

# Web Analytics and Natural Language Processing (WANLP)

# Chapitre 02 | 02

**Concepts avancée du NLP:** 

Analyse Lexicale, Syntaxique et Sémantique

#### Professeur BOURAMOUL Abdelkrim

Département IFA, Faculté NTIC

<u>abdelkrim.bouramoul@univ-constantine2.dz</u> <u>www.bouramoul.com</u>

#### Etudiants concernés

Faculté/Institut	Département	Niveau	Spécialité
NTIC	IFA	Master 1	SDIA

# Plan du Cours

# Analyse lexicale, syntaxique et sémantique dans le NLP

- Analyse Lexicale et Morphologique : identification des unités lexicales.
- Analyse syntaxique : identification et analyse de la structure grammaticale
- Analyse sémantique : compréhension du sens des mots et des phrases

# Sources et Ressources pour le NLP

- Importance des données pour le NLP
- Différence entre Sources et Ressources en le NLP
- Corpus
- Lexique
- Datasets Structurés
- Bases de Données Lexicales
- Réseaux Sociaux

#### **Conclusion**

- Récapitulatif du Cours
- Importance et Perspectives

# Section 1: Analyse lexicale, syntaxique et sémantique dans le NLP

# Sous Section 1 | 1 Analyse Lexicale et Morphologique

Définition et exemple

# L'analyse lexicale et morphologique

décomposer le texte en ses éléments constitutifs (mots, phrases, etc.) et à analyser la structure de ces éléments.

L'analyse lexicale se concentre sur l'identification et la classification des mots (tokens)

L'analyse morphologique examine la structure interne des mots pour comprendre leur formation et leur signification, y compris les racines, les préfixes, les suffixes et les flexions.

Définition et exemple

# Phrase: "Les dattes algériennes sont délicieuses."

Analyse lexicale Toutes la phrase Analyse morphologique Le mot "algériennes"

["Les", "dattes", "algériennes", "sont", "délicieuses", "."] Racine : "algérien", Suffixe : "nes" (indiquant le féminin pluriel)

# Fondements de l'Analyse Lexicale et Morphologique

#### **Structure des Mots**

• Les mots, pierres angulaires de la communication, sont composés de morphèmes, les unités de sens les plus élémentaires.

### Les Morphèmes

 Exploration des morphèmes : racines (ex. 'marcher'), préfixes (ex. 're-' dans 'remarcher'), et suffixes (ex. '-able' dans 'marchable'). Ces éléments forment la base de notre langage.

#### **Construction des Mots**

• Les mots sont formés par l'assemblage de morphèmes. Par exemple, 'inacceptable' se décompose en 'in-', 'accept', et '-able', illustrant comment différents morphèmes contribuent à la signification globale.

#### Rôle des Mots dans le Langage

• Les mots permettent de structurer le sens et de transmettre des informations. Ils sont essentiels pour exprimer des idées, des émotions et pour la communication efficace.

# La Démarche Analytique

# **Principe**

• La démarche analytique en NLP consiste à décomposer les structures complexes du langage en éléments simples et analysables.

#### **Tokenisation**

• Processus de division d'un texte en tokens - mots, symboles, ou phrases - permettant une analyse détaillée du langage. La tokenisation transforme les textes bruts en données structurées pour le TALN.

### Les Morphèmes Libres et Liés

• Exploration de la nature des morphèmes : les libres peuvent exister seuls (ex. 'livre'), tandis que les liés ajoutent du sens lorsqu'ils sont combinés avec d'autres (ex. '-ment' dans 'récemment'). Cette distinction est utile pour comprendre la formation des mots.

# **Application en TALN**

• La tokenisation et la reconnaissance des morphèmes facilitent des tâches comme l'analyse syntaxique et sémantique, en identifiant les composantes essentielles du texte.

Exemple de la Démarche Analytique,

#### **Phrase**

• "L'Algérie a remporté la coupe d'Afrique des nations."

#### **Tokens**

• ["L'Algérie", "a", "remporté", "la", "coupe", "d'Afrique", "des", "nations", "."]

# Les Morphèmes Libres et Liés

- "Algérie" : morphème libre (peut exister seul).
- "remporté" : morphème libre.
- "d'" (dans "d'Afrique") : morphème lié (préfixe lié au mot "Afrique").

#### **Application en NLP**

• Reconnaitre les entités nommées (identifier "L'Algérie" comme un lieu) et l'analyse de sentiment (positif dans ce cas).

#### Méthodes d'Identification Lexicale

### **Principe**

- C'est les techniques permettant d'identifier et de catégoriser les éléments de sens dans le texte.
- Exemple : "Le chat noir chasse la nuit« on dentifié : "chat", "noir", "chasse", "nuit" comme unités de sens principales.

#### Segmentation Textuelle

- Technique de division d'un texte en segments significatifs pour faciliter l'analyse lexicale.
- Diviser la phrase en tokens ["Le", "chat", "noir", "chasse", "la", "nuit", "."]

# **Utilisation des Expressions Régulières**

- Ce sont un outil puissant pour identifier des patterns spécifiques dans le texte.
- Utiliser une expression régulière pour identifier les mots contenant "n" "noir", "chasse", "nuit"

## **Avantages de l'Identification Lexicale**

• Grâce à l'identification des mots clés "chat", "chasse", "nuit", on peut mieux comprendre le thème central de la phrase

Exemple illustrant les Méthodes d'Identification Lexicale

#### **Phrase**

• "Le vent souffle fort sur la plage en été."

#### Mots clés identifiés

• "vent", "souffle", "fort", "plage", "été".

#### **Tokens**

• ["Le", "vent", "souffle", "fort", "sur", "la", "plage", "en", "été", "."]

# Expression Régulière pour identifier les patterns spécifiques

• "e": "Le", "vent", "souffle", "fort", "sur", "plage", "en", "été".

#### **Avantages de l'Identification Lexicale**

• L'identification des mots clés permet de mieux comprendre le thème central de la phrase, qui est l'impact du vent sur la plage en été.

Utilité de l'Analyse Lexicale pour Le NLP

# Impact sur la Compréhension du Texte

- La reconnaissance des formes morphologiques clarifie le sens des mots et améliore la compréhension globale des textes.
- Exemple : Identifier "marcher", "marchait", et "marchera" comme variantes de "marcher" aide à comprendre le temps des actions.

# Désambiguïsation dans le NLP

- L'analyse lexicale facilite la distinction entre les sens des mots en fonction du contexte.
- Exemple : "Banque" peut désigner une institution financière ou le bord d'une rivière. L'analyse contextuelle détermine le sens correct.

Illustration du processus d'analyse lexicale et morphologique

#### **Phares initiale**

• "Les enfants jouent joyeusement dans le parc."

### Segmentation Textuelle et Tokenisation

- Diviser la phrase en tokens (mots et signes de ponctuation).
- Tokens: ["Les", "enfants", "jouent", "joyeusement", "dans", "le", "parc", "."]

### **Identification des Morphèmes**

- Reconnaître les unités de sens les plus petites dans chaque token.
- "enfants" = enfant + s (morphème libre + morphème lié indiquant le pluriel)
- "jouent" = joue + ent (racine verbale + terminaison pour la 3e personne du pluriel)

# **Analyse Morphologique**

- Analyser la structure des mots pour comprendre leur formation et leur signification.
- "joyeusement" = joyeux + e + ment (adjectif transformé en adverbe)

# Sous Section 1 | 2 Analyse syntaxique : identification et analyse de la structure grammaticale.

Introduction à l'Analyse Syntaxique

# Qu'est-ce que l'Analyse Syntaxique?

- Explorer la structure grammaticale des phrases, et identifier comment les mots s'organisent pour former des phrases cohérentes et compréhensibles.
- Exemple: "Le chat dort sur le tapis.", une structure grammaticale cohérente ["Le chat" → Sujet, "dort" → Verbe, "sur le tapis" → Complément circonstanciel de lieu.]

# Objectifs de l'Analyse Syntaxique

• Elle vise à comprendre la hiérarchie et les relations entre les éléments d'une phrase pour faciliter l'interprétation du sens et la réalisation de tâches spécifiques en NLP.

#### Importance dans le NLP

• Elle permet la compréhension des structures linguistiques complexes, soutenant des applications comme la traduction automatique, la reconnaissance vocale, et l'extraction d'information.

# Composants de la Syntaxe

# **Objectif**

• Exploration des éléments structuraux du langage : phrases, clauses, et phrases nominales, qui forment le squelette de notre communication.

#### Phrases, Clauses, et Phrases Nominales

- La phrase est l'unité de base du discours. Les clauses, parties d'une phrase avec sujet et verbe, et les phrases nominales, groupes de mots agissant comme un nom.
- Exemple: "Le chat qui dort sur le tapis ronronne.", cette phrase illustre la clause "qui dort sur le tapis" (clause relative décrivant le chat) et la phrase nominale "Le chat qui dort sur le tapis" agissant comme sujet.

# **Fonctions Syntaxiques**

- Les éléments d'une phrase assument différentes fonctions : sujet (qui fait l'action), objet (qui reçoit l'action), et complément (informations supplémentaires).
- Exemple : "Amina donne un livre à Karim."
- Ici, "Amina" est le sujet, "un livre" est l'objet, et "à Karim" est le complément d'objet indirect, montrant comment chaque élément contribue à la structure de la phrase.

# Techniques d'Analyse Syntaxique

### **Parsing: Approches Ascendante et Descendante**

- Dans l'approche ascendante on construit l'arbre syntaxique à partir des feuilles (mots), tandis que l'approche descendante commence par la racine (phrase) et divise progressivement en composants plus petits.
- Exemple de phrase : "Le chat mange la souris."
- Approche ascendante : Identifier "chat" et "souris" comme noms, "mange" comme verbe, puis construire la relation sujet-verbe-objet.
- Exemple: Approche descendante: Commencer par la phrase entière comme action principale, diviser en sujet "Le chat", verbe "mange", et objet "la souris".

# **Arbre Syntaxique et Ses Composants**

• Visualisation de la structure d'une phrase à travers un arbre syntaxique, identifiant sujets, verbes, objets, et compléments, et montrant leurs relations hiérarchiques.

Exemple de phrase "L'enfant joyeux joue dans le parc.				
L'enfant joyeux	Joue	Dans le parc		
sujet	verbe	complément de lieu		

Outils et Technologies pour l'Analyse Syntaxique

#### **Les Parseurs Syntaxiques**

Des outils clés en analyse syntaxique, tels que spaCy et Stanford Parser,

# La librairie spaCy

- spaCy offre une interface intuitive pour l'analyse syntaxique, supportant de nombreuses langues avec des performances de haut niveau.
- Elle permet d'obtenir les dépendances syntaxiques d'une phrase, par exemple dans la phrase "Le chat dort sur le tapis", on peut 'identifier ['chat' → sujet, 'dort' → verbe principal, 'sur le tapis' → complément circonstanciel."

### Le Stanford Parser

- Stanford Parser, connu pour sa précision, utilise des modèles linguistiques complexes pour analyser la structure des phrases.
- En analysant la phrase 'Les enfants jouent dans le parc', Stanford peut construire un arbre syntaxique détaillé, montrant comment 'Les enfants' fonctionnent comme sujet, 'jouent' comme verbe, et 'dans le parc' comme complément de lieu, illustrant les relations complexes entre les mots.

# Sous Section 1 | 3 Analyse sémantique: compréhension du sens des mots et des phrases

Objectifs et importances

# Qu'est-ce que l'Analyse Sémantique?

• L'analyse sémantique examine le sens et l'interprétation des mots et des phrases, cherchant à comprendre leur signification dans différents contextes.

# Objectifs de l'Analyse Sémantique

• Elle vise à extraire et à interpréter le sens profond du texte, facilitant la compréhension automatique du langage naturel pour des applications telles que la traduction automatique et l'interaction homme-machine.

# Importance dans le NLP

 L'analyse sémantique est importante pour le NLP, permettant aux machines de traiter et de comprendre le langage humain de manière plus naturelle et intuitive.

Compréhension du Sens des Mots, Polysémie/ Homonymie/ Désambiguïsation

# Polysémie et Homonymie

 L'exploration de la polysémie (mots avec plusieurs sens liés) et de l'homonymie (mots identiques mais avec des sens différents), défis majeurs pour la compréhension du langage.

# Exemples

- Polysémie: Le mot "date" peut désigner soit le fruit très consommé en Algérie pendant le mois de Ramadan, soit un jour précis du calendrier. "J'ai acheté des dates pour le ftour" versus "La date de l'événement est fixée au samedi."
- Homonymie: "Son" peut signifier à la fois l'appartenance (possessif) et le bruit. Dans une phrase comme "Son de la musique est fort dans son appartement", sans contexte supplémentaire, il est difficile de distinguer les deux sens.

Compréhension du Sens des Mots, Polysémie/ Homonymie/ Désambiguïsation

# Désambiguïsation Sémantique

• Techniques et stratégies utilisées pour clarifier le sens des mots en contexte, permettant une interprétation précise du texte.

# Exemple

 Pour la phrase "Il est allé chercher du pain à la place", la désambiguïsation sémantique permet de comprendre si "place" se réfère à une place publique ou à une position dans une file d'attente, souvent clarifiée par le contexte environnant.

Analyse de la Signification des Phrases

# Relations Sémantiques entre les Mots

 Ce concept implique la manière dont les mots dans une phrase se rapportent les uns aux autres pour créer du sens. Par exemple, les synonymes, antonymes, et les relations hiérarchiques (hyperonymes/hyponymes) enrichissent la compréhension du texte.

# Exemple

- "Le drapeau vert et blanc avec un croissant et une étoile se lève"
- Ici, les relations sémantiques entre "drapeau", "vert et blanc", "croissant" et "étoile" évoquent l'identité nationale algérienne, illustrant comment la combinaison de mots crée du sens au-delà de leur signification individuelle.

Analyse de la Signification des Phrases

# **Pragmatique et Contexte**

 La pragmatique étudie comment le contexte influence la signification des phrases. Elle prend en compte le contexte situationnel, les connaissances partagées et les intentions des interlocuteurs pour interpréter le sens.

# Exemple

 "Il fait chaud à Tamanrasset." Sans contexte, cette phrase pourrait simplement indiquer la température. Cependant, dans le contexte de la planification d'un voyage, cela pourrait impliquer la nécessité de préparation appropriée, montrant comment le contexte influence la compréhension de la phrase

Outils et Technologies pour l'Analyse Sémantique

#### WordNet

• Une base de données lexicale pour l'anglais qui relie les mots à travers des relations sémantiques. Utilisé pour la compréhension du sens des mots et leur contextualisation.

#### **BERT**

• (Bidirectional Encoder Representations from Transformers, est un modèle de langage basé sur l'apprentissage profond, conçu pour prétraiter le langage naturel. BERT excelle dans la compréhension du contexte des mots dans une phrase.

#### GloVe

• **Glo**bal **Ve**ctors for Word Representation, est un modèle de représentation vectorielle des mots qui capture les significations à travers des espaces vectoriels globaux, utile pour comprendre les relations entre les mots.

#### **ELMo**

• Embeddings from Language **Mo**dels, offre des représentations de mots basées sur des contextes entiers des phrases, permettant une compréhension plus nuancée des sens en fonction du contexte d'usage.

Applications Pratiques en NLP

# Extraction d'information et résumé automatique

• À partir d'un article de journal sur un événement sportif en Algérie, un système peut extraire des informations clés (équipes, scores, moments importants) et générer un résumé concis de l'article.

# Reconnaissance d'entités nommées et classification de texte

• Dans les commentaires des utilisateurs sur un forum de discussion algérien, un outil d'analyse sémantique peut identifier des entités spécifiques comme les noms de lieux (ex. "Alger"), noms de personnes, ou organisations, et classifier les commentaires par sujet (ex. politique, sport, culture).

# Section 2: Sources et Ressources pour le NLP

Importance des données pour le NLP

# Importance des Données pour le NLP

• Les données sont indispensables pour comprendre et générer le langage humain, servant de base pour entraîner et évaluer les modèles d'intelligence artificielle (ML, et DL).

# Diversité et Qualité des Données

 La diversité et la qualité des données déterminent la performance des applications de NLP, soulignant la nécessité d'une sélection rigoureuse des sources et ressources pour des résultats précis et fiables.

Différence entre Sources et Ressources pour le NLP

# Les sources pour le NLP

- Sont les origines brutes des données textuelles, comme les livres, articles, et posts sur les réseaux sociaux, fournissant le matériau initial pour l'analyse.
- Exemple : Articles en ligne (articles de presse sur TSA), transcriptions de discours (conférences TEDx en Algérie), posts de réseaux sociaux (commentaires sur Facebook).

#### Les ressources

- Sont des ensembles de données déjà structurés ou annotés, tels que les corpus, lexiques, ou datasets, préparés spécifiquement pour faciliter les tâches de NLP. La principale différence réside dans le niveau de préparation et d'organisation des données pour leur utilisation directe en NLP.
- Corpus TALN (ex. Corpus de l'Arabe Algérien Moderne), lexiques (ex. WordNet pour l'anglais, LeFFF pour le français), datasets annotés (ex. GLUE Benchmark pour l'évaluation de modèles), bases de données lexicales (ex. Base de Données Lexicale pour le français).

# Les Corpus

# **Qu'est-ce qu'un Corpus?**

• Une collection structurée de textes utilisés pour l'entraînement et l'évaluation en NLP. Il fournit des données réelles pour modéliser le langage.

#### **Utilisation des Corpus**

• Les corpus alimentent l'apprentissage des modèles d'IA, permettant la reconnaissance vocale, la traduction automatique, et la génération de texte, grâce à une base de données riche et variée.

# Exemples de corpus utilisés en NLP

- Corpus de l'Université de Pennsylvanie (Penn Treebank) : Utilisé pour l'analyse syntaxique et la compréhension grammaticale.
- British National Corpus (BNC): Contient un large éventail de textes du XXe siècle en anglais britannique, utilisé pour l'étude de l'usage du langage.
- **Corpus Europarl** : Comprend des transcriptions des débats du Parlement européen, souvent utilisé pour la formation et l'évaluation des modèles de traduction automatique.
- **Google Ngrams** : Une collection massive de mots et de phrases extraites de livres numérisés, permettant l'analyse de tendances linguistiques sur de longues périodes.

# Les Lexique

### Qu'est-ce qu'un Lexique?

• Un lexique est un répertoire de mots accompagnés de leurs informations linguistiques, telles que les définitions, synonymes, et règles d'usage. Il sert de fondement à la compréhension sémantique en TALN."

#### Rôle des Lexiques en NLP

• Les lexiques aident à la désambiguïsation de sens des mots en fournissant un contexte sémantique clair, importanat pour des applications telles que la traduction automatique et l'analyse de sentiment.

## Exemples de lexiques utilisés en NLP

- WordNet: Un lexique sémantique pour l'anglais (et d'autres langues) qui organise les mots en ensembles de synonymes et fournit des relations sémantiques entre eux, telles que la hiérarchie et les associations conceptuelles.
- Lefff (Le Lexique des Formes Fléchies du Français): Un lexique morphologique et syntaxique pour le français, utilisé pour le traitement automatique du français.
- Sentiment140 Lexicon : Spécialisé dans l'analyse des sentiments, ce lexique fournit des polarités (positif/négatif) pour des mots fréquemment utilisés dans les tweets.

#### Les Datasets Structurés

# **Importance**

• Les datasets structurés jouent un rôle important dans le développement et l'évaluation des modèles de NLP, fournissant des données organisées pour entraîner et tester les algorithmes.

# **Exemples de Datasets Structurés**

- GLUE (General Language Understanding Evaluation) rassemble plusieurs tâches de compréhension de texte pour évaluer la compréhension globale du langage par les modèles.
- SQuAD (Stanford Question Answering Dataset), se concentre sur la capacité des modèles à répondre à des questions basées sur des passages de texte, testant ainsi la compréhension fine et la récupération d'informations précises.

# **Impact des Datasets Structurés**

 Ces ressources permettent de mesurer les progrès, de benchmarker les performances des modèles de manière objective, et d'identifier les défis futurs dans le domaine du NLP.

#### Bases de Données Lexicales

#### Importance des Bases de Données Lexicales

• Fondamentales pour le TALN, les bases de données lexicales stockent des informations détaillées sur les mots, leurs significations, et leurs relations.

#### Analyse Syntaxique avec Bases de Données Lexicales

• Ces bases sont utilisées pour structurer et analyser la grammaire des phrases, identifiant la fonction de chaque mot grâce à leurs caractéristiques linguistiques.

#### Analyse Sémantique et Bases de Données Lexicales

• Elles enrichissent la compréhension des textes en offrant un contexte sémantique, essentiel pour interpréter le sens précis des mots dans diverses situations.

#### Exemples de Bases de Données Lexicales Connuess en NLP

- WordNet : Un réseau sémantique pour l'anglais, facilitant la compréhension des relations entre mots.
- LeFFF (Lexique des Formes Fléchies du Français) : Riche ressource pour l'analyse du français.
- CELEX : Base de données pour l'anglais, l'allemand et le néerlandais, offrant des informations phonologiques, morphologiques et syntaxiques.
- BabelNet : Un réseau sémantique multilingue qui combine les éléments de WordNet et Wikipedia pour une couverture linguistique étendue.

#### **Exemples et Applications**

• WordNet permet l'exploration des relations entre les mots, tandis que LeFFF est spécifiquement conçu pour appuyer l'analyse du français, soutenant la désambiguïsation et la génération de langage.

#### Réseaux Sociaux

#### Réseaux Sociaux pour le NLP

 Les réseaux sociaux offrent un flux constant de langage naturel, reflétant des opinions, émotions, et interactions sociales réelles, essentielles pour l'analyse de sentiments et la reconnaissance d'entités nommées.

#### **Analyse de Sentiments**

• Utilisation des données des réseaux sociaux pour détecter des sentiments et des opinions dans le texte, permettant aux entreprises de comprendre les réactions des consommateurs et aux chercheurs d'étudier les dynamiques sociales.

#### Reconnaissance d'Entités Nommées

• Identification automatique des noms propres (personnes, lieux, organisations) dans les textes des réseaux sociaux, utile pour l'extraction d'information et la surveillance des mentions.

#### **Utilisation des Réseaux Sociaux en NLP**

- Les réseaux sociaux comme Twitter (x) et Facebook sont des mines de données
- Twitter, avec ses messages brefs, est idéal pour capturer les tendances et les sentiments en temps réel.
- Facebook permet une analyse plus profonde des préférences et comportements des utilisateurs grâce à des interactions détaillées.

# Section 5: Conclusion

# Récapitulatif du Cours

Introduction aux concepts avancés du NLP : analyse lexicale, syntaxique et sémantique.

Exploration des techniques d'identification lexicale, y compris la tokenisation et l'utilisation des expressions régulières.

Application de ces concepts dans diverses tâches du NLP, telles que l'analyse de sentiment et la reconnaissance d'entités nommées.

Compréhension de la structure des mots et des phrases à travers l'analyse morphologique et syntaxique.

# Importance et Perspectives

Importance de l'analyse lexicale, syntaxique et sémantique dans la compréhension approfondie des textes.

Impact sur le développement d'applications, telles que la traduction automatique, le résumé automatique et les assistants virtuels.

# Importance et Perspectives

Intégration de techniques d'apprentissage profond pour améliorer la compréhension du langage naturel et relever des défis plus complexes en NLP.

Exploration de l'interaction entre le NLP et d'autres domaines de l'intelligence artificielle, tels que la vision par ordinateur et la robotique



# Fin de Chapitre 02