Notions de base sur le texte, le document, leur transformation et leur analyse

Texte

- **Mot** : séquence comprise entre deux blancs, ou entre un **séparateur** et un blanc ou entre un blanc et un séparateur. Le blanc est une **espace** en typographie.
- Un texte se compte en mots, mais aussi en caractères, en lignes et en paragraphes
- **Séparateur** : ponctuations de toutes catégories
- **Occurrences** : un mot peut avoir des occurrences multiples.
- **Tokens** : occurrences de mots mais aussi de signes autres qu’on isole : ponctuation, etc.

Texte (2)

- Si on ne compte le mot qu'une fois, on ne compte que les **vocables** : dont l'ensemble constitue le **vocabulaire** d'un texte.
- Par **indexation** du texte : on obtient les mots du texte qui sont en fait des formes de mots (*chante, chantes, chantons, chantez, chantent, chanta, chantèrent, chantions, chanterions, chantiez*)
- Par **lemmatisation**, on réduit aux seuls **vocables**.
-

Lexique

- L'ensemble des **vocables** identifiables dans une langue : le **lexique** d'une langue
- constitué des **entrées** du dictionnaire

En lexicographie :

- Une **entrée** de dictionnaire comporte au moins trois choses :
- La **vedette**, la **partie du discours** (*pos*, syntaxe), les **acceptions** (sémantique)
- La liste des vedettes, c'est la **nomenclature** d'un dictionnaire.
- Autres types de dictionnaire. Voir le *Lexique juridique du droit fédéral canadien*.

Traitement de texte

- Distinction entre **texte** : séquence de caractères, **texte seul** et **texte seul codé** sont des formats de fichiers électroniques
- Et **document** : c'est du texte **mis en page** (format de page, en-tête, pied de page) et **mis en forme** (paragraphage, **styles**, **accidents** typographiques : **gras**, *italiques*, PETITES CAPITALES, ~~barré~~, corps de caractères, etc.)

Linguistique de corpus

- Texte, **collections** de textes, **archive**,
- **corpus** : corpus homogène (auteur, époque), ou plus ou moins homogène.
- corpus **de référence** (représentatif d'un état de langue, pour référence utile à la description linguistique, à la normalisation linguistique, et à la traduction)
- corpus massifs : 1Mwords – 1 billion words : richesse lexicale, constructionnelle, mémoires de traduction.

Texte et contexte

Qui dit texte dit **contexte** :

- Contexte à gauche et contexte à droite, que fait apparaître un **concordancier**.
- La concordance est fondée sur un **mot clef,** qu'on appelle en anglais *key word in context* (KWIC)
- Les concordanciers sont aussi **indexeurs**, parce qu'ils permettent de construire l'index de toutes les formes de mots d'un texte (*wordlist*)

Contexte, collocation, etc.

L’examen des formes de mots en contexte permet de dégager :

- Des **collocations**
- Des **colligations**
- Des **segments répétés** (*lexical bundles*)
- Des **expressions polylexicales** (*multiword expressions*)
- Des **idiomes**, figements de collocations