**Définitions :**

**Traitement du langage naturel :** Discipline étudiant l'ensemble des techniques permettant à une machine d'analyser et de comprendre le langage humain.

**Tokenisation :**

1. En monétique, la tokenisation est le processus de substitution de données bancaires (numéro de cartes, …) par des données jetables appelées « jeton » (token en anglais)
2. (Informatique) [Parsage](https://fr.wiktionary.org/wiki/parsage" \o "parsage) d'un texte en [tokens](https://fr.wiktionary.org/wiki/token" \o "token), par exemple : en mots et symboles.

Les mots, phrases et paragraphes sont détectés et regroupés

1. (Cryptographie) Procédé permettant de remplacer une donnée critique par un élément équivalent appelé [token](https://fr.wiktionary.org/wiki/token" \o "token), et qui n’aura aucune valeur intrinsèque ou signification exploitable une fois sortie du système.

**Racinisation :** La racinisation (en anglais : stemming) est un procédé qui consiste à regrouper les mots ayant la même racine, c’est-à-dire des mots ayant une sémantique proche.  
La racinisation designe la méthode permettant d’identifier et de regrouper les différentes formes d’un même mot autour d’une racine (racine d’un mot également appelé « stemme » ou « stem » ou « base »).  
La racine d’un mot (ou le stemme ) est identifiée par la suppression du préfixe, le suffixe d’un mot.  
  
Exemple :   
Les mots suivants ont pour racine «OPTIM» : Optimiser, Optimal, Optimisation, Optimalisation

**Corpus :**

1. Recueil de documents relatifs à une discipline, réunis en vue de leur conservation.
2. Ensemble fini d'énoncés écrits ou enregistrés, constitué en vue de leur analyse linguistique.

**Lemmatisation et Lemme:** La **lemmatisation** (en anglais : lemmatization) est l’analyse lexicale d’un texte dans le but de regrouper les mots d’une même famille. Les mots d’une même famille détectés dans un texte sont donc réduits en une unique entité que l'on appelle un « **lemme** » ou la « forme canonique d’un mot ».  
La lemmatisation consiste donc à regrouper ensemble toutes les formes que peut prendre un mot unique. Ses formes peuvent être le pluriel, le verbe à l’infinitif, le verbe conjugué à tous les temps, le nom …  
  
Par exemple, tous ces mots ont le même lemme "Définir" : Définir, définition, définitions, définissons

**Thésaurus :** Répertoire structuré de termes (mots-clés) pour l'analyse de contenu et le classement de documents.

**n-gamme**: Un n-gramme est une sous-séquence de n éléments construite à partir d'une séquence donnée. L'idée semble provenir des travaux de Claude Shannon en théorie de l'information. Son idée était que, à partir d'une séquence de lettres donnée (par exemple « par exemple ») il est possible d'obtenir la fonction de vraisemblance de l'apparition de la lettre suivante.

**Analyseur lexical:** Un programme réalisant une analyse lexicale est appelé un analyseur lexical, tokenizer ou lexer. Un analyseur lexical est généralement combiné à un analyseur syntaxique pour analyser la syntaxe d'un texte.

**Ontologie**: Partie de la philosophie qui traite de l'être indépendamment de ses déterminations particulières.

**Wordnet**: WordNet est une base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton depuis une vingtaine d'années. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Des versions de WordNet pour d'autres langues existent, mais la version anglaise est cependant la plus complète à ce jour.

**WOLF :** WOLF (Wordnet Libre du Français) est une ressource lexicale sémantique libre pour le français, construite automatiquement à partir du Princetown WordNet et de ressources libres, et en cours de validation manuelle.

**EuroWordNet :** EuroWordNet est un système de réseaux sémantiques pour les langues européennes, basé sur WordNet. Chaque langue développe son propre wordnet mais ils sont interconnectés avec des liens interlinguals stockés dans l'index interlingual.

**Analyse lexicale :** En informatique, l'analyse lexicale, lexing, segmentation ou tokenization (à différencier du terme tokenization utilisé en sécurité informatique) fait partie de la première phase de la chaîne de compilation. 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.

**Analyse syntaxique :** L'analyse syntaxique consiste à mettre en évidence la structure d'un texte, généralement une phrase écrite dans une langue naturelle, mais on utilise également cette terminologie pour l'analyse d'un programme informatique. L'analyseur syntaxique (parser, en anglais) est le programme informatique qui réalise cette tâche. Cette opération suppose une formalisation du texte, qui est vu le plus souvent comme un élément d'un langage formel, défini par un ensemble de règles de syntaxe formant une grammaire formelle. La structure révélée par l'analyse donne alors précisément la façon dont les règles de syntaxe sont combinées dans le texte. Cette structure est souvent une hiérarchie de syntagmes, représentable par un arbre syntaxique dont les nœuds peuvent être décorés (dotés d'informations complémentaires).

**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, par opposition aux analyses lexicales ou grammaticales qui décomposent le message à l'aide d'un lexique ou d'une grammaire.

**Analyse pragmatique :** Approche linguistique qui se propose d'intégrer à l'étude du langage le rôle des utilisateurs de celui-ci, ainsi que les situations dans lesquelles il est utilisé. (La pragmatique étudie les présuppositions, les sous-entendus, les implications, les conventions du discours, etc.)