| *Terme* | *Définition* |
|---|---|
| ***Analyse Factorielle*** | C'est un terme qui désigne plusieurs méthodes d'analyses concernant les grands tableaux de données, visant à extraire et déterminer des facteurs qui résument l'ensemble des informations contenues dans le tableau de départ. |
| ***Annotations morphosyntaxiques*** | L'annotation ou bien l'étiquetage morphosyntaxique est le processus qui consiste à associer aux mots d'un texte ou bien d'un corpus les informations grammaticales correspondantes comme la partie du discours, le genre, le nombre…etc. à l'aide d'un outil informatique. |
| ***Automatisation*** | Comme son nom l'indique, le TAL a pour objectif l'automatisation, c'est-à-dire l'élimination de la part de l'humain dans les traitements linguistiques. Autrement dit, l'utilisation de différents logiciels et outils qui ont pour but de traiter la langue automatiquement. |
| ***Cadre*** | Le cadre est relié au repérage de différents parties d'un texte ou bien d'un corpus donné. |
| ***Concordancier*** | Il affiche le contexte d'utilisation du mot dans le corpus. |
| ***Caractère*** | C'est un signe typographique utilisé pour l'encodage du texte sur un support lisible par l'ordinateur. |
| ***Cooccurrence*** | C'est la présence simultanée de deux ou de plusieurs mots (ou autres unités linguistiques) dans le même énoncé (la phrase, le paragraphe, l'extrait). Elle désigne l'apparition de deux mots en même temps et dans le même contexte. |
| ***Cooccurrent*** | Les cooccurrents sont les mots qui viennent souvent autour d'un mot choisi. |
| ***CQL*** | Corpus Query Language a pour objectif de définir des conditions spécifiques pour le mot ou l'expression recherchée. Autrement dit, donner des conditions à respecter quand on veut chercher un mot spécifique dans un corpus. Par exemple, si on veut ne chercher que les occurrences nominales, on doit taper une requête spécifique. |
| ***CQP*** | C'est l'acronyme de Corpus Query Processor, et c'est un module logiciel qui traite des requêtes. Autrement dit, c'est un moteur de recherche qui permet de trouver toutes les occurrences correspondant à une équation CQL dans un corpus donné. |
| ***Corpus*** | C'est un ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène dans plusieurs domaines notamment : études littéraires, linguistiques, scientifiques et philosophiques, etc.<br>Dans le domaine de la lexicométrie, le corpus désigne un ensemble de textes réunis à des fins de comparaison ; servant de base à une études quantitative. |
| ***DELA*** | Renvoie aux dictionnaires électroniques utilisés par UNITEX. Autrement dit, DELA est un formalisme qui existe sur UNITEX afin de pouvoir créer des dictionnaires électroniques. |
| ***Fichier binaire*** | C'est un fichier qui n'est pas un fichier texte. Il existe de nombreux formats notamment : les images, les vidéos, les sons ou bien les fichiers compressés. |
| ***Fichier texte brut*** | Un fichier texte brut est un fichier dont le contenu représente uniquement une suite de caractères. Ces caractères sont généralement considérés comme des caractères imprimables, d'espaces et des retours à la ligne. |
| ***Fréquence*** | Elle désigne le nombre des occurrences d'une unité textuelle dans un corpus donné. |
| ***Fréquence d'un segment*** | Elle désigne le nombre des occurrences de ce segment, dans l'ensemble du corpus. |
| ***Fréquence maximale*** | Elle désigne la fréquence de la forme la plus fréquente du corpus (en français, le plus souvent, la préposition "de"). |
| ***Hapax*** | C'est l'équivalent de "chose dite une seule fois". Elle désigne la forme dont la fréquence est égale à un dans le corpus (hapax du corpus) ou dans une de ses parties (hapax de la partie). |
| ***Identification*** | C'est la reconnaissance d'un seul et même élément à travers ses multiples emplois dans des contextes et dans des situations différentes. |
| ***Index*** | C'est une liste imprimée constituée à partir d'une réorganisation des formes et des occurrences d'un texte, ayant pour base la forme. |
| ***Lemmatisation*** | Action de lemmatiser, ça veut dire de donner à un mot la forme neutre canonique qu'il a. Par exemple, les entrées d'un dictionnaire. Elle vise à remplacer un mot par sa forme canonique. Par exemple, pour un verbe ; le lemme est la forme à l'infinitif. Pour un nom ou bien un adjectif, le lemme est la forme au masculin singulier. |

| *Terme* | *Définition* |
|---|---|
| ***Lemme*** | Une forme canonique (masculin, singulier, infinitif…etc.) d'un mot variable. |
| ***Lexicométrie*** | C'est un ensemble de méthodes permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire d'un corpus de textes. |
| ***Lexique*** | C'est un ensemble virtuel des mots d'une langue. |
| ***Longueur*** | La longueur d'un corpus, d'une partie de ce corpus, d'un fragment de texte, d'une tranche, d'un segment, etc. reliée au nombre des occurrences contenues dans ce corpus. Synonyme : taille. |
| ***Morphosyntaxique*** | Discipline qui regroupe l'étude des formes (morphologie) et celle des règles de combinaison des morphèmes (syntaxe), les considérant comme un tout indissociable. |
| ***Mot*** | Une suite de caractères séparées par un blanc (espace, ligne) et/ ou de la ponctuation, les apostrophes et les tirets. |
| ***Partie*** | Une partie d'un corpus de textes est reliée au fragment de texte correspondant aux divisions naturelles de ce corpus ou à un regroupement de ces dernières. |
| ***Partition*** | Il s'agit de la division d'un corpus de textes en parties constituées par des fragments de texte consécutifs, n'ayant pas d'intersection commune et dont la réunion est égale au corpus. |
| ***Progression*** | C'est une fonctionnalité de TXM. Elle nous permet d'évaluer la progression d'un mot dans un corpus. Par exemple, il est possible de voir la fréquence d' un mot dans le corpus sans faire recours à la fréquence de tous les mots. |
| ***Segment*** | Toute suite d'occurrences consécutives dans le corpus et non séparées par un séparateur de séquence. |
| ***Segmentation de données*** | Il s'agit d'organiser les données multiples et variées qu'on peut avoir et ceci dans le but de mieux les appréhender. |
| ***Séparateur de séquence*** | Un séparateur de séquence est relié à la ponctuation!. Les séparateurs les plus fréquents sont : la virgule, le point-virgule et les deux-points. |
| ***Token*** | Il s'agit d'une seule unité lexicale qui y a un sens particulier. |
| ***Tokenization*** | Il s'agit de séparer les propos en unités linguistiques (les tokens) en prenant en compte les ponctuations et toute autre marque qui montre la fin d'une phrase. On essaie ainsi de trouver l'unité minimal d'information et qui y a un sens particulier. |
| ***Trame*** | La Trame est reliée à la première segmentation d'un texte ou bien un corpus donné. Le texte + une segmentation = la Trame |
| ***Treetagger*** | C'est un outil qui permet d'annoter un texte avec des informations sur les parties du discours (genre de mots : noms, verbes, infinitifs et particules). Et des informations de lemmatisation. |
| ***Verbatim*** | C'est un terme venant du latin verbum « mot ». Ce terme veut dire « mot pour mot », « textuellement » ou « texto ». Autrement dit, verbatim désigne le compte rendu complet et fidèle rédigé par une partie dans une conversation diplomatique ou juridique à son seul usage mémoriel. |