Transformation, conversion calibrage de données sous Word

Soit le fichier Document source.rtf partiellement balisé pour Hyperbase.

• 1. Transformer la balise locuteur de telle sorte que le corpus soit exploitable par Lexico. Utiliser pour ce faire les fonctions de recherche et de remplacement de Word.

Rappel : Les clés utilisées par Lexico doivent respecter la syntaxe suivante :

<typeclé=nomclé>
Exemple : <locuteur=ferry> ou <date=2002/06/12>
- Aucun espace entre les chevrons. Les intitulés des types de clés doivent être rigoureusement identiques. Consulter le fichier « atrace.txt » en cas de problème.

• 2. Compter les balises locuteur.

• 3. Insérer une balise date en utilisant les indications notées sur la ligne suivant la balise locuteur. Exemple : **LCI, entretien du 12/06/2002**

Cette balise devra être précédée d'un saut de ligne.

• 4.Modifier la mise en forme des balises afin qu'elles apparaissent en caractères non gras.

• 5.Supprimer les commentaires, méta info et interventions des journalistes qui ne doivent pas être prises en compte lors de la segmentation. Cette opération pourra s'effectuer en deux temps.
• 6.Sauvegarder le fichier au format texte seul dans le répertoire de travail de Lexico.

Toutes ces manipulations sont à effectuer au moyen de la fonction « rechercher/remplacer » de Word. Certains cas de figure nécessitent d'utiliser les caractères génériques et expressions régulières ainsi que les caractères spéciaux.

Pour utiliser un opérateur comme simple caractère, on le fera précéder d'un anti-slash (\). Par exemple pour rechercher une balise ouvrante on saisira « \< »dans la Zone rechercher

- Quelques exemples d'opérateurs disponibles sous word :

Le passage à la ligne se note « ^p »
Pour rechercher n'importe quel caractère : ?
Pour rechercher de 1 à n caractères quelconques : ?*
Pour rechercher un chiffre : [0-9]
Pour rechercher une lettre : [A-Z]
Pour rechercher exactement n occurrences d'un chiffre : [0-9]{n}
Au moins n occurrences d'un chiffre : [0-9]{n ;}

…