Un exemple de corrigé de l'exercice 1.1 du 15 novembre 2003

Recueil, constitution, test d'un corpus.

(Site du Forum Social Européen 2003)

Rappel de l'énoncé :

Recueil, constitution, test d'un corpus (HYPERBASE, LEXICO, WEBLEX)

1. Récupérer sur la Toile (site du Forum Social Européen 2003) les listes annonçant les séances plénières, les ateliers, les séminaires. Les constituer en un corpus lexicométrique, qu'on nommera FSE03.txt, visant à étudier la formulation et les répartitions thématiques des titres selon la variable de genre.

2. A l'aide de CORDIAL, caractériser les formes syntaxiques principales et les thèmes globaux du corpus.

3. A l'aide d'HYPERBASE, contraster les trois partitions du corpus

4. Établir une synthèse et une comparaison des deux approches.

Avant de soumettre le corpus à l'analyse statistique plusieurs étapes sont nécessaires :

- Recueil des données (Recherche sur Internet, sauvegarde locale)
- Filtrage (Isoler l'information pertinente)
- Mise en forme (Transformer le tableau en texte)
- Nettoyage (Correction orthographique, ponctuation)
- Formatage (Balisage, enregistrement au format txt)

Plusieurs procédures permettaient de parvenir au résultat final : Une solution parmi d'autres.

Pour chaque document (séminaires, ateliers, séances plénières) :

- **Étape 1** : Ouvrir la page sur internet, sélectionner le tableau entier à l'aide de la souris, le copier (menu édition) et le coller dans un nouveau document word.

OU

- **Étape 1 bis** : Enregistrer la page au format HTML sur le disque local. Ouvrir le fichier sous Word (clic droit sur le fichier, « ouvrir avec »).

Dans les deux cas nous obtenons un tableau Word dont il s'agit de ne conserver qu'une partie de l'information, les énoncés français. Pour sélectionner la colonne voulue on choisira d'utiliser le tableur Excel.

- **Étape 2** : Copier le tableau et le coller dans Excel.

- Procédure : Dans Word cliquer sur une colonne du tableau. Dans le menu « tableau » activer la commande « sélectionner le tableau ». Dans le menu édition activer la commande « copier ».

Ouvrir un nouveau document Excel et y coller le contenu du presse papier. (Menu édition, commande coller). Notre tableau est désormais inséré dans un classeur Excel. On ne conservera que la colonne « titres français » qui nous intéresse ici.

- **Étape 3** : Isoler la colonne « titres français ».

- Procédure : Toujours dans Excel, sélectionner la première colonne du tableau en cliquant sur l'entête de colonne. (A). Le contenu de la colonne doit être en surbrillance. Copier cette colonne (menu édition, copier) puis la coller à nouveau dans un nouveau document word.

On obtient ainsi un tableau Word sur une colonne qu'il s'agira de transformer en texte.

- **Étape 4** : Transformer le tableau en texte.

- Procédure : Dans Word, sélectionner le tableau (menu tableau « sélectionner le tableau »). Dans ce même menu, activer la commande « convertir le tableau en texte. Choisir l'option « séparer le texte par des tabulations ».

Chaque titre se trouve ainsi est séparé par un passage à la ligne.

Réitérer l'opération pour les deux autres parties du corpus.

Fusionner les trois rubriques en un unique fichier. On doit avoir à la suite les ateliers, séances plénières et séminaires.

- **Étape 5** : Nettoyage des données

Dans le document source au format HTML, la ponctuation n'est pas toujours respectée, ce qui peut poser problème pour l'analyse au moyen des logiciels.

Ponctuation : On peut par exemple remplacer tous les doubles passages à la ligne par un passage à la ligne simple, suivi d'un point puis d'un espace.

- Procédure : dans le menu « édition » de Word, activer « rechercher remplacer ». Rechercher **^p^p** remplacer par **^p.** (suivi d'un espace).

Suppression des passages à la ligne superflus.

- Procédure : rechercher **^p**. Dans la fenêtre « remplacer » on ne saisira aucun caractère.

Éliminer les doubles ponctuations fortes :

- Procédure : Rechercher : **..** remplacer par **.**
- Rechercher **?.** remplacer par **.**

Correction orthographique

- **Étape 6** : Balisage.

Pour le passage Hyperbase, chaque partie sera dotée d'une partition. Cette partition sera insérée au début de chaque texte, précédée et suivie d'un passage à la ligne, et notée comme suit :

&&&FSE=sem03&&&

&&&FSE=ate03&&&

&&&FSE=ple03&&&

- **Étape 7** : Enregistrement au format texte seul avec sauts de ligne.