**Traitement Automatique du Langage Naturel** : discipline étudiant l'ensemble des techniques permettant à une machine d'analyser et de comprendre le langage humain. C’est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle.

**Tokénisation** : Parsage (analyse syntaxique) d'un texte en tokens (entité ou unité lexicale), par exemple en mots et symboles. La tokenisation est l'opération de segmenter un acte langagier en unités "atomiques" : les tokens. Les tokenisations les plus courantes sont le découpage en mots ou bien en phrases.

**Racinisation** : En linguistique, la racinisation ou désuffixation est un procédé de transformation des flexions (formes “conjuguées” ou “accordées” d’un mot) en leur radical ou racine (sans préfixe ni suffixe).

**Un corpus** est un ensemble de documents, artistiques ou non (textes, images, vidéos, etc.), regroupés dans une optique précise. Utilisation possible de corpus pour une étude linguistique.

**Lemme :** C'est une chaîne de signes formant une unité sémantique et pouvant constituer une entrée de dictionnaire. Les signes constituants peuvent par exemple être réalisés par des caractères, des sons, des gestes. Dans le vocabulaire courant, on parlera plus souvent de « mot ».

Lematisation => ontologie Lemmatisation : en linguistique, opération qui consiste à regrouper les occurrences d'un texte sous des adresses lexicales

Thesaurus ou Thésaurus : Liste alphabétique de mots standards utilisés pour le classement de la documentation.

n-gramme : Un n-gramme est une sous-séquence de n éléments construite à partir d'une séquence donnée. Comme leur nom l’indique, les n-grams sont des suites d’un nombre donné de caractères (bigrammes, trigrammes, 4 grams, etc.).Pour prendre un exemple, avec des 5 grams, la phrase suivante : Le chat mange la souris [oui, l’originalité des exemples est notre péché mignon] donne : Le ch, e cha, \_chat, chat\_, etc.

Analyseur lexical : Elle consiste à convertir une chaîne de caractères en une liste de symboles (tokens en anglais). Ces symboles sont ensuite consommés lors de l'analyse syntaxique. Un programme réalisant une analyse lexicale est appelé un analyseur lexical, tokenizer1 ou lexer. Un analyseur lexical est généralement combiné à un analyseur syntaxique pour analyser la syntaxe d'un texte.

Ontologie : En informatique et en science de l'information, une ontologie est l'ensemble structuré des termes et concepts représentant le sens d’un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances. L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts.

Wordnet : Base de données lexicale. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Le WordNet libre du français (WOLF) est une ressource linguistique généraliste libre pour le français issue d'une traduction automatique de WordNet 3. EuroWordNet est un système de réseaux sémantiques pour les langues européennes, basé sur WordNet. [1] Chaque langue développe son propre réseau de mots, mais ils sont interconnectés avec des liens interlingues stockés dans l'index interlingual (ILI).

Analyse lexicale : Elle consiste à convertir une chaîne de caractères en une liste de symboles (tokens en anglais).

**Analyse lexicale** : Appréhender un texte par son lexique, c'est essayer de comprendre son contenu en suivant les traces laissées par les locuteurs.

Analyse syntaxique : Elle consiste à mettre en évidence la structure d'un texte, généralement une phrase écrite dans une langue naturelle.

Analyse sémantique : L'analyse sémantique d'un message est la phase de son analyse qui en établit la signification en utilisant le sens des éléments (mots) du texte.

Analyse pragmatique : Elle est l’étude des signes dans leurs rapports avec leurs utilisateurs.