (LB)	Version française du manuel pour la 0.5
05/01/11	MD	Mise à jour pour la version de release 0.5
14/01/11	SH	Corrections
18/01/11	SH	Corrections, ajout de la section sur les modules d'importation
06/06/11	MD	Copie et mise à jour pour la version

13/03/10	Serge Heiden (SH)	Création
		0.6
09/04/12	SH	Màj 0.6 : garde, tdm
17/06/12	SH	Màj 0.6 préface, tdm
17/07/12	MD	Copie et mise à jour pour la version 0.7
06/08/12	SH	Màj , garde, préface, installation Windows, Mac et Linux
29-30/09/12	SH	Ajout sections « jeu d'étiquettes français » et « expressions régulières ». Déplacé « en cas de problème ». Correction de la section « préférences ».
05/03/2013	SH	Ajouté tampon « provisoire » Remplacé la section 6 par le « mémo CQL » de B. Pincemin
07/03/2013	SH	Fait une passe de relecture légère sur les modules d'import
17/05/2013	SH	Màj section regexp et lien cqpman.pdf
08/11/2013	MD	Ajout des sections « Macros », « Mise à jour automatique » et « installation d' extension»
08/11/2013	BP	Mise à jour de la section import Alceste
19/11/2013	MD	Ajout de la section « import Factiva XML »
28/11/2013	SH	Finalisation sections « Mise à jour automatique », « ajout d'extension », « Macros », réglages « Import Alceste »
10/12/2013	SH	Petits réglages des sections « Mises à jour automatiques », « Installer une extension », « En cas de problème »

13/03/10	Serge Heiden (SH)	Création
16/12/2013	SH	Refonte de la section « Piloter la plateforme TXM avec des macros et des scripts Groovy ou du code R », ajout de la section « Exécuter du code R depuis TXM ».
06/01/2014	SH	Màj, et reform. section 1 et réécr. part. section 6.3
06/02/2014	SH	Réglage des sections sur les mises à jour et extensions, màj TOC et indexes.
14/04/2014	SH	Màj titre courant
06/05/2014	SH	Màj installation Mac OS X + Configuration recommandée
19/06/2014	SH	Màj section 11.12 AFC
04/07/2014	Sébastien Jacquot (SJ)	Màj pour TXM 0.7.6 : graphiques
22/07/2014	MD	Màj pour TXM 0,7,6 : voir tickets #740, #747, #762, #777, #779, #782, #799, #914, #920
17/09/2014	MD	Maj pour ticket #1013
20/10/2014	SH	Ajout section 16.5.3.2 (CSS des éditions)
24/10/2014	SH	Accepted all SJ and MD modifications, began to rewrite import section
30/10/2014	SH	Transfered « specificity » section elements from Weblex manual
06/11/2014	SH	Updated installation sections

18 Table des matières