Préparation du corpus

1. Saisie

Vous l'effectuez - par frappe kilométrique ou au scanner - à partir d'un traitement de texte quelconque, pourvu qu'il ait une sauvegarde en mode texte. La présentation n'importe pas mais vous devez conserver la ponctuation, qui sera prise en compte.
Faites l'enregistrement dans un fichier unique pour l'ensemble du corpus à traiter, et n'oubliez pas d'effectuer la sauvegarde en mode "Texte seulement".

Si les paragraphes du document dépassent une vingtaine de lignes, nous vous conseillons de sauvegarder avec l'option "Texte seulement avec sauts de ligne".

Ceci fait, vous allez devoir effectuer un petit travail de "toilettage" de votre document afin qu'il soit conforme au formatage ALCESTE.

2. Majuscules

Par défaut, on utilise la règle de conversion suivante : La majuscule des mots en début de phrase est automatiquement transformée en minuscule. Par contre, les sigles ne le sont pas. Un mot retranscrit complètement en majuscules reste inchangé. Ces mots en majuscules sont placés dans une catégorie à part qui n'est pas prise en compte dans l'analyse.

3. Etoile (*)

L'étoile est un symbole réservé du logiciel ALCESTE. Il va jouer un rôle particulier de marquage. Vous devez donc dans un premier temps le faire disparaître complètement du document. Nous vous conseillons de le remplacer par un autre symbole, par exemple :

```
Marquise de ***
           * née en 1890
           * à Paris*
```

Devient

```
Marquise de XXX
           ◆ née en 1890
           ◆ à Paris1
```

4. Tiret haut (-) et tiret bas (_)

Le tiret haut est réservé par ALCESTE pour identifier les locutions. Vous n'avez pas besoin alors de vous en préoccuper. Mais si le logiciel ne reconnaît pas cette locution dans son dictionnaire des locutions, il supprimera le tiret haut et considèrera la locution comme deux mots. Ainsi, si vous voulez garder la forme composée d'un mot, ou bien l'imposer, vous remplacerez le tiret haut par le tiret bas. Vous pouvez aussi introduire cette forme dans le dictionnaire des locutions nommé ALC_LOC.

Exemples :

Si "savoir-faire" n'est pas dans le dictionnaire des locutions (ALC_LOC), on l'écrira "savoir_faire". Si vous voulez que "Général Boulanger" ou "Parti Radical" ou "Acte III, Scène 5" ou "Cat. soc. cult. 2" soient reconnus comme un seul mot, vous les écrirez alors : "Général_Boulanger", "Parti_Radical", "Acte_III_scène_5", "cat_soc_cult_2". Cependant si un couple (comme Parti Radical) apparaît plusieurs fois, ALCESTE vous en indiquera la fréquence (cf. : le dictionnaire des segments répétés).

5. Apostrophe

Bien sûr, dans le cas général, ALCESTE la reconnaît et vous n'avez pas besoin de vous en préoccuper. Mais attention au rôle particulier qu'elle peut jouer dans certains textes en transcription phonétique :
" Sur le boul'vard, déval' les loubards".
Il faudra écrire "boulevard" si on veut que ce mot soit reconnu comme tel, sinon "boul_vard".
Par contre, l'apostrophe de : déval', n'a pas besoin d'être retranscrite, ALCESTE fera lui même la séparation entre ce mot et le suivant.

6. Mots étoilés et lignes étoilées

Voici une rubrique essentielle parce qu'elle va vous permettre de "marquer" les mots qui vous sont indispensables en tant que repère ou comme information, mais que vous ne voulez pas faire intervenir dans l'analyse (en général simplement parce qu'ils ne figurent pas réellement dans le corpus étudié).

Généralement un corpus est composé de différents textes, chaque texte ayant sa spécificité de production : réponses à une question ouverte, chapitre d'un livre, etc.

Une ligne étoilée est précédée de quatre étoiles, ou d'un nombre entre 4 et 8 chiffres, par exemple :

**** *Sexe_m *Age_18 *Ville_Paris
ou bien
0001 *Sexe_m *Age_18 *Ville_Paris

Une ligne étoilée ne doit pas dépasser 255 caractères.
Les lignes étoilées permettent de séparer et reconnaître les énoncés naturels du corpus. Ainsi, par exemple, dans une question ouverte, chaque réponse sera précédée des informations concernant l'interlocuteur (âge, sexe, profession, ...) que nous appelons mots étoilés. Ces informations que l'on retrouve dans les résultats feront l'objet de questionnement, mais ne sont pas prise en compte dans l'analyse.

Les mots étoilés, d'une longueur maximum de 18 caractères, commencent par le symbole * (ex : *Sexe_m) et sont placés sur une ligne étoilée.

Par exemple :

```
**** *rep_3 *sex_masc *gr_soc_cult_2
J'ai profité de l'aide des pouvoirs_publics pour faire isoler ma maison et c'est à ce moment-là que j'ai choisi le tout_électrique...
```

Si ALCESTE rencontre un tiret bas (ex : pouvoirs_publics), il considère la locution comme un seul mot. En revanche, s'il rencontre un tiret haut (ex : moment-là), il vérifie si ce mot composé figure dans le dictionnaire des locutions. Cette locution sera scindée en deux si elle est absente du dictionnaire.