# CORPUS

## Quelques choix terminologiques

Le mot *corpus* signifie selon J. Singlair (1996) : « un corpus est une collection de données qui sont sélectionnées et organisées selon les critères linguistiques explicites pour servir d'échantillon du langage ».
On emprunte au québecois le terme *parsage* (parsing) pour désigner l'analyse syntaxique automatique et le mot *parseur* (parser) pour le programme qui effectue cette opération.
En recherche documentaire, la *précision* représente la proportion de réponses pertinentes données par rapport au total des réponses extraites. Le *rappel* est la proportion des réponses pertinentes extraites par rapport au total des réponses pertinentes possibles. Le *silence* correspond alors aux réponses pertinentes non extraites. Le *bruit* renvoie aux informations non pertinentes produites.

## Constitution d'un corpus

Pour le projet demandé pour ce cours, la constitution de corpus se fera par une grande collection de documents électroniques qui respecteront un regroupement adéquat à l'objectif défini par chacun. Beaucoup d'écrits professionnels existent sous forme éléctronique et le « captage » de textes est désormais aisé. Il faut préciser que des techniques d'échantillonage peuvent amener à briser la séquentialité des textes de départ: on peut extraire des fragments en plusieurs endroits d'un même texte pour éviter de sur-représenter ou sous-représenter certaines caractéristiques. J. Singlair ajoute : « un corpus est supposé contenir un grand nombre de mots. L'objectif fondamental de la constitution d'un corpus est le rassemblement de données en grandes quantités ».
Pour ce faire, il faut trouver un ensemble de mots qui définit de manière non exhaustive le périmètre de votre thématique sémantiquement. Ensuite, vous récupérez les textes (ou les parties de documents) qui correspondent le plus à votre thème, même si les documents ne sont pas pris de façon intégrale.

## Post-traitement

L'ensemble de textes ainsi recueillis seront mis dans un nouveau fichier texte qui sera « nettoyé » de balises de forme ou de contenu, (le plus souventr ce nettoyage a lieu « manuellement »). Vous allez ensuite « parser » l'ensemble de textes de façon à ajouter les balises aux endroits où l'information est pertinente pour le résultat que vous voulez avoir. Nous verrons pour chacun individuellement quel type d'information peut être intéressant d'extraire et donc de baliser au préalable. Les balises seront faites en XML en seront mises en place automatiquement grâce au parseur. Ensuite vous pouvez afficher soit des extraits pertinents suivant une information au l'ensemble de textes ainsi recueillis.