Séance 1 — Introduction au traitement automatique des langues

1. Introduction : pourquoi parler du TAL en linguistique ?

Nous produisons et recevons quotidiennement une quantité considérable de données linguistiques sous forme numérique.

Un étudiant peut aujourd'hui :

- envoyer des messages ;
- rédiger des documents ;
- effectuer des recherches sur Internet ;
- publier sur les réseaux sociaux ;
- consulter des articles scientifiques ;
- utiliser un correcteur automatique ;
- traduire un texte avec un outil numérique ;
- interagir avec un chatbot ;
- utiliser un assistant vocal.

Toutes ces activités produisent ou mobilisent du **langage humain sous forme numérique**.

Or, un ordinateur ne « comprend » pas spontanément une langue comme le fait un locuteur humain.

Lorsqu'un ordinateur reçoit la phrase :

> **« Le professeur observe les étudiants avec un microscope. »**

plusieurs interprétations peuvent être envisagées.

Le groupe **« avec un microscope »** peut être rattaché au verbe *observe* : le professeur utilise un microscope.

Mais, dans un autre contexte, une structure syntaxique similaire pourrait produire une autre interprétation.

Cette capacité à interpréter une expression en fonction de sa **structure**, de son **contexte** et de ses **connaissances linguistiques** constitue l'une des grandes difficultés du TAL.

Le TAL cherche précisément à développer des méthodes permettant aux ordinateurs de **traiter, analyser, représenter, rechercher, transformer ou générer des données linguistiques**.

---

2. Définition du Traitement Automatique des Langues

Le **Traitement Automatique des Langues**, généralement abrégé **TAL**, désigne l'ensemble des méthodes informatiques permettant de traiter automatiquement des données produites dans une langue naturelle.

En anglais, on rencontre principalement l'expression :

> **Natural Language Processing (NLP)**

Le terme « langue naturelle » permet de distinguer les langues humaines telles que :

- le français ;
- l'arabe ;
- l'anglais ;
- l'espagnol ;
- le chinois ;

des langages artificiels ou formels utilisés en informatique, comme Python, Java ou SQL.

Définition synthétique

On peut donc définir le TAL comme :

> **un domaine interdisciplinaire qui cherche à permettre aux ordinateurs de traiter et d'analyser des données exprimées en langue naturelle.**

Cette définition comprend plusieurs opérations :

Données linguistiques → traitement informatique → analyse / transformation / génération

Par exemple :

> Texte → segmentation → analyse morphosyntaxique → analyse syntaxique → interprétation → application

---

3. Que signifie «traitement automatique » ?

Le mot **automatique** ne signifie pas nécessairement que l'ordinateur fonctionne sans intervention humaine.

Il signifie principalement que certaines opérations qui pourraient être effectuées manuellement sont confiées à un système informatique.

Prenons un corpus contenant **500 000 mots**.

Un linguiste pourrait rechercher manuellement toutes les occurrences du mot *langage*, mais cette opération serait extrêmement longue.

Un programme peut rechercher automatiquement :

- toutes les occurrences ;
- leur fréquence ;
- leur position ;
- les mots qui apparaissent autour ;
- leur contexte ;
- leur catégorie grammaticale ;
- éventuellement leurs relations syntaxiques.

Le TAL permet donc notamment de passer :

> **d'une analyse manuelle de quelques exemples à une analyse assistée ou automatisée de très grandes quantités de données.**

---

4. Pourquoi le TAL est-il particulièrement important pour les linguistes ?

Le développement du numérique a profondément transformé la quantité de données linguistiques disponibles.

Avant l'informatisation massive des données, un chercheur pouvait travailler sur :

- quelques textes ;
- quelques dizaines d'ouvrages ;
- des enregistrements limités ;
- des exemples recueillis manuellement.

Aujourd'hui, il est possible d'étudier :

- des millions de textes ;
- des corpus journalistiques ;
- des corpus littéraires ;
- des conversations ;
- des commentaires sur les réseaux sociaux ;
- des archives numérisées ;
- des corpus multilingues.

Le TAL fournit donc aux linguistes des **méthodes permettant d'exploiter ces données à grande échelle**.

Exemple

Un chercheur souhaite étudier l'évolution de l'utilisation du mot **« peuple »** dans la littérature française.

Il pourrait constituer un corpus couvrant plusieurs siècles puis utiliser des outils de TAL pour étudier :

- la fréquence du terme ;
- ses contextes d'apparition ;
- les mots qui lui sont associés ;
- ses constructions syntaxiques ;
- son évolution sémantique ;
- sa distribution selon les périodes.

Le TAL ne remplace donc pas nécessairement l'analyse linguistique.

Il peut au contraire **augmenter les capacités d'observation du chercheur**.

---

5. Le TAL est un domaine interdisciplinaire

Le TAL se situe à l'intersection de plusieurs disciplines.

Linguistique

Elle apporte des connaissances concernant :

- la phonétique ;
- la phonologie ;
- la morphologie ;
- la syntaxe ;
- la sémantique ;
- la pragmatique ;
- l'analyse du discours.

Informatique

Elle fournit :

- les algorithmes ;
- les structures de données ;
- les langages de programmation ;
- les bases de données ;
- les systèmes de calcul.

Intelligence artificielle

Elle fournit notamment :

- l'apprentissage automatique ;
- les réseaux neuronaux ;
- les modèles de langage ;
- les méthodes de représentation des données.

Statistiques

Elles permettent notamment de :

- mesurer des fréquences ;
- calculer des probabilités ;
- modéliser des régularités ;
- évaluer les performances des systèmes.

On peut représenter cette interdisciplinarité ainsi :

Linguistique + Informatique + Statistiques + Intelligence artificielle → TAL

---

6. TAL et intelligence artificielle

Il est également important de ne pas confondre les deux notions.

Intelligence artificielle (IA) est un domaine beaucoup plus large.

Elle comprend notamment :

- la vision par ordinateur ;
- la robotique ;
- la reconnaissance vocale ;
- les systèmes experts ;
- l'apprentissage automatique ;
- le traitement du langage.

Le TAL constitue donc **l'un des domaines de l'intelligence artificielle**.

Quelques exemples d'application du TAL

🔍 1. Analyse stylistique et linguistique

• **Objectif :** Étudier les styles d'auteurs (ex. : Flaubert, Zola, Duras) en automatisant l'analyse des figures de style, de la syntaxe, ou du lexique.

• **Cas d'usage :** Comparer la syntaxe de Camus et de Sartre pour repérer des traits d'écriture propres à l'existentialisme.

❤️ 2. Analyse de sentiment et tonalité dans les romans

• **Objectif :** Étudier l'évolution émotionnelle dans un roman ou un corpus romanesque.

• **Cas d'usage :** Suivre l'évolution du sentiment de désespoir dans Les Misérables ou l'angoisse dans un roman noir.

📊 3. Cartographie des personnages et analyse des réseaux

• **Objectif :** Visualiser les relations entre les personnages d'un roman ou d'un corpus théâtral.

• **Cas d'usage :** Représenter les dynamiques sociales dans La Comédie humaine de Balzac.

🔄 4. Décryptage thématique de corpus

• **Objectif :** Identifier les thèmes majeurs dans un corpus d'ouvrages, d'articles ou de discours.

• **Cas d'usage :** Repérer les thèmes dominants dans la littérature francophone postcoloniale.

🧠 5. Exploration des idées philosophiques ou morales dans les textes

• **Objectif :** Analyser les positions philosophiques implicites dans un texte littéraire.

• **Cas d'usage :** Étudier les éléments stoïciens dans la poésie de Paul Valéry.

🗣️ 6. Analyse de discours ou de rhétorique

• **Objectif :** Étudier la rhétorique politique, littéraire ou philosophique.

• **Cas d’usage :** Comparaison des discours de leaders politiques algériens sur des thèmes similaires.