Projet de STATISTIQUE LEXICALE (Licence Informatique)

Gilles HUNAULT

Angers, 1997

1. Description du projet

Le projet a deux buts : faire programmer des comptages lexicaux simples pour que des non-linguistes découvrent les caractéristiques de certains textes, certaines langues et fournir aux linguistes un outil simple cumulatif. Le logiciel proposé devra en effet être capable de fournir des statistiques sur le texte proposé mais aussi de mettre à jour une base de données de comptages pour cumuler l'analyse du texte à celles déjà effectuées.

2. Corpus proposé

Le projet devra traiter 5 langues (français, anglais, allemand, espagnol, italien). Chaque langue sera analysée sur trois registres (texte littéraire, article journalistique, document technique). Le volume devra être d'un million de signes par langue, réparti équitablement entre les trois registres.

3. Traitement de lettres

Des tableaux de comptages d'unigrammes (1 lettre), bigrammes (2 lettres) et trigrammes (3 lettres) seront fournis. Les tableaux seront exhaustifs seulement pour les unigrammes. Pour les bigrammes et trigrammes on ne fournira que les plus significatifs (et les moins significatifs s'il y en a peu). Les lettres "7 bits" seront séparées des lettres accentuées. Les chiffres seront également présentés dans un tableau séparé, de même que tous les autres symboles de ponctuation. On traitera les unigrammes initiaux et finaux c'est à dire les lettres de début et de fin de mot à part.

4. Traitement de mots

Les structures traditionnelles de texte (paragraphe, phrase, mot) seront recensés et comptabilisés avec des indicateurs statistiques simples (nombre et longueur : moyenne, écart-type, coefficient de variation, mininum et maximum). On fournira aussi des dictionnaires par ordre alphabétique et par ordre décroissant de fréquence.

Aucun traitement de lemmatisation, syllabisation ou synthémisation ne sera effectué. Les lettres et mots seront pris "tels quels". Aucun indicateur de qualité, de lisibilité (style Flesch) ne sera calculé. Un projet plus "linguistique" pourra être envisagé dans un deuxième temps.

5. Disponibilité des résultats

Le contenu des bases de données des analyses sera disponible sur le Web. Il est envisageable qu'à terme un utilisateur quelconque d'Internet fasse analyser un de ses textes par le logiciel.