

## Recommandations pour la constitution de corpus lexicométriques

*Partie commune à l'ensemble des logiciels, document source*

1. Recueil des textes. Fusionner dans un fichier unique
2. Éliminer tous les caractères utilisés pour la segmentation (&, >, <, &, \$)
3. Vérifier l'orthographe. Accentuer toutes les majuscules et veiller à ce que les prépositions de début de phrase soient correctement orthographiées (à, À).
4. Enregistrer au format rtf

A voir: traitement de la méta-information, tours de paroles... Faut-il ou non conserver les questions pour un entretien, comment traiter les titres pour une série d'articles.....

*Pour Lexico3*

Insérer en début de paragraphe les balises matérialisant les partitions (contrastives ou chronologiques).

Exemple:

```
<locuteur=chirac>  
<année=1995>  
<source=lemonde>  
<genre=entretien>
```

Distinguer les noms propres si nécessaire (en les préfixant par exemple), traiter ou non les mots composés.

Changer la casse: tout en minuscules.

Enregistrer au format txt.

*Pour Alceste*

A partir du fichier initial en rtf:(Document source)

Conserver la casse originale

Conserver ou non la méta information

Insérer des variables sous forme de mots étoilés, sur une ligne

Exemple:

```
**** □ *loc_chirac □ *année_1995 □ *source_lemonde
```

où □ symbolise l'espace.

Enregistrer au format txt avec sauts de lignes.

• • •