Techniques du NLP

Pr T.RACHAD

QFM – 2024/2025

NLP?

- L'objectif général des techniques de la NLP (Natural Language Processing) est d'utiliser les connaissances générales sur un langage naturel pour doter un document d'un format plus structuré et qui est lisible par une machine.

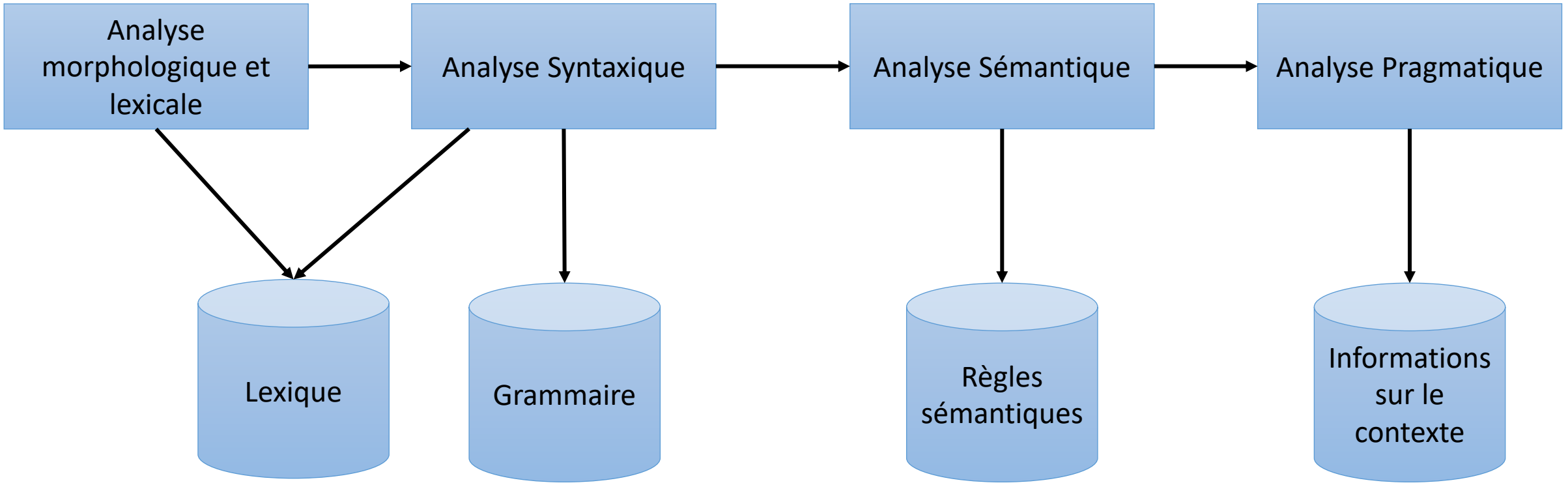
- Les représentations obtenues à l'issus des opérations de la NLP sont souvent transitoires, car elles ne permettent pas de résoudre des problèmes spécifiques.

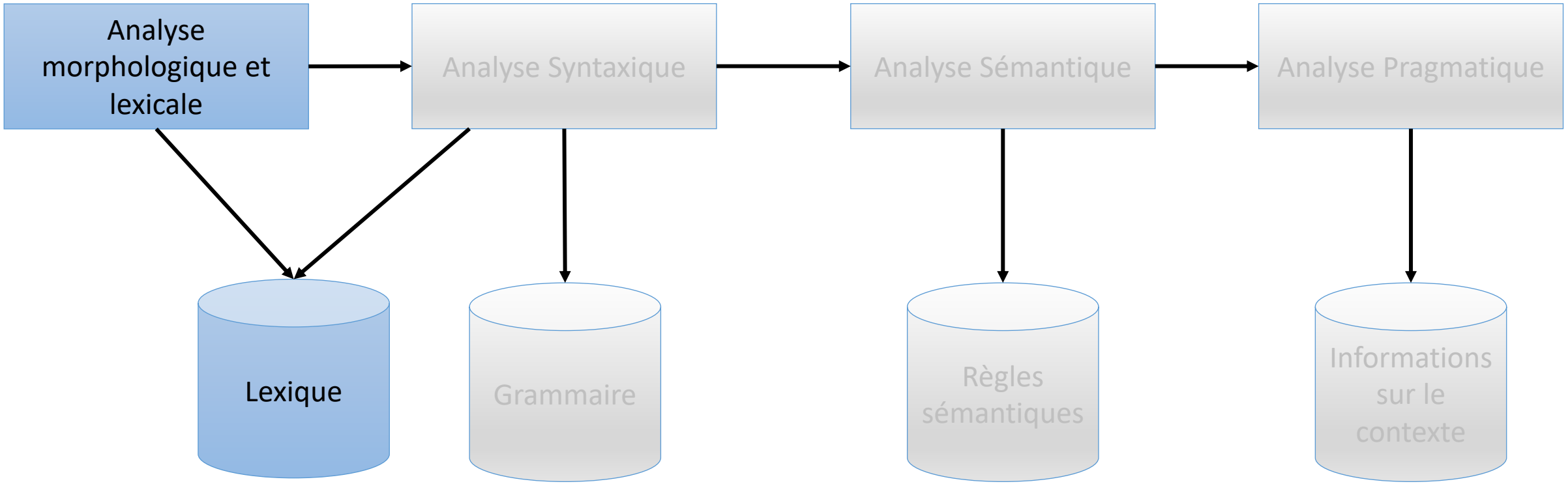
- Des traitements supplémentaires sont souvent sollicités pour avoir **un model représentatif** qui permet de réaliser aisément des opérations de découverte de patterns dans les documents.

Techniques du NLP

- Les techniques les plus courantes de la NLP sont:
  - La segmentation (Tokenization)
  - Les analyses morphologiques (racinisation et lemmatisation)
  - L’étiquetage morpho-syntaxique (POS tagging),
  - L’analyse syntaxique (syntactic parsing)
  - L’analyse sémantique

Processus du NLP




Analyse morphologique et lexicale
Analyse Syntaxique
Analyse Sémantique
Analyse Pragmatique
Lexique
Grammaire
Règles sémantiques
Informations sur le contexte

Analyse Morphologique

- L’analyse morphologique a comme objectif principal d’avoir une normalisation des mots contenus dans un texte en identifiant leurs formes canoniques.

- Une forme canonique est un mot qui correspond à la forme réduite commune de toutes ses dérivées.

- La **segmentation**, la **racinisation** et la **lemmatisation** sont les techniques déployées pour récupérer la forme canonique d’u mot.

Analyse Morphologique (Segmentation)

- La segmentation (tokenization) est le processus de subdivision des données texte en unités linguistiques plus simples à manipuler et qui sont appelées **tokens**(souvent des mots).

- D'autre termes qui sont reliés à la segmentation :
  - **Bigrams:** les tokens qui se composent de deux mots consécutifs.
  - **Trigrams:** les tokens qui se composent de trois mots consécutifs**.**
  - **Ngrams:** les tokens qui se composent d'un nombre «N» de mots consécutifs.

- Souvent les tokens obtenus doivent être nettoyés avant de passer aux opérations suivantes du prétraitement: enlever la ponctuation, enlever les termes non significatifs(stopwords), etc.

Analyse Morphologique (Racinisation)

- La racinisation ou la désuffixation (Stemming en anglais) est un algorithme qui permet d'éliminer le préfixe ou le suffixe d'un mot afin d'obtenir sa racine (ou radical)

- Il existe plusieurs algorithmes qui peuvent être utilisées pour réaliser la racinisation des mots:
  - En anglais: **Porter**, Lovins.
  - En français: Carry, **Porter**, Unine
  - En arabe: **Khoja**, Alkhalil….
  - Algorithmes génériques: Paice/Husk

Analyse Morphologique (Lemmatisation)

- Les algorithmes de racinisation sont peu efficaces, c'est pourquoi qu'il faut s'assurer que la forme résultante est un mot connu dans un dictionnaire.

- La **lemmatisation** permet d'extraire à partir d'un mot (verbe, adjectif…) sa forme canonique (***lemme***) enregistrée dans les **dictionnaires** de la langue.

- Plusieurs dictionnaires sont disponibles dans plusieurs langues pour réaliser les opérations de lemmatisation:
  - **SpaCy**
  - **WordNet**
  - TextBlob
  - Wordweb
  - …Pour plus de dictionnaires consulter http://www.nltk.org/nltk_data/

Analyse Morphologique (Exemples)

Racinisation		**Lemmatisation**	
• programmers	programm	• programmers	programmers
• programming	program	• Programming	program
• program	program	• Program	program
• because	becaus	• because	because
• its	it	• its	its
• like	like	• like	like
• english	english	• english	english
• people	peopl	• people	people

Analyse Morphologique (Exemples)

Initial Word	After removing	Root	Pattern	Pattern form	Lemma form
تنازل	تنازل	نزل	تفاعل	t1a23	تنازل
يستخرجون	يستخرج	خرج	يستفعل	yst123	استخرج
نحتاجهم	نحتاج	حوج	نفتعل	n1t23	احتاج
تندرج	تندرج	درج	تنفعل	tn123	اندرج

Tarek El-Shishtawy et al. , An Accurate Arabic Root-Based Lemmatizer for Information Retrieval Purposes

Analyse Lexicale

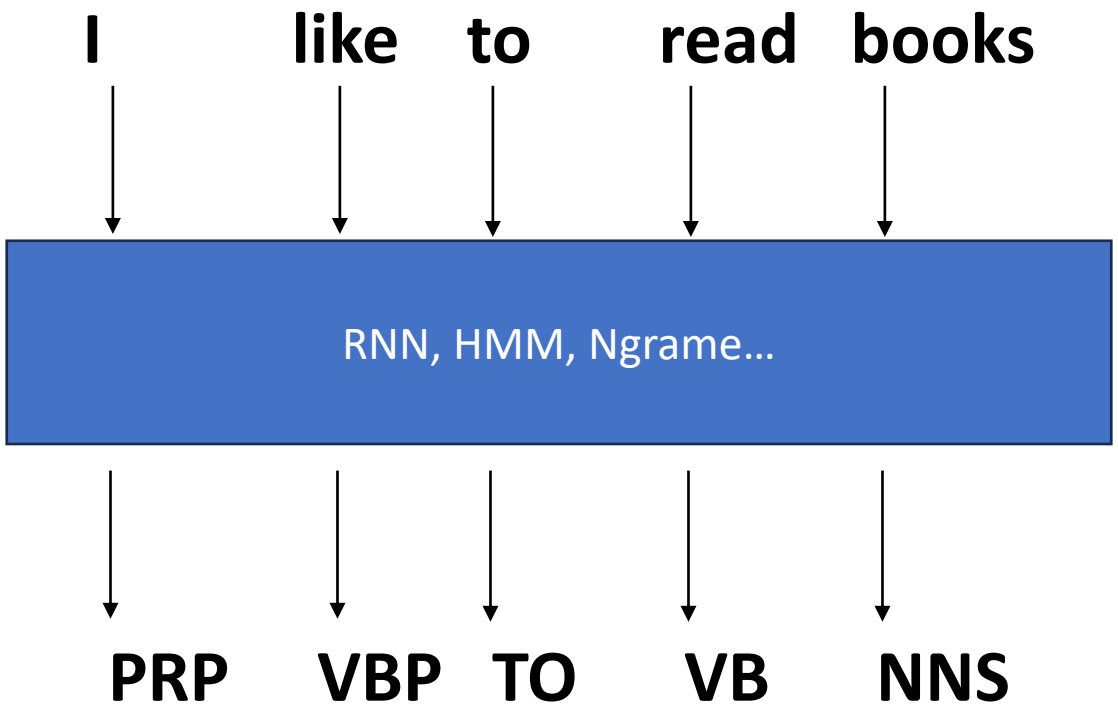
- L’analyse lexicale fait souvent référence à **l’étiquetage morpho-syntaxique** ou POS Tagging (Part of Speech Tagging)

- Le POS Tagging est l'annotation des mots en fonction du rôle qu'ils jouent dans la phrase dans laquelle ils apparaissent.

- Les étiquettes les plus courantes pour le pos tagging sont: article, nom, verbe, adjectif, préposition, nombre et nom propre.

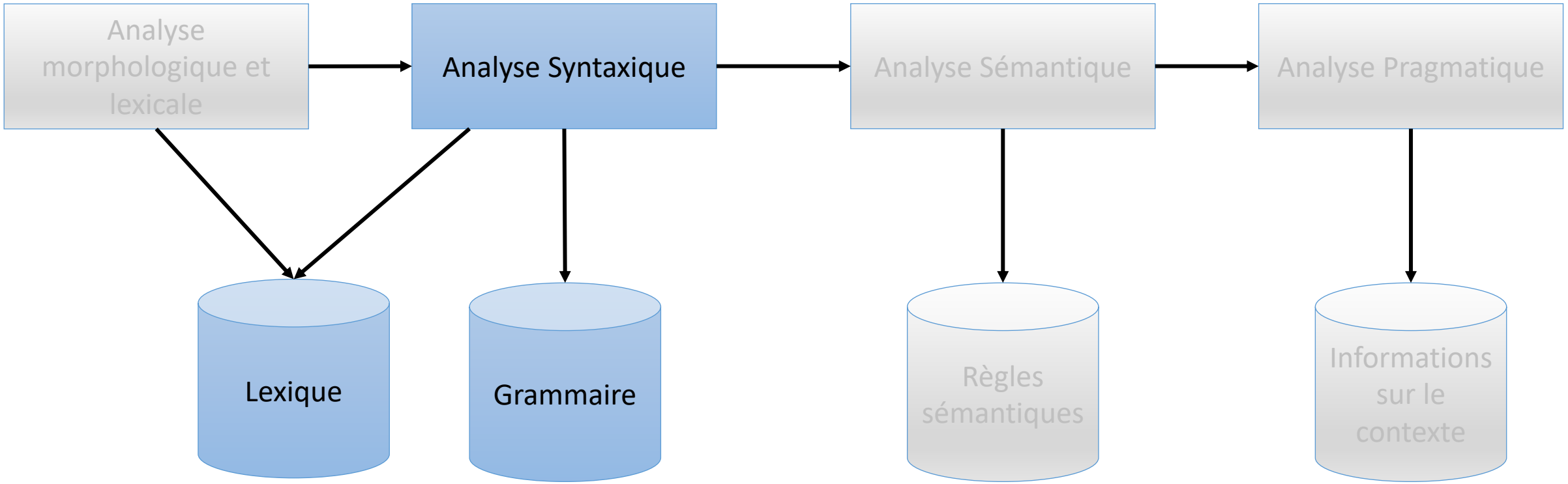
- Ça peut consister également à l’extraction du sens exacte de chaque mot dans une phrase.

Analyse Lexicale (Exemple)



Analyse Lexicale (Techniques)

- Plusieurs techniques peuvent être utilisées pour la réalisation de l'étiquetage morpho-syntaxique:
  - **Etiquetage à base de règles** qui utilise un dictionnaire ou une base de données lexicale afin de récupérer toutes les étiquettes possibles d'un mot. Par la suite un certain nombre de règles (expression régulières, Brill tagging …) seront exécutées pour identifier l'étiquette la plus adéquate en se basant sur le contexte du mot (**les n mots précédents et les n mots suivants**).
  - **Etiquetage stochastique** qui affecte à un mot l'étiquette qui occurre fréquemment avec le mot ou l'étiquette la plus adaptée aux étiquettes attribuées aux termes précédents (Ngrame, CRF, HMM, Baum-Welch…).
  - **Etiquetage à base de deep learning** (RNN)


Analyse morphologique et lexicale
Analyse Syntaxique
Analyse Sémantique
Analyse Pragmatique
Lexique
Grammaire
Règles sémantiques
Informations sur le contexte

Analyse Syntaxique

- L'analyse syntaxique est le processus d'analyse d’une expression texte conformément aux règles d’une **grammaire**.

- L’objectif est d’extraire une représentation structurelle des relations entre les composantes d’une expression texte (les phrases, les mots ou les symboles) souvent sous forme d’une arborescence.

- Applications:
  - NER
  - Question answering
  - Automatic Translation
  - Text summarization

Analyse Syntaxique (Exemple)

- L'exemple suivant illustre l'utilisation d'une petite grammaire pour l'identification de la structure de l'expression arithmétique$(8 + 6) \div 4$.

Expression

$(8 + 6) \div 4.$

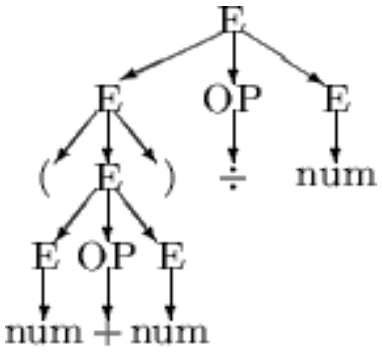
Grammaire

$$E \rightarrow E\,OP\,E \mid (E) \mid -E \mid \mathbf{num}$$
$$OP \rightarrow + \mid - \mid * \mid \div$$

Inférence

Rule Applied	*Sent. Form*
	E
$E \rightarrow E\,OP\,E$	$E\,OP\,E$
$E \rightarrow (E)$	$(\,E\,)\,OP\,E$
$E \rightarrow E\,OP\,E$	$(\,E\,OP\,E\,)\,OP\,E$
$E \rightarrow \mathbf{num}$	$(\mathbf{num}\,OP\,E\,)\,OP\,E$
$OP \rightarrow +$	$(\mathbf{num} + E\,)\,OP\,E$
$E \rightarrow \mathbf{num}$	$(\mathbf{num} + \mathbf{num})\,OP\,E$
$OP \rightarrow \div$	$(\mathbf{num} + \mathbf{num}) \div E$
$E \rightarrow \mathbf{num}$	$(\mathbf{num} + \mathbf{num}) \div \mathbf{num}$

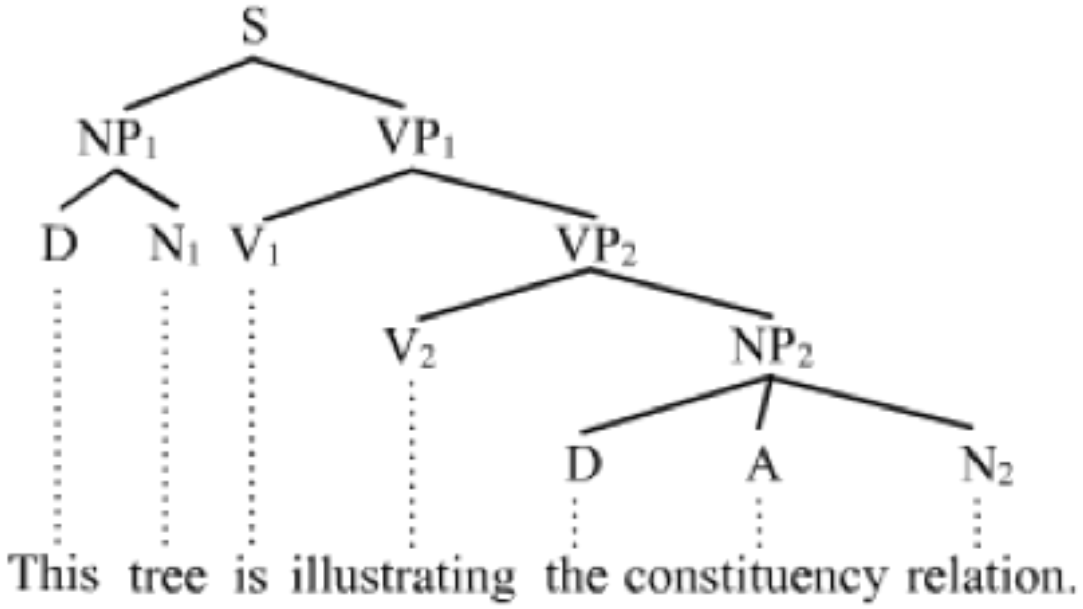
Représentation finale

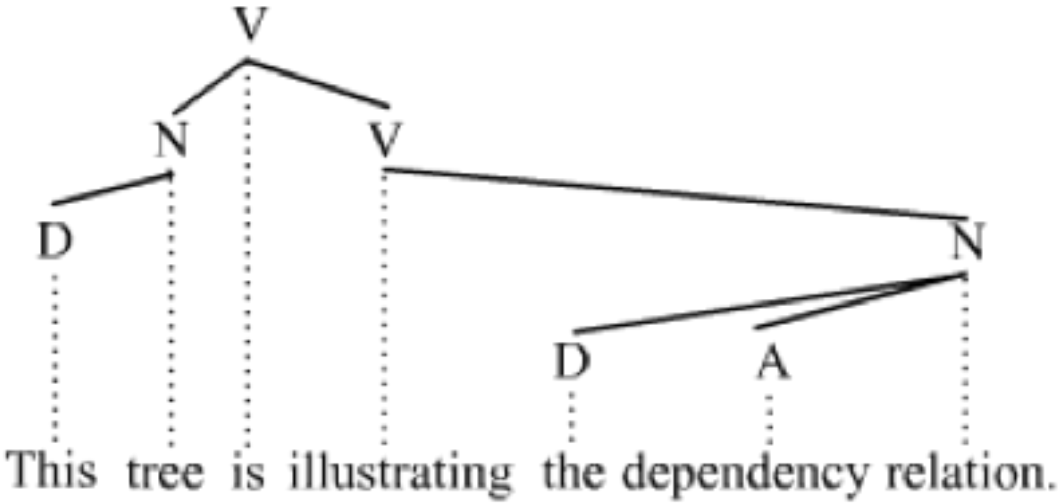


Analyse Syntaxique (Techniques)

- Approches basées sur des règles grammaticales.
  - **Grammaire syntagmatique**: PSG (Phrase Structure Grammar)
  - **Grammaire de dépendances**: Meaning-Text Theory, Link Grammar, Constraint Dependency Grammar, Extensible Dependency Grammar….
  - Plusieurs grammaires à base de règles sont proposées pour plusieurs langues:
    - Lexical Functional Grammar (LFG) ,
    - Head-Driven Phrase Structure Grammar (HPSG) LinGO Matrix framework,
    - Lexicalized Tree Adjoining Grammar XTAG.

Analyse Syntaxique (Techniques)





Analyse Syntaxique (Techniques)

- Approche statistique supervisée:
  - à partir d'un ensemble de couples (x, y) tel que x est une phrase et y est un arbre syntaxique, on doit découvrir les modèles permettant d'attribuer un arbre syntaxique à chaque nouvelle phrase.

  - Trois modèles sont possibles pour définir un analyseur syntaxique avec cette approche :
    - **Modèle à base de transitions**: l'analyseur syntaxique constitue graduellement l'arbre syntaxique en exécutant un certain nombre de transitions. À chaque étape, la transition qui a le plus grand score est choisie parmi toute les transitions possibles.
    - **Modèle a base de graphe**: trouver l'arbre syntaxique qui possède le score le plus élevé à partir d'un graphe qui représenté toutes les relations syntaxiques possibles entre les éléments d'une expression.
    - **L'approche hybride (ensemble**).

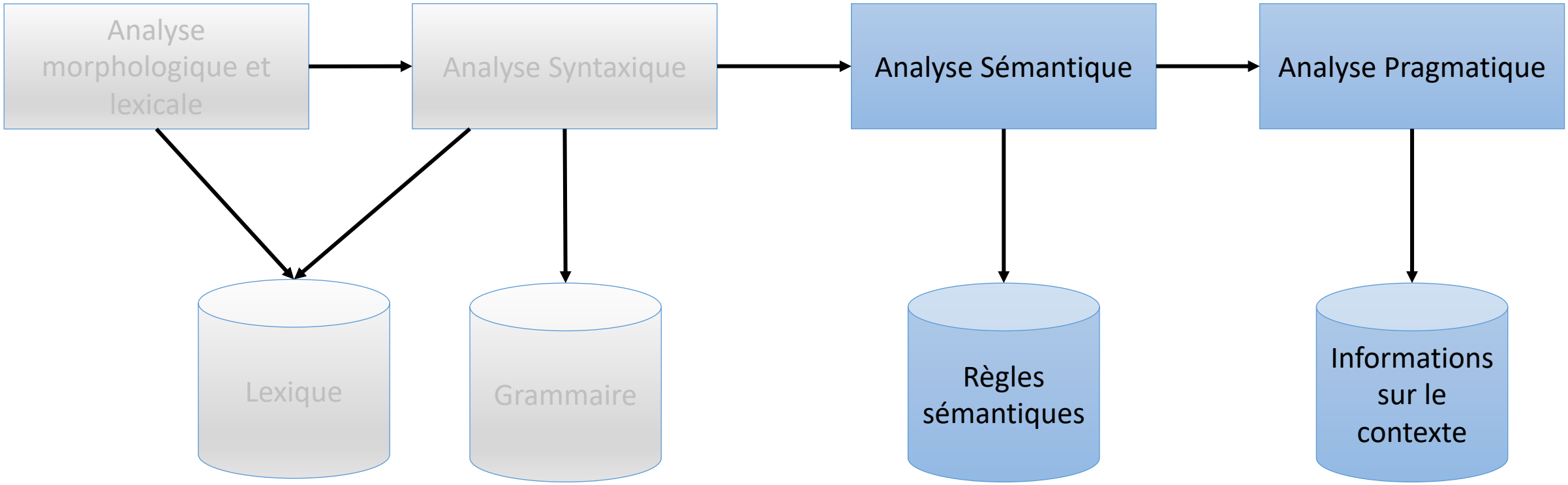
Analyse Syntaxique (Techniques)

- L’approche Statistique non supervisée:
  - Permet de résoudre le problème d'analyse en induisant des règles grammaticales cachées, leurs probabilités et des arbres syntaxiques à partir de données linguistiques non annotées.

  - Deux stratégies sont utilisées:
    - L’attraction lexicale des mots(Yuret, 1998)

    - Bootstrapping, c'est-à-dire guider l'apprentissage à partir des structures les plus simples et augmenter progressivement la complexité des phrases pour avoir une représentation complète(Spitkovsky et al., 2010).

Processus du NLP



Analyse Sémantique

- L’objectif de l’analyse sémantique est d’extraire le sens contenu (caché) dans des expressions texte.

- Ça doit commencer par une analyse sémantique lexicale pour extraire les sens « possibles » de chaque mot individuellement.

- Par la suite, découvrir les liaisons lexicales entre les mots pour extraire leur sens exacte.

- Finalement, constituer le sens exact d’une expression en fonctions des sens des mots qui la composent et en fonction du contexte.

Analyse Sémantique(Approche lexicale)

- L'analyse sémantique lexicale produit une représentation formelle du sens contenu dans une expression texte en se basant sur les notions d'entité, concept, relation et prédicat.
  - **Entités:** Des individus du monde réel(un emplacement (Casablanca), une personne (Bill gates), une date (01/01/2021)…)
  - **Concepts:** Catégorisation d'individus (Personne, Date, Ville, Pays….)
  - **Relations:** liaisons entre les concepts et les entités (Casablanca **est** une ville).
  - **Predicats** – la relation sémantique entre les entités dans une expressions texte: POS tagging, analyse syntaxique (Casablanca est une ville' ➔(sujet-predicat-objet))

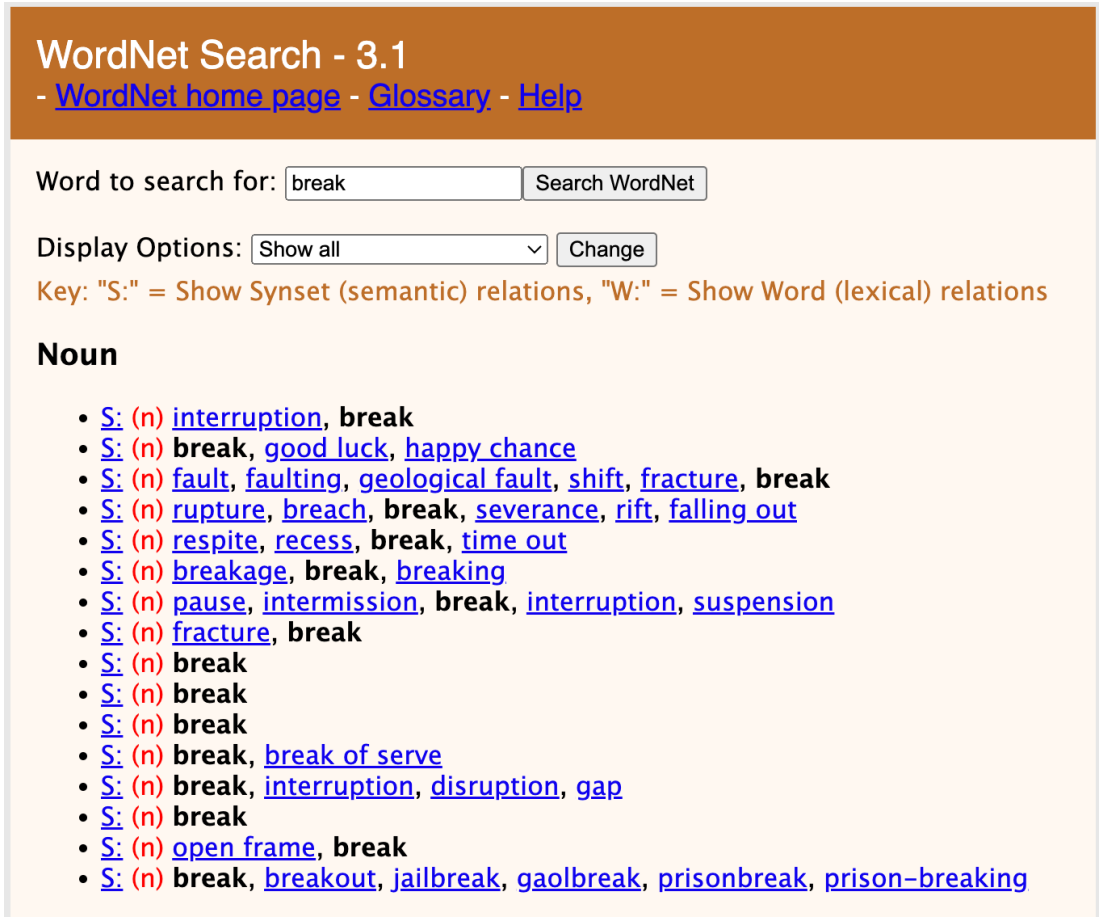
- Les labels (concepts + prédicats) sont prédéfinis dans des schémas lexicaux (WordNet, FrameNet, Openmind, Linkeddata…)

Analyse Sémantique(Approche lexicale)

- Les schémas lexicaux peuvent retourner plusieurs représentations formelles des sens des termes contenus dans une expression texte.

- Word Sens Disambiguation (WSD) permet de choisir à partir d'une base lexical le sens exact d'un mot selon le contexte de son utilisation.

- Plusieurs implementations du WSD



Analyse Sémantique(Approche lexicale)

```
function Simplified_Lesk(word, sentence)
  best-sense←most frequent sense for word
  max-overlap←0
  context←set of words in sentence
  for each sense in senses of word do
    signature←set of words in the gloss_examples of sense
      overlap←Compute_overlap(signature, context)
      if overlap > max-overlap then
                  max-overlap←overlap
                  best-sense←sense
  end
  return(best-sense){ returns best sense of word}
```

Analyse Sémantique (Approche statistique)

- Le processus général est le suivant:
  1. Tokenisation, (racinisation) stemming, lemmatisation.
  2. Extraction des éléments clés (Features) des données texte d'apprentissage, généralement sous format vectoriel:
     - Tokenized Document: Le format le plus naïf de la représentation vectorielle d'un document.
     - One hot vector.
     - bag-of-words (bow) / bag-of-n-grams (bong): Enregistre le nombre d'occurrences des termes dans un document.
     - matrix of word or n-gram (mow) : calculer (à partir de chaque document dans un corpus) le nombre d'occurrences de chaque terme d'un bow et représenter les résultats sous format matriciel.
     - TF-IDF (Term Frequency–Inverse Document Frequency) (tf-idf): permet d'evaluer l'importance d'un terme contenu dans un document par rapport aux autres documents du corpus. Les résultat sont représentés sous format matriciel..
     - Word2vec: cherche à apprendre les représentations vectorielles des mots composant un texte, de telle sorte que les mots qui partagent des contextes similaires soient représentés par des vecteurs numériques proches.
     - …

Analyse Sémantique (Approche statistique)

3. Optimiser la représentation des données vectorisées / réduction de la dimensionnalité
   - SVD: Singular value decomposition
   - LDA: Lattent Drichlet Allocation
   - SGD: Stochastic gradient descent
4. Découverte de modèles de classification, clustering…:
   - SVM: *Support Vector Machine*,
   - Multinomial classification (multiclass)
   - …

Analyse Sémantique(Applications)

- Analyse des similarités entre les documents
- Indexation de documents texte
- Analyse de topics
- Analyse de sentiments
- Analyse d’intentions
- Analyse d’opinions
- Réponse automatique aux questions: Chatbots
- Traduction automatique du texte