SYLED
SYstèmes Linguistiques, Énonciation et Discours
UPRES SYLED, EA 2290, Université de la Sorbonne nouvelle - Paris 3

ILPGA/Sorbonne nouvelle - Paris 3
19 rue des Bernardins, 75005 Paris
Tél : 01.44.32.05.75

Email : fleury@msh-paris.fr
Ou serge.fleury@univ-paris3.fr

Hypertoile SF :
http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/

Hypertoile TAL Paris 3 :
http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/

MkCorpus (prototype)

Outil de préparation et de manipulation de Corpus CLA2T-ILPGA@2000

Page Web MKCORPUS (téléchargement, documentation) :
www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/mkcorpusProject.htm

Document de travail (15/04/2001)

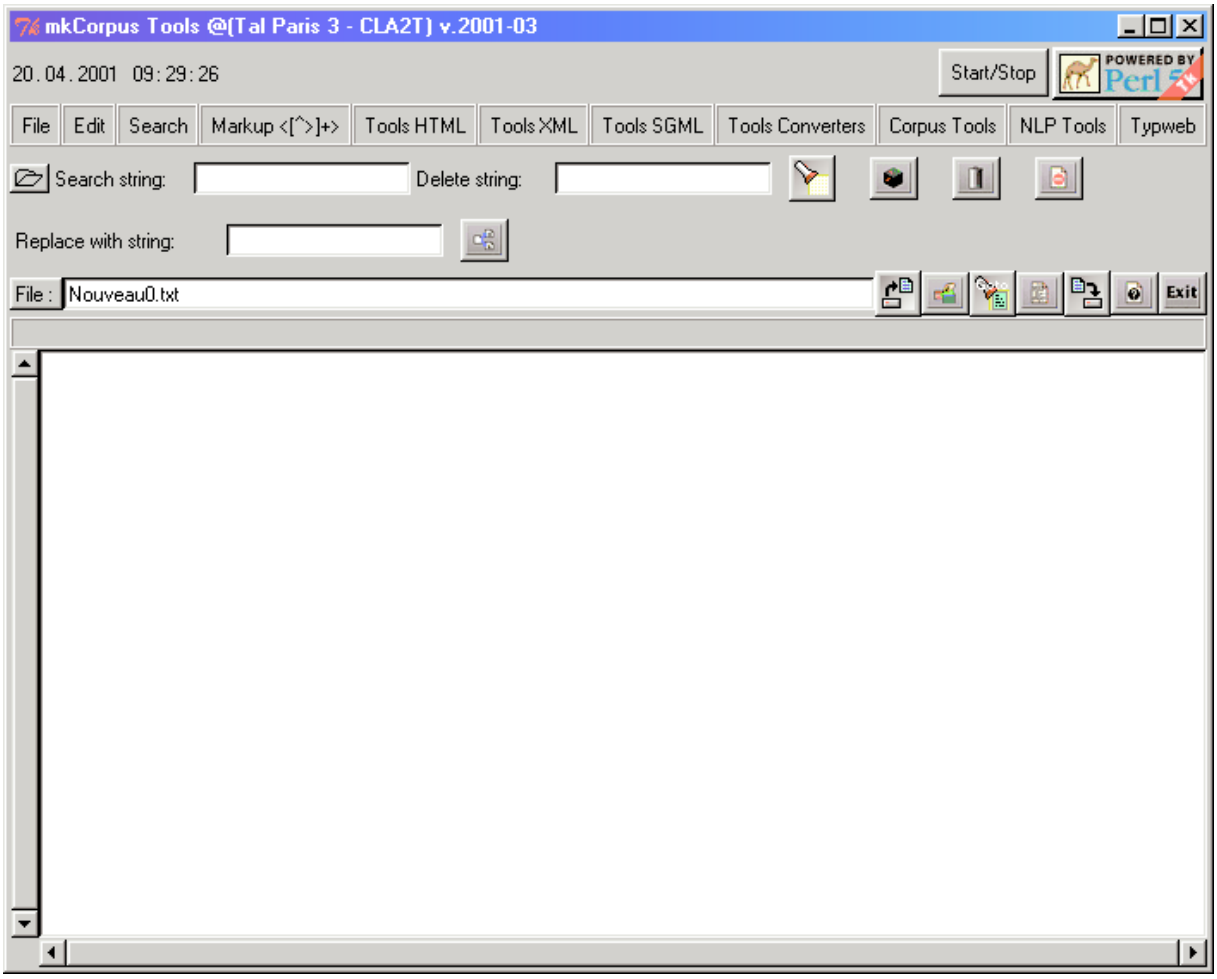
1 Préambule

Mkcorpus est un programme de préparation de corpus pour leurs analyses ultérieures via des outils traditionnels du TAL. Il est écrit en Perl/TK.

Ce programme permet :

- de visualiser le corpus,
- de manipuler via des outils idoines le contenu du corpus et de ses éléments pour les formater suivant les contingences imposées par les outils (suppression de balises, nettoyage…).

Cet outil se présente comme un éditeur traditionnel et les menus construits permettent de réaliser des opérations sur les fichiers visualisés dans la zone d'édition ou attachés aux programmes de traitement.



2 Installation

Pour utiliser MKCORPUS, il faut télécharger l'archive disponible sur la page :

www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/mkcorpusProject.htm

On trouve sur cette page deux versions de MKCORPUS.

- Une version exécutable : celle ci sera disponible dans une version complète dans des délais indéterminés. La version en ligne ne dispose pas de toutes les options présentées infra. Elle est utilisable sous PC-Windows. Il suffit de lancer le programme fournit dans l'archive.
- Une version en code source complète. Le plus simple pour l'installer est de "dézipper" l'archive dans un répertoire en respectant l'arborescence construite dans l'archive. Cette version a été testée sous Windows et Linux. Pour lancer MKCORPUS, il faut lancer le programme Perl nommé `mkCorpus-win32.pl`.

  Pré-requis : il faut disposer de Perl et de Perl/Tk. Il faut aussi installer les modules Perl contenus dans le répertoire `mkCorpusModules` contenu dans l'archive. Pour ces installations suivre les instructions contenues dans les `Readme` de chaque module :il faut en général déposer une bibliothèque dans le répertoire local `lib` de Perl (en général c:\Perl\lib sous windows, sous Linux, la procédure d'installation standard fait le travail parfaitement).

3 Modules de MKCORPUS

On donne ci-dessous un descriptif des différentes opérations actuellement disponibles via cet outil .

Les modules présentés ci-dessous sont disponibles actuellement :

- **Fichier** : ouverture, sauvegarde
- **Edition** : outils traditionnels d'édition
- **Search** : recherche (regexp), remplacement, (dans menu et via boutons)
- **Markups <[^>]+>**: liste, suppression, modification
- Ce module utilise la représentation de balise sous la forme d'expression régulière du type <[^>]+> : solution intéressante mais pas toujours concluante.
- Extraction de texte entre balises <[^>]+>.
- **Tools XML** : sur des documents xml valides : stat, comments, extraction de texte entre balises.
  - Tools XML/SGML/RTF : outils spécificiques pour XML, SGML et RTF (en cours de mise au point)
  - Incorporation d'un parser SGML/XML (*cf* programme balise)
- **Corpus Tools** :
  - préparation de corpus (recherche et nettoyage de caractères), avec sous-modules spécifiques pour Lexico, Cordial, Alceste
  - représentation graphique de corpus balisés : "parcours" dans l'arbre associé au texte balisé via une interface graphique
- **Typweb** : la chaine typweb est disponible (webxref, mktypo, mkstat). La phase d'aspiration de site est en cours de mise au point.
- **NLP tools** : concordance, bigrammes, collocation, programme `xword` (cf "Web programming with perl" O'Reilly)

3.1 Mise en œuvre et utilisation des modules

MKCORPUS se présente comme un éditeur de texte traditionnel : une fenêtre d'édition, des menus et des boutons. La fenêtre principale est donc destinée à abriter un fichier en édition. Les principales actions disponibles sont activées via les menus ou via les boutons (qui correspondent en fait à des raccourcis de fonctions disponibles dans les menus).

3.1.1 Les menus

3.1.1.1 Menu FICHIER

- *View dir* : affiche le contenu de l'arborescence du disque de travail sous une forme d'arbre.
- *Select File & open* : déclenche l'ouverture d'une boîte de dialogue pour l'ouverture d'un fichier afin de l'éditer.
- *Select File & not open* : déclenche l'ouverture d'une boîte de dialogue pour associer un fichier à des traitements sans ouvrir celui ci dans la fenêtre principale. Cette option est surtout destinée à être utiliser dans le cas de travail sur de gros fichiers. Cette option n'est pas encore disponible.

- *Open this file* : charge le fichier courant, celui dont le nom est affiché dans la boîte File.
- *Save this file* : sauvegarde le fichier courant, celui dont le nom est affiché dans la boîte File.
- *Save this file as* : sauvegarde le fichier courant, celui dont le nom est affiché dans la boîte File avec possibilité de le renommer. Cette option fait appel à une boîte de dialogue.
- *Clear* : cette option crée un nouveau fichier dans la fenêtre principale. Ce fichier n'existe pas physiquement. Il faut au préalable l'enregistrer.

3.1.1.2 Menu EDITION

Ce menu fait appel à des opérations traditionnelles disponibles dans les éditeurs : copier, coller, sélectionner...

3.1.1.3 Menu SEARCH

Ce menu fait appel à des opérations traditionnelles disponibles dans les éditeurs : recherche incrémentale, recherche globale, remplacement... Ces recherches utilisent la notions d'expression régulière. Le motif de recherche doit être au préalable intégré dans le champ "Search String" idem pour le motif de remplacement le cas échéant.

- Le bouton "highliht" est équivalent à une recherche globale du motif donné.
- Le bouton "delete" est équivalent à un remplacement global du motif donné par le motif de remplacement donné.
- Chaque recherche concluant provoque l'affichage coloré et clignotant du motif trouvé. Le bouton "no hightlight" inhibe ce formatage.

3.1.1.4 Menu MARKUP <[^>]+>

Ce menu permet de réaliser des traitements sur des fichiers balisés, avec des balises du type <BALISE>. Les opérations disponibles utilisent une représentation formelle de ce type de balise sous la forme d'une expression régulière suivante "<[^>]+>".

Remarque importante : la représentation d'une balise sous la forme de cette expression régulière n'est en aucun cas une solution optimale pour représenter une balise. Nous l'avons retenu dans la mesure où les fichiers que nous avons l'habitude de manipuler ne présentent pas de balises dont l'écriture est scindée sur deux lignes.

- *View all* : crée une nouvelle fenêtre contenant toutes les balises du fichier courant de la fenêtre principale.Il est possible de sauvegarder le fichier construit.
- *delete all* : supprime toutes les balises du fichier courant de la fenêtre principale.
- *global search* : recherche globale de toutes les balises.
- *I-search* : recherche incrémentale de motifs donnés dans le champ de recherche.
- *repeat I-Search* : nouvelle recherche du motif en cours de recherche.
- *Extract text in markup* : cette option active la génération d'une nouvelle fenêtre qui permet de sélectionnner des balises du texte courant pour : les modifier, les supprimer ou bien pour extraire le contenu situé entre la balise ouvrante et la balise fermante de la balise visée.

3.1.1.5 Menu HTML

- Module en cours de test et non documentés

3.1.1.6 Menu XML

Outils XML pour non documentés

- visualiser l'arbre XML associé au document XML contenu dans la fenêtre principale
- parser le document XML contenu dans la fenêtre principale
- afficher tous les commentaires du document XML contenu dans la fenêtre principale
- afficher des statistiques sur le document XML contenu dans la fenêtre principale

3.1.1.7 Menu SGML

- Module en cours de test.

3.1.1.8 Menu CONVERTERS

- Module en cours de test et non documentés.

3.1.1.9 Menu CORPUS

- Check Char

Cette option intègre des opérations de vérification et de remplacement de caractères (pris individuellement ou globalement).

1. Il est possible de visualiser tous les caractères du fichier.
2. On peut ensuite les modifier ou les supprimer interactivement et individuellement.

3. Il existe aussi une option qui permet d'appliquer des modifications sur tous les caractères du fichier sur la base d'une table de transcodage contenue dans le fichier nommée TableCharacter.txt. Il est possible de modifier cette table, chaque ligne est construite sous la forme suivante :

```
<caractèreInput><caractèreOutput><chiffre>
```

Le premier caractère sera remplacé par le second quand cette option est activée. Tous les caractères non contenus dans cette table seront remplacés par un blanc.

- PrepHtml2Txt

Ces items permet de transformer des pages HTML en fichiers texte.

Deux options sont disponibles :

1. la première permet de réaliser le transcodage sur un répertoire complet (de manière récursive). En sortie, on obtient un fichier réécrit pour chaque fichier lu dans l'arborescence parcourue et un fichier global concaténant l'ensemble des fichiers transcodés.
2. le seconde réalise l'opération de transcodage sur le fichier lu dans la fenêtre d'édition active.

Ces programmes ont été inspirés par ceux disponibles à l'adresse suivante : http://www.codearchive.com/home/jon/.

- Option Cordial : Make-Corpus Tag (For Lexico v1, v2)

Une opération disponible actuellement concerne le traitements des résultats issus de Cordial Analyseur. Il est possible de formater les résultats issus de Cordial et de les préparer pour être traitées par Lexico.

Le fichier à soumettre à Cordial doit être balisé sous la forme suivante :

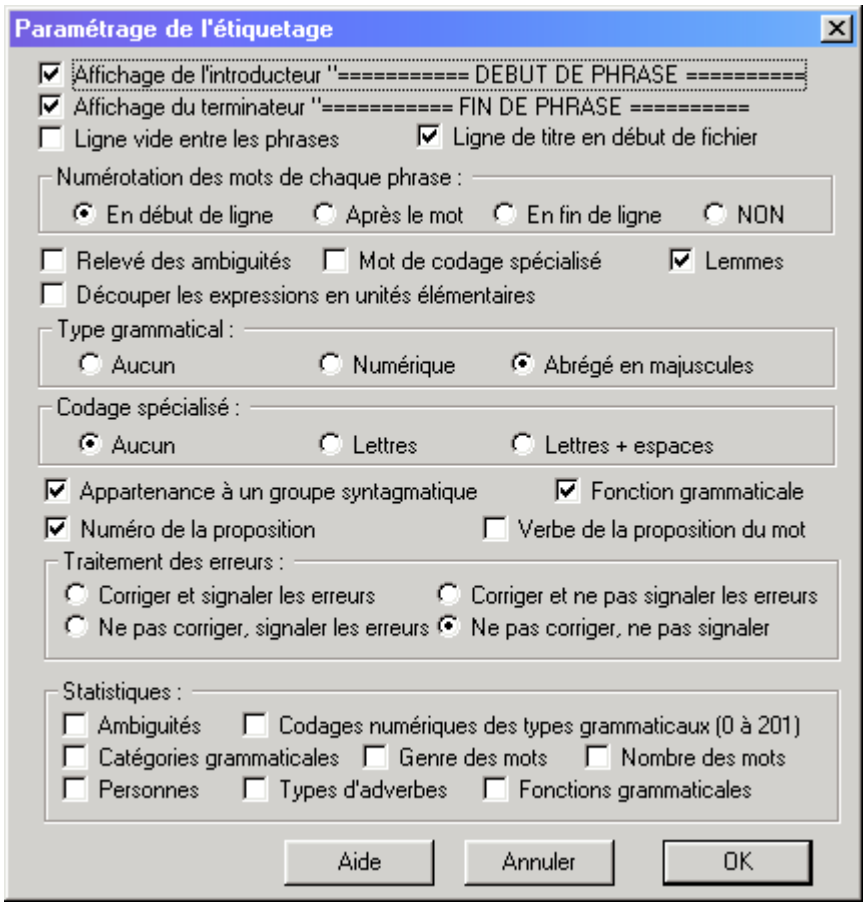
```
<balise1=valeur1>
zone textuelle 1
<balise2=valeur2>
zone textuelle 2
…
```

S'il y a des balises dans le texte, il faut qu'elles aient l'allure précédente, s'il n'y en a pas, aucun problème a priori. Le résultat créé par Cordial est ensuite reformaté pour Lexico.

Le résultat produit par l'étiquetage est de nouveau soumis à MKCORPUS qui se charge de remettre en forme ce résultat d'étiquetage pour qu'il soit de nouveau utilisable dans LEXICO. Ce travail de reformatage produit en fait 6 fichiers qui peuvent être tous traités par LEXICO :

- Un fichier contenant toutes les formes graphiques (i.e. le corpus initial)
- Un fichier contenant tous les lemmes
- Un fichier contenant toutes les étiquettes syntaxiques
- Un fichier contenant les couples (lemme, étiquette)
- Un fichier contenant les couples (forme, étiquette)
- Enfin, un fichier contenant la concaténation des trois premiers fichiers : i.e. une partition d'un même texte sous trois facettes complémentaires. On présentes un extrait de cet état du corpus ci-dessous :

Le paramétrage de cordial à utiliser est le suivant :



Les options "Cordial" de ce menu permettent ensuite de remettre en forme les résultats de Cordial. On obtient en fait plusieurs sorties : formes, lemmes, catégories, forme_catégorie, lemme_catégorie et une version du corpus qui contient une partition regroupant les lemmes, une partition regroupant les formes et une partition regroupant les catégories.

- Option MapXMLCorpus

Cette option s'inspire de la notion de "carte graphique de textes" disponibles dans Lexico3. L'enjeu est le suivant : il s'agit de donner une représentation graphique d'un document XML, sous la forme d'une carte de carrés colorés, sur la base d'une sélection d'un certain niveau de représentation (sélectionné par l'utilisateur) de ce document XML. On peut aussi considérer que cette option permet de se "promener" dans l'arbre XML du document visé en sélectionnant le niveau des nœuds de l'arbre à visualiser, on peut ensuite poursuivre la "descente de l'arbre" en sélectionnant un niveau de nœud plus profond. On donne une illustration de cette option dans les figures qui suivent.

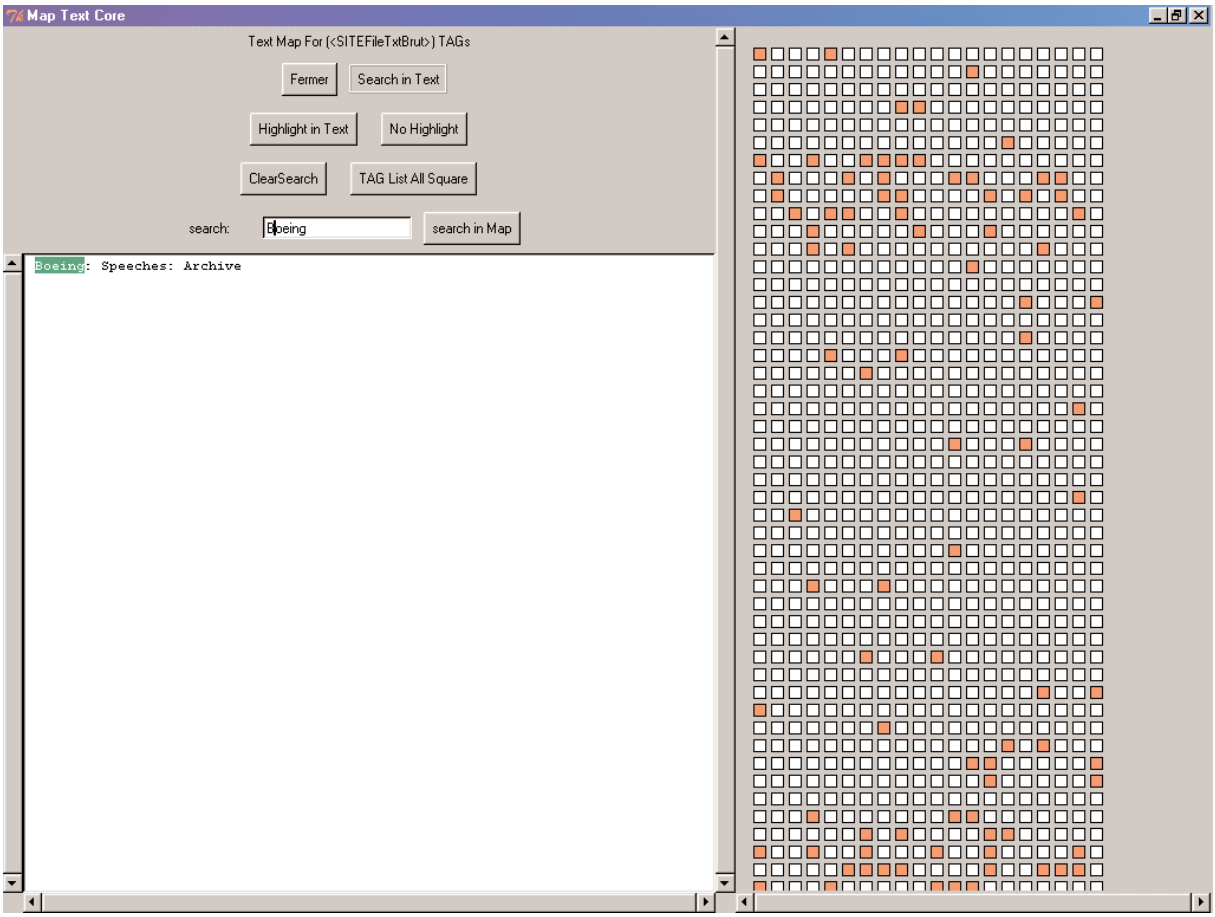
Dans cette figure, on suppose que MKCORPUS a chargé un corpus XML contenant des zones textuelles comprises entre les balises <SITEFILETXTBRUT> et </SITEFILETXTBRUT>. L'activation de l'option MapXMLCorpus déclenche la génération dans un menu déroulant de toutes les balises du corpus. On peut ensuite sélectionner la balise que l'on souhaite visualiser sous la forme d'une carte graphique. Dans notre exemple, la balise <SITEFILETXTBRUT> a été sélectionnée. MKCORPUS se charge ensuite de construire une carte de ces zones. Dans la figure, chaque carré construit correspond à une zone textuelle associée à la balise choisie (balises ouvrante et fermante). On peut ensuite réaliser différentes opérations :

(1) Rechercher des éléments textuels dans la carte ; pour cela, il convient de donner une chaîne de caractère dans la zone de recherche puis d'activer le bouton "search in map", les zones textuelles rouges contiennent la chaîne visée, les autres, en blanc, ne le contiennent pas. Dans notre figure, le mot Boeing a été mis en valeur dans la carte, les carrés rouges contiennent donc ce mot.

(2) Afficher le contenu textuel d'un carré de la carte ; en cliquant sur le bouton gauche de la souris au dessus du carré, le contenu textuel apparaît dans la zone d'édition.

(3) Mettre en valeur une chaîne de caractères dans la zone d'édition : mécanisme Highlight déjà vu plus haut. Dans notre figure, le mot Boeing a été mis en valeur dans la zone d'édition.

(4) Sélectionner une balise fils de l'un des carrés de la carte et la sélectionner pour générer une nouvelle carte : le clic droit sur un carré de la carte déclenche la génération de toutes les balises présentés dans cette zone textuelle. La sélection de balises peut ensuite déclencher la génération d'une nouvelle carte sur la base de cette nouvelle sélection.



3.1.1.10 Menu NLP

Ce menu regroupe différents programmes de manipulation pour les textes manipulés (documentation complète à venir) :

- Concordance
- Collocation
- Bigramme
- Fourgramme
- Wordcount
- xword

3.1.1.11 Menu TYPWEB

Les outils Typweb sont intégrés dans ce menu. Pour un descriptif de ces outils : http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/typweb.htm

Chaîne Typweb version 036

- *webxref*

On peut activer webxref sur un site aspiré localement et contenu dans un réperoire donné.
On peut activer webxref sur un fichier index.htm sauvegardé localement et contenu dans un répertoire donné. Il faut au préalable chargé le fichier dans la fenêtre principale.

- *mktypo*

On peut activer le programme mktipo présenté supra dans ce menu.

- *mkstat*

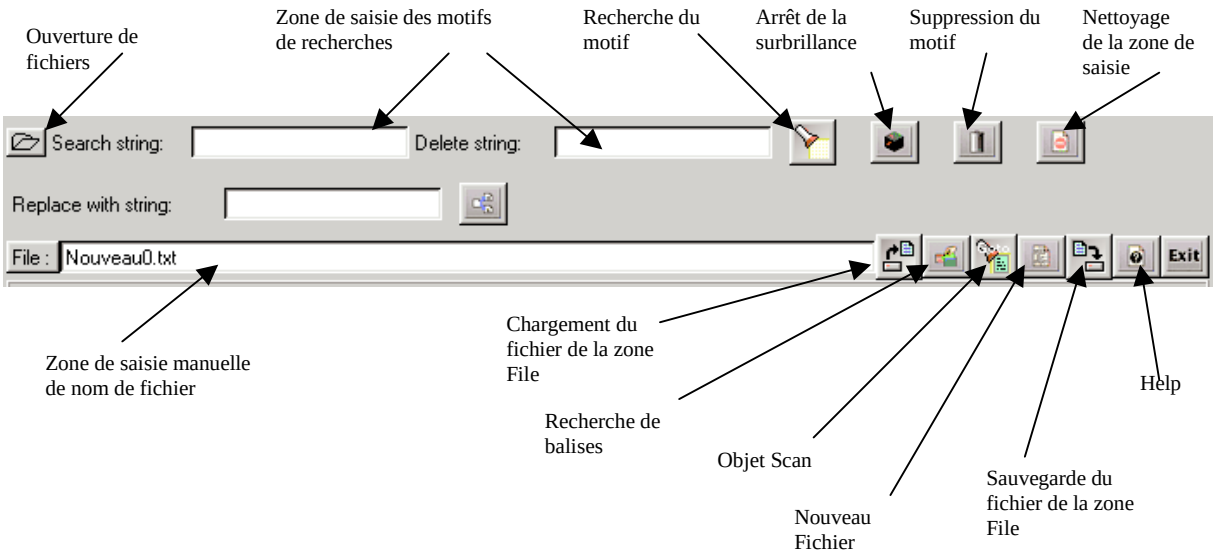
On peut activer le programme ExtAndStatFrCorpTwb présenté supra dans ce menu.

Chaîne Typweb version 037/38

- *webxref*

On peut activer webxref sur un site aspiré localement et contenu dans un réperoire donné.
On peut activer webxref sur un fichier index.htm sauvegardé localement et contenu dans un répertoire donné. Il faut au préalable chargé le fichier dans la fenêtre principale.

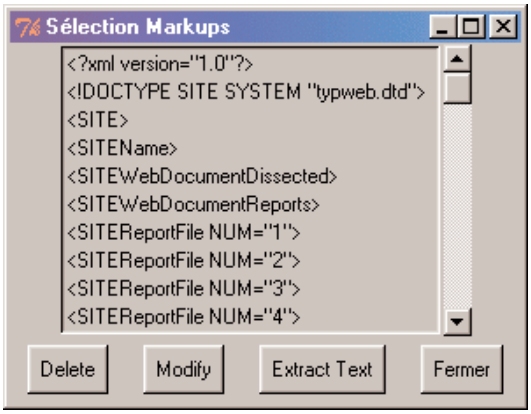
3.1.2 Les boutons



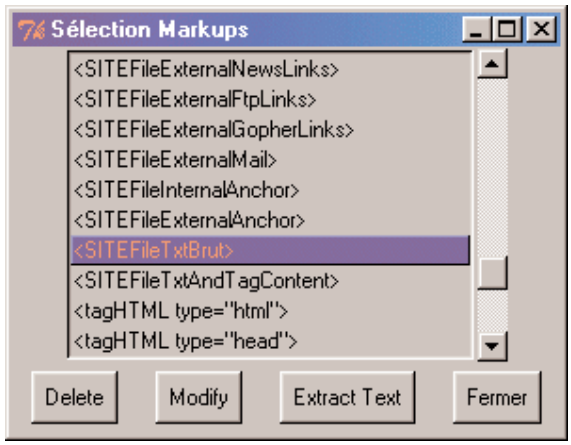
4 Traces graphiques

Dans les figures sui suivent on présentent certaines des opérations décrites précédemment.

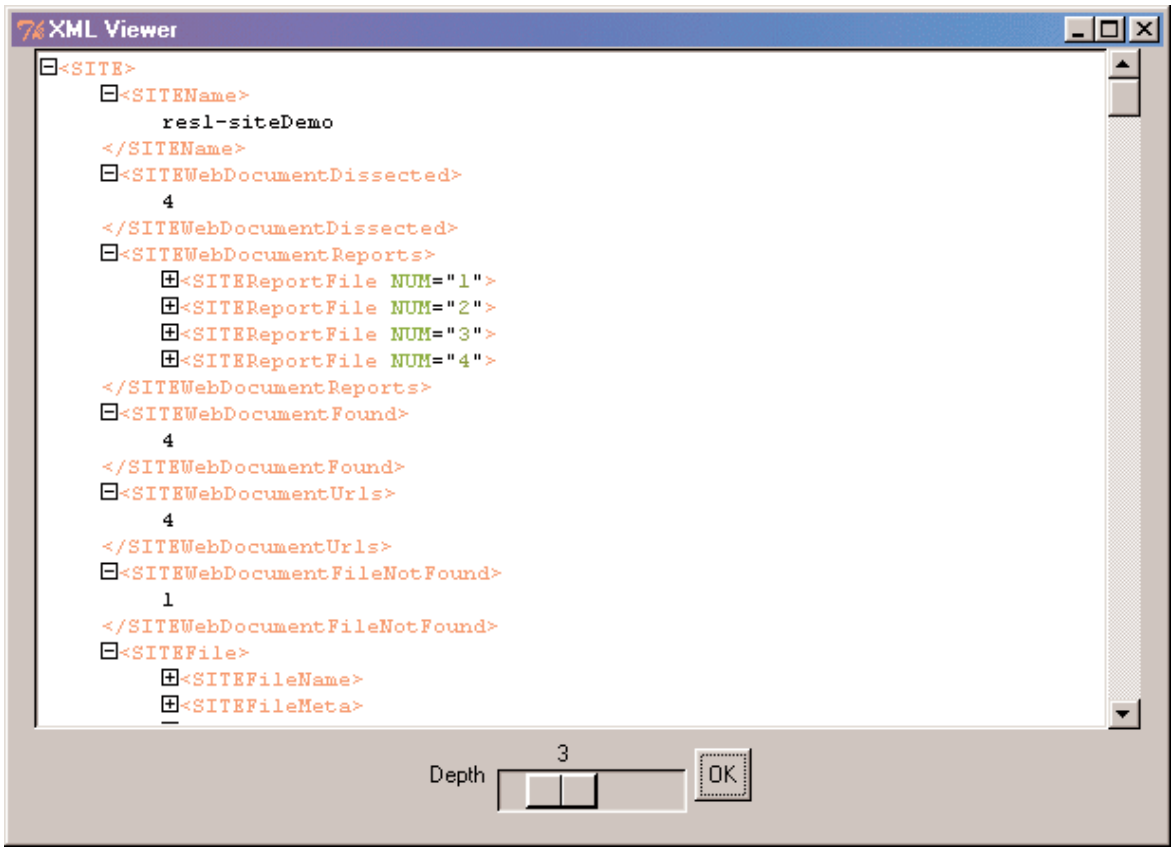
Dans la première figure, le programme de préparation a chargé le corpus démo et a produit une liste de toutes les balises de ce corpus :



On peut ensuite opérer certaines manipulations sur le corpus (supprimer les balises, les modifier, extraire le texte situé entre deux balises...). Dans la figure qui, on présente le texte brut contenu dans toutes les pages du site et extrait via l'activation du bouton d'extraction sur une balise sélectionnée.



Dans l'exemple ci-dessous, on a utilisé le menu "Balises <[^<]+>" pour réaliser le travail de détection de balises et d'extraction du contenu des balises sélectionnées. On peut aussi travailler avec le parser XML sur des documents XML valides. Dans la figure qui suit, on donne une présentation de cette démarche de travail sur le même corpus associé au site démo.



Dans cette première figure, on donne une représentation de l'arbre du document XML associé au site démo. Dans la seconde, on présente le résultat d'extraction du contenu des éléments structurels et textuels sur le même site.

5 Références bibliographiques

- Introduction à Perl/Tk, **Nancy Walsh**, O'Reilly
- Programmation en Perl, **L. Wall & al.** , Traduction française, 2ème édition, O'Reilly
- Perl cookbook, **Tom Christiansen & Nathan Torkington**, O'Reilly
- Perl Annotated Archives, **Martin Brown** , Ed. Osbourne/Mc Grawhill
- Perl 5 how-to, **Glover Mike, Humphreys, Ed Weiss,** The Wait Group, Inc.
- Web Programming with Perl, **Clinton Wrong**, O'Reilly
- Bien Débuter avec GNU Emacs, **Frédéric Pierrestéguy**, Masson