LEXICO

Lexico est un outil puissant pour le traitement lexicometrique de grands corpus basé sur l'analyse statistique des séquences. Ce programme est développé par l'équipe LEXICO à l'université de la Sorbonne nouvelle (Paris 3) dirigée par André Salem (http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/). Il permet le calcul de fréquences, du pourcentage (taux) de réalisation des formes, donne une classification hiérarchique des termes (par fréquence, ordre alphabétique, rang).

Le répertoire où se trouvent les exécutables est sur winnt : z:\taln\lexical\lexico2. Utilisez l'aide en ligne ainsi que le fichier fbrochure.html que vous trouvez sur le site web de lexico.

On dispose de quatre outils :

1. la segmentation
2. la recherche contextuelle
3. la comparaison par parties
4. la comparaison chronologique

Manipulations

1. Préparation du texte

Choisissez un petit texte .txt ; ouvrez-le à partir de démarrer⇒programmes⇒accessoires : word pad.

L'analyse statistique d'un texte se base sur l'unité textuelle qu'est la forme pour estimer ses occurrences, délimiter les séquences de mots où elle apparaît et comparer ses fréquences dans une partie donnée du corpus par rapport à une ou plusieurs autres. Pour rendre possible et fiable cette segmentation il est nécessaire de définir des normes de saisie qui assurent la cohérence du texte et de son codage.

Le texte ne doit comporter que des minuscules ; pour désambigüiser les noms propres et noms communs homographes on accolera au nom propre un signe de distinction (par exemple : `~barre`). Toutefois, ce signe ne doit alors pas figurer dans la liste des délimiteurs. ***Ponctuations, délimiteurs*** : afin que le programme d'analyse LEXICO2 puisse reconnaître les formes du corpus que l'on segmente, on affecte un statut à chaque signe présent dans le texte. On distinguera d'une part les délimiteurs : `--_:;/.,?¿!¡*$"+=(){}` ainsi que l'espace. (Cette liste est donnée à titre indicative et peut être modifiée). D'autre part, sont appelés non-délimiteurs tous les caractères restants contenus dans la police : `a, b, c,...` Toute suite de caractères non-délimiteurs bornée à ses deux extrémités par des délimiteurs est considérée comme une occurrence, une forme à relever et à classer. On pourra également insérer des ***clés*** et des marques de ***paragraphes*** si nécessaire.

Enregistrer les modifications opérées et fermer le texte.

2. Segmentation

Lancer l'exécutable segment.exe ; ouvrir le texte que vous venez de modifier. Vérifier les délimiteurs. lancer le traitement.

Pour un rapport du traitement effectué ouvrir via word pad le fichier atrace.txt

Les fichiers .dic et .par on été crées dans votre répertoire ; ouvrir le fichier .dic (résultat du calcul de fréquence des formes) via word pad.

3. Recherches contextuelles

Lancer le programme de concordance context.exe. Explorer les nombreuses options proposées : les tris possibles, les multiples requêtes, etc. utilisez le menu contextuel de la souris en cliquant sur les mots trouvés.

4. Comparaison par parties

Lancer le programme parties.exe et utiliser le fichier démo DUTCH.par

Quels résultats obtenez-vous ?

5. Comparaison chronologique

Lancer chrono.exe et utiliser le fichier démo DUTCH.par. Quelles fenêtres apparaissent, quels fichiers sont créés ? a quoi servent-ils ?