

Université Mohammed V- Rabat
Ecole Mohammadia d'Ingénieurs
Département Génie Informatique
Filière Génie Informatique et Digitalisation



Text Analytics

Dr. H. SEBBAQ

h.sebbaq@gmail.com

Pr. N. EL FADDOULI

nfaddouli@gmail.com

2023-2024

CC-BY NC SA

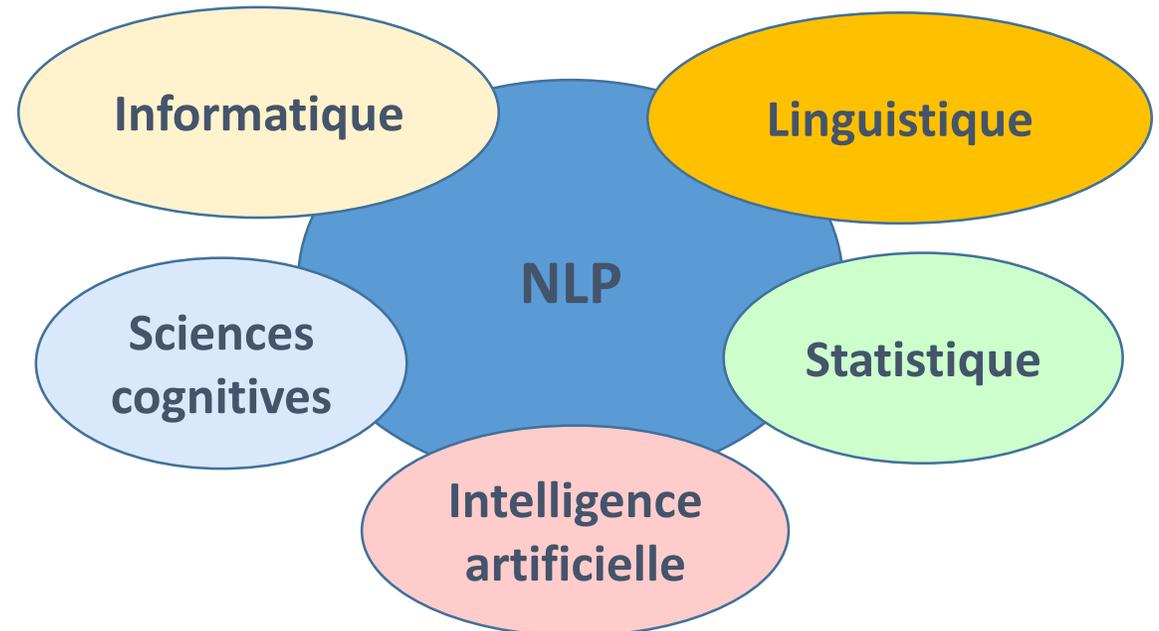
NLP : Définition

On regroupe sous le vocable Natural Language Processing (NLP) l'ensemble des recherches et développements visant à modéliser et reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication

- ❑ Le traitement du langage naturel (NLP) vise à **extraire une représentation complète du sens d'un texte libre** en répondant aux questions sur qui, quoi, quand, où, comment et pourquoi.
- ❑ Il utilise des **concepts linguistiques** tels que les parties du discours (noms, verbes, adjectifs, etc.) et la structure grammaticale (phrases nominales, prépositionnelles, relations de dépendance).
- ❑ Le NLP s'appuie sur **diverses représentations de connaissances**, comme un **lexique des mots** avec leurs significations, **des propriétés grammaticales**, des règles de grammaire, et des ressources comme des ontologies d'entités/actions ou des thésaurus de synonymes/abréviations.

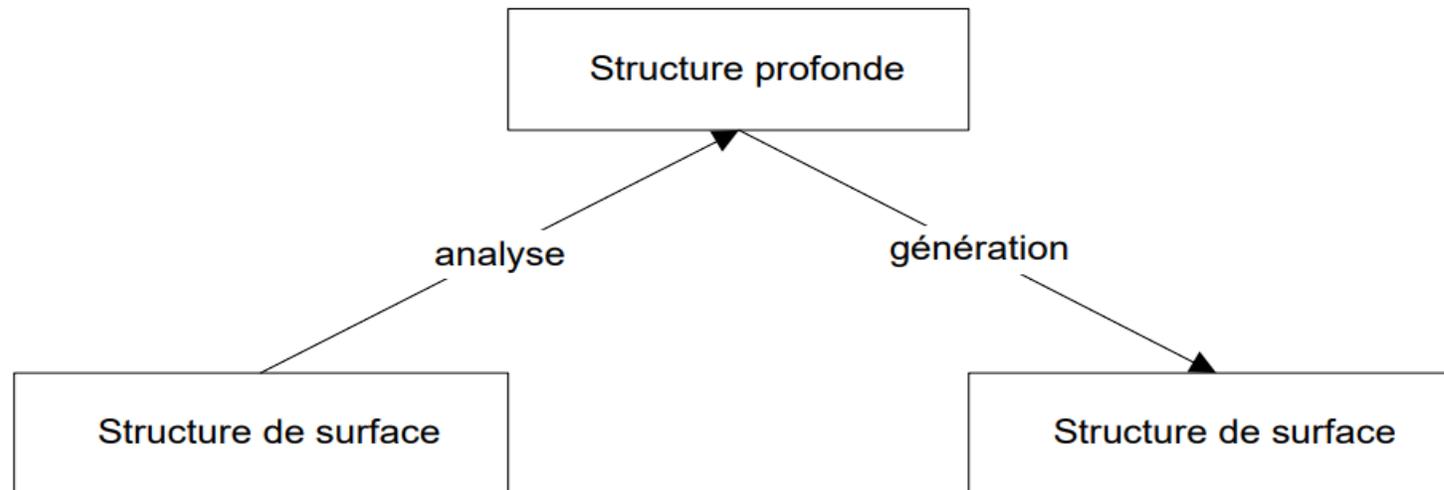
NLP : Pourquoi c'est complexe?

- ❑ Domaine pluridisciplinaire
- ❑ Discipline à cheval entre la linguistique et l'informatique
- ❑ Entretient des liens étroits avec les sciences cognitives et les statistiques
- ❑ Possède des zones de recouvrement avec l'Intelligence Artificielle
- ❑ Données complexes, naturelles, hétérogènes et multidimensionnelles, de grande dimension, d'une grande variété
 - non formel
 - ambiguïté, variabilité
 - implicite, redondance
 - grand nombre d'événements rares



NLP : L'ANALYSE VS LA GENERATION

- ❑ En fonction des objectifs attendus d'un système NLP, on peut le classer dans l'un des domaines suivants : analyse et génération.
- ❑ Le processus d'analyse consiste à démarrer de la **structure de surface (locution ou texte écrit)** pour arriver à la structure profonde équivalente.
- ❑ Le processus de génération est le processus inverse : il consiste à passer de la structure profonde à la structure de surface.
- ❑ Il est possible qu'une application de NLP appartienne en même temps aux deux domaines : analyse et génération; c'est le cas par exemple d'une application de traduction automatique.



NLP : BREF HISTORIQUE

- ❑ **Années 1950** : Les **premières recherches** en traitement automatique des langues commencent, avec le **projet de traduction automatique de Georgetown-IBM**.
- ❑ **Années 1960-1970** : Le système **SHRDLU de Terry Winograd (1972)** permet à un ordinateur de **manipuler des blocs dans un univers virtuel** en utilisant un langage naturel simplifié, démontrant les capacités du langage naturel dans un contexte restreint.
- ❑ **Années 1980** : L'avènement des **techniques statistiques** comme la méthode du maximum d'entropie et les modèles probabilistes dans le NLP, conduisant à une approche plus data-driven.
- ❑ **Années 1990** : **L'essor des systèmes de recherche d'informations** basés sur des algorithmes de traitement du langage naturel, permettant de trouver des informations pertinentes dans de vastes corpus textuels.
- ❑ **Années 2000** : L'émergence de **la traduction automatique statistique** et des premières applications commerciales comme les systèmes de recommandation basés sur le traitement du langage naturel.
- ❑ **Années 2010** : **L'essor du Deep Learning** dans le NLP, marqué par des avancées majeures dans les modèles neuronaux tels que les réseaux récurrents et les réseaux de neurones à convolution, menant à des améliorations significatives dans la traduction automatique, la reconnaissance vocale, etc.
- ❑ **Années 2020 à aujourd'hui** : La montée en puissance des **modèles de langage pré-entraînés** comme BERT, GPT et d'autres, révolutionnant le domaine du NLP grâce à leur capacité à comprendre et à générer des textes avec une précision et une cohérence sans précédent.

NLP : Les données textuelles sont superficielles

An iceberg is a large piece of freshwater ice that has broken off from a snow-formed glacier or ice shelf and is floating in open water.

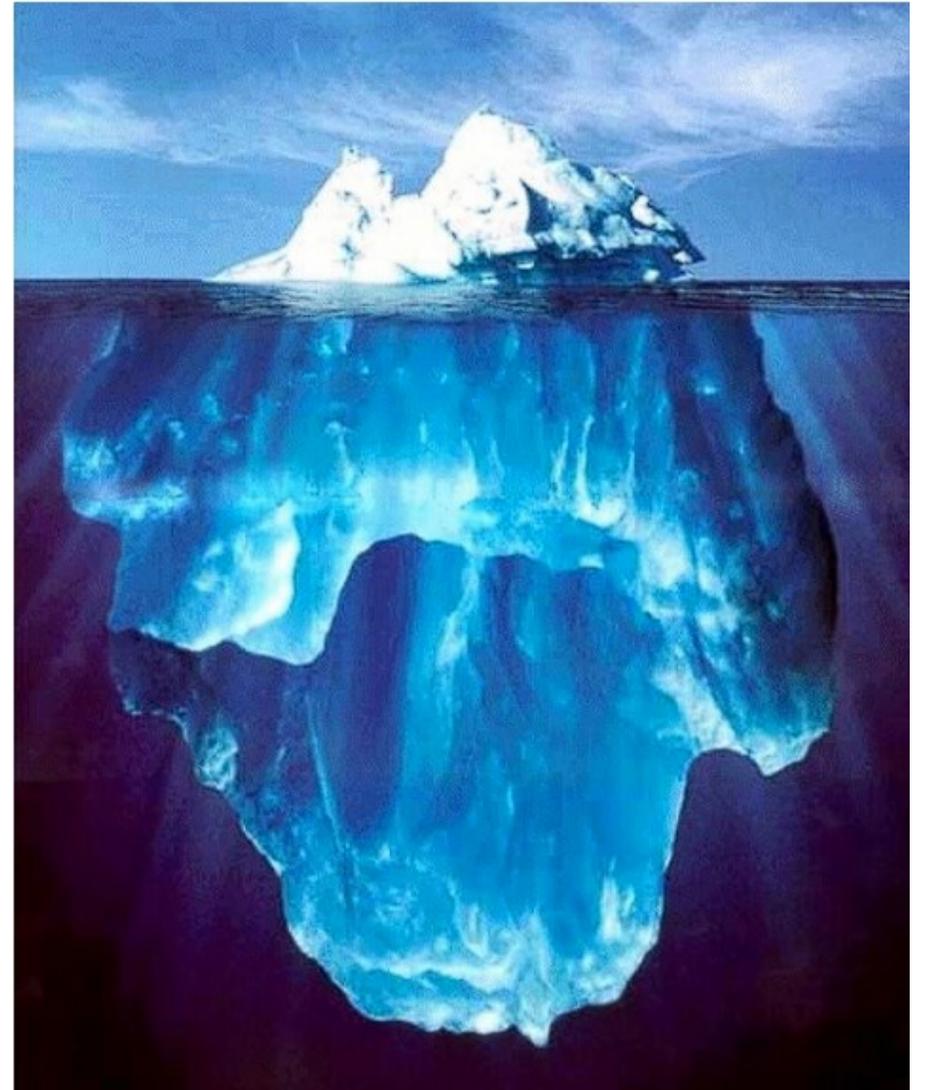
Un iceberg est un gros morceau de glace d'eau douce qui s'est détaché d'un glacier ou d'une plate-forme de glace formé par la neige et qui flotte en eau libre.



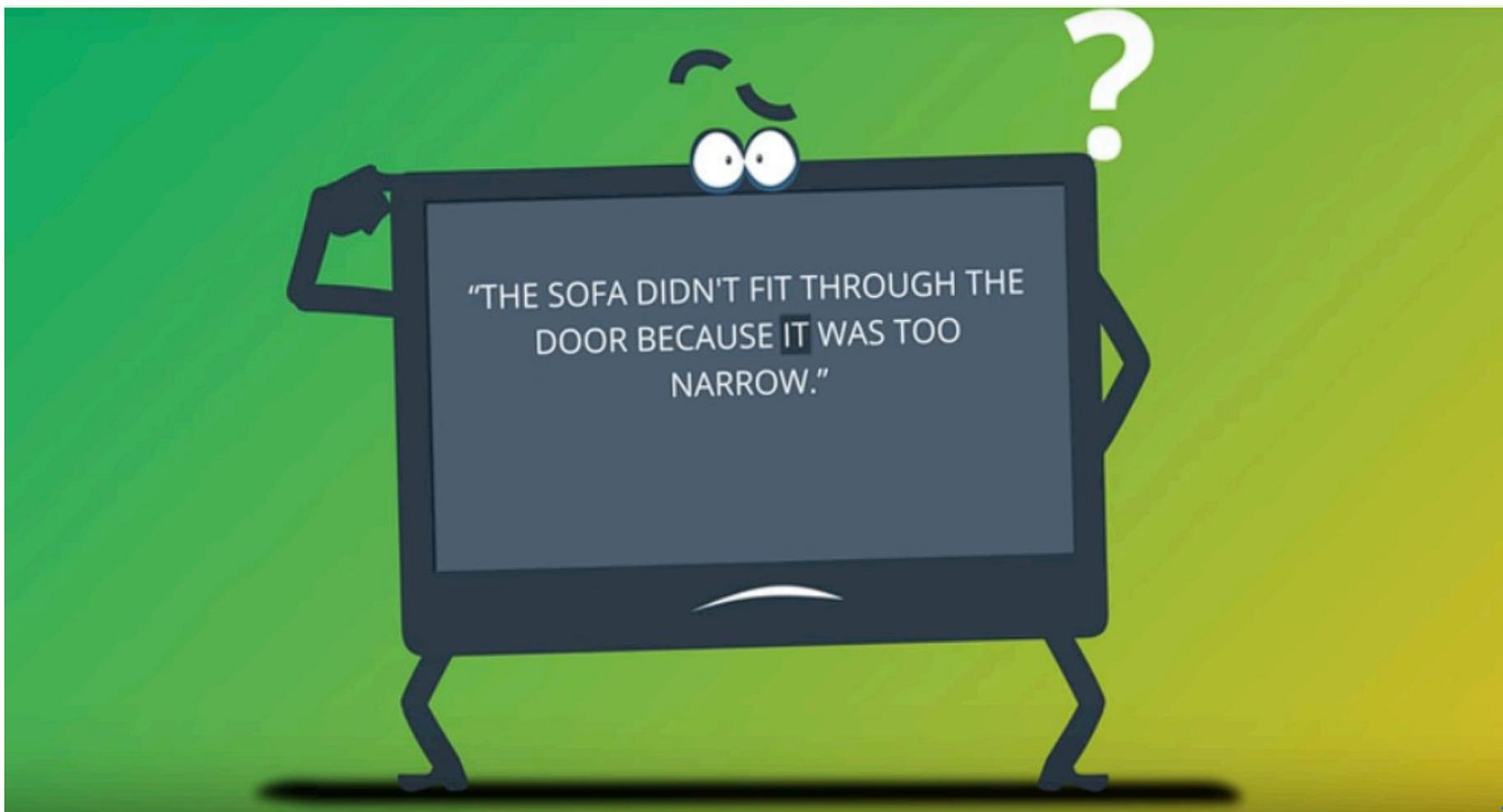
NLP : ... mais la langue est complexe

An iceberg is a large piece of freshwater ice that has broken off from a snow-formed glacier or ice shelf and is floating in open water.

Un iceberg est un gros morceau de glace d'eau douce qui s'est détaché d'un glacier ou d'une plate-forme de glace formé par la neige et qui flotte en eau libre.



NLP : Le contexte est essentiel



NLP : Le contexte est essentiel



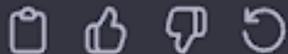
NLP : Le contexte est essentiel

 You

The sofa didn't fit through the door because it was too narrow.

 ChatGPT

It seems like you're mentioning a situation where a sofa couldn't fit through a door due to the door being too narrow. If you need assistance or want to discuss this situation further, feel free to provide more details or let me know how I can help you with this.



NLP : Le contexte est essentiel

 **You**

The sofa didn't fit through the door because it was too wide.

 **ChatGPT**

It appears there might be a contradiction in the statement. If the sofa didn't fit through the door because it was too wide, it means that the sofa was larger than the door frame. This situation suggests that the sofa's width exceeded the available space of the door's width, preventing it from passing through.

NLP : Approches

❑ Approche Linguistique :

- Cette approche se concentre sur l'utilisation des connaissances linguistiques et des structures grammaticales pour analyser et comprendre le langage.
- Elle utilise des règles, des schémas syntaxiques et des principes linguistiques pour décomposer et interpréter le texte.
- **Exemples** : analyse grammaticale, identification des parties du discours, traitement des règles syntaxiques.

❑ Approche Rule-Based (Basée sur des Règles) :

- Cette approche repose sur des ensembles de règles préétablies pour effectuer des tâches de traitement du langage.
- Elle utilise des règles spécifiques pour interpréter les éléments linguistiques et déduire le sens des phrases.
- **Exemples** : analyse syntaxique basée sur des règles, extraction d'entités nommées à l'aide de patrons définis.

NLP : Approches

❑ Approche Statistique :

- Repose sur des méthodes probabilistes et statistiques traditionnelles pour comprendre et analyser le langage, telles que l'utilisation de modèles de langue et de probabilités conditionnelles.
- **Exemples** : Modèles de Markov cachés (HMM), modèles de n-grammes, probabilités conditionnelles, modèles de langue statistiques.

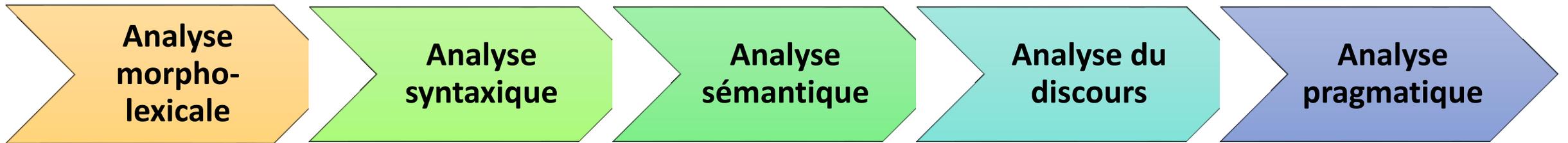
❑ Approche Machine Learning (ML) :

- Exploite des techniques d'apprentissage automatique pour entraîner des modèles à partir de données, permettant aux systèmes de détecter des schémas et des relations sans avoir besoin de règles explicites.
- **Exemples** : Réseaux de neurones récurrents (RNN), réseaux de neurones convolutionnels (CNN), modèles de Transformers, algorithmes de classification et les méthodes de clustering.

En résumé, bien que l'approche statistique soit souvent utilisée dans l'apprentissage automatique, elle diffère de l'approche ML par sa focalisation sur des modèles probabilistes et des méthodes **statistiques plus traditionnelles** pour l'analyse linguistique, tandis que **l'approche ML repose davantage sur des algorithmes d'apprentissage** pour extraire des schémas à partir des données.

NLP : Niveaux

- ❑ Pour traiter le langage naturel, on a besoin d'informations coordonnées et pertinentes sur des niveaux divers.
- ❑ Le plus souvent on a recours à cinq niveaux de connaissances sur une langue. Cela nous donne la hiérarchie suivante :



NLP : Niveaux - Analyse morpho-lexicale

- ❑ **Morphologie** : Branche de la linguistique étudiant la structure et la formation des mots dans une langue à partir d'unités plus petites appelées **morphèmes**
- ❑ **Morphème** : Unité de sens la plus petite d'une langue. Les morphèmes sont l'élément fondamental de la morphologie.
 - Par exemple, 'Lapin' est constitué d'un seul morphème, tandis que 'Lapins' en a deux : 'Lapin' et le morphème "-s".
- ❑ **Radical (Stem)**: Partie d'un mot à laquelle un affixe est attaché.'
- ❑ **Différence entre morphèmes et mots** : Les morphèmes peuvent être indépendants ou non, tandis que les mots sont généralement des unités autonomes dans une phrase.
 - Par exemple, "chat" est un morphème autonome, mais lorsqu'on considère "chats", le suffixe "-s" indique que le morphème est la forme plurielle de "chat".

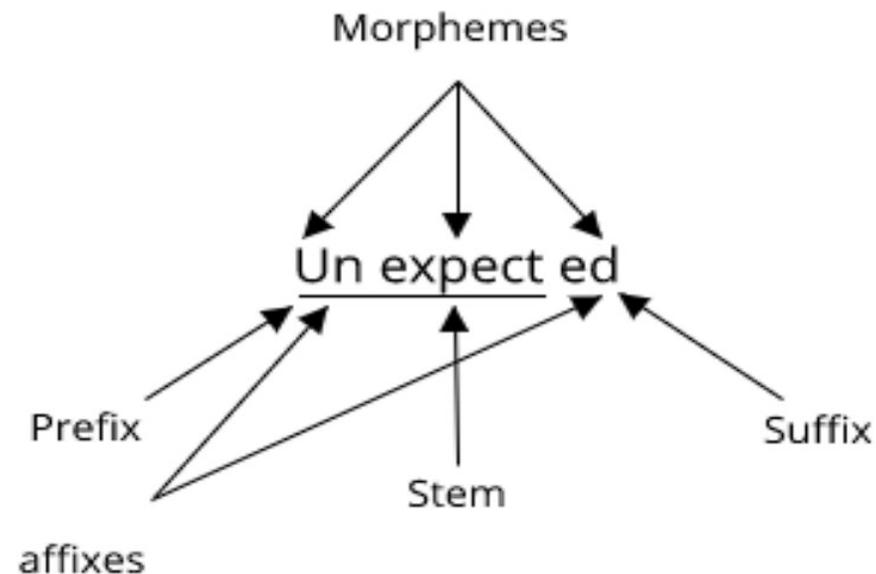
NLP : Niveaux - Analyse morpho-lexicale

Utilisation de l'analyse morphologique :

- ❑ Cette étape est la première si l'entrée du système est textuelle (Si Voix alors une étape de traitement phonétique est la première)
- ❑ Cette étape consiste à déterminer les informations grammaticales associées à chaque mot considéré isolément (traitement morphologique)
- ❑ Elle est employée dans la segmentation des mots et l'étiquetage de parties du discours (POS). Cette dernière sera expliquée plus tard.
- ❑ Dans un système de NLP, l'analyse morphologique a pour objectif de :
 - Reconnaître la catégorie syntaxique et les propriétés grammaticales des mots
 - Proposer une lemmatisation
 - Reconnaître les entités nommées (noms de personnes, d'organisations, d'entreprises, de lieux, quantités, distances, valeurs, dates...)

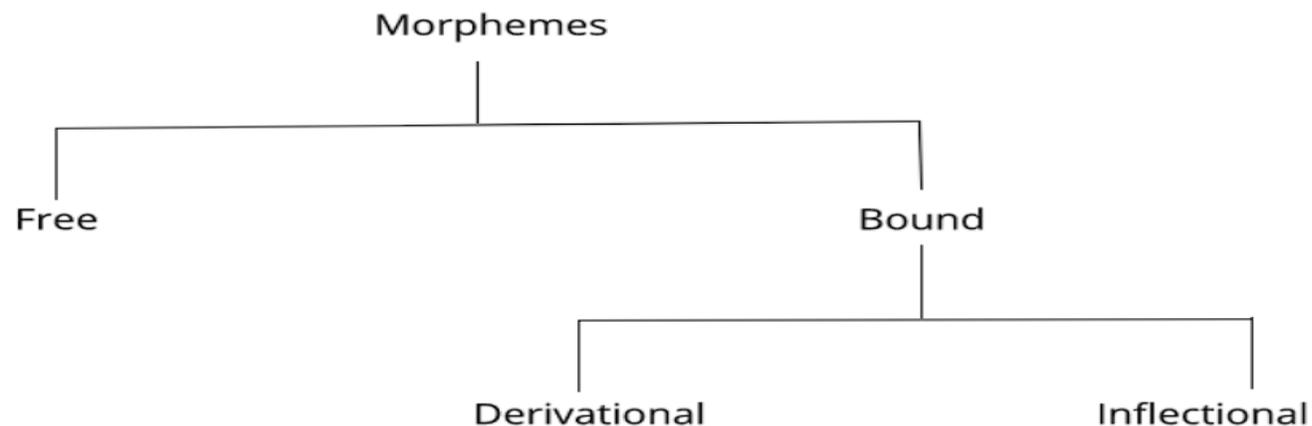
NLP : Niveaux - Analyse morpho-lexicale

- ❑ **Analyse morphologique** : L'analyse grammaticale examinant la formation des mots en utilisant les morphèmes, les plus petites unités de signification.
- ❑ **Les morphèmes** sont généralement des affixes et peuvent être classés en quatre types :
 - Affixes : Préfixes, Suffixes, Infixes (Apparaissent à l'intérieur du radical **bumili** (buy en Tagalog, une langue des Philippines))
 - Circonfixes : Entourent un mot en étant attachés au début et à la fin du radical, Exemple : « encourageant », le préfixe « en » et le suffixe « ant » encadrent le radical « courage »



NLP : Niveaux - Analyse morpho-lexicale

- ❑ **Classification des morphèmes** : Ils sont classifiés en deux grandes catégories :
 - **Morphèmes libres** : Peuvent exister seuls et agir comme un mot autonome.
 - Exemple : Dog, cats, town, and house
 - Doghouse, town hall. Ici, le sens de 'doghouse' est différent du sens de 'dog' et 'house'. Même chose pour 'Town hall'.
 - **Morphèmes liés (Bound)**: Généralement accompagnés d'affixes. Divisés en deux sous-catégories :
 - Morphèmes dérivationnels : Modifient le sens du radical, comme 'un' dans 'kind'.
 - Morphèmes flexionnels (Inflectional) : Ajoutent des informations grammaticales au mot, comme "-s" pour indiquer le pluriel, ou "-ed" pour le passé.



NLP : Niveaux - Analyse morpho-lexicale

❑ L'analyse lexicale :

- Le processus de décomposition d'un texte;. **en lexèmes significatifs** : mots, phrases et autres éléments significatifs.
- Basée sur l'analyse au niveau des mots. Dans ce type d'analyse, nous nous concentrons également sur le sens des mots, des phrases et d'autres éléments, tels que les symboles.

❑ Qu'est-ce qu'un jeton (Token) ?

Les jetons sont définis comme des éléments significatifs générés à l'aide de techniques d'analyse lexicale.

❑ Qu'est-ce que un POS tag?

Une partie du discours est une catégorie de mots ou **d'éléments lexicaux** qui ont **des propriétés grammaticales similaires**. Les mots appartenant à la même catégorie de partie du discours (POS) ont un comportement semblable dans la structure grammaticale des phrases. Exemples : le verbe, le nom, l'adjectif, l'adverbe, le pronom, la préposition,

NLP : Niveaux - Analyse morpho-lexicale

❑ Tokenisation des phrases

- Processus qui consiste à diviser un corpus de texte en phrases qui constituent le premier niveau de jetons du corpus.
- Ce processus est également connu sous le nom de segmentation des phrases, puisque nous essayons de segmenter le texte en phrases significatives.
- Tout corpus de texte est un ensemble de textes dont chaque paragraphe comprend plusieurs phrases.
- Il existe plusieurs façons d'effectuer la tokenisation des phrases. Les techniques de base consistent à rechercher des délimiteurs spécifiques entre les phrases, tels qu'un point (.) ou un caractère de nouvelle ligne (\n) et parfois même un point-virgule (;).

❑ Tokenisation des mots

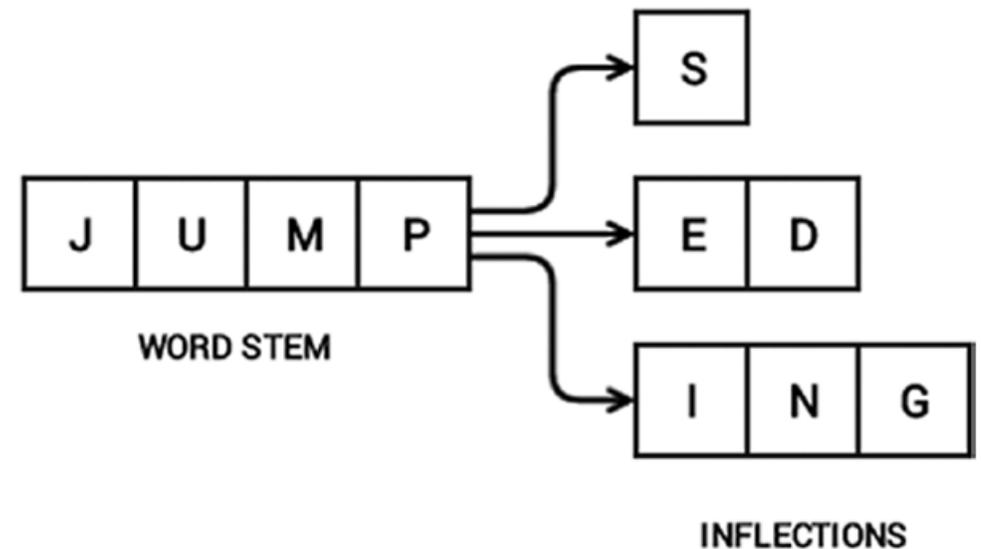
- Processus de division ou de segmentation des phrases en leurs mots constitutifs.
- Une phrase est un ensemble de mots et la tokenisation permet essentiellement de diviser une phrase en une liste de mots qui peuvent être utilisés pour reconstruire la phrase.
- La tokenisation des mots est très importante dans de nombreux processus, en particulier pour le nettoyage et la normalisation des textes, où des opérations telles que le stemming et la lemmatisation travaillent sur chaque mot individuel sur la base de ses stems et de ses lemma respectifs.

NLP : Niveaux - Analyse morpho-lexicale

Le stemming consiste à obtenir la forme de base d'un mot à partir de sa forme infléchie.

Exemple : Le mot "JUMP" peut former des mots comme "JUMPS", "JUMPED", "JUMPING" en ajoutant des affixes, mais sa forme de base reste "JUMP".

- Le stem du mot est présent dans toutes ses flexions, puisqu'il constitue la base sur laquelle chaque flexion est construite à l'aide d'affixes.
- Le stemming nous aide à normaliser les mots en fonction de leur racine de base, indépendamment de leurs inflexions → facilite de nombreuses applications telles que la classification ou le regroupement de textes ou même la recherche d'informations.
- Les moteurs de recherche utilisent largement ces techniques pour obtenir des résultats plus précis, quelle que soit la forme du mot.



NLP : Niveaux - Analyse morpho-lexicale

Lemmatisation

Le processus de lemmatisation est très similaire au stemming, qui consiste à supprimer les affixes des mots pour obtenir une forme de base du mot.

- Toutefois, dans ce cas, cette forme de base est également connue sous le nom de mot-racine (root), mais pas sous celui de radical.
- La différence entre les deux est que le radical n'est pas toujours un mot lexicographiquement correct, c'est-à-dire qu'il peut ne pas être présent dans le dictionnaire, alors que le mot racine, également connu sous le nom de lemme, sera toujours présent dans le dictionnaire.
- Le processus de lemmatisation est considérablement plus lent que le stemming parce qu'une étape supplémentaire est impliquée où la forme de la racine ou le lemme est formé en supprimant l'affixe du mot si et seulement si le lemme est présent dans le dictionnaire.

NLP : Niveaux - Analyse morpho-lexicale

Comparaison du stemming et de la lemmatisation

Stemming	Lemmatisation
Opère généralement sur un seul mot sans connaissance du contexte	Prend généralement en compte le contexte du mot dans la phrase
ne prend pas en compte les POS	Prend en compte les étiquettes POS
Regroupe les mots ayant un sens de base similaire	Le concept de lemmatisation est utilisé pour créer un dictionnaire ou un dictionnaire de type WordNet.

NLP : Niveaux - Analyse morpho-lexicale

❑ **Suppression des mots vides**

- Les mots vides sont des mots qui n'ont que peu ou pas de signification et qui sont généralement supprimés du texte lors de son traitement afin de conserver les mots ayant le plus de signification et de contexte.
- Les mots vides sont généralement les plus fréquents si l'on regroupe un corpus de textes basé sur des jetons singuliers et que l'on vérifie leur fréquence.
- Des mots comme "a", "le", "et", etc. sont des mots vides.
- Il n'existe pas de liste universelle ou exhaustive de mots vides et, souvent, chaque domaine ou langue possède son propre ensemble de mots vides.

NLP : Niveaux - Analyse syntaxique

Syntaxe : étude des règles qui régissent la combinaison des mots en phrases. L'objectif de ces règles est de prédire :

- la nature des constituants de la phrase
- la structure hiérarchique de ces constituants leurs fonctions syntaxiques

❑ Utilisée pour vérifier la grammaire, l'agencement des mots et montre la relation entre les mots.

La phrase telle que **'Run the quickly dog'** est rejetée par l'analyseur syntaxique anglais.

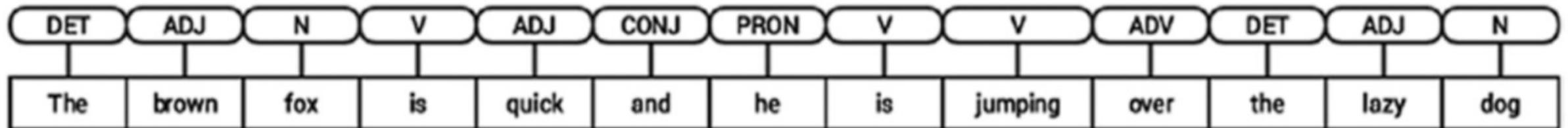
❑ Formes d'analyse syntaxique :

- Parts of speech (POS) tagging
- Shallow parsing or chunking
- Dependency parsing
- Constituency parsing

NLP : Niveaux - Analyse syntaxique

❑ POS tagging

- L'étiquetage des parties du discours est crucial en NLP → permet une analyse syntaxique plus approfondie des textes → analyses plus ciblées et des filtrages basés sur des catégories spécifiques. → Cela aide à comprendre la structure grammaticale et à extraire des informations clés.
- Les étiquetages POS permettent de filtrer les mots par catégories grammaticales spécifiques.
- Cela peut être utilisé pour simplifier l'analyse en réduisant le nombre de mots à considérer ou pour effectuer des analyses spécifiques basées sur des parties du discours particuliers
- Chaque POS tag, tel que les noms (N), peut avoir des subdivisions comme noms singuliers (NN), noms propres singuliers (NNP), noms pluriels (NNS), etc.



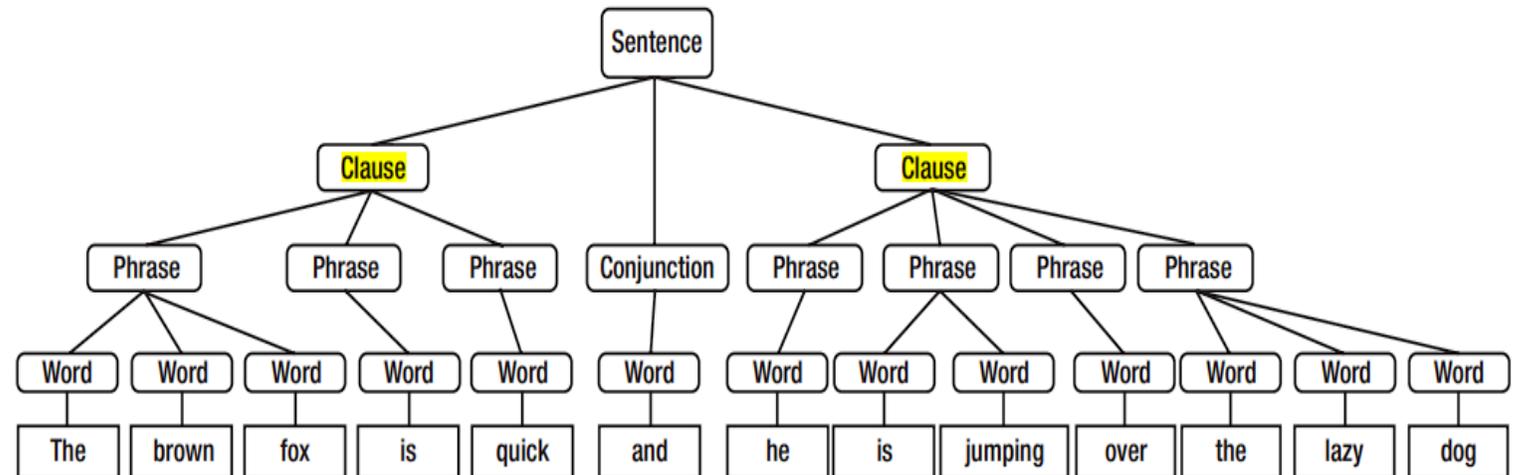
https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

NLP : Niveaux - Analyse syntaxique

- Les mots se combinent pour former d'autres unités constitutives.
- Ces unités constitutives comprennent les mots, les syntagmes (phrases en anglais), les clauses et les phrases.
- Tous ces constituants existent ensemble dans n'importe quel message et sont liés les uns aux autres dans une structure hiérarchique. **Sentence** → **clause** → **phrase** → **word**
- En outre, une phrase (sentence) est un format structuré représentant une collection de mots, à condition qu'ils respectent certaines règles syntaxiques comme la grammaire.
- Prenons un exemple de phrase, “The brown fox is quick and he is jumping over the lazy dog”

La grammaire et l'ordre des mots donnent un sens à une phrase. Et si nous mélangions les mots ? La phrase aurait-elle encore un sens ?

dog the over he
lazy jumping is the fox
and is quick brown



NLP : Niveaux - Analyse syntaxique

En effet, les langues ne sont pas simplement **constituées d'un sac ou d'un tas de mots non structurés**. **Les syntagmes dotées d'une syntaxe appropriée** → Donner une structure adéquate aux mots + les mettre en relation + aider à transmettre un sens en fonction de leur ordre ou de leur position

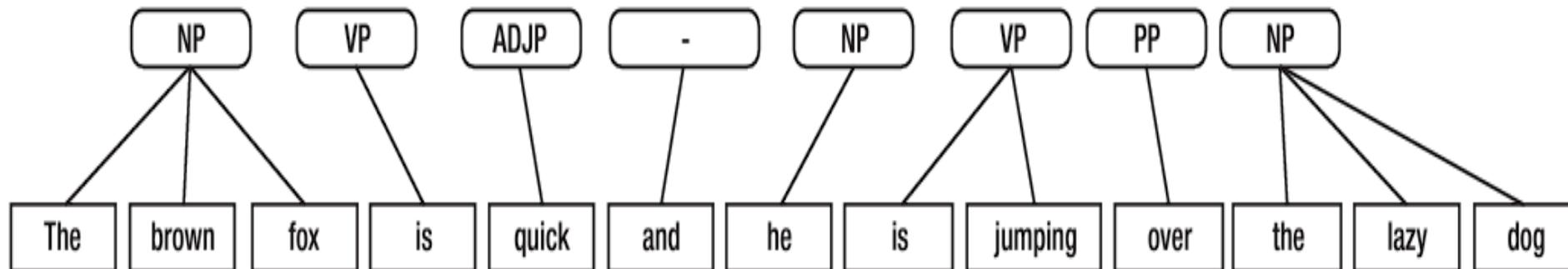
- ❑ Les syntagmes (phrase en anglais) sont des groupes de mots dans une hiérarchie arborescente de syntaxe.
- ❑ Les syntagmes peuvent être composées d'un seul mot ou de plusieurs mots en fonction de leur position dans une clause ou une phrase.
- ❑ Il existe cinq catégories principales de syntagmes : nominales (NP), verbales (VP), adjectivales (ADJP), adverbiales (ADVP), et prépositionnelles (PP).
- ❑ **Les phrases nominales** agissent comme sujet ou objet d'un verbe, remplaçables par un pronom sans rendre la phrase incorrecte.
- ❑ **Les phrases verbales** ont un verbe comme mot principal et peuvent varier selon la forme du verbe et les composants associés.
- ❑ **Les phrases adjectivales et adverbiales** qualifient respectivement les noms/pronoms et les verbes/adverbes dans une phrase.
- ❑ **Les phrases prépositionnelles** contiennent une préposition comme mot principal et décrivent d'autres mots ou phrases.

NLP : Niveaux - Analyse syntaxique

- ❑ **Shallow parsing or Chunking** → Identification de groupes de mots liés

L'analyse syntaxique superficielle est une technique de traitement du langage naturel qui consiste à identifier et à extraire de courtes phrases ou des morceaux de mots d'une phrase sur la base de certains modèles ou règles.

Contrairement à l'analyse syntaxique complète, qui vise à créer un arbre d'analyse complet en analysant l'ensemble de la structure grammaticale de la phrase, l'analyse syntaxique superficielle se concentre sur l'identification de morceaux spécifiques sans analyser leur structure interne en détail.



NLP : Niveaux - Analyse syntaxique

- Les clauses peuvent agir soit comme des phrases indépendantes soit être combinées pour former une phrase.
- Les clauses sont composées de **mots liés** et comprennent généralement **un sujet et un prédicat**.
- On peut les classifier en deux catégories principales : la clause principale (indépendante) et la clause subordonnée (dépendante).
- La clause principale** peut exister seule en tant que phrase, tandis que **la clause subordonnée** dépend de la principale pour son sens.
- Elles peuvent être catégorisées selon leur structure syntaxique : **déclaratives, impératives, relatives, interrogatives et exclamatives**.
- Les clauses déclaratives** présentent des déclarations neutres et factuelles.
- Les clauses impératives** sont des ordres, demandes ou conseils.
- Les clauses relatives** dépendent d'une autre partie de la phrase, faisant référence à un élément antécédent.
- Les clauses interrogatives** prennent la forme de questions affirmatives ou négatives.
- Les clauses exclamatives** expriment la surprise, le choc ou des compliments.

NLP : Niveaux - Analyse syntaxique

- ❑ **Analyse syntaxique en linguistique informatique** : Tâche consistant à créer un arbre syntaxique à partir d'une phrase donnée pour représenter sa structure syntaxique selon une grammaire formelle.
- ❑ **Définition de l'arbre syntaxique** : Représentation visuelle illustrant les relations entre les mots ou les sous-phrases, mettant en évidence la structure syntaxique de la phrase.
- ❑ **Influence de la grammaire** : Le type de grammaire utilisé détermine les caractéristiques de l'arbre syntaxique résultant.
- ❑ **Analyse de constituency et dépendance** : Deux méthodes utilisant des grammaires différentes, produisant des arbres très différents en raison d'assomptions distinctes.
- ❑ **Objectif commun** : Malgré les structures d'arbres divergentes, l'objectif reste l'extraction d'informations syntaxiques des phrases.

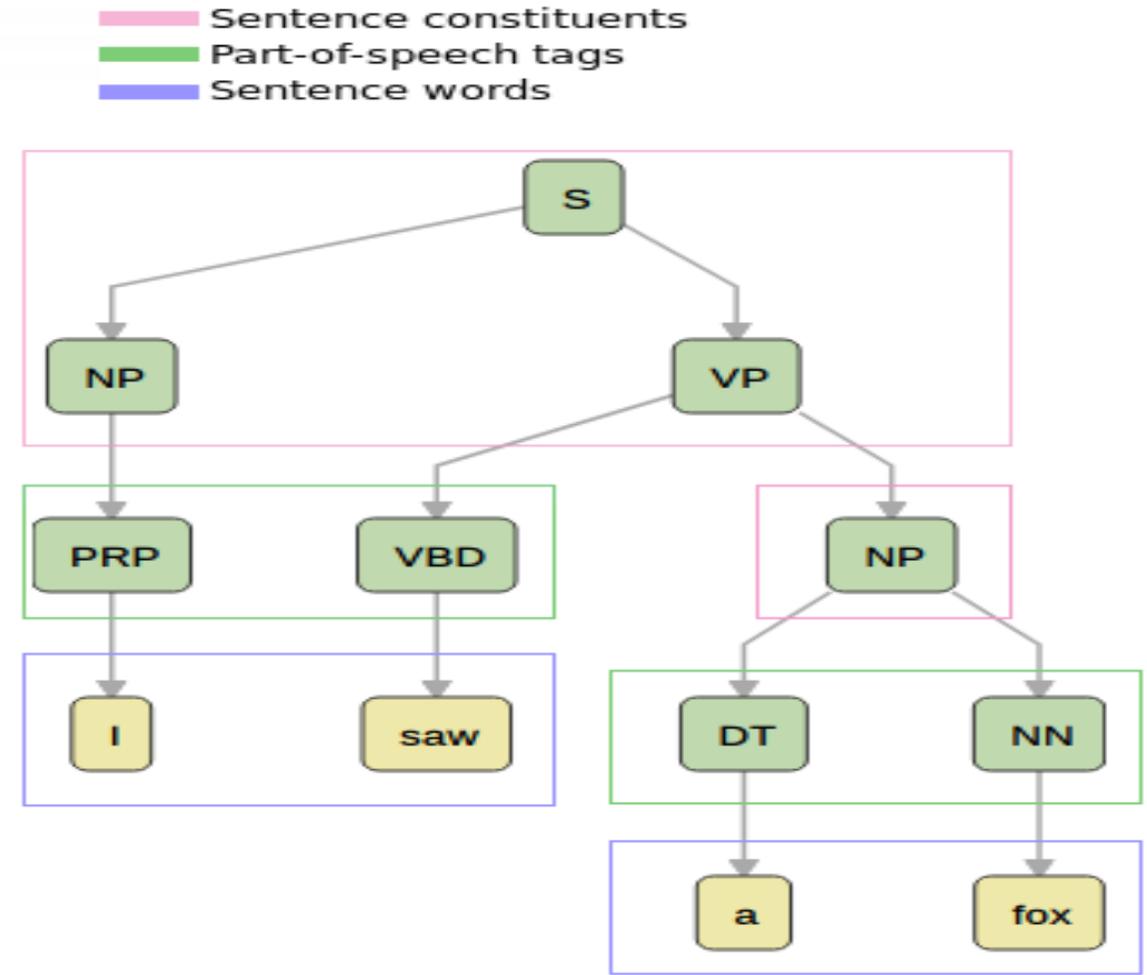
NLP : Niveaux - Analyse syntaxique

❑ Grammaire de constituance

- **Arbre syntaxique de constituance** : Basé sur les grammaires sans contexte, divise la phrase en constituants (sous-phrases appartenant à des catégories grammaticales spécifiques).
- **Spécification des règles grammaticales** : Les règles déterminent la formation des phrases (par ex., $VP \rightarrow V NP$ pour les phrases verbales).
- **Méthodologie d'analyse** : Utilise des règles pour générer des phrases valides et extraire la structure syntaxique des phrases données.
- **Structure Hiérarchique** : Organise les phrases en constituants phrastiques, créant une représentation hautement hiérarchique basée sur la grammaire sans contexte (**Context-free grammar**).

NLP : Niveaux - Analyse syntaxique

- Dans cet exemple, au premier niveau sous la racine, notre phrase a été divisée en un syntagme nominal, composé du seul mot "I", et un syntagme verbal, "saw a fox".
- Cela signifie que la grammaire contient une règle comme $S \rightarrow NPVP$, qui permet de créer une phrase par la concaténation d'un syntagme nominal et d'un syntagme verbal.
- De même, le syntagme verbal est divisé en un verbe et un autre syntagme nominal \rightarrow correspond également à une autre règle de la grammaire.



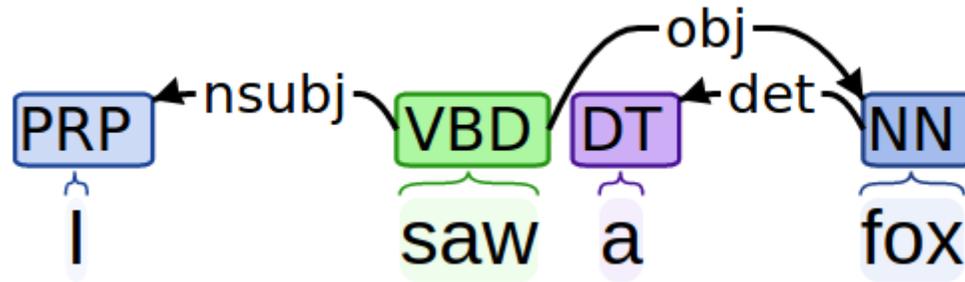
NLP : Niveaux - Analyse syntaxique

❑ Grammaire de dépendance

- **Contrairement à l'analyse syntaxique de constituants**, l'analyse syntaxique de dépendances **n'utilise pas de constituants phrastiques ou de sous-phrases**. Au lieu de cela, la syntaxe de la phrase est exprimée en **termes de dépendances entre les mots**, c'est-à-dire des arêtes dirigées et typées entre les mots d'un graphe.
- Suppose que tous les mots sauf un ont des dépendances, le mot racine (souvent le verbe) n'ayant pas de dépendances.
- Ne donne pas la priorité à l'ordre des mots, mais met l'accent sur les relations entre les mots.
- Décrit des fonctions syntaxiques spécifiques par le biais d'arêtes dirigées, indiquant les relations de dépendance.
- La liste des types de dépendance universels fournit des informations complètes sur ces relations.

<https://universaldependencies.org/u/dep/>

NLP : Niveaux - Analyse syntaxique



- Le résultat est complètement différent. Avec cette approche, la racine de l'arbre est le verbe de la phrase, et les arêtes entre les mots décrivent leurs relations.
- Par exemple, le mot "saw" a une arête sortante de type nsubj vers le mot "I", ce qui signifie que "I" est le sujet nominal du verbe "saw". Dans ce cas, nous disons que "I" dépend de "saw".

NLP : Demo
