CHAPITRE V

Traitement automatique du langage naturel

Natural Language Processing

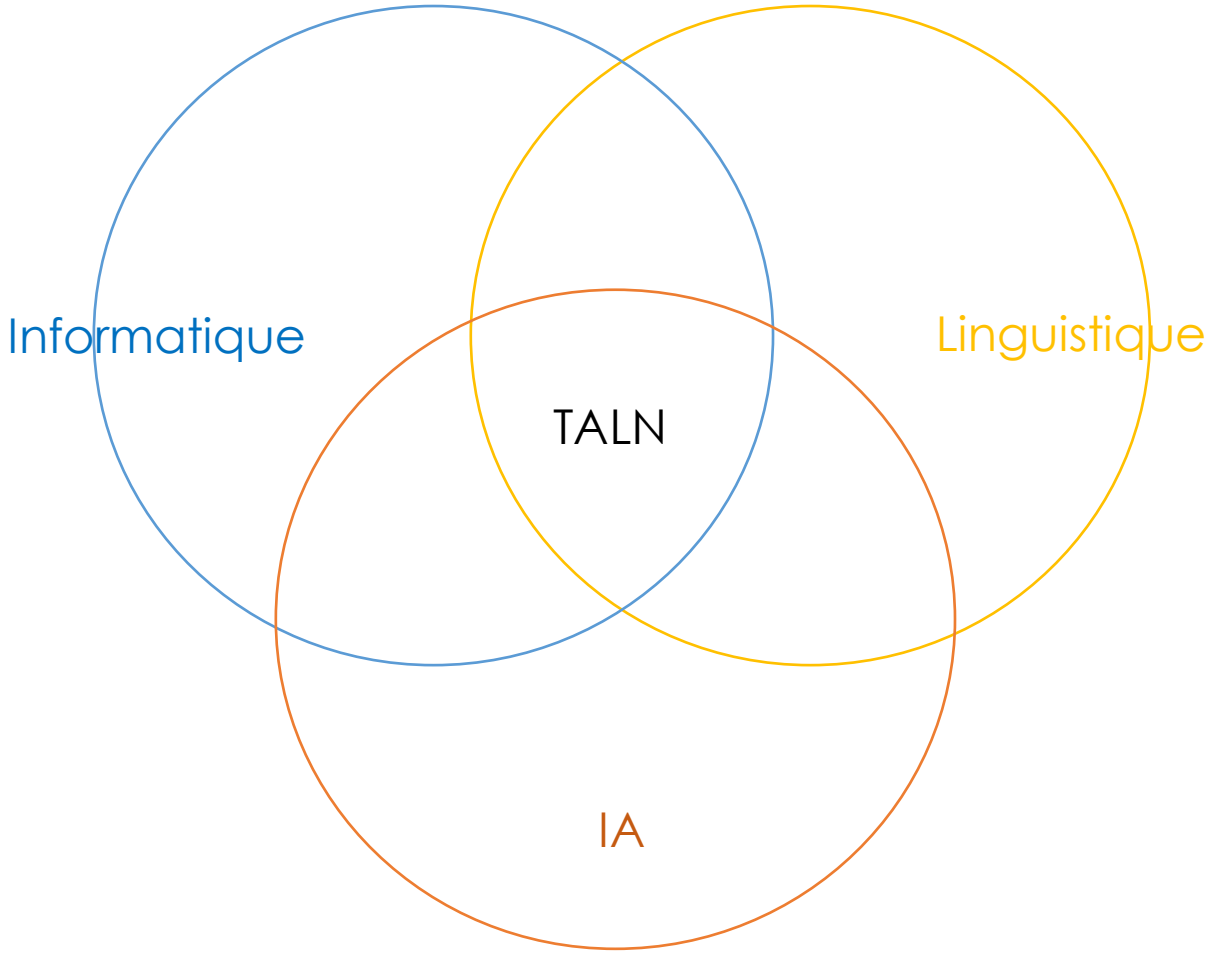
Histoire

- L'histoire du TALN commence dans les années 50s.
- Les premiers travaux concernaient la traduction automatique qui en fut l'une des premières applications informatiques
- Traitement automatique des conversations
- Tester l'intelligence des machines (test de Turing)
- 1954, traduction automatique des phrases russes (politique) vers l'anglais
- SHRDLU (1960)
- ELIZA (1964-1966)

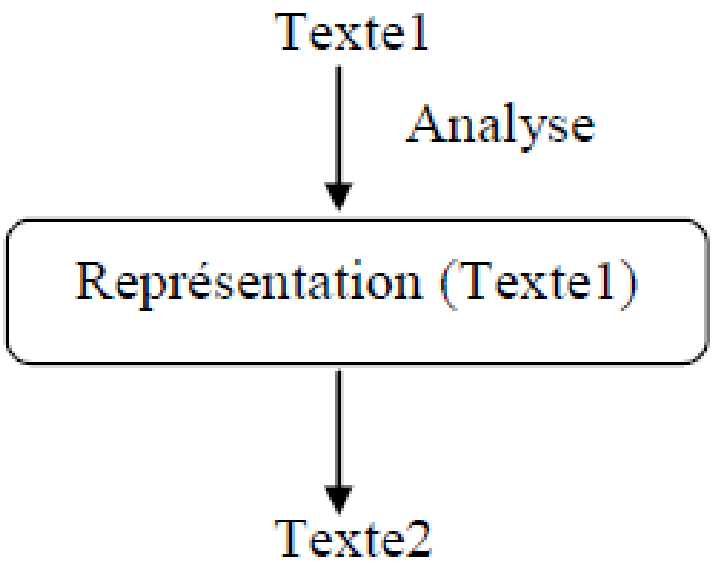
Définition

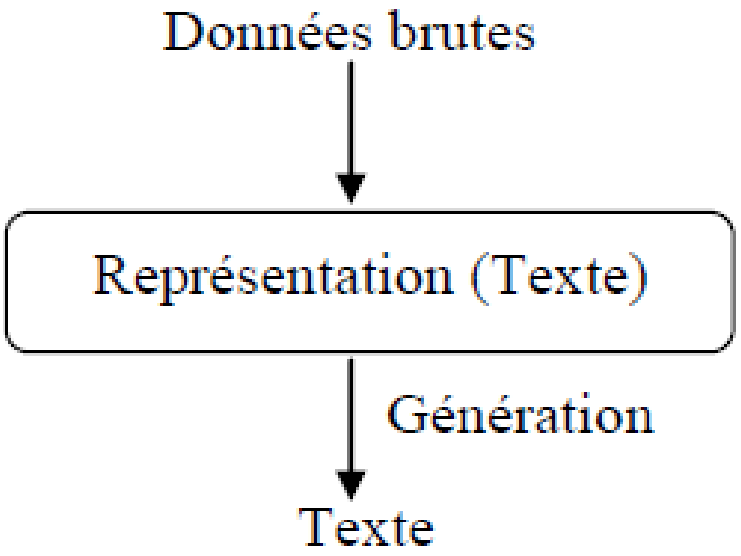
- Le TALN à pour objet la création de programmes informatiques capables de traiter automatiquement les langues naturelles
- La langue naturelle désigne la langue parlée ou écrites par les êtres humains
- Traitement des données linguistiques :
  - Textes écrits (paragraphes, phrases, mots,..)
  - Données orales (phonèmes, prosodies,..)

TALN : Domaine multidisciplinaire



Analyse vs Génération





Domaines d'application

	Texte	Parole
Analyse	▪ Correction orthographique ▪ Aide à la reformulation ▪ Recherche d'information fouille textuelle ▪ Reconnaissance d'entités nommées ▪ Classification et catégorisation de documents ▪ Reconnaissance de l'écriture manuscrite ▪ Annotation morpho-syntaxique / sémantique	▪ Reconnaissance vocale ▪ Identification du locuteur
Génération	▪ Génération automatique de textes ▪ Résumé automatique	▪ Synthèse de la parole

Outils TALN

- **Linguistiques :** théories linguistiques qui décrivent les différentes connaissances relatives à la langue

- **Formels :** expriment les connaissances linguistiques dans un formalisme qui convient à un traitement automatique

- **Informatiques :** utilisent la description formelle des connaissances dans une application informatique concrète

Niveaux de traitement

Traitement phonétique

Prétraitement

Traitement morphologique

Extraction des caractéristiques

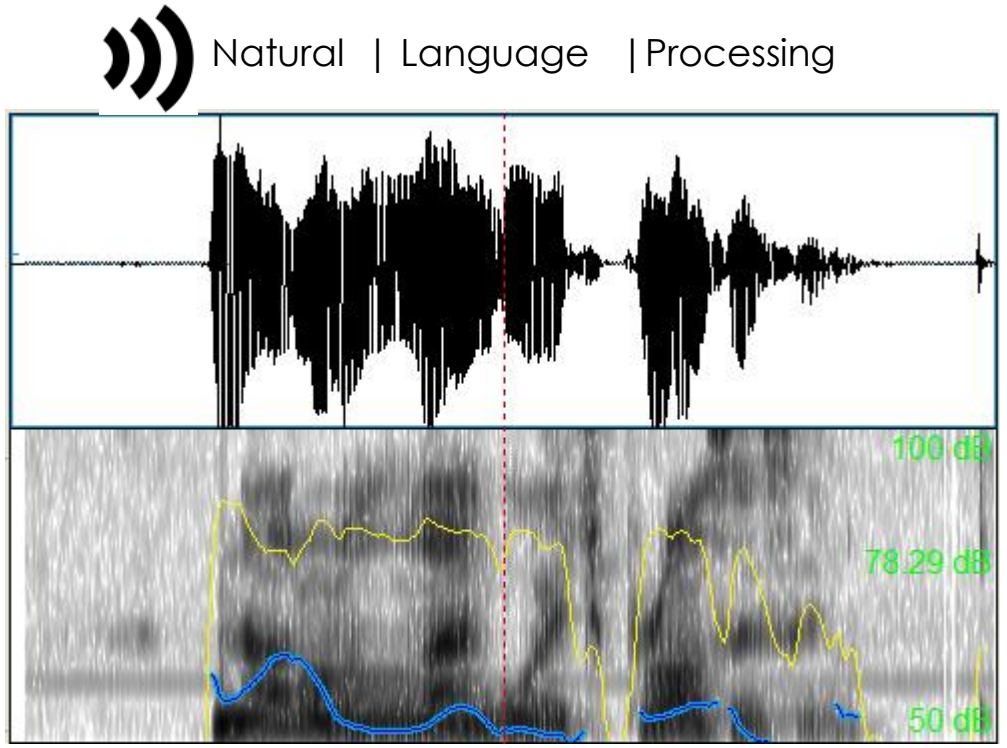
Traitement syntaxique

Traitement sémantique

Traitement pragmatique

Niveaux de traitement

- **Niveau phonétique (phonologique)**
  - Traitement de la langue orale
  - La machine doit reconnaître des signaux acoustiques et les identifier en temps que mots via une interface vocale
  - Identifier les prosodies



ˈnæʧrəl ˈlæŋgwɪʤ ˈprəʊsɛsɪŋ

Niveaux de traitement

- **Prétraitement**
  - Segmentation
  - Tokenisation
  - Suppression de mots vides
  - Normalisation
  - Encodage des caractères

يذهب محمد إلى المسجد كل يوم

يذهب، محمد، إلى، المسجد، كل، يوم

يذهب، محمد، المسجد، يوم

Niveaux de traitement

- **Niveau morphologique (morpholexical)**

  - Etudier la formation des mots et leur variation de formes.

  - Flexion vs Dérivation
    - Flexion : modifications que subit un mot dans sa terminaison (ou son début)
      - Ajout d'un affixe qui ne crée pas un nouveau lexème
      - Exemple : cheval / chevaux œuf / œufs
    - Dérivation : formation de mots nouveaux par addition, suppression ou remplacement d'un élément grammatical d'un mot simple
      - création d'un nouveau lexème par l'ajout d'un affixe
      - Exemple : délavé, intolérable, illisible

Niveaux de traitement

- **Niveau morphologique (morpholexical)**
  - Racinisation (stemming)
    - petit, petite, petits, petites ⇒ petit
  - Lemmatisation
    - Pêche, pêcher, pêcheur ⇒ pêche
    - eu, avions ⇒ avoir

Niveaux de traitement

- **Extraction des caractéristiques**
  - Text vectorization
    - Bag of words
    - N-grams
  - TF (term frequency)
  - TF-IDF (term frequency-inverse document frequency)

Variantes de TF

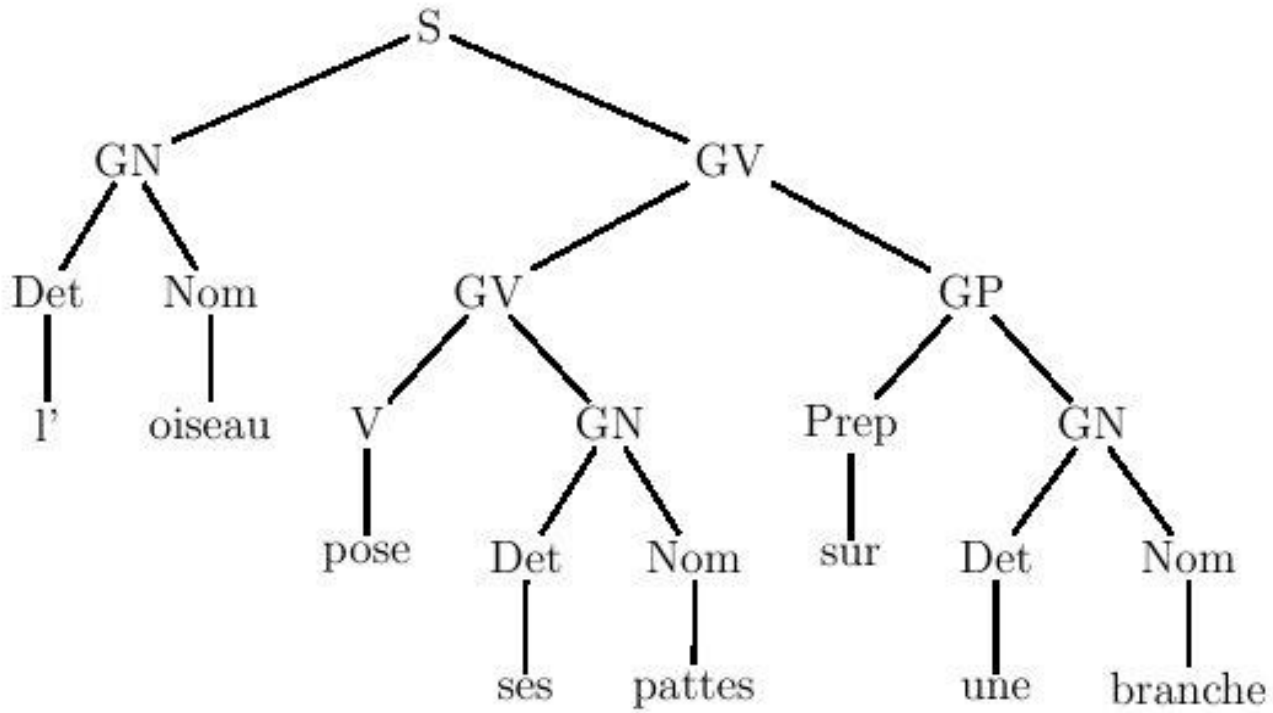
Schéma de pondération	formule du TF
binaire	$0, 1$
fréquence brute	$f_{t,d}$
normalisation logarithmique	$1 + \log(f_{t,d})$
normalisation « 0.5 » par le max	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
normalisation par le max	$K + (1 - K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

$$\mathrm{idf_i} = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

$$\mathrm{tfidf_{i,j}} = \mathrm{tf_{i,j}} \cdot \mathrm{idf_i}$$

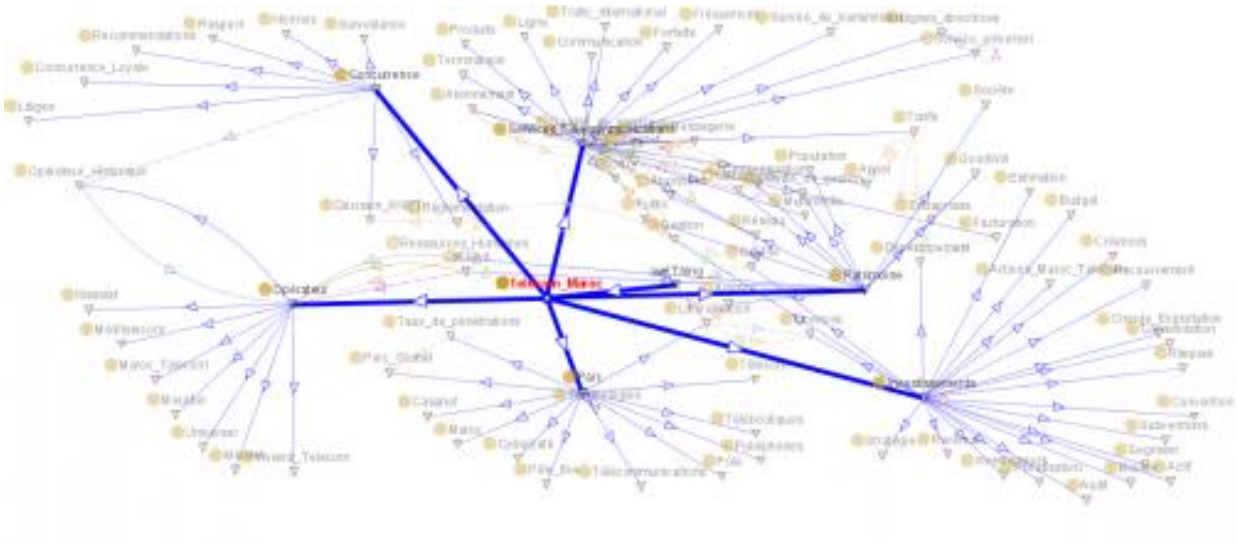
Niveaux de traitement

- **Niveau syntaxique**
  - Catégorisation des mots
  - Etudier l'agencement des mots et leurs relations structurelles
  - Détection et formation des phrases correctes (en utilisant des grammaires par exemple)
  - Lever les ambiguïtés grammaticales



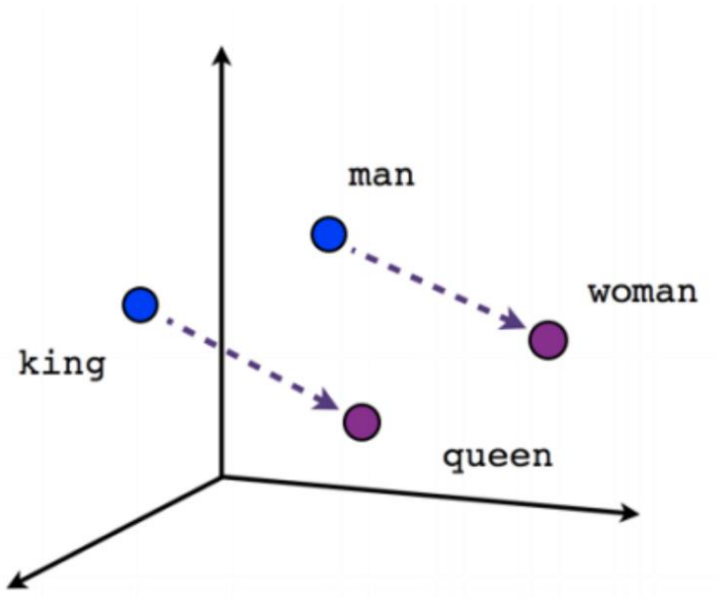
Niveaux de traitement

- **Niveau sémantique**
  - Associer un sens aux mots, phrases, énoncés
  - Détecter les structures n'ayant pas de sens
  - Utilisation des ressources lexicales et des ontologies pour :
    - Chercher des synonymes (WordNet)
    - Identifier les entités nommées
    - Voyéllation automatique (Arabe)

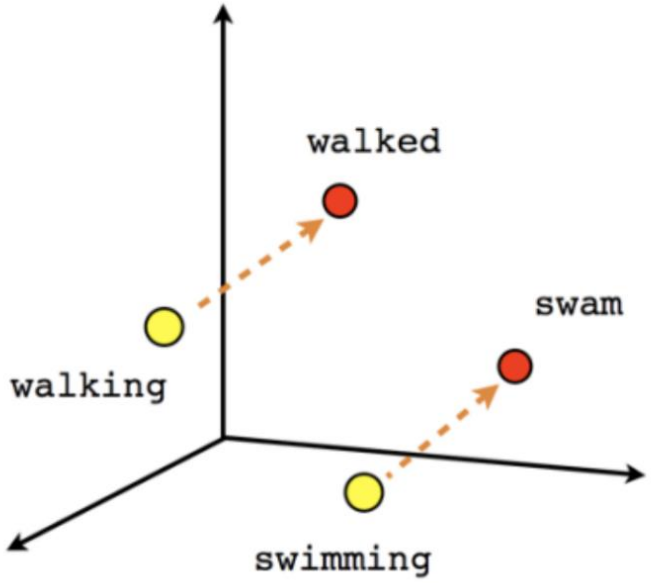


Niveaux de traitement

- **Niveau pragmatique**
  - Situer les mot et les énoncés dans les bons contextes



Male-Female



Verb tense

Evaluer un système TALN

- Connaitre la sortie souhaitée **?**
  - Recherche d'information : connaitre les documents pertinents
  - Résumé automatique : connaitre les phrases correctes
  - Traduction automatique : connaitre les traductions correctes
  - Classification de documents : connaitre les documents bien classés

Mesures d'évaluation d'un système TALN

- Rappel vs Précision
  - Classification de documents : connaitre les documents bien classés

$$Rappel_i = \frac{Documents\ correctement\ attribués\ à\ la\ classe_i}{Nombre\ de\ documents\ attribués\ à\ la\ classe_i}$$

$$Précision_i = \frac{Documents\ correctement\ attribués\ à\ la\ classe_i}{Nombre\ de\ documents\ appartenant\ à\ la\ classe_i}$$

Mesures d'évaluation d'un système TALN

- Rappel vs Précision
  - Classification de documents : connaitre les documents bien classés

Sujet réel	Sport	Sport	Autre	Autre	Autre	Sport	Autre	Sport	Autre	Sport
Document	1	2	3	4	5	6	7	8	9	10
Sortie système	Sport	Autre	Sport	Sport	Autre	Sport	Autre	Sport	Sport	Sport

Réponses correctes

$$Rappel_i = \frac{Documents\ correctement\ attribués\ à\ la\ classe_i}{Nombre\ de\ documents\ attribués\ à\ la\ classe_i}$$

  - Rappel et Précision pour la classe sport **?**

$$Précision_i = \frac{Documents\ correctement\ attribués\ à\ la\ classe_i}{Nombre\ de\ documents\ appartenant\ à\ la\ classe_i}$$