Instructions élémentaires **pour** le recueil des textes et **la constitution de corpus lexicométriques** (HYPERBASE, LEXICO, WEBLEX)

1. *Saisie des données.* Réaliser un scannage des textes si ceux-ci n'existent pas sous un format informatique;

2. *Recueil des textes dans un format homogène.* Récupérer le résultat du scannage ou de la saisie (téléchargement de fichiers pdf ; exportation de CD-ROM) à l'intérieur d'un logiciel de traitement de textes, WORD, format .doc, de préférence;

3. *Corrections orthographiques.* Corriger les fautes de saisie, coquilles, fautes d'accentuation, grâce au correcteur orthographique; pour les langues à accents orthographiques autoriser les majuscules accentuées (Menu outils /options /édition /majuscules accentuées) puis sauvegarder

4. *Suppression des caractères et lignes inutiles.* En cas de téléchargement à partir de la Toile, le texte est souvent rempli de caractères graphiques et de sauts de ligne indésirables. Il faut les supprimer ou les remplacer par la fonction *Remplacer* de Word. La démarche à suivre pour supprimer les interlignes est la suivante :
- veiller à ce que les paragraphes soient séparés par une ligne blanche;
- aller dans la fonction word "Rechercher-Remplacer" dans le menu Edition;
- remplacer les doubles sauts de lignes (code suivant : ^p^p) par $$;
- remplacer les simples sauts de lignes (^p) par un espace;
- remplacer $$ par ^p. Votre texte normalement est désormais "propre".
(On peut faire précéder les paragraphes du signe §, balise utile pour reconnaître les paragraphes)

5. *Fusion des textes en un seul fichier et partition du corpus.* Rassembler tous les textes en un seul fichier ; à l'intérieur du fichier, séparer chaque texte différent par une clé du type:

-Pour LEXICO (ou WEBLEX)

<loc= CGT1> : à gauche du signe = : le nom de la clé, trois lettres qui rappellent la nature de la partition; à droite du signe = : trois lettres qui rappellent l'émetteur du texte ou le thème de travail et un chiffre qui numérote le nombre de textes différents qui formeront la base, classés par exemple par ordre chronologique.

-Pour la version HYPERBASE MAC la clé est de la forme
A la ligne
$$$gov=3D01$$$
A la ligne
Après le dernier texte du fichier, aller à la ligne et taper
&

-Pour la version HYPERBASE WIN la clé sera de la forme
A la ligne
&&&gov=3D01&&&
A la ligne

6. *Homogénéisation casse, police, lignes.* Sélectionner tout le texte puis

-aller dans le menu "format", sélectionner la commande "changer la casse", sélectionner la sous-commande "tout en minuscule";

-Changer la police de caractères, choisir la police "Courrier";

-Réduire la marge du texte pour que celui-ci ait des lignes de 65 caractères environ;

7. *Sauvegarde finale.* Sauvegarder le fichier en un format "texte seul (ou brut) avec saut de ligne" avec nom bref rappelant le contenu du corpus : exemple : CGT.txt

8. *Réalisation d'une note documentaire.* Deux parties

- l'une explique le choix des textes, les hypothèses de travail, les sources, les caractèristiques extralinguistique

- l'autre recense les références de chacun des textes de la base et le sens des abréviations utilisées dans les clés (locuteur du discours, récepteur du discours, lieu, date, références NB. Cette note explicative comportant les méta information sur le corpus, ses conditions de saisie de recueil et de constitution peut être stockée sous hyperbase dans un fichier biblio, au format txt.