Séminaire doctoral Informatique textuelle

Exercice 1.2 : Préparation du corpus Bush Kerry pour traitement lexico

- Télécharger le corpus Bush Kerry sur Textopol Dynamique. (Recherche par mots du titre uniquement).

En prenant appui sur le protocole de saisie (séance 1, documents), convertir le corpus BK6.txt actuellement encodé pour Hyperbase aux formats Lexico puis Alceste.

On utilisera les fonctions rechercher/remplacer de MS Word en cherchant à effectuer ces conversions de la façon la plus automatisée possible.

On rappelle que :
- la casse originale peut être conservée pour Alceste, ainsi que pour Hyperbase
- qu'il est préférable de convertir le texte en minuscules sous Lexico.

✓ Conversion du balisage Hyperbase vers Lexico

- Ouvrir le document BK6.txt sous Word.
Ainsi que la notice détaillée le précise (cf base dynamique) le corpus est muni de partitions qui matérialisent les trois interventions de chaque locuteur. (Bush1, Bush2, Bush3, Kerry1, Kerry2, Kerry3).

Le balisage attendu sous Lexico est du type <loc=Bush1> (pas d'espace dans les clés)

Le balisage actuel est du type &&&loc=BUSH1&&&
On pourrait commencer par remplacer les caractères de fin « &&& » par une balise fermante (>). Puis remplacer le motif **&&&loc** par **<loc**

Cette manipulation a surtout pour objectif de montrer que la conversion des textes n'est pas une chose triviale. On attire aussi l'attention sur le fait que les régularités structurelles, ou de mise en page d'un document permettent bien souvent de mettre en œuvre des procédures automatisées qui peuvent être un gain de temps considérable sur des corpus dotés de partitions nombreuses.

✓ Nettoyer le corpus

Les tours de parole sont matérialisés par **$** suivi d'un nombre à deux chiffres.
Il peut être utile de conserver cette information sous lexico afin de marquer les différentes sections. (Cf séance de janvier)

Toujours sou word, on recherchera alors le motif suivant : **$** suivi de deux chiffres quelconques que l'on remplacera par **$** suivi d'un espace.

✓ Homogénéiser la graphie

Modifier la casse et convertir l'ensemble du texte en minuscules. Il faut bien souvent autoriser les majuscules accentuées, lancer une correction orthographique avant de procéder à la modification de la casse. (Cf. protocole de saisie)

Enregistrer le document sous un nom différent, au format texte seul, avec sauts de lignes.

✓ Pour aller plus loin…Enrichir le balisage

Réfléchir à une procédure qui permettrait de passer d'un balisage en textes à une partition plus complète. Si sous Hyperbase, un état du corpus doit correspondre à une partition et une seule, sous lexico, plusieurs divisions (locuteur, date, genre, source…) peuvent co-exister.

On pourrait envisager une partition binaire qui permettrait de confronter le lexique des locuteurs Bush et Kerry.
<locuteur=Bush>
<locuteur=Kerry>

Autre possibilité une partition matérialisant les débats :
<Débat=1>
<Débat=2>
<Débat=3>

On obtiendrait alors trois niveaux de partition : texte, locuteur, débat et passer de l'une à l'autre selon les éclairages à porter sur le corpus.

✓ Préparation du corpus pour Alceste

Reprendre l'état original du corpus et transformer le balisage afin d'obtenir les motifs suivants

****_*Loc=Bush1
****_*Loc=Bush2
****_*Loc=Bush3
****_*Loc=Kerry1
****_*Loc=Kerry2
****_*Loc=Kerry3

(Où _ matérialise l'espace)

L'état du corpus Alceste sera enregistré au format texte seul avec sauts de lignes.

• • •